

Dynamiques sociales et sémantiques dans les communautés de savoirs, morphogenèse et diffusion

Jean-Philippe Cointet

▶ **To cite this version:**

Jean-Philippe Cointet. Dynamiques sociales et sémantiques dans les communautés de savoirs, morphogenèse et diffusion. Sciences de l'Homme et Société. Ecole Polytechnique, 2009. Français. NNT : . tel-02812231

HAL Id: tel-02812231
https://hal.inrae.fr/tel-02812231

Submitted on 6 Jun 2020

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

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

tel-00464316, version 1 - 16 Mar 2010

Ecole Polytechnique

T H E S E

pour obtenir le titre de

Docteur en Sciences

Mention : HUMANITÉS ET SCIENCES SOCIALES

présentée par

Jean-Philippe COINTET

Dynamiques sociales et sémantiques dans les communautés de savoirs

---

Morphogenèse et diffusion

Thèse co-dirigée par Paul BOURGINE et Pierre-Benoît JOLY

préparée au CREA (Ecole Polytechnique & CNRS) et à TSV (INRA)

soutenue le 9 Octobre 2009

devant le jury composé de :

Président :	Michel CALLON	-	Mines Paris Paristech (CSI)
Rapporteurs :	Jean-Benoît ZIMMERMANN	-	CNRS (GREQAM)
	Michel GROSSETTI	-	Toulouse II (LISST-CERS)
Examinateurs :	Natalie GLANCE	-	Google
	Matthieu LATAPY	-	CNRS (LIP6)
Directeurs :	Paul BOURGINE	-	CNRS (CREA)
	Pierre-Benoît JOLY	-	INRA (INRA/SenS)

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Résumé de la thèse:

Les *communautés de savoirs* en tant qu'espaces hybrides, mettent en jeu, d'une part, des individus qui interagissent au sein d'un *réseau social* et qui produisent des contenus localement, et d'autre part, des entités sémantiques, liées au sein d'un *réseau sémantique* doté d'une structure autonome. Cette thèse vise à explorer les dynamiques sociales et sémantiques qui animent ces communautés de savoirs, ainsi que leur couplage structurel.

Elle s'appuie essentiellement sur deux types d'objets empiriques : les communautés scientifiques et les blogosphères politiques, dont nous suivons l'activité grâce à l'observation *in-vivo* de leurs *traces* digitales. Notre ambition est de réaliser une reconstruction phénoménologique réaliste des dynamiques socio-sémantiques de ces communautés de savoirs. Pour ce faire, nous distinguons entre les dynamiques liées à la morphogenèse puis à la phylogenèse des réseaux qui composent ces systèmes et les processus dynamiques de diffusion qu'ils supportent.

Les méthodes de caractérisation de la *morphogenèse* des communautés de savoirs permettent, d'une part, de repérer et représenter les structures et propriétés émergentes de ces communautés de savoirs, et d'autre part, de mettre en évidence les comportements individuels réguliers qui les animent. La *phylogenèse* des communautés de savoirs consiste en la reconstruction des dynamiques de ces structures de haut-niveau émergeant au cours de la morphogenèse. L'analyse des *processus de diffusion* nous amène à interroger les déterminants structurels micro et macro qui conditionnent la circulation d'information dans ces communautés. L'ensemble de ces approches nous permettent de décrire les communautés de savoirs comme des systèmes dont les dimensions sociale et sémantique sont, par essence, en co-évolution.

tel-00464316, version 1 - 16 Mar 2010

Remerciements

Une thèse consacrée aux nouveaux espaces de production collective des savoirs ne serait pas complètement fidèle à son propos si elle avait ressemblé à une traversée solitaire. Ce travail est avant tout le fruit de nombreuses rencontres dont certaines remontent bien avant le commencement de ma thèse. Je tiens, en préambule, à remercier collectivement toutes celles et tout ceux auprès de qui j'ai pu mûrir ce projet et plus généralement mon goût pour la recherche.

J'aimerais, en premier lieu, exprimer ma gratitude à deux personnes dont les conseils précieux et le soutien indéfectible m'ont permis de m'engager dans cette aventure. Merci à *Claude Millier* qui a su prolonger mes doutes et mes questions plutôt que de les éteindre, et à *François Taddéi* pour sa disponibilité et sa très grande liberté d'esprit. Merci également à tous ceux qui ont contribué à ma cause lors des délibérations estivales ayant précédé mon entrée en thèse. Et merci au corps du GREF bien sûr de m'avoir finalement fait confiance.

Je remercie sincèrement mes deux directeurs de thèse pour leur confiance. Merci *Paul*, de m'avoir toujours maintenu dans un état de curiosité scientifique. Merci pour tes intuitions fulgurantes, ton regard pétillant de jour comme de nuit, et ton exigence. Merci *Pierre-Benoît* d'avoir également accepté de diriger cette thèse. Merci pour la patience dont tu as fait preuve pour tenter de m'initier aux sciences sociales, et à la sociologie des sciences en particulier. Merci pour le tact dont tu as usé, et dont tu risques de devoir user encore...

Je remercie vivement *Natalie Glance*, *Michel Callon*, *Matthieu Latapy* d'avoir accepté de prendre part à mon jury de thèse ainsi que *Michel Grossetti* et *Jean-Benoît Zimmermann* pour avoir accepté d'en être les rapporteurs, malgré les délais plutôt serrés que je leur ai imposés.

Le travail quotidien de recherche peut parfois vous transformer en ermite, merci tout particulièrement à *Camille* et *David*, pour m'avoir pris sous leur aile à mon arrivée au laboratoire et m'avoir évité cet écueil. J'ai énormément appris à votre contact ; nos collaborations et projets communs, de tout ordre, ont donné une véritable orientation à mon travail, j'espère qu'elles se prolongeront encore longtemps.

Je remercie tout particulièrement mes comparses de bureau pour avoir supporté avec tant de bonhomie mes humeurs, ma faculté d'occupation de l'espace et ma tendance chronophage à partager certains *écrans* à haute voix. Merci à *Emmanuel* pour sa décontraction naturelle, à *Benoît* pour sa gentillesse, à *Thierry* pour ses trois bises et pour bien d'autres choses encore, ainsi qu'à Marc et Pipo, pour les prétextes qu'il n'ont cessé d'offrir. Sans oublier le sens de l'orientation en milieu urbain de *Masa*, la bonne humeur de *Carla*, les déjeuners avec *Antoine*, les oursins de *Louise*, les pauses pelouse avec *Lionel*, la belle époque du Brelan avec *Thomas*, *Alex* et *Julien*, le Têtu de Marc, nos ballades cartographiques avec Jean-Paul et Christophe.

tel-00464316, version 1 - 16 Mar 2010

Merci aux “gens d’en haut” de l’Institut des Systèmes Complexes Paris-Île de France, qui m’a hébergé pendant deux années, Daniel, Romain, Pierre et bien d’autres pour m’avoir honoré d’un titre prestigieux bien que mensuel. Je n’oublie pas *Tam Kien* dont j’ai tant apprécié l’éclectisme. Merci également à *René* au charme anglo-saxon et au sens de l’ordonnancement inimitable.

Je remercie l’ensemble du personnel administratif qui a toujours été d’une aide précieuse pour m’offrir un environnement de travail confortable, et des moments de respiration bienvenus. Merci beaucoup à vous : *Stéphanie*, *Geneviève*, *Noemi*, *Nadiège*, *Marie-Jo*, *Yamina* et *Danielle*.

Je tiens également à remercier l’ensemble des stagiaires avec qui j’ai eu le plaisir de travailler, et qui, je l’espère, auront autant appris que moi à leur contact. Je tiens particulièrement à remercier *Pierre* pour sa gentillesse et son courage. Mais aussi, *Olivier*, *Jean-Charles*, *Hugo*, *Richard*, *Flavien*, *Matthieu* et *Adrien*.

Je souhaite également remercier et tirer mon chapeau à celles et ceux qui ont accepté de se confronter à mon orthographe et à ma ponctuation toutes personnelles durant la délicate phase de correction de cette thèse. Mes remerciements (ainsi que mes excuses) à, *Adrien* pour son enthousiasme ; *Benoît* pour sa minutie ; *Michael* pour son haut degré de tolérance, *Tanguy* pour sa science du point virgule ; *Marie*, pour m’avoir appris à ne pas me satisfaire des “formes admises”, etc.

Je terminerai en remerciant tous ceux qui m’ont toujours entouré, que ce soit dans mon travail ou dans la vie de tous les jours. Je remercie mes parents, sans l’aide desquels je n’aurais pu réaliser ce travail, et ma sœur, si loin si proche (un coucou à Hugo). Mais aussi mes amis qui ont supporté mon rythme estudiantin, *Léonard*, *Mathieu*, *Hadia*, *Franck*, *Vincent*, *Antonin*, *Bastien*, *Corentin*, *Benjamin*, et les autres...

Merci à toi *Jessica* d’être à mes côtés.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Table des matières

Introduction **7**

I Suivre les communautés de savoirs **11**

1 Communautés de savoirs **13**
1.1 Propriétés des communautés de savoirs 13
1.1.1 Communautés de blogueurs politiques et communautés scientifiques : des communautés de savoirs 14
1.1.2 Hétérogénéité des engagements et frontières des communautés de savoirs 17
1.1.3 Un système socio-cognitif distribué 20
1.1.4 Un mode de coordination stigmergique 23
1.2 Des réseaux sociosémantiques 24
1.2.1 Dualité socio-sémantique 24
1.2.2 Réseaux épistémiques 28
1.2.3 Formalisme 31

2 Dynamiques multi-échelles des communautés de savoirs **35**
2.1 Analyse longitudinale des dynamiques des communautés de savoirs 36
2.1.1 Limites de l'approche statique 36
2.1.2 Formalisme dynamique 38
2.2 Articuler les niveaux micro et macro 38
2.2.1 Boucle émergence immergence 38
2.2.2 Individus et réseaux 42
2.3 Observation *in-vivo* des dynamiques 44
2.3.1 Suivre les acteurs à l'ère digitale 44
2.3.2 Reconstuction phénoménologique des traces 46
2.3.3 Des traces textuelles au réseau épistémique 48
2.3.4 Un échantillon de la blosphère politique française 50
2.3.5 Un multi-réseau dynamique 53
2.3.6 Caractérisation sémantique 55
2.3.7 Blogosphère américaine 58
2.4 Une approche par faces 59

II Morphogenèse dans les réseaux de savoirs 65

3 Dynamiques locales 67
3.1 Dynamiques locales dans le réseau social 69
3.1.1 Attachements préférentiels 69
3.1.2 Attachement préférentiel aux degrés : capital social et capital sémantique 71
3.1.3 Attachement préférentiel à la distance sociale et sémantique . 76
3.1.4 Motifs cohésifs locaux 81
3.1.5 Capitaux, homophilie sociale et sémantique, découpler les effets 84
3.2 Dynamiques locales dans le réseau socio-sémantique 86
3.2.1 Similarité et interaction 86
3.2.2 Cohésion socio-sémantique locale 91
3.3 Dynamiques locales dans le réseau sémantique 93
3.3.1 Mesures d'occurrences 93
3.3.2 Mesures de co-occurrences 95

4 Structures émergentes 101
4.1 Communautés thématiques et communautés structurelles 104
4.1.1 Une portion du web social français 104
4.1.2 Détection des communautés structurelles 105
4.1.3 Hétérogénéité des topologies 107
4.1.4 Conclusion 107
4.2 De l'analyse de l'activité scientifique à la cartographie des sciences . 108
4.2.1 Les mutations contemporaines de l'activité scientifique 108
4.2.2 Les bases de données de publications scientifiques, une opportunité pour la cartographie des sciences 110
4.2.3 Un modèle de l'activité scientifique 110
4.2.4 un modèle multi-échelle de la connaissance 112
4.2.5 Méthodes scientométriques de cartographie des sciences . . . 113
4.3 Cartographier les sciences 114
4.3.1 Jeux de données 114
4.3.2 Une mesure asymétrique de proximité entre termes 116
4.3.3 Construction du réseau lexical 122
4.3.4 Echelle microscopique : voisinages locaux 122
4.4 Echelle mésoscopique : la notion de champ épistémique 123
4.4.1 Définitions 123
4.4.2 Identifier les champs épistémiques 125
4.4.3 Plongement des clusters dans un espace bi-dimensionnel . . . 127
4.4.4 Qualifier les clusters 129
4.4.5 Représentation macroscopique 130

tel-00464316, version 1 - 16 Mar 2010

4.4.6 Reconstruction multi-échelle . . . 132
4.4.7 Procédures de validation . . . 136
4.5 Méthode de reconstruction dynamique . . . 137
4.5.1 Dynamiques de voisinage . . . 138
4.5.2 Dynamique d'un champ épistémique . . . 138
4.5.3 Vers les dynamiques macroscopiques . . . 141
4.5.4 Reconstruction de la phylogénie des sciences . . . 144
4.6 Trajectoires des individus au sein des paysages sémantiques. . . . 153
4.6.1 Opérateur de projection . . . 153
4.6.2 Rétroaction macro-micro . . . 156
4.6.3 Se déplacer dans un espace mouvant . . . 160

III Diffusion dans les réseaux sociaux 165

5 Corrélations intertemporelles entre sources 167
5.1 Création des catégories de blogs . . . 169
5.1.1 Définition des profils sémantiques instantanés des blogs . . . 169
5.1.2 Catégorisation des blogs selon leur sensibilité politique . . . 170
5.2 Diagramme de corrélations intertemporelles . . . 172
5.2.1 Contexte . . . 172
5.2.2 Machines à états causaux . . . 173
5.2.3 Alphabet des concepts . . . 174
5.2.4 Définition d'une dynamique symbolique . . . 175
5.2.5 Resultats . . . 176
5.2.6 Perspectives . . . 178

6 Du rôle de la topologie des réseaux sur la diffusion 183
6.1 Protocole de simulation . . . 186
6.1.1 Protocole de simulation . . . 186
6.1.2 Topologies de réseaux . . . 188
6.2 Dynamiques de diffusion . . . 191
6.2.1 Résultat des simulations . . . 191
6.2.2 Interprétation . . . 193
6.3 Rôle des règles de transmission . . . 196
6.3.1 Directionalité de la transmission . . . 196
6.3.2 Hypothèses de transmission réalistes . . . 199
6.3.3 Modèles de transmission stylisés . . . 201
6.3.4 Résultats des simulations . . . 202

tel-00464316, version 1 - 16 Mar 2010

7 Cascades informationnelles **207**
7.1 Cascades informationnelles et diffusion . . . 208
7.1.1 Jeu de données empirique . . . 208
7.1.2 Distance attentionnelle . . . 209
7.1.3 Sous-graphes de diffusion . . . 212
7.2 Relais d'information et attention . . . 214
7.2.1 Premières transmissions . . . 214
7.2.2 Petits-Fils . . . 215
7.2.3 Secondes transmissions et attention . . . 216
7.3 Courts-circuits informationnels . . . 217
7.3.1 Secondes transmissions et edge range . . . 217
7.3.2 Effets couplés . . . 219
7.3.3 Conclusion . . . 220

Conclusion **225**

A Liste des termes associés à la blogosphère politique française **231**
A.1 Liste des 190 syntagmes utilisés définir le bagage sémantique des blogs politiques français : . . . 231
A.2 Liste des termes associés à la blogosphère politique américaine . . . 232

B Corpus de termes des bases des domaines scientifiques explorés **233**
B.1 Systèmes complexes . . . 233
B.2 La métaphore réseau en biologie . . . 235
B.3 Développement durable - CAB . . . 238

C Requête développement durable **242**

Bibliographie **243**

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Introduction

AVEC le développement des nouvelles technologies de l'information et de la communication les processus de production et de partage de la connaissance subissent une transition majeure. Le savoir se retrouve maintenant archivé de façon systématique au sein de bibliothèques digitales "librement" consultables et largement distribuées à travers la planète, il fait système au sens où les nouvelles briques de connaissance s'inscrivent au sein d'une topologie de la connaissance qu'elles contribuent à modifier de façon infinitésimale. Cet espace, pourtant bâti par l'activité cognitive des hommes, est doté d'une dynamique largement autonome, au sens où elle donne lieu à l'émergence d'un certain nombre de structures endogènes pérennes. Parallèlement, les régimes de production et de régulation de la connaissance s'inscrivent dans des architectures toujours plus résiliaires : les individus réunis au sein de *communautés de savoirs* manipulent, échangent, et produisent ces connaissances au sein de réseaux évolutifs mêlant des acteurs hétérogènes se déployant dans des espaces sociaux interconnectés. Certaines structures émergentes stables caractérisent également ces réseaux sociaux. En ce sens, la sphère sociale peut, comme la sphère culturelle, être décrite comme un *système complexe autonome*.

Couplage structurel des systèmes social et sémantique. Néanmoins, les dynamiques qui animent les communautés de savoirs ne sont jamais purement sociales ou purement culturelles. ***Notre thèse consiste à affirmer que les dynamiques respectives de la sphère sociale et de la sphère culturelle ne peuvent être saisies sans envisager les couplages structurels qu'elles entretiennent à différents niveaux***.

Nous parlerons dans cette thèse de *systèmes sociosémantiques* mettant en jeu d'une part des acteurs en interaction au sein d'un réseau social, et d'autre part des entités sémantiques dotées d'une organisation propre et décrivant une topologie de la connaissance à part entière. En suivant l'hypothèse d'une autonomie à un certain niveau des dimensions sociale et sémantique, nous tâcherons donc, non pas, de traiter ces dynamiques sociosémantiques d'un seul bloc, mais, de proposer une modélisation aussi symétrique que possible des deux dimensions sociale et sémantique. Dans un second temps, nous évaluerons la façon dont le réseau sociosémantique, qui lie les individus aux entités sémantiques qu'ils mobilisent, produit des couplages plus ou moins forts entre, d'une part, les dynamiques individuelles

et la distribution des connaissances au sein du réseau social, et, symétriquement, les dynamiques conceptuelles et la distribution des individus au sein du réseau sémantique. À un plus haut niveau, les structures du réseau social, que décrivent les motifs plus ou moins stables émergeant des comportements individuels, font écho aux structures organisant le réseau sémantique.

Observer les traces de la vie sociale. La numérisation des contenus et des interactions entre individus, dont la production et l'accessibilité s'est largement accélérée ces dernières années par les technologies de l'Internet, offre la possibilité d'une observation quasi-complète de ces dynamiques sociales et sémantiques. La digitalisation des traces (aussi bien sociales que sémantiques) produites par les interactions des individus avec leur environnement permet de suivre les acteurs au plus près et de mettre en œuvre un observatoire *in-vivo* des dynamiques sociocognitives du monde social. Néanmoins, se doter d'outils d'observation des traces digitales textuelles et d'une méthodologie permettant de gérer de larges corpus de données hétérogènes afin de collecter des données sur les dynamiques sociales et sémantiques dans ces espaces ne constitue qu'une première étape vers la compréhension de ces systèmes sociosémantiques. Peut-on mettre en évidence des lois ou, *a minima*, des régularités qui dirigent ces dynamiques sociales et sémantiques, quels motifs émergents remarquables structurent ces réseaux, etc ? À cette étape de collecte, il faut donc ajouter le développement de méthodes quantitatives de reconstruction et de représentation dans le perspective de réaliser une *reconstruction phénoménologique* formelle des dynamiques multi-échelles qui caractérisent ces systèmes sociosémantiques sans en réduire la complexité intrinsèque.

Protocole d'observation. Cette thèse s'attache donc à décrire, analyser et reconstruire des communautés de savoirs mettant en jeu des individus qui interagissent au sein d'un réseau social et qui produisent des contenus liés à un domaine spécifique, ces contenus s'inscrivant dans un espace sémantique doté d'une structure et d'une dynamique propre. Elle s'appuie essentiellement sur deux objets empiriques. En premier lieu, les *communautés scientifiques* mettent en rapport, d'une part, des chercheurs en interaction au sein de différents réseaux (collaborations, citations, etc.) et, d'autre part, des réseaux de proximité entre concepts construit à partir de leurs publications. Notre objectif consistera dans un premier temps à reconstruire la structure du réseau sémantique produit par la juxtaposition de l'ensemble de la production d'une communauté scientifique donnée afin de proposer un représentation pertinente de son organisation et de ses dynamiques. Dans un second temps nous examinerons la façon dont ces structures sémantiques sont susceptibles d'agir sur les comportements individuels des chercheurs dans la perspective plus large d'une co-évolution entre les dynamiques sociales et sémantiques qui animent ces communautés. Le second objet abordé est un espace public nou-

tel-00464316, version 1 - 16 Mar 2010

veau et très réactif : les blogosphères politiques propres à chaque pays (nous nous concentrerons précisément sur les deux blogosphères politiques française et américaine). Il est à nouveau interprété de façon duale : d'une part, comme un système d'interactions sociales multiples entre blogueurs, d'autre part, comme un système de production de contenus distribué sur un ensemble de sources. Le substrat essentiellement numérique des dynamiques d'interaction et de production de contenu au sein de ce territoire virtuel en fait un lieu d'expérimentation particulièrement pertinent pour étudier les processus de diffusion d'information dont il est le siège. L'observation *in-vivo* des dynamiques de ces communautés nous permettra également de caractériser certains couplages existant ente les comportements micros (liés aussi bien aux nouvelles interactions entre individus, qu'à la production de nouveaux contenus) et les propriétés structurelles observées dans les réseaux sociaux, sémantiques, et socio-sémantiques.

Plan des parties à venir. L'objectif de cette thèse est donc de développer des outils d'analyse quantitative des dynamiques et des processus en jeu au sein de ces communautés de savoirs en adoptant une double perspective sociale et sémantique. Après avoir, dans la partie I, défini ce que recouvre le concept de communauté de savoirs et présenté le formalisme dans lequel nous envisageons leur analyse, nous découpons les interrogations portées sur ces systèmes socio-sémantiques en deux grandes catégories : (i) morphogenèse de réseau (partie II) : mise en évidence de motifs structurels globaux non triviaux (structures sémantiques ou/et sociales remarquables, détection de champs épistémiques, etc.) couplée à l'étude des comportements locaux (propension d'un acteur à s'attacher préférentiellement à des acteurs présentant certaines propriétés structurelles (par exemple homophilie sociale ou sémantique)) susceptibles de faire émerger ces motifs mésoscopiques ou macroscopiques, puis phylogenèse : évaluation des dynamiques de certaines structures émergentes du système (ii) étude des dynamiques des processus à l'œuvre sur ces réseaux, en particulier la diffusion (partie III) : à un niveau global, en appréhendant les motifs d'influence existant entre des groupes de sources de contenus, puis au niveau du réseau social, en interrogeant le rôle de la topologie du réseau sur la vitesse de circulation d'une information sur l'ensemble du réseau, et enfin au niveau individuel, en mettant en évidence l'importance de certaines propriétés structurelles locales sur la dynamique de cascades informationnelles empiriques.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Première partie

Suivre les communautés de savoirs

CETTE première partie vise à définir notre objet d'étude : les *communautés de savoirs*, à travers l'examen de deux cas d'étude : les communautés scientifiques et les communautés de blogueurs politiques. Sans avoir prétention à décrire de façon analytique ce qu'entend recouvrir l'appellation communauté de savoirs, nous tâcherons, dans le premier chapitre, de saisir certaines de leurs caractéristiques saillantes en dégageant les propriétés communes et les spécificités respectives de ces deux terrains empiriques. Nous essaierons également de remettre la notion de communauté de savoirs en perspective par rapport aux différents modèles de communauté existants. Nos communautés de savoirs seront formalisées comme des *réseaux épistémiques* qui permettent d'articuler dans un même cadre les dimensions sociale et sémantique. Cette formalisation nous accompagnera tout au long de cette thèse.

Dans le second chapitre, nous reviendrons plus en détail sur les avantages d'une modélisation de ces systèmes sous forme de réseaux. Nous insisterons notamment sur la possibilité d'intégrer au sein d'un même cadre les dynamiques microscopiques des individus, et les motifs macroscopiques émergents. Dans ce même chapitre, nous présenterons et justifierons notre méthodologie de recueil longitudinal des traces de l'activité des communautés de savoirs afin de reconstruire aussi fidèlement que possible la phénoménologie de leurs dynamiques.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

CHAPITRE 1

Communautés de savoirs

Sommaire

1.1 Propriétés des communautés de savoirs 13

1.1.1 Communautés de blogueurs politiques et communautés scientifiques : des communautés de savoirs 14

1.1.2 Hétérogénéité des engagements et frontières des communautés de savoirs 17

1.1.3 Un système socio-cognitif distribué 20

1.1.4 Un mode de coordination stigmergique 23

1.2 Des réseaux sociosémantiques 24

1.2.1 Dualité socio-sémantique 24

1.2.2 Réseaux épistémiques 28

1.2.3 Formalisme 31

Cette thèse s'appuie essentiellement sur deux types d'objets empiriques : d'une part, les communautés scientifiques dont les membres interagissent notamment au travers de collaborations ou de citations pour produire de la connaissance, d'autre part, des communautés de blogueurs politiques formant un espace public virtuel cohérent au sein de la blogosphère. Nous entreprendrons dans ce chapitre de donner une définition et d'énumérer les propriétés de ce que nous appellerons des communautés de savoirs à travers les spécificités de nos deux cas d'étude. Nous introduirons également le formalisme des réseaux épistémiques qui est un cadre d'analyse privilégié pour appréhender la dualité socio-sémantique intrinsèque des communautés de savoirs.

1.1 Propriétés des communautés de savoirs

Nous partons de nos deux cas d'étude empiriques : *communautés de blogueurs politiques* et *communautés scientifiques* pour illustrer les propriétés que recouvrent la notion de communauté de savoirs tout en tâchant de contextualiser cette notion par rapport aux différents modèles connexes de communautés existants.

1.1.1 Communautés de blogueurs politiques et communautés scientifiques : des communautés de savoirs

Nous nous sommes intéressés à deux terrains empiriques durant cette thèse : des communautés scientifiques dont les membres interagissent notamment au travers de collaborations pour produire des énoncés scientifiques au sein de publications, ainsi que des communautés de blogueurs "citoyens" formant une arène de discussion virtuelle : la blogosphère politique qui se définit à la fois comme un territoire d'expression, de prise de position sur la chose publique et comme un espace d'échange et de mises en relation entre blogueurs.

Ces exemples de communautés de savoirs ne sont pas uniques, on peut également citer : l'encyclopédie ouverte que constitue Wikipedia, ou nombre de wikis ouverts au public, les communautés de logiciel libre, etc. Il faut d'ailleurs noter que la plupart des des communautés de savoirs existantes ont récemment vu le jour essentiellement grâce à Internet, et qu'elles se déploient généralement sur le Web. D'après Conein (2004), "les technologies cognitives à base Internet et les infrastructures Open Source faciliteraient en même temps l'accroissement de la connaissance et la coordination sociale" au sein de systèmes qu'il appelle, par ailleurs, des réseaux socio-cognitifs.

Nos deux exemples présentent *a priori* de nombreuses différences, autant du point de vue des normes qui en régissent l'activité, des modalités de coopération qui les animent ou encore plus simplement de la nature des énoncés qui y circulent. Sans souci d'exhaustivité, et en écho à la liste qui vient d'être donnée, on peut détailler deux différences majeures qui apparaissent entre nos deux systèmes. En ce qui concerne les *régimes de régulation de la production de contenus*, un blogueur peut publier ses billets librement sans avoir à recevoir l'aval de ses pairs, *a contrario*, de nombreux processus d'expertise des articles scientifiques en régulent la publication. L'exigence d'originalité et d'innovation au-delà de l'état de l'art, qui prévaut dans les communautés scientifiques, perd de sa pertinence dans la blogosphère politique, où les emprunts, copies, voire plagiats parfois revendiqués sont fréquents. Ces différences induisent un certain nombre de conséquences quant à la vitesse d'évolution des contenus produits : les communautés de blogueurs politiques traitent souvent de thématiques faisant écho à l'actualité et sont susceptibles d'y réagir rapidement alors que les communautés scientifiques ont une évolution qui pourra paraître plus lente et qui se trouve parfois déphasée par rapport aux événements extérieurs.

Concernant, les *modalités de coopération*, lorsque des chercheurs collaborent, leur travail donne lieu à un texte (généralement sous la forme d'un article) qu'ils cosigneront, la coopération ou plutôt la coordination dans la blogosphère n'implique pas le même degré d'engagement entre agents, elle se manifeste plutôt comme la participation d'un ensemble d'intervenants à une "conversation" ou à un "forum" sans frontière précise. Ainsi, si les billets de blogs se répondent les uns aux autres,

tel-00464316, version 1 - 16 Mar 2010

chaque blog édite néanmoins seul ses billets, il n'y a pas de partage du statut d'auteur (à moins de considérer les commentaires d'un billet comme une forme d'excroissance de celui-ci) ; ce qui est produit collectivement, c'est l'ensemble de la conversation, et non pas les contenus de base que constituent les billets (dont le pendant serait l'article dans le monde scientifique). Ce mode de coordination propre aux blogueurs peut par contre être rapproché d'un modèle de construction incrémentale de la connaissance scientifique lorsque des auteurs proposent un nouvel énoncé en s'appuyant sur, et en faisant référence à, un certain nombre d'articles déjà publiés.

Malgré ces différences majeures, nous faisons l'hypothèse qu'il est pertinent de les appréhender dans un même cadre. Communautés scientifiques et communautés de blogueurs politiques sont toutes deux modélisables comme des systèmes sociosémantiques de production de contenus et d'interactions distribués.

Nous définissons les communautés de savoirs comme un ensemble d'individus en interaction au sein d'un réseau social et qui manipulent, échangent et produisent de l'information liée à un domaine d'intérêt ou d'expertise partagé. À ce titre, les communautés de savoirs se caractérisent conjointement comme un système social complexe *et* comme un ensemble de sources de contenus distribuées. Ainsi concernant nos deux cas d'étude, on retrouve dans chacun des réseaux d'interaction entre agents (par exemple réseaux de collaboration ou de citation entre auteurs dans le premier cas, réseaux de citation ou de commentaire entre blogueurs dans le second cas) tissés au gré de leurs actions locales. Les contenus sont également produits de façon distribuée (par chaque chercheur ou groupe de chercheurs, ou par les blogueurs ou leurs commentateurs) et *s'inscrivent* dans des espaces situés.

La notion de communauté de savoirs entretient une forte proximité avec les concepts de *communauté d'intérêt* et de *communauté de pratique*. La communauté d'intérêt se définit comme un groupe composé d'individus qui partagent un intérêt, des expériences ou des préoccupations communes. Elle avait été prophétisée par Licklider and Taylor (1968) comme le mode de regroupement que les individus privilégieraient à l'ère informatique, imaginant par là-même le concept à venir de communauté virtuelle :

> "[...] geographically separated members, sometimes grouped in small clusters and sometimes working individually. They will be communities not of common location, but of common interest."

Les communautés d'intérêt ont pour objectif essentiel de disséminer une information *a priori* diffuse. Ainsi, les membres d'une communautés d'intérêt tels qu'un ensemble de patients victimes d'une maladie (comme la sclérose en plaques chez Dillenbourg et al. (2003)) sont à la recherche d'informations sur les symptômes de leur maladie ou les traitements existants que d'autres membres pourraient détenir. L'objectif n'est pas de produire une connaissance nouvelle ou d'élaborer un savoir

tel-00464316, version 1 - 16 Mar 2010

collectif mais d'instaurer une structure d'échanges entre ses membres permettant la circulation de connaissances ou de témoignages.

La communauté de pratique est tout aussi informelle et spontanée que la communauté d'intérêt. Elle a néanmoins des objectifs distincts et décrit des groupes d'une autre nature. La notion de communauté de pratique est attribuée à Etienne Wenger (Wenger and Snyder, 2000), elle peut désigner une grande variété de situation d'apprentissage collectif (Amin and Roberts, 2006) figurant "des groupes d'individus qui partagent un intérêt, un ensemble de problèmes, ou encore une même passion à propos d'un sujet, [et qui] approfondissent leur connaissance et leur expertise en interagissant continûment" [1](Wenger and Snyder, 2000).

Les analyses des processus de production de connaissances dans les communautés de pratique se sont essentiellement portées sur la nature sociale du processus d'apprentissage. Si nous nous référons aux premières études de terrain, l'apprentissage de marins novices (Hutchins, 1996) ou d'apprentis bouchers (Lave and Wenger, 1991) au sein de communautés de pratique tend vers un objectif cognitf relativement bien balisé (par exemple, dans le premier cas, être à même de réaliser des manœuvres navales complexes). La communauté de pratique est généralement bâtie sur un système de connaissances *unique* (quel que soit la nature de ces connaissances) dans lequel les nouveaux venus, partant de la périphérie de la communauté, doivent cheminer en son centre au gré des expériences qu'ils partagent avec les membres plus anciens et plus expérimentés de la communauté. C'est donc un mode d'organisation essentiellement centré sur les compétences de ses membres. Une communauté de pratique est certes généralement créée spontanément à un moment donné et en réponse à un problème apparu au sein d'une organisation, mais une fois ses objectifs atteints (*i.e.* le problème résolu), elle se dissout aussi rapidement qu'elle est apparue.

Si les communautés de savoirs s'appuient à la fois sur les notions d'un domaine d'intérêt partagé entre ses membres et d'un apprentissage situé capitalisant sur le collectif, qui sont des caractéristiques premières des communautés d'intérêt et de pratique, elles se différencient néanmoins de ces dernières dans le sens où elles sont réellement orientées vers la création de nouvelles connaissances. *A contrario*, les communautés d'intérêt sont focalisées sur la dissémination d'informations existantes même si distribuées sur un ensemble d'agents tandis que les communautés de pratique, dans leur acception originale [2], tendent vers la réalisa-

1. "[Communities of practice are] groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise by interacting on an ongoing basis."

2. La littérature sur les COPs (communautés de pratique) a été foisonnante ces dernières années, ce foisonnement ayant également donné lieu à une diversification définitionnelle, les COPs recouvrent maintenant une grande diversité de structures caractérisées par différents rapports à la connaissance et à l'apprentissage dans l'action , ainsi Amin and Roberts (2006, 2008) proposent une typologie rendant compte de cette variabilité en distinguant entre les modes de relation au savoir du type : épistémique, créatif, professionnel ou virtuel.

tel-00464316, version 1 - 16 Mar 2010

tion d'une activité bien définie (Fischer, 2001; Cohendet et al., 2003) dont l'accomplissement marque la fin de la communauté.

Le terme de communauté épistémique est également parfois mobilisé pour définir notamment les communautés de création de nouvelles connaissances (Amin and Roberts, 2008; Cowan et al., 2000), telles que les communautés de logiciel libre (Conein, 2003). Cet usage nous semble néanmoins problématique pour deux raisons. En premier lieu, la définition la plus commune des communautés épistémiques introduite par Haas (1992) englobe des groupements d'experts, souvent transnationaux, articulés en réseau et partageant un ensemble de croyances ou de normes et dont l'autorité dans leur domaine d'expertise les rend légitimes pour formuler des recommandations à destination des décideurs publics. Une communauté scientifique en son entier ne saurait donc pas relever, *a priori*, de cette définition restreinte. Mais même en évacuant les rapports qu'entretiennent les communautés épistémiques à la décision politique, cette définition nous paraît encore trop restrictive pour englober l'ensemble des communautés rentrant dans le "répertoire" des communautés de savoirs. Ainsi, il paraîtrait abusif de dire d'une communautés de blogueurs politiques (fût-elle en partie peuplée de journalistes professionnels, ou de militants politiques) qu'elle est composée "d'experts". Dans une moindre mesure, un telle notion de communauté interdit le renouvellement de ses membres par l'intégration progressive de "novices", nouveaux entrants en position d'apprentissage (comme l'étudiant rentrant dans une communauté scientifique tout en faisant l'apprentissage de ses problématiques principales et de ses principaux acteurs).

C'est pour l'ensemble des ces raisons que nous préférons parler de communauté de savoirs dans le but d'y intégrer une population de membres plus hétérogène sans lui attribuer un objectif ou un pouvoir de prescription *a priori*. Il pourrait paraître prétentieux de parler de production de savoirs quand on envisage un ensemble de programmeurs travaillant au développement d'un logiciel libre ou de blogueurs chroniquant l'actualité politique. La notion de "savoirs" doit naturellement être entendue dans son acception la plus large possible ; le dénominateur commun de l'ensemble de ces systèmes résidant dans l'existence de processus cognitifs donnant lieu à la production d'artefacts culturels variés (lignes de codes, argumentaire en faveur d'une baisse des impôts ou énoncé scientifique) distribués, discutés et négociés au sein d'un réseau social.

1.1.2 Hétérogénéité des engagements et frontières des communautés de savoirs

Malgré le contexte local et situé de la production de contenus au sein des communautés de savoirs, une forme de continuité perdure dans l'ensemble du système et lui confère sa cohérence. Cette cohérence d'ensemble n'est pas seulement thématique : les membres d'une communauté de savoirs *s'engagent* dans celle-ci

tel-00464316, version 1 - 16 Mar 2010

à travers leur activité éditoriale mais aussi à travers les liens qu'il entretiennent avec d'autres membres. Les régularités inhérentes à ces mises en relation stabilisent les représentations mentales de ses membres et les motifs relationnels qui les réunissent.

Ainsi, un scientifique spécialisé dans la recherche sur les cellules souches a un sentiment d'appartenance à la communautés scientifique associée, non seulement, parce que ses problématiques sont partagées par la communauté mais également au travers des *mises en relations* qui le lient à divers éléments de la communauté. Ces liens sont variés : ses publications peuvent faire référence à (ou être citées par) d'autres travaux issus de la communauté, il peut collaborer avec des chercheurs de la communauté, ou encore, entrer régulièrement en interaction avec d'autres membres ou se tenir au courant des dernières avancées de sa spécialité lors des conférences du domaine. De la même façon, un blogueur politique, dont on considère fréquemment qu'il s'adonne à une pratique "individualiste" (voire narcissique), s'engage dans une activité de chronique de la vie politique non seulement en réagissant à l'actualité politique mais aussi en "se liant" par une grande variété de modalités de mise en contact à une "conversation" plus large à laquelle participent d'autres membres de la communauté des blogueurs politiques. Ces engagements sont toujours locaux et limités. Néanmoins, ce sont ces mises en relation qui permettent aux membres d'une communauté de savoirs de développer un sentiment d'appartenance, d'en apprendre les règles et les valeurs et de définir le rôle qu'ils y jouent, tout en donnant corps, dans un même temps (et on retrouve alors le double processus d'individuation individuelle et collective de Simondon que nous détaillerons dans le chapitre suivant), à la communauté en question.

Pour résumer, nous somme en présence d'un système d'action collective guidé par des actions purement individuelles. il y a construction d'une forme de structure collective (sans que quiconque n'est nécessairement cette construction comme objectif ni même conscience de l'opération de construction en cours), chaque nouvel élément (et ce quel que soit sa nature : une personne ou un texte) s'appuyant sur certaines entités antérieures pour se définir au sein de la structure globale et la modifier dans un même mouvement.

Naturellement, il existe autant de formes d'engagement dans ces communautés de savoirs que de membres et ces engagements peuvent revêtir des intensités diverses. Ainsi, certaines personnes auront une relation épisodique avec un communauté scientifique uniquement au travers de collaborations avec un membre actif de la dite communauté, tandis que d'autres auront un haut degré d'implication et s'investiront par exemple dans l'organisation de conférences ou dans une activité d'édition pour une revue spécialisée dédiée à la communauté. L'appartenance à une communauté de savoirs n'est pas exclusive ; les individus participent *a priori* à un large nombre de *cercles sociaux* pour reprendre la terminologie de Simmel. C'est l'intersection des diverses "affiliations" (Simmel, 1955) d'un individu qui en définit l'identité (Breiger, 1974; Kadushin, 1966). C'est également pour cette

tel-00464316, version 1 - 16 Mar 2010

raison que les communautés de savoirs présentent des degrés contrastés d'engagement de la part de leurs membres, dont l'investissement dépend notamment de leur disponibilité et donc du temps et des ressources déjà investis par ailleurs dans d'autres activités et éventuellement dans d'autres communautés de savoirs (que l'on songe simplement à un scientifique impliqué dans la vie de plusieurs communautés scientifiques, ou un blogueur technophile animant également un blog politique). L'investissement dépend également du degré d'expertise des membres. Ainsi, les blogs politiques peuvent aussi bien être animés par des personnalités politiques nationales ou locales, des journalistes (travaillant pour le compte de leur journal ou entretenant un blog indépendant), des conseillers en communication, ou encore des citoyens "ordinaires" (Flichy, 2000), chacun ayant des motivations différentes vis-à-vis de son activité de blogueur. Il en va de même dans les communautés scientifiques ou dans d'autres communautés de savoirs, les ingénieurs se mêlent aux chercheurs, eux-mêmes plus ou moins expérimentés vis-à-vis d'une spécialité de recherche. Dans le cas de la Wikipedia, Bryant et al. (2005) décrivent les mécanismes de transformation progressive des modes de participation des contributeurs de la communauté.

L'hétérogénéité dans l'engagement de leurs membres (que nous illustrerons quantitativement à travers la notion d'activité dans le cas la blogosphère politique américaine dans le chapitre 3) que les communautés de savoirs autorisent et encouragent est également une conséquence de la perméabilité de leurs frontières. Les membres d'une communauté de savoirs peuvent être renouvelés à un rythme très important sans que ne soit nécessairement mise en péril la dynamique collective de la communauté. Nous reviendrons, également dans le chapitre 3, sur la stabilité de certaines structures caractérisant ces communautés. L'ouverture des frontières de la blogosphère politique paraît même constitutive de son essence. Cardon and Delaunay-Teterel (2006), dans sa typologie des blogs, définit la blogosphère politique par son rapport à l'espace public. Ces blogs se caractérisent (contrairement aux autres types de blogs : journaux intimes, blogs de "tribus", ou blog de "fans") par un énoncé détaché de la personne de l'énonciateur afin d'afficher "les marques de distanciation indispensables à la prolifération d'une opinion, d'un jugement, ou d'une critique dans un espace public". Dès lors, la blogosphère politique sort d'un logique de fermeture communautaire afin d'élargir le débat public aux non-professionnels de la politique en un immense forum de discussion.

Au delà de la variété et de la mobilité des acteurs de ces communautés, le domaine d'intérêt qui réunit ses membres peut être sujet à glissements. Contrairement aux cas des communautés de pratique classiques ou des communautés d'intérêt , la nature même de ce qui réunit les individus est soumise à une négociation collective permanente. Ainsi, la définition de ce que doit être une encyclopédie collaborative comme Wikipedia ainsi que les processus de régulation y étant observés ont largement évolué au cours de son développement (Viegas et al., 2007b). Les problématiques autour desquelles se retrouve dans une conférence annuelle

tel-00464316, version 1 - 16 Mar 2010

l'ensemble des participants d'une communauté scientifique donnée subissent des évolutions en fonction d'avancées récentes ou de la découverte de nouvelles "aires d'ignorance" (Mulkay, 1976).

1.1.3 Un système socio-cognitif distribué

L'une des caractéristiques premières des communautés de savoirs est que leur dynamique n'est dirigée par aucun organe de centralisation définissant pour l'ensemble un objectif commun ou des règles de fonctionnement internes. Les interactions ainsi que les contenus produits sont entièrement distribués sur l'ensemble du système. Ainsi c'est l'ensemble des comportements individuels rassemblés qui définit la dynamique de la communauté de savoirs en son entier.

La somme des contenus produits par une communauté de savoirs est distribuée sur l'ensemble des agents du système (billets écrits par un blogueur ou articles rédigés par un chercheur). Ces agents sont liés à travers un réseau social composé de la somme des mises en relation de chacun. Ces mises en relation peuvent mettre en jeu des dyades (un blogueur écrit un commentaire sur le blog d'un autre blogueur) ou, plus largement, un ensemble d'agents réunis au sein d'un hyperlien, lorsque des chercheurs co-publient un article.

Morris and der Veer Martens (2008) s'appuie sur Storer (1966) pour remarquer que le système social de la Science diffère de celui d'autres formes d'organisations, formelles ou informelles, au sens où, passée la phase de recrutement, les rôles occupés par leurs membres sont bien moins hiérarchisés et différenciés qu'ils ne le sont classiquement. Ce dernier parle de *canaux d'implication* qui sont essentiellement dictés par la pertinence et la complémentarité des approches d'un point de vue purement cognitif [3]. Les relations entre blogueurs politiques suivent le même trait, et apparaissent comme largement supportées par des critères cognitifs ; le chapitre 3 illustre d'ailleurs de façon quantitative l'importance d'un critère d'alignement cognitif entre deux blogs sur la propension d'un couple de blogs à entrer en relation.

Les dynamiques à l'œuvre dans les communautés de savoirs sont donc d'abord le fait des comportements individuels de ses membres ; néanmoins, cet aspect distribué n'empêche pas l'existence d'une structuration d'ordre macroscopique particulière. Ainsi, aussi bien au sein de la blogosphère que dans le monde académique, on observe des effets de hiérarchie, relatifs à l'autorité dont sont pourvus certains agents dans le système. Cette hiérarchie émergente peut induire un certain nombre d'asymétries dans le système, les ressources disponibles par chacun ne sont pas nécessairement les mêmes (attention exercée sur la communauté, opportunités de collaborations, etc.), mais cette distribution hétérogène des ressources

3. Naturellement les institutions d'appartenance des chercheurs ainsi qu'un certain nombre de règles tacites sont autant d'effets exogènes qui appellent à nuancer cette vision idéalisée.

tel-00464316, version 1 - 16 Mar 2010

est en permanence renégociée et est construite par cristallisation progressive d'actions d'ordre exclusivement microscopique.

Les communautés de savoirs se caractérisent également par des phénomènes de structuration à un niveau que l'on pourrait qualifier de mésoscopique regroupant des "sous-groupes" de participants attachés à des "sous-domaines" de spécialité de la communauté. Ces agrégats locaux se distinguent par les sujets débattus en leur sein (à titre d'exemple, on a observé au sein de la blogosphère politique française un ensemble de sites spécialisés dans le droit, ou encore dans les affaires européennes ; dans les sciences, ces sous-groupes renvoient à divers niveaux de spécialisation de la communauté scientifique envisagée) mais également par la forte densité de relations liant ses participants, formant une micro-communauté au sein de la communauté au sens large. Ainsi l'espace peut se structurer selon des critères hiérarchiques, mais aussi selon des sensibilités plus ou moins partagées entre membres. Un exemple classique est celui mis en évidence par Adamic and Glance (2005a) de découpage de la blogosphère politique américaine selon des critères partisans - ce découpage politique s'accompagnant d'une forme d'isolation *structurelle* des clans démocrates et républicains. Ces questions de structuration dans la blogosphère politique sont primordiales car elles peuvent mettre à mal les conditions d'un débat démocratique controversé (Goldman, 2008). Nous verrons néanmoins comment la variété des modalités de mises en relation autorisées peut permettre de contrecarrer la balkanisation de l'espace et des opinions publics (Flichy, 2008).

Cette structuration de l'espace ne se limite pas à la seule dimension sociale. Les contenus qui circulent dans les communautés de savoirs s'inscrivent dans des agrégats sémantiques cohérents dont l'articulation forme des espaces cognitifs à part entière. On peut parler d'une organisation de la connaissance qui peut également structurer les dynamiques individuelles. Nous explorerons plus précisément dans le chapitre 4 la nature et le rôle de ces agrégats dans l'organisation d'une communauté de savoirs.

Le caractère distribué et local des communautés de savoirs est également permis grâce aux technologies de communication et de gestion de la connaissance offertes par Internet (Origgi, 2006). En effet, la plupart de ces systèmes sont intimement liés avec le medium sur lequel ils se déploient. Les interactions entre des individus dispersés à travers le globe sont rendues possible grâce aux technologies de communication (comme le mail, ou les outils du web 2.0). L'archive universelle que constitue le Web autorise l'accessibilité et la visibilité des contenus produits par chacun, tandis que les moteurs de recherche garantissent la possibilité de naviguer à travers ces contenus.

Les communautés scientifiques ont sans doute été parmi les premières à avoir bénéficié des technologies du Web, à la fois comme espace d'échange, par le mail ou les listes de discussion, et de partage, d'abord grâce à la multiplication des bases de données de publications (de plus en plus souvent ouvertes et gratuites)

tel-00464316, version 1 - 16 Mar 2010

et maintenant à travers la mise à disposition croissante de données expérimentales communes (telles que le proposent les grands équipements déployés en physique atomique et sub-atomique ou la biologie systémique par exemple). Les blogs sont également caractéristiques d'un usage des outils d'Internet croisant dispositif d'auto-publication *et* outil de communication collective (Cardon and Delaunay-Teterel, 2006).

Malgré l'utilisation que font les blogueurs et les scientifiques des technologies d'Internet, il faut néanmoins apporter des éléments de nuance qui distinguent très fortement l'usage qui est fait de ces outils dans les deux cas. Dans le cas des scientifiques, les échanges sur Internet apparaissent comme le prolongement d'une communication qui a pris et peut prendre d'autre voies par ailleurs. Ainsi, une rencontre initiée au détour d'un café durant une conférence se poursuivra par un échange de mails. Les technologies d'Internet apparaissent alors comme une voie de communication parmi d'autres. La communication par production de liens hypertextes au sein de billets de blogs ou par le biais de commentaires, est d'une autre nature. Un blogueur n'a *a priori* d'autre choix pour communiquer avec un autre blogueur que d'utiliser l'espace du blog [4] (que ce soit le sien, en rédigeant un billet en réaction au billet d'un blog tiers, ou en commentant directement le billet d'un blog tiers). En bref, la communication inter-individuelle est nécessairement médiatisée par l'espace du blog qui est un espace ouvert et public. Ainsi une singularité de la communication dans la blogosphère réside dans le fait que le public du blog constitue l'invité permanent de toutes les formes d'interaction entre deux individus, une certaine forme "d'exhibitionnisme" étant de rigueur dans ce dispositif panoptique [5].

Ainsi, toute intervention sur la blogosphère (production d'un billet ou d'un commentaire) est *exposée* au regard public des internautes [6]. Dans le cas des communautés scientifiques, seuls les traces de certaines interactions prolongées entre chercheurs apparaissent publiquement telles qu'une liste de coauteurs ayant collaboré à l'élaboration d'une publication. Même si l'on peut reconstruire, grâce à ces traces (cas de figure très particuliers mis à part, on peut raisonnablement supposer que les coauteurs d'un papier ont été amenés à échanger les uns avec les autres), les réseaux sociaux qui structurent ces communautés, la majorité des échanges entre chercheurs restent néanmoins privés.

4. Il est parfois possible de prendre contact directement avec le blogueur lorsque son adresse mail est publique mais ces pratiques sont sans doute minoritaires.

5. Cette exhibitionnisme doit néanmoins également être nuancé. Même si la technologie du blog garantit une forme publicisation de l'ensemble des contenus produits, tous les blogs ne sont pas nécessairement dotés de la même visibilité, comme on l'a vu il y a également des effets de hiérarchisation entre sources, l'hétérogénéité des engagements de chacun se dédoublant d'une forme d'hétérogénéité dans la visibilité ou l'audience des sites(Hindman et al., 2003).

6. Mieux, les billets des blogs peuvent être commentés par tout un chacun, que le commentateur possède un blog ou non.

tel-00464316, version 1 - 16 Mar 2010

1.1.4 Un mode de coordination stigmergique

La discussion sur les usages des technologies de communication employées dans nos deux types de communautés de savoirs nous amène à discuter une autre propriété essentielle de ces systèmes, le caractère *stigmergique* de la coordination entre agents. En effet, dans les deux cas qui nous occupent, l'intégralité des contenus produits (que la publication soit soumise à un certain nombre de règles dans le cas de la production scientifique, ou qu'elle soit entièrement libre avec les outils d'auto-publication que constituent les blogs) sont publiquement consultables sur le Web (parfois de façon payante concernant l'accès aux bases de données de publications)[7].

L'accessibilité de ces contenus est essentielle pour la coordination entre les participants des communautés de savoirs. Ainsi, un chercheur peut grâce à cette littérature archivée et structurée par des liens de citation, saisir l'état de la connaissance à un instant donné dans son domaine de spécialité[8]. La disponibilité et l'accessibilité de ces textes lui permet de "converser" à distance (spatialement et temporellement) avec d'autres chercheurs. À la manière de la communication entre les fourmis, signalant leur passage par tel ou tel lieu en laissant une trace dans leur environnement sous la forme d'un dépôt de phéromones, on dit de cette coordination qu'elle est stigmergique (Heylighen et al., 2004) car les articles sont *publiquement* consultables et que cette visibilité permet un mode d'interaction *indirecte* et *diachronique* entre les chercheurs. Cette propriété est également vérifiée dans le cas de la blogosphère. Les contenus publiés par un blogueur sont en permanence consultables car automatiquement archivés par la plate-forme d'édition. Un blogueur peut alors aisément mobiliser en s'y référant un billet rédigé parfois plusieurs mois plus tôt.

Dans les deux cas, l'accessibilité des contenus déjà produits joue un rôle fondamental dans le production de nouveaux contenus. Il ne s'agit pas uniquement pour un acteur de "s'informer sur l'état de l'art" d'un domaine, mais d'être à même d'y contribuer en proposant sur un mode plutôt incrémental (en sciences), plutôt polémique (dans la blogosphère), de nouveaux contenus qui participent de la création d'une "œuvre collective". Il y a coordination stigmergique à chaque fois qu'un agent lie un de ses textes à un autre texte de façon à prolonger une "discussion" dont les frontières spatiales et temporelles restent ouvertes.

7. Nous parlons ici d'une certaine portion de la production scientifique, l'intégralité (incluant notamment la recherche privée) n'étant pas publique, ou ne l'est qu'à à un certain degré (pour une discussion de la notion de la Science comme bien public, voir (Callon, 1994)).

8. La possibilité de mettre en relation des textes épars peut mener aux plus grandes "révolutions" scientifiques comme l'illustre Latour (1987) à propos de la du travail de recomposition du "puzzle épars des textes adultérés" de l'almageste de Ptolémée réalisé par Copernic.

tel-00464316, version 1 - 16 Mar 2010

1.2 Des réseaux sociosémantiques

Nous décrivons et justifions dans cette section le formalisme des *réseaux sociosémantiques* ou *réseaux épistémiques* que nous adoptons pour rendre compte des dynamiques duales des communautés de savoirs.

1.2.1 Dualité socio-sémantique

Du fait de la nature distribuée des interactions inter-individuelles, nous nous sommes naturellement orientés vers le formalisme offert par *l'analyse des réseaux sociaux* pour donner un cadre à l'étude de nos communautés de savoirs. L'analyse des réseaux sociaux propose un certain nombre de méthodes formelles appliquées à des données de terrains (Freeman, 2004; Wasserman and Faust, 1994) ; sa popularité s'est accrue récemment notamment grâce à certains apports venus de la physique statistique (Barabási and Albert, 1999; Strogatz, 2001), mais c'est avant tout à cause des hypothèses sur lesquelles elle se fonde, à savoir la primauté des interactions dans la compréhensions du monde social, que nous nous tournons vers ce type d'approche.

L'analyse traditionnelle des réseaux sociaux s'appuie essentiellement sur des hypothèses structuralistes. Elle place au centre de l'analyse non pas les individus et leurs attributs ou des catégories censées agir les individus en fonction d'une appartenance de classe, mais des relations inter-personnelles et les structures relationnelles entre nœuds comme l'unité première de compréhension et de modélisation du social.

Comme le résument Smith-Doerr and Powell (2005) en s'appuyant sur une analyse antérieure de Granovetter (1985),

> "In contrast to deterministic cultural (oversocialized) accounts, network analysis afforded room for human agency, and in contrast to individualist, atomized (undersocialized) approaches, networks emphasized structure and constraint [...] Network studies offered a middle ground, a third way [...]"

Le trait saillant de l'approche réseau est donc de prendre comme unité première d'analyse les relations inter-individuelles. Dans une revue récente, Marin and Wellman (2010) affirment :

> "Social network analysis takes as its starting point the premise that social life is created primarily and most importantly by relations and the patterns formed by these relations."

Comme le citent Marin and Wellman (2010), la primauté du lien social sur les entités trouve notamment ses racines dans la sociologie de Simmel :

> "The significance of these interactions among men lies in the fact that it is because of them that the individuals, in whom these driving

tel-00464316, version 1 - 16 Mar 2010

> impulses and purposes are lodged, form a unity, that is, a society. For unity in the empirical sense of the word is nothing but the interaction of elements. An organic body is a unity because its organs maintain a more intimate exchange of their energies with each other than with any other organism ; a state is a unity because its citizens show similar mutual effects." (Simmel, 1971)

Cette troisième voie comme l'appellent Smith-Doerr et Powell est célébrée par la plupart pour sa fidélité à un impératif anti-catégorique. Ainsi, White et al. (1976) écrivent : "We would like the reader to entertain instead the idea that the presently existing largely categorical descriptions of social structure have no solid theoretical grounding ; furthermore, network concepts may provide the only way to construct a theory of social structure" ou encore dans le même article : "[f]irst social structure is regularities in the patterns of relations among concrete entities ; it is *not* a harmony among abstract norms and values or a classification of concrete entities by their attributes. Second, to describe social structure, we must aggregate these regularities in a fashion consistent with their inherent nature as networks".

Néanmoins cette perspective qui postule que les structures sociales relèveraient exclusivement des motifs relationnels entre individus est critiquée par certains auteurs aussi bien au sein de la communauté proche de l'analyse des réseaux sociaux qu'à l'extérieur pour sa tendance à oblitérer les autres dimensions participant de l'action sociale. Ces auteurs reprochent à l'analyse des réseaux sociaux de négliger les opérations cognitives portées par les agents qui animent les systèmes étudiés ainsi que les autres formes que revêt la vie sociale au-delà de ses manifestations purement structurelles.

C'est la critique principale qu'Emirbayer and Goodwin (1994) adressent à l'analyse de réseaux lorsqu'ils écrivent :

> "We have also suggested, however, that despite its powerful conceptualization of social structure, network analysis as it has been developed to date has inadequately theorized the causal role of ideals, beliefs, and values and of the actors that strive to realize them ; as a result, it has neglected the cultural and symbolic moment in the very determination of social action."

Ainsi dans un article interrogeant les pré-supposés théoriques de l'analyse de réseau vis-à-vis des notions de structure social, d'"agency" et de culture et à travers une revue critique des principaux travaux d'analyse de réseaux ayant une ambition socio-historique, Emirbayer and Goodwin (1994) renvoient dos à dos les trois approches qui, selon eux, animent la recherche sur les réseaux sociaux : l'approche déterministe (qui néglige les causes pouvant émaner des croyances ou des valeurs des individus), l'orientation instrumentale (qui envisage les acteurs comme de simples agents maximisant une fonction d'utilité en niant leur capacité d'innovation), et enfin, la perspective constructiviste (la plus apte à leur sens

tel-00464316, version 1 - 16 Mar 2010

à conceptualiser les transformations potentielles de l'action sociale induites par les idiomes culturels ou les engagements normatifs mais qui échoue néanmoins à interroger pleinement la co-évolution existant entre les deux dimensions).

Archer (1996) a introduit les notions d'*upward conflation* et de *downward conflation* pour désigner deux positions extrêmes tendant à fonder les régimes d'interprétation sur des bases relevant uniquement de la structure sociale ou uniquement du domaine culturel. Selon Emirbayer and Goodwin (1994), les deux approches font l'erreur de réduire la dualité entre dimensions sociale et sémantique à un simple épiphénomène alors même qu'elle sont toutes les deux dotées d'autonomie et se co-déterminent :

> "[...]agency and structure interpenetrate with one another in all individual units (as well as complexes) of empirical action, and that all historical processes are structured at least in part by cultural and political discourses, as well as by networks of social interaction."

Même si l'approche privilégiée dans cet article est celle de la compréhension des processus historiques, la critique s'étend naturellement à l'analyse de tout système social que l'on tenterait de réduire à sa seule composante relationnelle, en faisant l'impasse sur un point capital : l'autonomie des structures culturelles qui prévalent dans le systèmes contraignent et rendent possibles (ou énactent) l'action sociale aussi bien que ne le feraient des structures relationnelles pré-existantes. En bref, "[network analysis] has negected the cultural and symbolic moment in the very determination of social action." ibid.

Callon (2006) développe un argumentaire convergent à l'encontre de l'analyse des réseaux sociaux (voir également Callon and Ferrary (2006)). Même s'il souligne l'attrait épistémologique d'une analyse *ex post* des catégories s'opposant à une approche adossant un certain nombre d'attributs pré-définis aux individus, il critique l'analyse des réseaux sociaux dans sa tendance à donner de l'espace social une représentation unifiée composée d'acteurs homogènes dotés généralement d'un seul régime d'action et limitant de par la même les modes de "mise en relation". Un des résultats principaux de la *sociologie de la traduction* consiste justement à introduire dans le cadre d'analyse des dynamiques liées aux sciences et aux techniques une pluralité d'acteurs hétérogènes (les *actants* regroupant aussi bien des humains que des non-humains), et une diversité de régimes d'action (Boltanski and Thévenot, 1991) mettant en relations ces acteurs (selon des modalités de natures différentes : scientifique, économique, politique, etc.). Selon (Callon, 2006), la composition des réseaux égocentrés hétérogènes [9] ainsi formés offrent les conditions d'existence *d'espaces* sociaux locaux et non nécessairement unifiés.

Le cadre théorique et méthodologique développé par l'analyse des réseau so-

9. Ces réseaux sont même doublement hétérogènes. D'une part, leurs constituants peuvent aussi bien désigner des individus, des collectifs, que des non-humains, d'autre part, ces entités sont reliées les unes aux autres par une grande variété de mises en relation.

tel-00464316, version 1 - 16 Mar 2010

ciaux rend selon lui impossible la prise en compte de cette complexité. C'est la raison pour laquelle il lui préfère une approche fondée sur la cartographie de réseaux égocentrés hétérogènes recomposés. Seule la *représentation* de ces réseaux est, toujours selon (Callon, 2006), à même de rendre compte de la diversité des régimes d'action, des acteurs, et des espaces. Elle constitue de plus une interface pratique pour partager avec les acteurs du système envisagé le travail du sociologue.

La critique portant sur l'hypothèse d'un espace social unifié s'appuie essentiellement sur une certaine forme d'analyse des réseaux sociaux que l'on pourrait qualifier de positionnelle [10] et dont l'objet consiste à définir la position d'un nœud du réseau en comparant son environnement relationnel à celui de l'ensemble des autres nœuds. Cette approche vise alors à définir le rôle, le statut, ou la position d'un individu au sein d'un réseau en fonction de l'ensemble des nœuds et des liens du réseau et en ignorant la pluralité potentielle des espaces sociaux. Il est vrai que cette méthode est susceptible d'assigner un même rôle ou une même position à un ensemble de nœuds parfaitement déconnectés et suppose donc en premier lieu que ceux-ci appartiennent à un espace social partagé.

Pour autant, elle ne saurait recouvrir l'ensemble des modes d'analyse des réseaux existants. Emirbayer and Goodwin (1994) opposent ainsi l'analyse relationnelle (ou de type "social cohesion") à l'analyse positionnelle. L'analyse relationnelle s'appuie directement sur les connections directes ou indirectes reliant les acteurs entre eux. La pertinence de ces deux grands modes d'analyse des réseaux a ainsi, notamment, été comparée à propos des méthodes de division des réseaux en sous-groupes, ainsi qu'à propos du rôle respectif des positions et des voisinages locaux pour saisir la dynamique d'une diffusion (Burt, 1978,?). Parmi ces approches relationnelles, on retrouve des notions telles que les liens faibles (Granovetter, 1973) ou les trous structurels (Burt, 2004). D'autres s'attachent, encore, à repérer dans le réseau certains motifs (Holland and Leinhardt, 1976) ou groupes cohésifs (Moody and White, 2003). L'ensemble de ces approches invite à la prise en compte de groupes sociaux locaux différenciés au sein d'une structure plus large. Ces exemples nous éloignent d'une analyse des réseaux sociaux cantonnée à un espace social unique et semblent nous rapprocher un peu plus d'un objectif de représentation des agencements sociotechniques (Callon, 2006) : "[...] the most faithful *representation of the configurations* resulting from the eternally unfinished work of composition and combination of egocentric networks."

La question de l'unicité des régimes d'action peut être partiellement résolue par la prise en compte de multi-réseaux intégrant l'ensemble des modalités de mises en relation entre acteurs. Au sein de la théorie des réseaux sociaux, le lien ou la relation entre nœuds peut être de différents types. Borgatti et al. (2009) dé-

10. Plus précisément Callon (2006) prend comme exemple la notion d'équivalence structurelle (Lorrain and White, 1971) qui permet de regrouper dans un même "bloc" (White et al., 1976) les nœuds du réseau ayant un rôle structurellement équivalent vis-à-vis du réseau (même si ceux-ci peuvent ne pas être voisins des mêmes nœuds).

tel-00464316, version 1 - 16 Mar 2010

nombrent quatre grands types de relations généralement abordées par l'analyse de réseaux sociaux : relations de similarité (deux entités du réseau sont connectées lorsqu'elles partagent un certain nombre d'attributs exogènes), relations sociales (on entend par relations sociales tout type de relation continue entre individus : ex : amitié, collaboration, connaissance, etc.), interactions (dans cette catégorie se retrouvent l'ensemble des liens associés à un comportement ou à une action spécifique et *a priori* délimité dans le temps tels qu'apporter un conseil, citer le travail d'autrui, etc.) et enfin, les relations de type flux (elles incluent l'ensemble des réseaux d'échange de type flux financiers, ou flux informationnels) [11]

La diversité de ces catégories laisse entrevoir la variété des modes de mise en relation possibles entre entités que l'analyse des réseaux sociaux a comme ambition d'étudier. La prise en compte d'une pluralité de modes de mise en relation au sein d'un même réseau (qui forme alors un multi-réseau), permet d'intégrer l'ensemble des processus pertinents en fonction d'une problématique donnée. Les outils d'analyse des multi-réseaux peuvent sembler encore assez primitifs, mais il n'en demeure pas moins possible, dans une perspective comparative, d'étudier la façon dont la dynamique de deux réseaux est guidée par tel ou tel processus (nous nous attacheront ainsi, chapitre 3, à comparer la dynamique d'un réseau de commentaire et d'un réseau de citation), ou dans une perspective de compréhension des couplages entre réseaux, de rechercher les corrélations existant entre les observables d'un réseau vis-à-vis d'un autre réseau (par exemple, afin de mieux comprendre les dynamiques des cascades informationnelles, chapitre 7, nous analyserons les corrélations entre la structure des réseaux formés par les flux d'information et certaines propriétés du réseau d'interaction social sous-jacent).

Concernant la dernière critique adressée au manque d'hétérogénéité des acteurs engagés dans la modélisation du réseau ("endowing human and non-human actors with larger and richer competencies"), elle rejoint en partie les objections soulevées par Emirbayer and Goodwin (1994). Nous tentons d'y répondre en introduisant dans la section suivante un formalisme original : *les réseaux épistémiques* déjà introduit par Roth (2006, 2008a,b) qui étend la perspective classique couverte par l'analyse des réseaux sociaux en rajoutant une dimension sémantique au profil relationnel des acteurs (qui sont alors dotés d'un profil de relation social *et* sémantique), cette dimension sémantique étant elle-même dotée d'une dynamique autonome qui co-évolue avec le réseau social décrivant les relations entre acteurs.

1.2.2 Réseaux épistémiques

Au-délà des critiques opposées à l'analyse de réseaux sociaux discutées ci-dessus, le contexte actuel d'une "science des réseaux complexes" dans laquelle des

11. Naturellement, ces différents types de relation peuvent s'avérer redondants. Par exemple, on s'attend à ce qu'un réseau d'email entre étudiants d'une même université, qui est un réseau d'interaction, soit fortement corrélé au réseau d'amitié entre ces mêmes étudiants.

tel-00464316, version 1 - 16 Mar 2010

méthodes génériques issues des sciences physiques (et principalement la physique statistique) ou de l'informatique des graphes ont été massivement appliquées, rend à nos yeux encore plus pertinente la nécessité de rediscuter les hypothèse sous-tendant ces modélisations. Le développement et l'utilisation d'outils sur de très grands réseaux d'interaction indifférenciés (réseaux d'infrastructure, métaboliques, sociaux ou de communication sont souvent simplement considérés comme autant d'instance d'un même objet), a certes permis d'interroger de façon transversale certains invariants propres à l'ensemble des réseaux réels (comme la distribution de degré de ces réseaux de terrain qui semble être généralement hétérogène), mais cette recherche d'universalité s'est également parfois réalisée en ignorant la spécificité des systèmes étudiés et des questionnements qu'ils appellent.

Ainsi, dans le cas qui nous occupe, l'entreprise de modélisation et de caractérisation de la morphogenèse des communautés de savoirs à partir des seules observables liées au réseau social nous semble problématique ; il paraît judicieux de s'interroger sur le rôle des contenus sémantiques dans le développement des comportement individuels : la distribution des représentations culturelles étant un facteur d'influence capital des dynamiques sociales. Les réseaux de savoirs sont en effet le théâtre d'interactions à fort contenu sémantique, soulignant ainsi la pertinence de l'utilisation de motifs structurels ou de mécanismes de mise en relation qui ne soient pas d'ordre strictement sociaux (au sens où ils engageraient uniquement des humains avec des humains). Plus exactement, les nouvelles interactions produites au sein d'un système complexe *sociosémantique* sont, au moins partiellement, déterminées par la structure des interactions passées et par les affinités conceptuelles entre agents.

L'étude des communautés de savoirs requiert donc une double attention aux dynamiques sociales et aux dynamiques de production de savoirs ou de diffusion d'information qui les traversent. Ainsi le formalisme des réseaux épistémiques permet l'intégration d'une dimension sémantique autonome couplée aux réseaux d'interaction entre individus classique en analyse de réseaux sociaux [12].

La dimension sémantique n'est pas intégrée comme une variable exogène supplémentaire adossée aux caractéristiques individuelles mais bien comme une source de dynamiques à part entière traduisant aussi bien la distribution des "connaissances" sur l'ensemble des sources que l'organisation des connaissances spécifiques à la communauté de savoirs étudiée.

Comme l'appelle de ses vœux Emirbayer and Goodwin (1994), cette dimension sémantique est, *a priori*, dotée de sa propre autonomie : " These symbolic formations [Cultural discourses, narratives, and idioms] have emergent properties - an internal logic and organization of their own - that require that they be conceptua-

12. Le réseau social peut, en fonction de l'objet d'étude et de la question posée sur le système, être constitué de liens d'attribution d'autorité (par exemple, un lien de citation dans une publication scientifique), de liens d'interactions créés par exemple dans la blogosphère via les commentaires, ou de différents types de liens si l'on envisage des multi-réseaux.

tel-00464316, version 1 - 16 Mar 2010

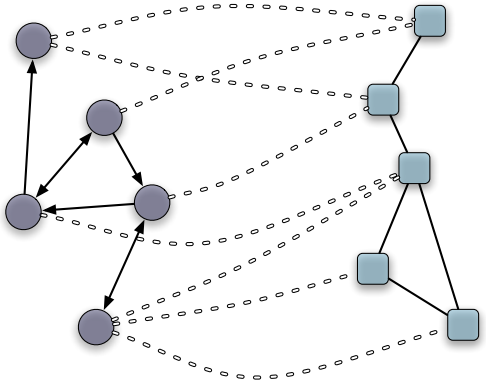


FIGURE 1.1: Réseau épistémique réunissant deux types d'entités, un ensemble d'*acteurs* (matérialisés par des cercles), en interaction au sein d'un réseau social (liens dirigés), un ensemble d'*attributs sémantiques* (matérialisés par des carrés), reliés par des relations de similarité (lignes continues non dirigées) au sein d'un réseau sémantique, et enfin un réseau biparti socio-sémantique (lignes en pointillés) caractérisant l'usage des entités sémantiques par les acteurs

lized as "cultural structures"". Elle n'est donc pas réductible à une "sur-couche" qui doterait les agents de certains attributs exogènes, elle forme au contraire une dimension autonome à part entière des dynamiques humaines qui animent la communauté de savoirs. Bien que dotée d'autonomie, elle est co-déterminée par le réseau social et le co-détermine en retour ; on peut dire que les dimensions sociales et sémantiques sont en *co-évolution* l'une avec l'autre.

Le schéma figure 1.1 offre une vision schématique à la fois de notre approche théorique et de notre formalisation pratique des communautés de savoirs. Un réseau cognitif est en fait la combinaison de trois réseaux : *un réseau social* qui traduit les interactions inter-individuelles, un *réseau socio-sémantique* qui décrit la façon dont les entités sémantiques sont mobilisées par les acteurs, et enfin un *réseau sémantique* rendant compte de la structuration de l'ensemble des entités sémantiques. Cette modélisation peut être interprétée comme un premier pas vers une formalisation de l'ensemble des "mises en relation" (Latour, 2005; Callon, 2006) d'entités humaines ou non-humaines possibles.

Ces différents réseaux sont en co-évolution les uns avec les autres. Ainsi, une de nos hypothèses est que les interactions entre individus décrites par le réseau social sont fortement liées aux attributs cognitifs de ces agents que l'on modélise grâce au réseau socio-sémantique biparti reliant les agents aux entités sémantiques du système (nous préciserons plus tard la façon dont ces entités sont définies). De façon symétrique au réseau social, les entités sémantiques sont également reliées en fonction de leur similarité au sein d'un réseau sémantique. Ce réseau est construit en fonction des fréquences de co-occurrences des couples d'entités sémantiques observées dans l'ensemble du système.

Concernant la blogopshère, par exemple, un blogueur qui rédige un billet en

tel-00464316, version 1 - 16 Mar 2010

faisant référence au billet d'un blog tiers, met en relation son propre blog avec le blog cité au sein d'un réseau social. Les occurrences de termes au sein de ce billet définissent un ensemble de relations liant le blog aux contenus qu'il produit (réseau socio-sémantique). Enfin, l'ensemble des billets produits par l'ensemble des blogueurs de la portion de blogosphère considérée forme un corpus à partir duquel on peut construire un réseau de similarité qui relie les entités sémantiques fréquemment mises en relations au sein d'un même billet et qui représente la structure des "connaissances" mobilisées dans la communauté de blogueurs considérée (réseau sémantique).

L'ensemble de ces réseaux traduisent donc tous, à leur façon, des opérations de co-présence ou de co-apparition d'entités hétérogènes. La perspective adoptée se situe donc dans la lignée des approches de la sociologie de la traduction pronant la prise en compte de réseaux hybrides. Néanmoins nous avons à faire à différents types de cooccurrences (correspondant à autant de modes de mise en relation différents) en fonction des types d'entités rencontrées (individu/individu, individu/entité sémantique, entité sémantique/entité sémantique). C'est pourquoi, même si notre objectif est bien de saisir la façon dont les trois réseaux se co-déterminent, nous respecterons les spécificités de ces réseaux en proposant un cadre d'analyse distinct pour chacun d'entre eux. Ce choix méthodologique est également dicté par l'hypothèse que nous suivons d'autonomie, à un certain niveau d'émergence, des structures sociales et sémantiques. À nouveau , les trois réseaux ne sont nullement réductibles les uns aux autres [13].

Pratiquement, nous définissons les entités sémantiques qui caractérisent l'état cognitif de chaque agent directement à partir des traces textuelles qu'ils produisent. Plus précisément, nous ferons l'hypothèse qu'un ensemble limité de termes et d'expressions que nous appellerons *concepts* par commodité suffit à définir le bagage sémantique des individus. Nous détaillerons par la suite les hypothèses qui président au choix de ces ensembles.

1.2.3 Formalisme

Formellement, nous définissons les trois réseaux qui composent *le* réseau épistémique de la façon suivante (les trois réseaux sont représentés figure 3.1) : On note $G^{\mathbf{S}} = (\mathbf{S}, R^{\mathbf{S}})$ le réseau social où $\mathbf{S}$ désigne l'ensemble des agents et $R^{\mathbf{S}} \subset \mathbf{S} \times \mathbf{S}$ représente la liste des liens du réseau social : un lien (éventuellement orienté) $l = (s, s') \in R^{\mathbf{S}}$ signifie que s est relié à s'.

13. Il faut noter ici que la "projection" du réseau socio-sémantique biparti liant les agents à leurs attributs sémantiques sur la dimension sociale (ce qui correspondrait à relier les agents partageant les mêmes attributs sémantiques) est *a priori* différente du réseau social (construit par l'agrégation des interactions ou des relations entre individus indépendamment de toute production de contenu), symétriquement, le réseau sémantique est distinct de la projection du réseau socio-sémantique sur la dimension sémantique

tel-00464316, version 1 - 16 Mar 2010

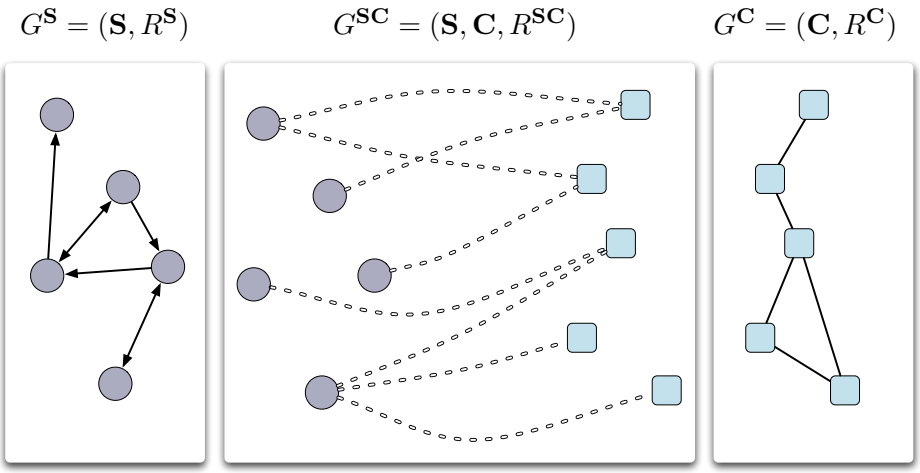


FIGURE 1.2: Les trois réseaux distincts : social ($G^{\mathbf{S}}$), socio-sémantique ($G^{\mathbf{SC}}$) et sémantique ($G^{\mathbf{C}}$) dont est composé un réseau épistémique. L'ensemble des agents (désignés pas $\mathbf{S}$) sont figurés par des cercles gris, tandis que l'ensemble concepts (désignés pas $\mathbf{C}$) sont figurés par des carrés bleus. Les trois réseaux sont tous bien différents les uns des autres, ainsi le réseau social n'est pas une projection du réseau socio-sémantique.

Du côté sémantique, nous appellons concepts, les termes et expressions qui peuvent composer le bagage sémantique des agents, $\mathbf{C}$ désigne l'ensemble des concepts. Le réseau socio-sémantique $G^{\mathbf{SC}} = (\mathbf{S}, \mathbf{C}, R^{\mathbf{SC}})$ est constitué des agents $\mathbf{S}$, de l'ensemble des concepts $\mathbf{C}$ et des liens entre ces éléments $R^{\mathbf{SC}}$, soit l'usage de certains concepts par certains agents : dans sa forme la plus simple, un agent est lié aux concepts qu'il mentionne. Ainsi, un lien $l = (s, c) \in R^{\mathbf{SC}} \subset \mathbf{S} \times \mathbf{C}$ indique que s mentionne c.

Ces deux réseaux sont de nature différente, le réseau social est un réseau monoparti reliant uniquement des acteurs les uns avec les autres, le réseau socio-sémantique est un réseau biparti (appelé également réseau 2-mode) [14] mélangeant acteurs et concepts ce qui revient pratiquement à doter les agents d'attributs cognitifs.

Enfin, on définit le réseau sémantique $G^{\mathbf{C}}$ de façon similaire au réseau social. $G^{\mathbf{C}} = (\mathbf{C}, R^{\mathbf{C}})$ désigne le réseau sémantique où $\mathbf{C}$ est l'ensemble des concepts et $R^{\mathbf{C}} \subset \mathbf{C} \times \mathbf{C}$ représente la liste des liens du réseau sémantique : un lien $l = (c, c') \in R^{\mathbf{C}}$ signifie que c est relié à c' (nous précisons ultérieurement la façon dont le lien est modélisé dans le réseau sémantique).

14. Un graphe est dit biparti s'il existe une partition de son ensemble de nœuds en deux sous-ensembles S et C telle que chaque lien ait une extrémité dans S et l'autre dans C. Dans notre cas, les ensembles S et C désignent alors respectivement l'ensemble des acteurs et l'ensemble des concepts.

tel-00464316, version 1 - 16 Mar 2010

Résumé du chapitre:

Dans ce premier chapitre nous avons tenté de délimiter notre objet d'étude, *les communautés de savoirs*, de façon inductive à travers l'examen des propriétés de nos deux cas d'étude : blogosphères politiques et communautés scientifiques. Nous nous sommes également efforcés de les situer par rapport aux notions théoriques existantes pour qualifier des collectifs produisant de la connaissance. Nous définissons une communauté de savoirs comme un ensemble d'individus, en interaction au sein d'un réseau social, qui manipulent, échangent et produisent de l'information liée à un domaine d'intérêt ou d'expertise.

Les communautés de savoirs se construisent comme un système sociocognitif distribué : l'ensemble des dynamiques première(sociales ou sémantiques) qu'elles abritent sont, par essence, locales. La coordination entre les éléments d'une communauté est de type stigmergique, ainsi la production de nouveaux contenus s'appuie sur la mise en réseau de textes d'auteurs distincts écrits à différents moments (réseau de citation scientifique ou de billets de blogs). Ses frontières sont par nature ouvertes à la circulation d'entités sociales ou sémantiques qui sont susceptibles, par leur agrégation d'en redéfinir l'identité.

Ces communautés se caractérisent également par la dualité entretenue entre les *dynamiques sociales* qui affectent le réseau de relations interindividuelles et les *dynamiques sémantiques* traduisant la production de contenus. Un cadre d'analyse, celui des réseaux épistémiques a été introduit. Il permet de modéliser une communauté de savoirs comme un ensemble de trois réseaux : le réseau social (liant les individus entrant en interaction), le réseau socio-sémantique (traduisant l'usage des concepts par les individus), et enfin le réseau sémantique (liant les concepts proches). Ce cadre nous semble pertinent et nécessaire pour rendre compte de la dynamique de ces communautés, car il respecte le mode d'organisation décentralisée de ces communautés, et permet d'avoir une approche symétrique vis-à-vis des dynamiques d'interactions sociales et sémantiques.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

CHAPITRE 2

Dynamiques multi-échelles des communautés de savoirs

Sommaire

2.1 Analyse longitudinale des dynamiques des communautés de savoirs 36
2.1.1 Limites de l'approche statique 36
2.1.2 Formalisme dynamique 38
2.2 Articuler les niveaux micro et macro 38
2.2.1 Boucle émergence immergence 38
2.2.2 Individus et réseaux 42
2.3 Observation *in-vivo* des dynamiques 44
2.3.1 Suivre les acteurs à l'ère digitale 44
2.3.2 Reconstuction phénoménologique des traces 46
2.3.3 Des traces textuelles au réseau épistémique 48
2.3.4 Un échantillon de la blosphère politique française 50
2.3.5 Un multi-réseau dynamique 53
2.3.6 Caractérisation sémantique 55
2.3.7 Blogosphère américaine 58
2.4 Une approche par faces 59

Les communautés de savoirs dont nous avons tâché de cerner les propriétés dans le chapitre précédent se saisissent essentiellement comme des objets évolutifs : formation de nouvelles interactions, ou déformation de configurations passées, production de nouveaux contenus, départ de certains acteurs remplacés par de nouveaux arrivants, émergence de "normes", processus d'apprentissage collectif ou de diffusion portés par ces communautés, etc... L'originalité des communautés de savoirs est qu'elles sont par nature largement auto-organisées, leur fonctionnement étant en partie lié à la stratification progressive et auto-entretenue de certaines structures stables émergeant des dynamiques individuelles. Cette dualité micro-macro nécessite donc, pour la compréhension du fonctionnement de ces communautés, d'appréhender aussi bien les dynamiques microscopiques (qui doivent être envisagées en regard de l'espace de contraintes ou de possibles laissé vacant par les motifs et les propriétés de haut-niveau) que macroscopiques (dont la morphogenèse est directement liée aux premières).

La dualité sociosémantique de nos communautés de savoirs s'enrichit donc de cette division par niveaux. Nous devons être attentifs aux dynamiques sociales et sémantiques et à leur couplage, simultanément, à un niveau local, *microscopique*, en suivant l'activité des membres de la communauté et les dynamiques d'usage liées à tel ou tel sujet d'intérêt, et à un niveau *mésoscopique* ou *macroscopique*, en suivant l'évolution des propriétés et motifs émergents du réseau épistémique qui ne peuvent pas être attribués à un simple élément de la communauté mais qui sont tributaires de l'ensemble de ses constituants.

Ce chapitre vise à expliciter le type de couplage dynamique entre niveaux à l'œuvre dans nos communautés de savoirs. Or la compréhension de ce couplage et de la morphogenèse d'ensemble de la communauté requiert un suivi longitudinal des dynamiques individuelles et collectives. A cet effet, nous enrichirons le formalisme des réseaux épistémiques déjà introduit dans le chapitre précédent par l'ajout d'une dimension temporelle. Nous préciserons enfin notre approche empirique, fondée sur un recueil de données longitudinales, qui fonde une sociologie des traces tout en soulignant les limites et défis expérimentaux afférents.

2.1 Analyse longitudinale des dynamiques des communautés de savoirs

Dans cette section nous introduisons le cadre dynamique dans lequel nous appréhendons nos communautés de savoirs. Nous énumérons l'ensemble des avantages que ce cadre confère par rapport à une approche statique.

2.1.1 Limites de l'approche statique

On peut critiquer l'analyse des réseaux sociaux à cause de sa tendance à restreindre son champ d'application à des analyses transversales (appelées également "cross-sectional" dans la littérature) des systèmes, limitées à l'observation de systèmes en un seul point dans le temps, notamment parce que ce type d'étude ne permet d'interroger que partiellement la morphogenèse du réseau et ne permet pas d'en saisir les processus de régulation. Néanmoins, cette limite n'est pas due à des limites théoriques inhérentes mais bien à la difficulté pratique que constitue une entreprise de collecte longitudinale de données. Dans les cas où ces données sont accessibles, les outils de l'analyse de réseaux permettent d'aborder la question de la causalité sans ambiguïté (*a contrario* d'une analyse statique qui ne permet *a priori* que d'inférer des corrélations entre variables). De nombreuses études récentes ont montré l'intérêt d'une perspective dynamique s'appuyant sur des séries temporelles décrivant les états successifs d'un réseau (Gulati, 1995; Padgett and Ansell, 1993; Christakis and Fowler, 2007, 2008; Steglich et al., 2006; Jones and Handcock, 2003; Gotz et al., 2009). La principale innovation à venir tiendra sans doute à la

tel-00464316, version 1 - 16 Mar 2010

granularité temporelle des données empiriques issues de l'observation des systèmes sociaux. Les données collectées dans les communautés en ligne permettent de rendre toujours plus précise cette granularité. Certaines analyse dynamiques récentes s'appuient sur une granularité temporelle de l'ordre du jour (site de réseau social : (Holme et al., 2004)) de l'heure (réseaux de communication par email (Kossinets et al., 2008)) et tendent maintenant vers le temps réel (production de contenus de la blogosphère et des media sociaux (Leskovec et al., 2009)).

Effectuer un suivi temporel de l'activité de nos communautés de savoirs offre trois avantages par rapport à une vision purement statique. L'analyse de données temporelles permet de procéder à un suivi longitudinal des motifs structurant les communauté de savoirs, elle est également une opportunité pour saisir les régularités guidant les dynamiques individuelles, enfin, elle offre la possibilité d'interroger les processus dynamiques, que supportent les communautés de savoirs, tels que la diffusion.

En premier lieu, ces données permettent naturellement d'apprécier la façon dont certains motifs se stabilisent ou se reconfigurent en effectuant un simple suivi longitudinal des variables qui les caractérisent. Les interactions entre deux agents sont-elles entretenues dans le temps ? Est-ce que le nombre de motifs cycliques évolue dans le réseau ? L'usage de certains concepts augmente-t-il dans le temps ?

Au delà d'une analyse longitudinale de certaines observables décrivant l'état du système à différents pas de temps, un jeu de données dynamique doté d'une granularité temporelle suffisante permet également d'interroger les régularités guidant les comportements des entités du système. On peut ainsi chercher les corrélations existant entre les structures du réseau épistémique entre deux pas de temps successifs. Certaines propriétés structurelles d'un nœud ou d'un couple de nœuds (on parle alors de dyade) rendent-elles plus ou moins probable la création d'un nouveau lien s'arrimant à ce nœud ou connectant les nœuds ce cette dyade ? Le suivi des communautés de savoirs dans le temps permet ainsi de saisir les régularités éventuelles dans les comportements individuels qui resteraient inaccessibles si l'on considérait un réseau statique, le chapitre 3 s'attachera notamment à définir la façon dont la probabilité de création de nouveaux liens dans le réseau social ou socio-sémantique, peut être estimée sous la forme d'un attachement préférentiel à certains paramètres structurels du réseau épistémique.

Enfin le suivi dynamique de l'activité des communautés de savoirs offre l'opportunité d'observer les processus dynamiques qui s'y déploient tels que les phénomènes de diffusion (cf. partie III). Ces processus peuvent à nouveau être interprétés comme des phénomènes émergeant des dynamiques microscopiques. Ils se différencient des motifs émergents statiques (tels qu'une distribution de degrés ou une structuration modulaire, qui sont déduits de l'observation du réseau épistémique à un moment donné) par leur nature exclusivement dynamique, au sens où leur description appelle naturellement un cadre dynamique. À la manière de la houle ridant la surface de l'océan que nous sommes intuitivement plus enclin

tel-00464316, version 1 - 16 Mar 2010

à interpréter comme la succession d'ondes que comme le déphasage progressif d'un ensemble des particules d'eau soumises à des oscillations locales, la diffusion d'une information au sein d'un réseau social se conçoit naturellement comme un processus dynamique de propagation d'une information de proche en proche.

2.1.2 Formalisme dynamique

Par extension du formalisme déjà introduit dans le précédent chapitre, nous définissons, dans le cadre dynamique, le réseau social $\mathcal{G}^{\mathbf{S}} = (\mathbf{S}, \mathcal{R}^{\mathbf{S}})$ où $\mathbf{S}$ désigne l'ensemble des agents et $\mathcal{R}^{\mathbf{S}} \subset \mathbf{S} \times \mathbf{S} \times \mathbb{N}$ représente la liste des liens du réseau social : un lien $l = (s, s', t) \in \mathcal{R}^{\mathbf{S}}$ signifie que s se lie à s' au temps t.

Le réseau socio-sémantique dynamique $\mathcal{G}^{\mathbf{SC}}$ est constitué des agents $\mathbf{S}$, de l'ensemble des concepts $\mathbf{C}$ et des liens entre ces éléments $\mathcal{R}^{\mathbf{SC}}$, soit l'usage de certains concepts par les agents à un moment donné. Ainsi, $\mathcal{R}^{\mathbf{SC}} \subset \mathbf{S} \times \mathbf{C} \times \mathbb{N}$, et un lien $l = (s, c, t) \in \mathcal{R}^{\mathbf{SC}}$ indique que s mentionne c au temps t.

Enfin, on définit le réseau sémantique dynamique $\mathcal{G}^{\mathbf{C}}$ de façon similaire au réseau social dynamique, $\mathcal{G}^{\mathbf{C}} = (\mathbf{C}, \mathcal{R}^{\mathbf{C}})$ désigne le réseau sémantique où $\mathbf{C}$ est l'ensemble des concepts et $\mathcal{R}^{\mathbf{C}} \subset \mathbf{C} \times \mathbf{C} \times \mathbb{N}$ représente la liste des liens du réseau sémantique : un lien $l = (c, c', t) \in \mathcal{R}^{\mathbf{C}}$ signifie que c est relié à c' au temps t.

2.2 Articuler les niveaux micro et macro

Dans cette section, nous souhaitons détailler plus avant le cadre multi-niveau dans lequel les dynamiques des communautés de savoirs se déploient. Nous montrerons ensuite comment une modélisation par des réseaux permet d'appréhender ces dynamiques multi-échelles.

2.2.1 Boucle émergence immergence

Le caractère naturellement distribué de la communication et de la production de contenus au sein des communautés de savoirs contraste avec un mode classique de régulation hiérarchique des organisations. Les interactions qui animent ces systèmes qu'elles soient d'ordre social ou sémantique sont essentiellement locales, et nulle entité centralisatrice fonctionnant sur un mode d'organisation vertical ne contrôle l'activité de leurs membres.

Pour autant, ces communautés ne sont pas dénuées de propriétés de haut niveau, que l'on songe aux nombreuses règles et normes, tacites ou non, qui en régulent l'organisation, ou les processus de différenciation des membres que l'on retrouve dans de nombreuses communautés de savoirs : wikis comme la Wikipedia (Bryant et al., 2005; Lih, 2003; Weiss and Moroiu, 2007; Viegas et al., 2007a), communautés de logiciel libre (Conein, 2003). Les communautés scientifiques ne

tel-00464316, version 1 - 16 Mar 2010

dérogent pas à la règle, malgré l'existence d'un contexte institutionnel fort (instituts de recherche, division en départements des universités, laboratoires, institutions de financement, etc.), les dynamiques de production de connaissance sont souvent le fait de collaborations locales mettant en jeu un ensemble d'individus et de ressources (Lazega et al., 2007). C'est bien l'ensemble de ces événements de production de connaissance, aussi locaux soient-ils, qui oriente, collectivement, la dynamique d'une communauté scientifique. De façon similaire, le tissu relationnel observé au sein de la blogosphère présente des motifs caractéristiques d'une structuration par couleur politique (Adamic and Glance, 2005a) ou plus simplement thématique (Uchida et al., 2007). Ces propriétés macroscopiques sont néanmoins organiquement liées à l'activité des individus qui constituent ces communautés, elles *émergent* des dynamiques individuelles distribuées sur l'ensemble du système.

Ces motifs et ces propriétés de haut niveau émergent des dynamiques individuelles mais ne sont pas pour autant inopérants par rapport à ces dernières. En continuité avec l'idée d'un *individualisme méthodologique complexe* (Dupuy, 2004), nous qualifierons d'immergentes les "causes induites" par les dynamiques de haut niveau sur les dynamiques de bas niveau. Dans cette perspective, Il serait vain de chercher lequel des deux niveaux précède l'autre dans une chaîne causale : les deux niveaux se *co-déterminent* l'un l'autre. Une telle modélisation permet également de résoudre le débat entre l'hypothèse d'une détermination du macro par le micro (hypothèse atomiste) ou du micro par le macro (hypothèse holiste) en les rendant compatibles au sein d'une boucle émergence/immergence qui lie de façon ciculaire dynamiques de bas et de haut niveau.

La figure 2.1 schématise cette boucle émergence/immergence en la saisissant à l'intersection des dynamiques de bas et de haut-niveau respectivement décrites par l'évolution des fonctions $x(t)$ et $X(t)$. L'état macro $X(t)$ du système dépend des comportements micros $x(t)$ à travers la fonction d'émergence, tandis que les transformations de l'état micro du système $x(t) \rightarrow x(t+dt)$ sont conditionnées par la fonction d'immergence [1].

Micro et macro-sociologies se rejoignent donc en un même modèle qui intègre simultanément le rôle des motifs de haut niveau dans la détermination des dynamiques individuelles (ces motifs ou catégories ne sont pas données comme une variable exogène immanente, mais émergent des actions individuelles de bas-niveau) et la capacité des acteurs à agir les phénomènes collectifs de haut-niveau. Ainsi, les acteurs sont bien dotés d'un "libre-arbitre", ils ne sont pas des producteurs conscients d'un ordre "supérieur" ("The result of human action but not of human

1. Ce diagramme est de plus commutatif, *i.e.* l'état du macro $X(t+dt)$ du système peut être décrit (voire prédit dans une perspective de reconstruction) à partir de l'état micro du système $x(t)$ soit par la composition de la fonction d'émergence suivie de la fonction guidant la dynamique de haut niveau $x(t) \rightarrow X(t) \rightarrow X(t+dt)$, soit inversement en composant les fonctions dirigeant la dynamique micro puis la fonction d'émergence $x(t) \rightarrow x(t+dt) \rightarrow X(t+dt)$ (Roth, 2006).

tel-00464316, version 1 - 16 Mar 2010

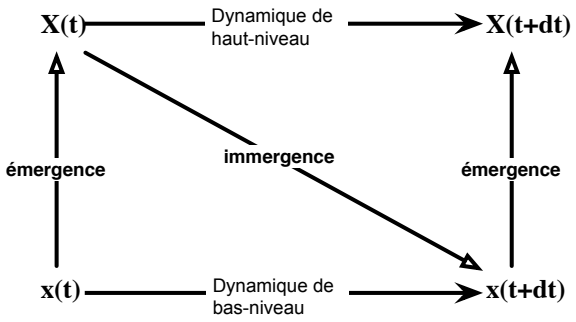


FIGURE 2.1: Schéma de la boucle émergence/immergence, sur un système dont la dynamique se déploie à deux niveaux : micro et macro

design" selon la formule d'Adam Ferguson rappelée par Dupuy (2004)) mais y contribuent par un effet de composition de l'ensemble de leurs actions ne s'exprimant que dans les limites induites par une forme d'ordre social pré-existant. L'existence d'un "point fixe endogène", que chacun contribue à constituer, réconcilie la possibilité d'une conception de la société produite par l'ensemble des individus qui la compose et néanmoins dotée d'une autonomie propre (émergence de propriétés auto-organisées dotées de dynamiques autonomes).

Les niveaux micro et macro que nous nous efforcerons de décrire ne sont donc soumis à aucune forme de prééminence d'un niveau sur l'autre distinguant entre un niveau absolu et un niveau relatif d'explication du social. Notre approche prend néanmoins comme point de départ le niveau micro, les traces que nous suivrons étant essentiellement le fait d'inscriptions locales. L'ambition consiste alors à reconstruire les émergences de haut niveau associées à ces dynamiques. La démarche inductive que nous adoptons rejoint finalement le principe de reconstruction du "grand" par le "petit" prôné par Tarde (1898) et sur lequel Latour (2001) s'appuie pour opposer à une logique d'explication des comportements individuels par quelque macro-structure sociale surplombante une perspective attentive aux "actions élémentaires" des acteurs :

> "[..]au lieu d'expliquer tout par la prétendue imposition d'une loi *d'évolution* qui contraindrait les phénomènes d'ensemble à se reproduire, à se répéter identiquement dans un certain ordre, au lieu d'expliquer le petit par le grand, le *détail* par le *gros*, j'explique les similitudes d'ensemble par l'entassement de petites actions élémentaires, le grand par le petit, le gros par le détail. Cette manière de voir est destinée à produire en sociologie la même transformation qu'a produite en mathématiques l'introduction du calcul infinitésimal " Tarde (1898), p. 21

Il ne s'agit pas, à nouveau, de nier l'existence de règles ou de propriétés macroscopiques propres au système en son entier et s'appliquant sur l'ensemble des

tel-00464316, version 1 - 16 Mar 2010

individus, mais de prendre au sérieux la dialectique de co-construction des dynamiques individuelles et de la "structure" d'ensemble du système. Sewell (1992) part de la dualité intrinsèque de la notion de structure chez Giddens pour réinscrire la genèse des structures dans les actions sociales. Dans l'acception qu'il leur confère, les structures ne sont pas statiques mais dynamiques car sans cesse entretenues par les agents du système :

> "Structures shape people's practices, but it is also people's practices that constitute (and reproduce) structures. In this view of things, human agency and structure, far from being opposed, in fact presuppose each other." Sewell (1992)

Ces niveaux ne relèvent pas non plus d'ontologies nécessairement discordantes. Ce qui rend possible la boucle émergence/immergence, c'est notamment la nature des différentes formes observées à différents niveaux : quel que soit le niveau envisagé, et on a toujours à faire à des "motifs" sociaux, socio-sémantiques ou sémantiques, dont les manifestations sont perceptibles, au moins partiellement par les agents.

Il n'y a donc pas d'émergence transcendantale mais une circulation des motifs à travers toutes les échelles, puisque dans notre cadre sociosémantique, il est, *a priori*, possible de mettre en relation ces motifs hétérogènes les uns avec les autres. Considèrons un exemple simple, tel que la pérennisation d'un champ de recherche émergent à travers la création d'une nouvelle revue ou d'une nouvelle conférence. On peut, si l'on reprend la définition des macro-acteurs de (Callon and Latour, 1981), suivre la façon dont ce champ grandit en enrôlant de nouvelles volontés et en s'adossant à un certains nombre de boîtes noires. On peut également suivre la circulation des macro-acteurs dans le monde des micro-acteurs (Latour, 2006), les macro-acteurs sont par nature "publicisés", car ils s'appuient sur leur rôle de représentants d'entités plus petites. Ainsi la notion d'immergence, qui, en attribuant un pouvoir causal à des entités macroscopiques sur les constituants mêmes qui lui donnent vie, pourrait sembler paradoxale, redevient naturelle pourvu que les entités émergentes circulent dans le monde "d'en-bas" aussi librement qu'un " appel à contribution pour une conférence" diffusé à l'ensemble des abonnés d'une "mailing list".

L'existence de structures sociales et culturelles, émergeant des dynamiques locales, renforce encore la dualité socio-sémantique que nous avions introduit au chapitre précédent (section 1.2.1). Bien que fortement couplées au niveau individuel (ce sont bien les acteurs qui mobilisent des concepts dans le réseau biparti socio-sémantique $G^{\mathbf{SC}}$), les dimensions sociale et sémantique n'en demeurent pas moins largement autonomes au sens où elles se caractérisent toutes deux par la présence de structures pérennes (au moins par rapport à l'échelle de temps caractéristique des dynamiques micros du système) : des *organisations* sociales ou symboliques, qui sont toutes deux à même de conditionner les actions des individus et

tel-00464316, version 1 - 16 Mar 2010

la dynamique à venir du système.

2.2.2 Individus et réseaux

La prise en compte de l'histoire relationnelle d'un acteur au moment de la création de nouvelles interactions est également un moyen de donner une certaine "épaisseur biographique" aux entités. Ainsi, les agents ne recréent pas *ex-nihilo* leur *identité* à chaque nouvelle interaction comme sils se retrouvaient plongés dans une scène goffmanienne durant laquelle leur identité était éternellement redéfinie. La dimension temporelle assure une forme de continuité biographique en offrant un contexte et une antériorité au delà du lieu et du moment précis de l'interaction dans laquelle les acteurs s'engagent [2]. Lahire (1998) cite Montaigne pour illustrer la pluralité de l'identité humaine construite à travers ses expériences :

> "Voilà pourquoi, pour juger d'un homme, il faut suivre longuement et curieusement sa trace." (Montaigne, Essais, Livre second), extrait de (Lahire, 1998), p. 348

Notre objectif est de rester fidèle à cet impératif de suivi longitudinal des actions individuelles. Que celles-ci traduisent un comportement changeant ou une extrême stabilité, nous souhaitons bien reconstruire la phénoménologie de toute la diversité des dynamiques individuelles.

Un des principaux attraits de la modélisation sous forme de réseaux est de proposer un cadre d'analyse qui ne place pas *a priori* l'individu ou ses attributs au centre du dispositif. L'analyse de réseaux sociaux prend comme objet d'étude les relations inter-individuelles. Mais l'attention qu'elle porte à l'interaction sociale ne saurait être réduite à une théorie ou même une approche purement *interactionniste*. L'interaction ou la relation inter-individuelle se trouve toujours plongée dans un contexte plus large, un tissu relationnel. On peut donc dire que la modélisation sous forme de réseau décrit les interactions locales des agents les uns avec les autres tout en intégrant un contexte, méso- ou macroscopique incluant une partie ou l'ensemble des agents et de leurs interactions.

Pour reprendre les termes de Granovetter (1985) l'approche "réseau" propose une alternative aux "conceptions sur-socialisées ou sous-socialisées de l'action humaine en sociologie ou en économie" [3]. Dans ce texte, "l'homo-economicus des économistes" et l'"individu sur-socialisée des sociologues" sont renvoyés dos à dos au motif qu'ils se fondent sur le postulat d'un acteur atomisé, dont les comportements ne sauraient être influencés qu'à la marge par le "contexte social local". La notion "d'embededness" permet de remplacer les approches classiques d'explication des comportements humains soit par un "calcul" strictement utilitaire de

2. S'appuyer sur ces interactions antérieures est d'autant plus aisé, dans le cas qui nous occupe ici, que les inscriptions textuelles accessibles offrent des éléments de contexte très riches sur "l'histoire" des individus. L'individu est ainsi, *de facto*, mis en position de se constituer en sujet réflexif.

3. "over- and undersocialized conceptions of human action in sociology and economics", p. 483

tel-00464316, version 1 - 16 Mar 2010

l'acteur en fonction de ses intérêts personnels définis de manière restrictive, soit par la subordination d'un acteur à un ensemble de normes sociales intériorisées. Une fois les comportements individuels plongés ("enlitement") dans un système empirique de relations sociales, les effets de "traductions" locales, propres à un environnement donné, apparaissent comme les motifs premiers de détermination des comportements des individus.

Cette vision de l'individu comme saisi entre un contexte local et un horizon plus large entretient certaines similitudes avec la notion d'individuation, défendue par Simondon. Dans Simondon (1989), l'individu est par essence en "devenir". L'individu, d'être pré-individuel devient donc un *individu de groupe* à travers une double individuation : *psychique* puis *collective*. Ce dernier écrit

> 'Il n'est donc pas juste de parler de l'influence du groupe sur l'individu ; en fait, le groupe n'est pas fait d'individus réunis en groupe par certains liens, mais d'individus groupés, d'*individus de groupe*. Les individus sont individus de groupe comme le groupe est groupe d'individus. [...] On ne peut pas dire que le groupe exerce une influence sur les individus, car cette action est contemporaine de la vie des individus et n'est pas indépendante de la vie des individus" Simondon (1989)

L'individuation de groupe, tout comme l'individuation psychique ne muselle pas l'autonomie de l'agent qui reste dans un état métastable empli de *potentialités*. Aussi bien que le groupe advient grâce aux individus, les individus sont eux-mêmes transformés par la structure qu'ils ont produite. C'est un phénomène similaire que Gilbert (2003) analyse sur une configuration minimale : deux individus adaptant le rythme de leur pas pour "marcher ensemble".

La sociologie des organisations épouse cette même attention à la dualité entre l'acteur et le système (Friedberg, 1997). L'ordre est sans cesse construit et reconstruit dans le champ des interactions sociales. À nouveau, l'affirmation de l'existence d'un système dans lequel les acteurs développent leur stratégie n'est pas contradictoire avec leur autonomie. Au contraire, "Système et acteur sont co-constitutifs, ils se structurent et restructurent mutuellement". Les acteurs participent donc activement à la régulation du système, ce qui chez Friedberg garantit également une certaine hétérogénéité des ordres locaux :

> "Ici les acteurs n'existent ni dans un vaccum social, ni dans un champ social homogène et unifié, mais bien dans un système social fractionné par l'enchevêtrement désordonné d'une multiplicité de régulations locales hétérogènes - leurs actions et leur rationnalité ne peuvent être analysées que replacées dans ce jeu global." Friedberg (1997)

De façon similaire, Giddens (1981) refuse d'identifier le concept de structure avec celui de grille ou de contrainte. Selon lui, cette définition nous condamnerait

tel-00464316, version 1 - 16 Mar 2010

au dualisme entre théories institutionnelles et théories de l'action (ou plus précisément de "l'agency"). Plutôt que de définir les systèmes sociaux comme des structures, Giddens renverse l'assertion : ce sont "les systèmes sociaux qui ont des *propriétés structurelles*". Au dualisme entre action et institution, il oppose la dualité du concept de structure :

> "[...] the structure is both the medium and outcome of the social practices it recursively organizes."

Structure et actions locales se retrouvent alors entremêlées, simultanément causantes et causées, la notion de récursivité nous invitant à examiner les dynamiques sociales locales de reproduction ou de transformation des structures.

Les réseaux apparaissent donc comme une façon pertinente de modéliser les systèmes sociaux en rendant compte dans un même cadre des actions locales et des structures sociales de plus haut niveau dans lesquelles elles se déploient, pourvu que cette modélisation permette simultanément : (i) de suivre fidèlement les actions individuelles, et plus précisément les dynamiques de mise en relation (ii) de décrire les ordres locaux émergents des dynamiques individuelles. L'importance de la place des processus dynamiques (qu'ils rendent compte des comportements individuels ou de la dynamique des structures de haut-niveau) appelle à l'intégration d'une dimension temporelle à notre formalisme.

2.3 Observation *in-vivo* des dynamiques

Un parti-pris important de cette thèse tient également à l'approche *empirique* de l'analyse des dynamiques des communautés de savoirs que nous adoptons. Ce choix méthodologique est notamment dû à l'impératif que nous nous fixons de restituer les dynamiques des entités qui constituent les communautés de savoirs dans leur contexte d'action locale afin de ne pas imposer de cadre d'analyse ou de catégories *a priori*[4]. Nous présentons également dans cette section la stratégie de collecte que nous adoptons depuis la collecte des traces digitales sur Internet jusqu'à la reconstruction phénoménologique de ces traces.

2.3.1 Suivre les acteurs à l'ère digitale

La multiplication des échanges sur Internet et des traces digitales qui en résultent est une opportunité sans précédent pour "suivre les acteurs à la trace". Pour prendre l'exemple de la blogosphère, l'activité des blogueurs laisse des marqueurs textuels au cœur même des blogs sous la forme d'un billet archivé ou d'un

4. Au delà des modalités d'analyse, cet impératif peut être problématique au moment de la délimitation pratique des frontières des systèmes considérés (selon quel critère peut-on décréter que telle ou telle entité appartient à une communauté de savoirs donnée ou non), on verra néanmoins par la suite comment cette délimitation peut-être conçue en limitant autant que possible la prise en compte de critères exogènes.

tel-00464316, version 1 - 16 Mar 2010

ensemble de commentaires. Dans ce cas, la trace laissée par l'action de l'agent s'apparente à l'action en tant que telle. La technologie impose ici à l'agent comme unique moyen d'expression et de communication avec son environnement d'y apposer une trace textuelle. Le caractère panoptique du medium Internet qui s'appuie sur l'archivage quasi-systématique des contenus (historiques des éditions dans les wikis (Viegas et al., 2007a), archives des mailing-listes de communautés de développeurs (Dorat et al., 2007) ou enfin, plus classiquement, archives en ligne de publications scientifiques (Newman, 2004b)) ouvre de nouvelles perspectives pour les sciences sociales. Comme l'affirme Latour (2007), "[les sciences sociales] ont enfin accès à des masses de données du même ordre de grandeur que celles de leurs grandes sœurs, les sciences de la nature." [5]

C'est grâce à cette visibilité et au caractère public des échanges sur Internet que l'on peut parler d'observation *in-vivo* des dynamiques sociales et sémantiques dans ces espaces. Pour reprendre l'analyse de Latour and Woolgar (1986) les manipulations opérées dans le cadre du laboratoire sont ici directement assurées par le medium même de communication qui fonctionne comme un instrument d'enregistrement systématique des traces laissées par les activités humaines. Un autre avantage de ce mode d'interrogation du social est qu'il garantit une "discrétion" quasiment absolue vis-à-vis de l'objet d'étude. Les traces étant produites dans le cadre de l'action, on peut alors également faire l'économie des questionnements méthodologiques traditionnels qui accompagnent habituellement les enquêtes de terrain. Cette disponibilité permet de lire sous un jour nouveau l'introduction du premier chapitre des *Lois de l'imitation* dans laquelle Tarde écrit :

> "En matière sociale, on a sous la main, par un privilège exceptionnel, les causes véritables, les actes individuels dont les faits sont faits, ce qui est absolument soustrait à nos regards en toute autre matière."Tarde (1890), p. 20

Les "actes sociaux" s'écrivent maintenant par écrans interposés et sont gravés dans des mémoires en silicium, ce qui facilite considérablement leur accumulation (avant de s'adonner, selon le programme de sociologie quantitative de Tarde (1890), au travail de repérage des ressemblances nécessaire pour "nombrer" et "mesurer" les faits sociaux)

Ces traces constituent donc un véritable laboratoire d'expérimentation sociologique et ethnographique permettant d'étudier les dynamiques sociales et cognitives à des échelles inédites (Shalizi, 2007) :

> "The precise forces that mould our subjectivities and the precise characters that furnish our imaginations are all open to inquiries by the social sciences. It is as if the inner workings of private worlds have been

5. "[Social Sciences] can finally have access to masses of data that are of the same order of magnitude as that of their older sisters, the natural sciences".

tel-00464316, version 1 - 16 Mar 2010

> pried open because their inputs and outputs have become thoroughly traceable." (Latour, 2007)

Mais au-delà de l'accès à de nouvelles sources de données, la publicisation et l'accélération des échanges dans ces espaces nées, entre autres, de l'avènement du Web 2.0, constituent également un enjeu de recherche à part entière. Ce sont bien de nouveaux modes de sociabilité, d'expression publique, d'élaboration de la connaissance ou d'action collective, qui sont à l'œuvre au sein de ces nouveaux "conduits pour la vie sociale contemporaine" [6] (Hine, 2005), et dont la compréhension constitue un véritable défi pour les années à venir (Chateauraynaud and Trabal, 2003). Ce changement de medium a également pour effet de faire émerger de nouveaux acteurs au sein du jeu social ou politique traditionnel, les pouvoirs sont susceptibles d'être redistribués très rapidement, des effets de loupe pouvant subitement faire augmenter la "taille" d'un acteur au sein du forum mondial que constituent ces nouveaux espaces.

2.3.2 Reconstuction phénoménologique des traces

La disponibilité à grande échelle de ces traces est également une condition nécessaire à leur traitement statistique. L'idée tardienne de faire *entrer le monde social en statistique* semble de plus en plus crédible grâce à ces données, mais aussi grâce aux outils de traitement de ces données dont nous disposons. Il est utile à ce titre de revenir plus précisément sur les "conquêtes" auxquelles la statistique sociologique, considérée comme "l'étude appliquée de l'imitation et de ses lois" pouvait prétendre selon Tarde.

Pour illustrer sa vision d'une statistique sociologique, Tarde (1890) compare la courbe des récidives criminelles ou correctionnelles avec les brusques relèvements du vol d'une hirondelle. L'originalité de la métaphore de Tarde réside dans le fait qu'il ne considère pas que les deux tracés (d'un côté, les courbes construites par un bureau de statistiques et, d'un autre côté, celles que le mouvement d'un oiseau dessine sur une rétine) reflètent deux représentations du monde de natures différentes. Pourtant, la première est classiquement "réputée symbolique" tandis que la seconde est usuellement jugée comme "une réalité inhérente à l'être même qu'elle exprime". Tarde réfute cette distinction. Selon lui, les seules différences qu'entretiennent les "courbes graphiques des statisticiens" et les "images visuelles" se résument essentiellement à la difficulté que l'on peut rencontrer lorsqu'il s'agit de tracer et d'interpréter les premières alors que les secondes seraient captées et interprétées de façon continue [7].

6. "There is then a considerable will to research and understand technologically mediated interactions [...] as an important conduit for contemporary social life"

7. "La différence la plus saisissable qui subsiste dès lors entre les courbes graphiques des statisticiens et les images visuelles, c'est que les premiers coûtent de la peine à l'homme qui les trace et même à celui qui les interprète, tandis que les secondes se font en nous et sans nul effort de notre

tel-00464316, version 1 - 16 Mar 2010

La comparaison de la statistique sociologique avec *l'observation rétinienne* d'un objet mouvant révèle bien ici l'ambition tardienne. Dans les deux cas, qu'il s'agisse du monde des "figures mobiles en mouvement", ou de celui des "faits sociaux", nous somme confrontés au même besoin de *reconstruction phénoménologique* des traces que la réalité nous permet d'observer. La reconstruction du vol d'une hirondelle est rendue possible et nous semble "naturelle" grâce et à cause de l'extrême perfectionnement de notre rétine et de notre système cognitif transformant un processus d'inscription photochimique en "l'illusion" d'un vol d'oiseau. En matière sociale, il faudra nous équiper d'un certain nombre de prothèses sensorielles [8] avant d'être à même de *reconstruire* les dynamiques sociales avec une précision telle que "de chaque fait social en train de s'accomplir, s'échappera pour ainsi dire automatiquement un chiffre".

Or, on sait que la vision, ainsi que l'ensemble de nos sens, n'agissent pas comme de simples instruments de captation d'une réalité extérieure, mais comme des machines adaptatives sophistiquées. À partir de signaux relativement frustres de cette réalité, ces machines reconstruisent un certain nombre de structures remarquables telles une texture, une forme, ou un corps un mouvement, qui rassemblées composent la *phénoménologie*. Les faits sociaux appellent au même travail de reconstruction. À partir d'une captation de la réalité, aussi bruitée et limitée soit-elle, il s'agit non pas simplement de compter et d'énumérer mais de reconstruire la phénoménologie de ces faits en décrivant et en détectant leurs motifs caractéristiques, leurs régularités remarquables, les différentes formes, couleurs, et textures qu'ils revêtent également.

Latour and Woolgar (1986) décrivent le travail de recherche comme fondé sur les inscriptions produites par des instruments à même d'identifier et d'enregistrer des traces laissées par des entités invisibles. Mais collecter ces traces n'est pas suffisant. Comme l'affirme Callon (2006), les chercheurs, de toutes disciplines, travaillent à partir d'inscriptions produites par des instruments qu'ils doivent par la suite traduire en énoncés [9]. Cette dernière opération de traduction n'est pas toujours la plus aisée, et c'est celle que nous décrirons sous le terme de reconstruction phénoménologique. Ce que nous recherchons à travers ces traces, ce sont l'ensemble des traits et motifs réguliers nous permettant de *reconnaître* dans ces traces la manifestation de tel ou tel processus social, aussi sûrement que les traces que l'évolution d'un oiseau dans le ciel laissent sur nos rétines nous permettent de les reconnaître comme le vol d'un oiseau.

part, et se laissent interpréter le plus facilement du monde ; c'est encore que les premières sont tracées longtemps après l'apparition des faits et la production des changements qu'elles traduisent de la manière la plus intermittente, la plus irrégulière aussi bien que la plus tardive, tandis que les secondes nous révèlent ce qui vient de se faire ou ce qui est même en train de se faire, et nous le révèlent toujours régulièrement, sans interruption"(Tarde, 1890), p101.

8. Tarde (1890) compare d'ailleurs ses "bureaux de statistiques" à l'œil ou à l'oreille.

9. "Researchers, irrespective of their discpline, work on inscriptions produced by instruments, which they have to decipher and put into statements."

tel-00464316, version 1 - 16 Mar 2010

Internet nous promet un accès de plus en plus facilité et détaillé aux traces de l'activité sociale telles qu'elle se déploie dans les communautés virtuelles et ailleurs, néanmoins, le prochain défi pourrait consister à être capable à reconstruire, à grande échelle, et de façon satisfaisante, la phénoménologie des faits sociaux cachés derrière ces traces.

2.3.3 Des traces textuelles au réseau épistémique

La construction du réseau épistémique d'une communauté de savoirs revient à modéliser la dynamique du système à partir de ses traces. Il s'agit de conceptualiser de façon aussi fidèle que possible la richesse originale des dynamiques des traces textuelles produites par l'activité de ces communautés tout en simplifiant suffisamment sa description de façon à la rendre interprétable par les outils d'analyse à notre disposition. Nous décrivons dans cette partie cette opération de modélisation qui, à partir d'un corpus de textes $\mathcal{T}$, permet de construire les trois réseaux qui composent notre réseau épistémique : le réseau social $\mathcal{G}^{\mathbf{S}}$, le réseau sémantique $\mathcal{G}^{\mathbf{C}}$, et le réseau socio-sémantique $\mathcal{G}^{\mathbf{SC}}$.

Notre matériau de départ est constitué d'un ensemble de textes, des billets dans le cas de la blogosphère, des articles scientifiques dans le cas des communautés scientifiques ; chaque texte combine des éléments sociaux et sémantiques, à savoir des acteurs et des concepts. Nous avons représenté figure 2.2 le schéma conceptuel d'un texte dans cet espace socio-sémantique (dont la séparation est signalée par un trait en pointillés et, de part et d'autre duquel, on retrouve l'ensemble des acteurs ($\mathbf{S}$) et des concepts ($\mathbf{C}$) de la communauté). Le texte est représenté par une surface rouge qui, à une date de publication donnée (dans notre exemple le texte a été produit à la date D_2), rassemble un auteur qui initie la frontière de la surface (S_1) [10], un ensemble de concepts mobilisés (C_1, C_3), et un ensemble de relations liant ce texte à d'autres entités (dans notre exemple, nous avons représenté des liens de citation vers d'autres auteurs (S_2, S_3) [11]). Formellement un texte est donc un arc (ou plutôt un hyper-arc) dans un hyper-réseau réunissant deux types d'entités (des acteurs et des concepts) liés selon différentes références (un auteur faire référence à d'autres auteurs (citations) ou à différents concepts (usages)). On peut donc décrire un texte T comme l'objet doté des attributs suivants :

$$T : \mathit{Texte}(\mathit{Auteur}(T) : \mathbf{S}, \mathit{Usages}(T) : 2^{\mathbf{C}}, \mathit{Citations}(T) : 2^{\mathbf{S}}, \mathit{Date}(T) : \mathbb{N})$$

Dans notre schéma, on a ainsi représenté le texte suivant :

10. Dans le cas des billets de blogs, le nombre d'auteurs est naturellement limité à 1, il en va autrement dans le cas des publications scientifiques, mais par mesure de simplicité, nous considérerons dans la suite que chaque texte a un seul auteur.

11. Néanmoins, ces relations peuvent être de différents types, elles peuvent aussi bien désigner des références dans un article ou dans un billet ou des commentaires de blogueurs. Elles peuvent également pointer vers un auteur ou un autre texte, etc. Dans notre exemple nous nous sommes limités pour simplifier la représentation à des relations de citation pointant vers d'autres acteurs.

tel-00464316, version 1 - 16 Mar 2010

Texte$(S_1, (C_1, C_3), (S_2, S_3), D_2)$). En toute rigueur un texte est donc un hyper-lien orienté dans un hyper-réseau complexe. Les fonctions utilisées pour définir l'objet texte sont les suivantes :

- *Auteur* : *Texte* $\rightarrow \mathbf{S}$, dans notre exemple *Auteur*$(T) = S_1$
- *Usages* : *Texte* $\rightarrow 2^{\mathbf{C}}$, dans notre exemple *Usages*$(T) = (C_1, C_3)$
- *Citations* : *Texte* $\rightarrow 2^{\mathbf{S}}$, dans notre exemple *Citations*$(T) = (S_2, S_3)$
- *Date* : *Texte* $\rightarrow \mathbb{N}$, dans notre exemple *Date*$(T) = D_2$

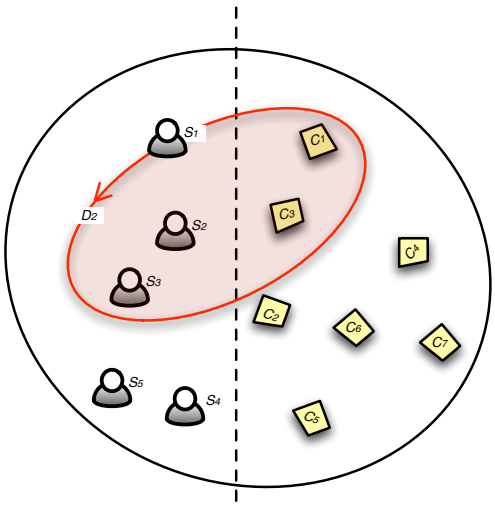


FIGURE 2.2: Représentation schématique d'un texte daté (D_2) (pouvant figurer aussi bien un article scientifique qu'un billet de blog) regroupant un auteur (S_1), les concepts mobilisés (C_1, C_3) et l'ensemble des acteurs cités (S_2, S_3)

À partir de ces opérateurs de projection on peut définir les réseaux dynamiques qui constituent notre réseau épistémique. Ainsi, le réseau social $\mathcal{G}^{\mathbf{S}}$ est composé des liens datés suivants :

$$\mathcal{R}^{\mathbf{S}} = \{(S_1, S_2, D) \mid \exists T \in \mathcal{T}, S_1 = \textit{Auteur}(T) \wedge S_2 \in \textit{Citations}(T) \wedge D = \textit{Date}(T)\}$$

De la même façon le réseau $\mathcal{G}^{\mathbf{SC}}$ est constitué de l'ensemble des triplets :

$$\mathcal{R}^{\mathbf{SC}} = \{(S, C, D) \mid \exists T \in \mathcal{T}, S = \textit{Auteur}(T) \wedge C \in \textit{Usages}(T) \wedge D = \textit{Date}(T)\}$$

On définit enfin les liens $\mathcal{R}^{\mathbf{C}}$ dans le réseau sémantique $\mathcal{G}^{\mathbf{C}}$ comme l'ensemble des triplets défini tel que :

$$\mathcal{R}^{\mathbf{C}} = \{(C_1, C_2, D) \mid \exists T \in \mathcal{T}, \{C_1, C_2\} \subset \textit{Usages}(T) \wedge D = \textit{Date}(T)\}$$

Nous serons amenés par la suite à affiner ces différentes définitions (et en particulier celles ayant trait à la définition du réseau sémantique) afin de les rendre plus à même de restituer de façon réaliste les phénomènes sous-jacents.

tel-00464316, version 1 - 16 Mar 2010

Néanmoins, ce qu'il importe de noter ici, c'est que nous avons transformé notre hyper-réseau de départ dont chaque texte formait un arc en trois réseaux de types dyadiques (deux monopartis et le troisième biparti). La réduction est double, nous avons, d'une part, séparé les entités sociales et sémantiques en les considérant isolément dans les réseaux social et sémantique, ou en les regroupant au sein d'une topologie de réseau bien particulière, celle d'un réseau biparti dans le cas du réseau socio-sémantique. Cette première hypothèse de modélisation s'appuie sur notre hypothèse de départ d'autonomie des sphères sociales et sémantiques, tout en garantissant l'existence d'un couplage entre les deux dimensions.

La seconde réduction importante que nous avons opérée relève plus d'une nécessité pratique que d'une hypothèse de modélisation. Même en supposant qu'il est pertinent de modéliser l'activité des communautés de savoirs comme la combinaison de nos trois réseaux, en toute rigueur, les réseaux ainsi décrits devraient en fait être des hyper-réseaux. Par exemple, le réseau sémantique devrait mettre en relation au sein d'un même hyperlien l'ensemble des concepts ayant été conjointement utilisés dans un même texte, de la même façon, un réseau de collaboration (non décrit dans notre cadre mais formellement équivalent) devrait théoriquement être constitué d'hyperliens de tailles variables regroupant l'ensemble des auteurs d'un texte. Nous avons pourtant fait le choix de modéliser nos réseaux à l'aide de relations purement dyadiques plutôt que n-adiques. Les outils et la conceptualisation même des hyper-réseaux sont encore embryonnaires [12], aussi nous privilégierons une modélisation par des réseaux exclusivement dyadiques. L'enjeu est alors de garder à l'esprit cette opération de réduction au cours de la modélisation et de rester attentif aux biais éventuels introduits par cette dénaturation des traces originales.

2.3.4 Un échantillon de la blosphère politique française

Dans cette section et les suivantes, nous décrivons en détail le protocole de collecte de données qui a été employé pour décrire les jeux de données analysés dans cette thèse. La structure d'un jeu de données décrivant une communautés de savoirs est la même pour l'ensemble de nos cas d'étude, néanmoins, la réalisation pratique de cette collecte varie énormément selon que l'on s'applique à "crawler" la blogosphère ou à interroger une base de données de publications scientifiques. On peut grossièrement décrire la procédure générale comme la succession de trois étapes : (i) délimitation d'un ensemble de textes ou d'acteurs qui définit les limites de notre communautés de savoirs (cette phase peut aller de la défini-

12. On peut néanmoins citer l'approche hypergraphique mise en œuvre dans Ruef et al. (2004) pour caractériser la formation de nouvelles équipes d'entrepreneurs ou le cadre théorique plus large actuellement conçu par Johnson (2006) pour rendre compte de la dynamiques des systèmes complexes à travers un formalisme original fondé sur les hyper-réseaux sans oublier les hypothèses premières de nature typiquement hypergraphique que Simmel (1955) mettait en œuvre.

tel-00464316, version 1 - 16 Mar 2010

tion d'une requête pour interroger une base de données, à un travail d'enquête pour identifier les acteurs d'une communauté de savoirs), (ii) collecte du corpus de textes correspondant et extraction des informations pertinentes ("parsing" d'un corpus de pages web ou d'un ensemble de publications scientifiques) et enfin (iii) modélisation des informations extraites pour construire notre réseau épistémique. Cette dernière phase correspond à la première étape de "reconstruction phénoménologique" des données textuelles originales en une description plus réduite mais également plus facile à appréhender par les outils d'analyse.

Nous illustrerons ces trois phases de façon détaillée sur l'un de nos cas d'étude : celui constitué d'une portion de la blogosphère politique française. La constitution de ce jeu de données a nécessité de trouver un compromis entre la richesse des information extraites de l'observation de l'activité de la blogosphère et le nombre de blogs sélectionnés pour figurer dans cette collection.

Concernant la première phase de délimitation de notre système, nous avons appliqué un processus de sélection de type "boule de neige" (Herring et al., 2005) dont l'initiateur a été un blog choisi parmi les 5 blogs les plus influents de la blogosphère des commentateurs politiques français [13]). Partant de ce blog, on a réuni tous les blogs mentionnés dans son "*blogroll*" (le blogroll d'un blog est classiquement interprété comme ses "favoris" d'un blog, on considère généralement que le blogroll réunit les blogs régulièrement lus par l'auteur du blog). Ces blogs situés dans le voisinage de notre blog initial ont ensuite été sélectionnés ou rejetés sur la base de leur activité (leur activité moyenne devait être au moins égale à un billet hebdomadaire) et en fonction des thématiques traitées (nous avons écarté les blogs qui n'étaient pas focalisés sur le commentaire de la vie politique française) ; certaines considérations techniques nous ont également parfois contraint à écarter certains sites mais ce type de cas s'est finalement limité à quelques blogs seulement. Cette première couronne, constituée par les blogs politiques actifs situés à un clic du blogroll de notre site initial, est constituée de 23 blogs. Elle a ensuite été complétée par une opération similaire à partir des blogs figurant dans le blogroll de cette première couronne. Au final, notre jeu de données contient un ensemble de 120 blogs notés $\mathcal{B}$. Ces blogs se trouvent tous, au plus, à deux clics du blog de départ (si l'on se limite aux seuls liens de blogroll). Cependant, les liens de blogroll étant orientés, deux blogs de cet ensemble peuvent très bien, par exemple, se trouver à une distance 5 l'un de l'autre ; le réseau de blogroll résultant n'est d'ailleurs naturellement pas nécessairement connexe. La phase de sélection est résumée par la représentation du processus de boule de neige figure 2.3, c'est le réseau de blogroll qui est représenté de façon à ce que l'agencement spatial des nœuds (organisation en deux cercles concentriques) corresponde aux deux couronnes construites en s'éloignant à distance 1 puis 2 du blog initial.

13. http ://versac.net fait partie du top 5 du classement des blogs de commentaires politiques publié par le moteur de recherche spécialisé dans la blogosphère Technorati (http ://www.technorati.com

tel-00464316, version 1 - 16 Mar 2010

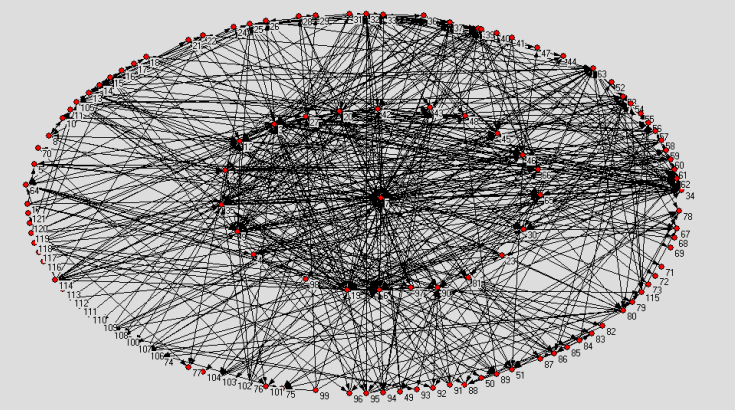


FIGURE 2.3: Ensemble des 120 blogs constituant notre jeu de données, représentés selon leur appartenance à la première ou à la seconde couronne autour du blog de versac (au centre du schéma).

La seconde phase de collecte et d'extraction d'un corpus textuel pertinent a été menée sur cette portion de la blogosphère politique française en "collectant" systématiquement les traces textuelles produites pendant 6 mois d'activité, du 1^{er} janvier 2007 jusqu'au 30 juin 2007, une période durant laquelle des débats nourris ont accompagné le déroulement des élections présidentielles (1^{er} et 2^{nd} tour au mois d'avril et mai) puis législatives (en juin) françaises.

L'activité de chacun des blogs de notre sélection a été "crawlée" grâce à des scripts développés de façon ad-hoc pour chaque plate-forme de blog et si nécessaire pour chaque blog en réalisant une exploration "rétrospective" de l'historique d'édition des blogs. Nous avons ainsi extrait, sur l'ensemble de la période d'observation, 11 552 billets, leur date de publication (telle qu'elle apparaît dans le billet) et l'ensemble des commentateurs d'un billet donné (pourvu que le commentaire ait été signé par un des blogs de notre échantillon). Nous avons également extrait l'ensemble des liens de blogrolls entre blogs de notre échantillon $\mathcal{B}$. [14]

Restreindre notre sélection à un nombre relativement faible de blogs, nous a ainsi permis d'employer une méthodologie d'extraction certes coûteuse (chaque blog a bénéficié *a minima* d'une adaptation du script de collecte à ses spécificités techniques), mais permettant d'obtenir une description très précise des dynamiques d'interaction entre blogueurs (sur les liens de commentaire et de citation) et de production de contenu. D'autres méthodologies ont été mises en œuvre pour collecter des données de blogs à grande échelle. L'une d'elle consiste à employer les flux *rss* des blogs (Shi et al., 2007) afin d'aider à la reconstruction du contenu des billets à partir d'heuristiques simples, une autre méthode consiste à effectuer

14. Techniquement, la technique de "web scrapping" employée était fondée sur le repérage de *DOM* associés aux champs d'intérêt (date du billet, titre, liens de commentaires, etc.) au sein du code de la page d'un blog. Ce travail de collecte a été réalisé avec le concours précieux de Hugo Lebrun, Richard Norton, et Charles Cizel.

tel-00464316, version 1 - 16 Mar 2010

des crawls quotidiens de la page principale d'un blog et de détecter les différences entre deux états successifs du site pour tâcher d'en isoler les nouveaux contenus (Glance et al., 2004). De par leur robustesse, ces méthodes sont bien adaptées à la collecte de jeux de données massifs, mais elles sont exposées à certaines limitations quant à l'exhaustivité et la fiabilité de l'opération de collecte : absence de flux rss pour certains blogs, difficulté à séparer les changements issus d'une modification de la charte graphique ou de l'ajout de commentaires dans la méthode différentielle, les réseaux de commentaires restant généralement le parent pauvre de ces analyses.

Notre méthode "rétrospective" présente l'avantage de donner une représentation plus détaillée de l'activité des blogueurs, incluant notamment le réseau des commentaires. Elle permet également un "parsing" des données homogène dans le temps, le crawl étant effectué "en une passe" sur l'ensemble des 6 mois d'activité des blogueurs (on élimine ainsi le bruit induit, par exemple, par des modifications de chartes graphiques). De façon plus générale, la reconnaissance automatisée de la structure des textes au sein d'un blog ou d'une page web constitue un champ de recherche à part entière (Bar-Yossef and Rajagopalan, 2002) qui se développe très rapidement. Dans un contexte plus industriel, il semble désormais possible de construire des jeux de données beaucoup plus larges et, sans doute bientôt, plus détaillés que celui sur lequel nous nous penchons.

Enfin, la troisième et dernière phase de modélisation de notre corpus de textes est décrite dans la section suivante.

2.3.5 Un multi-réseau dynamique

Les blogueurs peuvent entrer en relation les uns avec les autres de trois manières différentes. Ainsi un blog peut se lier à un autre blog : (i) en l'incluant dans son *blogroll* (ii) en *commentant* un de ses billets (iii) en le *citant* au sein d'un billet (la citation peut pointer vers un de ses billet ou directement vers l'adresse du blog). Ces trois types de liens sont illustrés figure 2.4 (*à gauche*) à partir d'une capture d'écran du premier site de notre échantillon. Ces relations définissent trois réseaux de blogs dont une représentation schématique est donnée sur la même figure (*à droite*). Ces liens sont de différentes natures. On peut grossièrement les distinguer de la manière suivante, les liens de blogroll sont des liens correspondant à des relation d'autorité, les liens que l'on trouve au sein des billets sont des liens de citation, et enfin, nous considérerons les liens de commentaire comme des interactions entre blogs.

L'analyse de l'activité de cette portion de la blogosphère pendant 6 mois a permis de reconstruire l'intégralité du multi-réseau social composé du réseau de citation, du réseau de commentaire et du réseau de blogroll. On définit formellement ce multi-réseau de la façon suivante. Le quadruplet $(\mathcal{B}, R, P, C)$ forme le *multi-réseau* de blogs —- R, P et C désignent respectivement les liens de blogroll, de

tel-00464316, version 1 - 16 Mar 2010

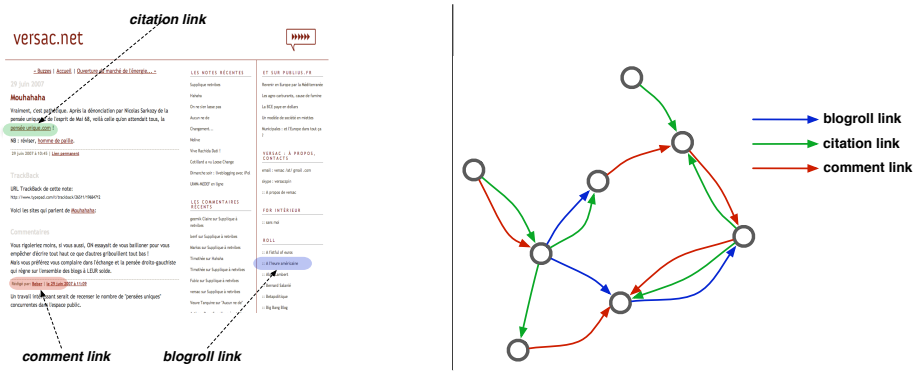


FIGURE 2.4: *À gauche* Exemple de blog (http ://versac.net) illustrant les trois types de liens possibles, *à droite*, schéma du multi-réseau de blogs, on distingue les liens de blogroll (en bleu), les liens de citation (en vert), et enfin les liens de commentaire (en rouge). Ces deux derniers sont de plus dynamiques, les liens étant datés, le premier réseau de blogroll est considéré comme statique.

citation de billet, et de commentaire, dont les matrices d'adjacence associées sont toutes de taille $|\mathcal{B}| \times |\mathcal{B}|$. Ces données sont également *dynamiques*, avec une granularité temporelle de l'ordre du jour, on notera P_t et C_t les réseaux de citation et de commentaire considérés au temps t (t variant entre 1 et 181). Ainsi $P_t(i,j) = 1$ si i cite j dans un billet au temps t ; $C_t(i,j) = 1$ lorque i commente un billet sur le blog j à t (en laissant l'URL de son blog dans la signature du commentaire qu'il laisse à la suite d'un billet de j).

La dynamique du réseau de blogroll est particulière à au moins deux titres : (i) son temps caractéristique d'évolution est bien plus lent que celui des autres réseaux (Qazvinian et al., 2007) — une grande majorité des liens restant inchangés durant les six mois — et (ii) ce n'est pas un réseau croissant comme les autres réseaux puisque certains liens peuvent disparaître. Par mesure de simplicité, nous considérerons que le réseau de blogroll est statique : $R_t(i,j) = R(i,j)$. Ce réseau ne sera de toute façon employé que de façon très marginale dans l'ensemble des analyses qui vont suivre, et nous aura principalement aidé dans la conception du processus de collecte de données.

Dans la suite nous omettrons l'indice t quand la dépendance dans le temps est implicite. Nous avons respectivement extrait 725, 2 299 and 3 396 liens orientés et datés dans dans R, P et C, formant au total seulement 1 568 liens distincts (les liens de commentaire et de citation étant datés, l'interaction entre certaines dyades peut être répétée dans le temps). La figure 2.5 represente le recouvrement entre les trois réseaux en nombre de liens uniques. Ces différents liens ne sont que très partiellement recouvrants. On remarque, qu'environ deux tiers des liens sont propres à un seul des trois réseaux. Cette observation est cohérente avec une étude antérieure réalisée par Ali-Hasan and Adamic (2007). Cette spécificité des liens de chaque réseau semble indiquer que ce sont bien trois espaces relationnels distincts

tel-00464316, version 1 - 16 Mar 2010

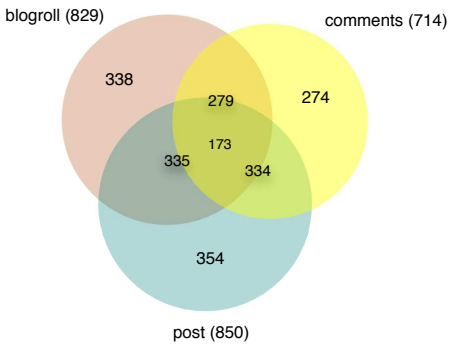


FIGURE 2.5: Proportion de liens appartenant à chaque réseau et recouvrements respectifs en nombre de liens. Par exemple, 279 liens sont communs au réseau de blogroll et au réseau de commentaires, tandis que 173 apparaissent dans les trois réseaux.

qui co-existent, même si des corrélations entre réseaux sont sûrement possibles.

La grande diversité dans le nombre d'occurrences de chaque lien nous pousse à introduire pour les réseaux de citation, et de commentaire leur version pondérée et agrégée définie comme : $\mathbf{P}_t = \sum_{t'=1}^{t} P_{t'}$ et $\mathbf{C}_t = \sum_{t'=1}^{t} C_{t'}$ [15]. Le blogroll étant considéré comme statique, nous écrirons $\mathbf{R}_t = R$. Le réseau social $\mathcal{G}^{\mathbf{S}}$ dynamique décrivant la dimension sociale de cette communauté de savoirs est donc défini comme le multi-réseau dynamique agrégeant les réseaux de citation, de commentaire et de blogroll : $(\mathcal{B}, \mathbf{P}_t, \mathbf{C}_t, \mathbf{R}_t)$.

2.3.6 Caractérisation sémantique

Au-delà, de la structure sociale liant ces blogs, nous souhaitons caractériser ce système en introduisant une dimension sémantique liée aux contenus produits et échangés par les blogueurs. Les billets traitent généralement d'un sujet ou d'une question en particulier, parfois en s'appuyant sur certaines ressources extérieures. Nous faisons l'hypothèse que nous pouvons caractériser le contenu des billets des blogueurs à partir des sujets et thématiques qu'ils discutent ; ceux-ci sont repérés dans le texte des billets à partir d'un ensemble de syntagmes jugés pertinents vis-à-vis de la communauté de savoirs. En ce qui concerne les ressources extérieures sur lesquelles les blogueurs peuvent s'appuyer, nous faisons également l'hypothèse que les URLs qui ne sont pas des liens de citation permettent de définir un ensemble de ressources digitales qu'un blogueur est à même de diffuser dans son environnement. Ces ressources serviront spécifiquement à définir des entités atomiques diffusant dans la blogosphère.

Par conséquent nous distinguerons par la suite :

– un ensemble de sujets de haut-niveau $\mathcal{W}$, que nous appellerons également

15. Les notations en gras désigneront de façon générale dans le texte des mesures agrégées temporellement.

tel-00464316, version 1 - 16 Mar 2010

concepts, relatif à l'activité de chronique de la vie politique dans notre contexte. $\mathcal{W}$ est constitué de 190 syntagmes, dont la liste est donnée en annexe A.1, allant de noms de figures politiques à des questions de sociétés qui ont animé la dernière campagne présidentielle comme " *changement climatique*", "*impôt sur la succession*", "*débat public*", "*prévention de la délinquance*", "*referendum sur la Constitution Européenne*", "*heures supplémentaires*", etc.

– un ensemble d'URLs, noté $\mathcal{U}$, distinctes de liens dans le réseau de citation — celles-ci sont simplement des ressources extérieures : vidéos en ligne, articles de media, billets d'autres blogs extérieurs à notre sélection, etc. $\mathcal{U}$ consiste en une sélection de 3 140 URLs (dont le nombre de caractère est supérieur à 10[16]).

Munis de l'ensemble $\mathcal{W}$ des syntagmes pertinents, nous procédons ensuite à l'indexation de ces concepts au sein de l'ensemble des billets extraits. Cette indexation, dans le cas de la blogosphère politique française, a été réalisée avec l'aide de Didier Bourigault et Franck Sajous concepteur de Leximedia 2007[17]. Le logiciel d'analyse syntaxique Syntex (Bourigault et al., 2005) a été utilisé de façon à intégrer certains traitements linguistiques tels que le repérage des types grammaticaux de chaque occurrence de notre ensemble de concepts (ainsi le logiciel différencie par exemple durant l'indexation le terme "Royal" entre son emploi en tant qu'adjectif ou en tant que nom propre). Cette tâche d'indexation permet de connaître le ou les concepts employés par tel ou tel blogueur un jour donné. On introduit ainsi la matrice temporelle W_t qui retrace les contenus publiés par les blogueurs : $W_t(i, w)$ vaut 1 si le terme $w \in \mathcal{W}$ apparaît dans un billet publié par le blog i au temps t, 0 sinon.

On définit le *profil sémantique* d'un agent i au temps t comme l'agrégation des sujets qu'il a abordé jusque là. Ce profil est un vecteur de dimension $|\mathcal{W}|$ noté $\mathbf{W}_t(i)$. Il est défini comme étant égal à la somme des vecteurs $W_{t'}(i)$ pour $t' \leq t$. Le profil sémantique de chaque agent est donc représenté par un vecteur dans un espace dont les termes forment les dimensions.

Cette matrice dynamique peut s'interpréter de façon équivalente comme le réseau socio-sémantique dynamique $\mathcal{G}^{\mathbf{SC}}$ liant les agents du système aux concepts qu'ils mobilisent. On définit simplement l'ensemble des liens $\mathcal{R}^{\mathbf{SC}}$ de $\mathcal{G}^{\mathbf{SC}}$, dans sa version non pondérée, comme l'ensemble des couples (i, w) d'agents et de concepts vérifiant $\mathbf{W}_t(i, w) > 0$. Dans sa version pondérée, les liens (i, w) du réseau $\mathcal{G}^{\mathbf{SC}}$ sont simplement dotés d'un poids égal à $\mathbf{W}_t(i, w)$. Le réseau socio-sémantique, comme le réseau social est par définition croissant. En pratique nous n'emploierons que de façon mineure cette formalisation (dans le chapitre 3 exclusivement), et utiliserons essentiellement l'expression des profils sémantiques des agents soit pour les plonger dans une structure sémantique plus large, soit pour construire une distance sémantique entre deux agents.

16. les chaînes de caractères "http ://" et "www" mises à part
17. http ://erss.irit.fr :8080/LexiMedia2007/

tel-00464316, version 1 - 16 Mar 2010

Pour comparer les contenus produits par deux agents i et j au temps t, nous adopterons une mesure classique de similarité basée sur un calcul de corrélation soit le cosinus de leur profil sémantique $\mathbf{W}_t(i)$ et $\mathbf{W}_t(j)$. Mais avant de réaliser cette mesure, nous appliquons d'abord une procédure de normalisation des termes en fonction de leur fréquence respective en suivant l'approche du "tf·idf" Salton et al. (1975). Ce type de traitement est largement appliqué dans l'ingénierie documentaire.

La procédure de normalisation consiste à pondérer la "fréquence des termes", "tf" , ou fréquence du terme au sein de la production textuelle d'une source, avec la "fréquence inverse de document", "idf", soit l'inverse de la fréquence du terme dans l'ensemble du corpus des sources porté au log. Cette méthode permet de donner plus de poids aux termes rares dans les profils sémantiques des agents. Les profils $\mathbf{W}_t(i)$ sont donc remplacés par des profils ajustés par tf·idf : $\hat{\mathbf{w}}_t(i)$ définis de la façon suivante :

$$\hat{\mathbf{w}}_t(i,w) = \frac{\mathbf{W}_t(i,w)}{\sum_{w=1}^{|\mathcal{W}|} \mathbf{W}_t(i,w)} \cdot \log \frac{|\mathcal{B}|}{|\{j, \mathbf{W}_t(j,w) > 0\}|}$$

où la partie droite de la formule correspond au ratio inverse du nombre de sources mentionnant le terme w.

Nous obtenons ensuite une expression de la similarité entre deux agents i et j comme le produit scalaire de leur profil sémantique normalisé divisé par le produit de leur norme. La dissimilarité sémantique entre deux agents i et j, $\delta_t(i,j)$ peut s'interprèter comme une "dissonance cognitive" et s'exprime sous la forme :

$$\delta_t(i,j) = 1 - \frac{\hat{\mathbf{W}}_t(i) \cdot \hat{\mathbf{W}}_t(j)}{\|\hat{\mathbf{W}}_t(i)\| \|\hat{\mathbf{W}}_t(j)\|}$$

La dissimilarité $\delta_t(i,j)$ vaut 0 si les deux agents partagent exactement le même profil sémantique à t, et vaut 1 s'ils n'ont jamais mobilisé les mêmes concepts jusque là.

Le réseau sémantique $\mathcal{G}^{\mathbf{C}}$ nécessite généralement un traitement particulier lié à la définition d'une mesure de proximité entre concepts. Nous n'en donnons donc pour l'instant qu'une description simplifiée. Le réseau sémantique reflète la structure des concepts tels qu'ils sont mobilisés dans l'ensemble de la communauté de savoirs à un moment donné. Dans sa forme la plus simple, ses liens $\mathcal{R}^{\mathbf{C}}$ sont définis au temps t comme l'ensemble des couples de concepts (w_1, w_2) qui co-apparaissent dans un même billet publié à la date t. Nous ne décrivons pas, pour le moment, les formes plus évoluées que peuvent prendre ce réseau ; nous aurons l'occasion d'approfondir sa description dans la section 3.3.2 et plus largement au chapitre 4 lorsque nous aborderons les questions de cartographie des dynamiques scientifiques. Notons néanmoins que, contrairement aux deux réseaux précédents, le réseau sémantique n'est pas croissant.

tel-00464316, version 1 - 16 Mar 2010

2.3.7 Blogosphère américaine

Malgré la richesse du jeu de données de la blogosphère politique française, sa faible étendue peut parfois rendre son observation quantitative peu concluante. Les questions concernant le processus d'échantillonnage employé sont également à même de compliquer l'interprétation de nos résultats. Aussi, nous avons souhaité, au prix d'une richesse descriptive un peu moindre, adosser nos analyses à une seconde base de données de même nature mais de taille plus importante : les blogs politiques américains suivis durant la dernière compagne présidentielle américaine par RTGI [18], une entreprise spécialisée dans le monitoring de conversations sur le web social.

Cette second jeu de données est construit à partir du suivi de l'ensemble des billets publiés par une sélection de $1,066$ blogs politiques américains dans le contexte de l'élection présidentielle américaine ; ces données, originellement déployées au sein du portail *Presidential Watch '08* [19], ont été collectées du 1^{er} novembre 2007 au 29 février 2008. La sélection des blogs politiques américains actifs a été réalisée "manuellement" par des documentalistes, ce qui garantit la pertinence et l'exhaustivité de cette sélection.

Au total, un ensemble de $71\ 376$ billets datés a été collecté par RTGI. À partir de ce corpus de billets, nous avons ensuite suivi la même procédure de construction des réseaux social, socio-sémantique et sémantique que celle qui a été décrite à propos de la blogosphère politique française.

A partir des URLs extraites de ces billets, nous avons ainsi construit le réseau de citation dynamique que nous noterons également $\mathbf{P}_t$. $229\ 736$ liens de citation datés entre blogs de notre sélection ont été extraits dont à peine $15\ 032$ sont des liens uniques. Dans le cas de la blogosphère américaine, le réseau social $\mathcal{G}^{\mathbf{S}}$ est limité à ce seul réseau de citation. Une sélection de 79 syntagmes (dont la liste est fournie dans l'annexe A.2) a servi à indexer les billets de ces blogs et à définir leur profil sémantique $\hat{\mathbf{w}}_t(i)$ en suivant la même procédure que précédemment. Le réseau socio-sémantique biparti $\mathcal{G}^{\mathbf{SC}}$ a également été construit dans sa version non pondérée. Nous avons représenté, figure 2.6 l'évolution du nombre total de liens dans le réseau social $\mathcal{G}^{\mathbf{S}}$ et dans le réseau socio-sémantique $\mathcal{G}^{\mathbf{C}}$ (les liens répétés ne sont pas comptabilisés, ce qui explique la diminution progressive de la croissance de ces courbes).

Nous nous sommes également appuyé sur une sélection de $96\ 637$ URLs pour construire la matrice U_t qui définit l'usage de ces ressources par les blogueurs. Enfin le réseau sémantique $\mathcal{G}^{\mathbf{C}}$ a été construit en suivant la même méthodologie que précédemment.

L'ensemble de nos études touchant à la blogosphère politique ont ainsi été menées de paire sur les deux continents, à la fois en ce qui concerne la morphogenèse

18. http ://linkfluence.net
19. http ://presidentialwatch08.com

tel-00464316, version 1 - 16 Mar 2010

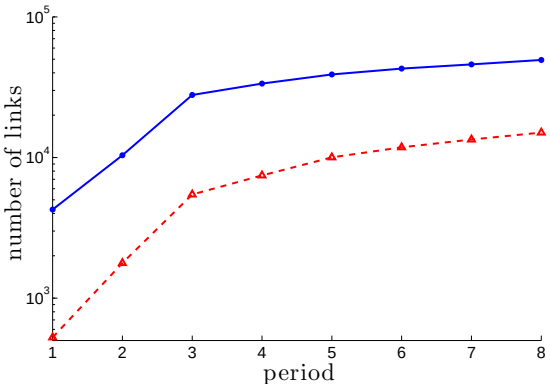


FIGURE 2.6: Evolution hebdomadaire du nombre de liens dans le réseau social ($\mathcal{G}^{\mathbf{S}}$) (en bleu) et dans le réseau socio-sémantique ($\mathcal{G}^{\mathbf{SC}}$) (en rouge) — une période correspond à une semaine d'activité.

(chapitre 3) et la diffusion (chapitre 7). Concernant la morphogenèse, par souci de clarté, la plupart des résultats ne sont présentés que pour le continent américain, sauf lorsqu'une différence qualitative apparaissait entre les deux jeux de données, ou lorsque la richesse descriptive de la blogosphère politique française permettait d'approfondir l'analyse. Concernant l'étude de la diffusion, nous avons systématiquement présenté les résultats sur nos deux jeux de données, d'une part, pour illustrer la stabilité des phénomènes mesurés sur deux cas d'étude, mais aussi pour illustrer le gain en significativité qu'un jeu de données plus large nous a permis d'obtenir.

2.4 Une approche par faces

La dualité sociosémantique doublée de notre hypothèse de couplage des niveaux micros et macros enrichit le schéma conceptuel dans lequel nous formalisons les communautés de savoirs. Ce schéma, représenté figure 2.7, fait apparaître, au premier plan, les dynamiques du réseau social inter-individuel $\mathcal{G}^{\mathbf{S}}$ (que l'on désignera par abus de langage comme la dimension sociale), à l'arrière plan, les dynamiques du réseau sémantique $\mathcal{G}^{\mathbf{C}}$ (dimension sémantique). On distingue également, sur le plan inférieur, le niveau micro des dynamiques individuelles (matérialisées, au niveau social, par une flèche temporelle transformant l'état $s(t)$ du réseau social à l'instant t en son état au pas de temps suivant : $s(t + dt)$; les dynamiques micros du réseau sémantique sont représentées par la transformation $c(t) \rightarrow c(t + dt)$) et, sur le plan supérieur, le niveau macro correspondant ($S(t)$ et $C(t)$ décrivant respectivement les structures émergentes de haut-niveau propres à chaque dimension). Les dynamiques du réseau socio-sémantique $\mathcal{G}^{\mathbf{SC}}$ — mettant en relation les agents et les concepts qu'ils mobilisent — sont figurées par les

tel-00464316, version 1 - 16 Mar 2010

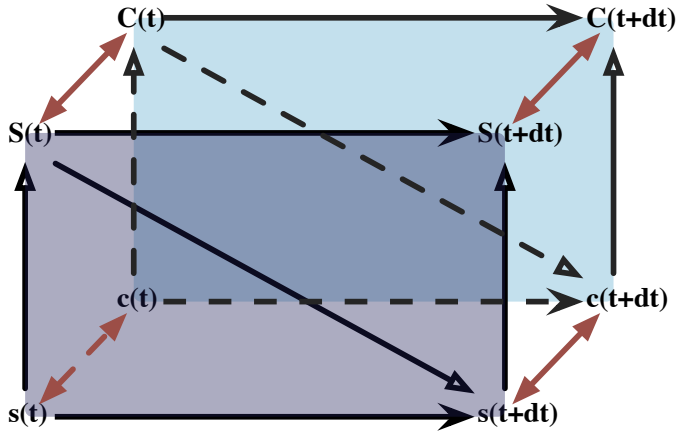


FIGURE 2.7: Schéma du couplage entre niveau micro et niveau macro figurant la nature sociosemantique de nos communautés de savoirs. Les dynamiques sociales guident la dynamique de $s(t)$ (niveau micro) et $S(t)$ (niveau macro) ; les dynamiques sémantiques, $c(t)$ et $C(t)$. Les flèches ascendantes correspondent aux fonctions d'émergence du niveau micro vers le niveau macro, tandis que les rétroaction macro/micro sont représentées par les flèches descendantes (par souci de clarté, toutes les rétroactions possibles n'ont pas été représentées (ex : rétroaction des dynamiques sémantiques macro ($C(t)$) sur les dynamiques sociales individuelles $s(t)$). Les mises en relation socio-sémantiques sont figurées par des flèches rouges ($s(t) \leftrightarrow c(t)$ au niveau micro, $S(t) \leftrightarrow C(t)$ concernant les structures socio-sémantiques émergentes).

flèches rouges à double sens (nous n'avons pas nommés ces dernières, mais en toute rigueur, on peut également noté $sc(t)$ et $SC(t)$ les états micros et macros du réseau socio-sémantique au temps t).

Enfin, le couplage émergence/immergence entre niveau micro et niveau macro est représenté par des flèches ascendantes et descendantes. Nous n'avons pas représenté l'ensemble des fonctions d'immergence possibles mais celles-ci doivent *a priori* toutes être envisagées. Ainsi un agent est immergé dans des structures sociales, socio-sémantiques, et sémantiques de haut niveau et sa dynamique micro est *a priori* susceptible d'être modifiée par l'ensemble de ces structures.

Dans la suite, nous tâcherons à partir de l'observation *in-vivo* des dynamiques de nos communautés de savoirs d'éclairer les différentes *faces* de ce parallélogramme afin de caractériser les couplages entre dimensions sociale et sémantique et entre niveaux micro et macro. Nous découperons notre analyse en deux parties : (i) morphogenèse du réseau épistémique (partie II) (ii) étude des processus à l'œuvre dans le réseau épistémique (partie III).

L'étude de la morphogenèse du réseau épistémique des communautés de savoirs peut se découper en une analyse couplant les deux niveaux : — *face supérieure du parallélogramme* — mise en évidence de motifs structurels de haut niveau non triviaux dans les trois réseaux (hiérarchie, émergence de communautés socio-

tel-00464316, version 1 - 16 Mar 2010

sémantiques, stabilité de structures conceptuelles émergentes, etc.) — *face inférieure du parallélogramme* — étude des comportements locaux réguliers (homophilie, transitivité, influence sociale, etc.) susceptibles de faire émerger ces motifs macroscopiques.

Dans le chapitre 3, nos communautés de blogueurs politiques nous serviront de terrain empirique pour interroger la morphogenèse des réseaux social, socio-sémantique, et sémantique en adoptant une approche partant essentiellement des dynamiques microscopiques, mais en restant attentif à certains motifs émergents et à leur dynamique. Nous montrerons également la façon dont certains processus guidant les dynamiques micro couplent la dynamique d'un réseau à l'état d'un autre réseau. Dans le chapitre 4, nous faisons l'opération contraire, en partant d'une reconstruction des dynamiques scientifiques permettant d'obtenir une représentation de haut niveau du réseau sémantique. Une fois ces dynamiques émergentes reconstruites (phylogenèse), nous interrogerons la façon dont ces structures de haut-niveau rétroagissent, par immergence, sur les dynamiques individuelles (et plus précisément sur les dynamiques micro du réseau socio-sémantique)

Enfin, nous traiterons séparément les motifs émergents de nature dynamique comme la diffusion. Les processus de diffusion sont en effet interprétés comme des émergence d'une autre nature au sens où on ne peut les caractériser de façon statique (alors qu'on peut très bien définir une distribution de degré comme une observable de haut niveau calculée en fonction de l'état du réseau à un moment donné [20]). Un épisode de diffusion constitue une émergence d'ordre purement dynamique de l'évolution de réseau épistémique (et plus précisément, d'une certaine évolution du réseau socio-sémantique étant donnée la structure du réseau social sous-jacent). C'est pourquoi nous traiterons séparément le cas de la diffusion au sein des communautés de savoirs en distinguant entre des approches plus ou moins agrégées et plus ou moins locales (depuis les influences croisées entre groupes de sources à un niveau macro jusqu'à la caractérisation fine de l'influence d'un agent en fonction de son environnement relationnel, en passant par l'influence de la structure topologique du réseau en son entier sur la vitesse d'un processus de diffusion).

L'ensemble des chapitres suivants s'appuient sur l'analyse de données empiriques. Notre objectif ne consistant naturellement pas à réaliser une analyse comparative exhaustive de nos deux types de communautés de savoirs, nous nous sommes tantôt appuyés sur des communautés scientifiques [21] (chapitres 4 et 6) tantôt sur nos blogosphères politiques (chapitres 3, 5 et 7). Les critères qui ont pré-

20. Le caractère statique de ces observables n'interdit naturellement pas d'en suivre l'évolution dans le temps, cependant cette évolution est reconstruite à partir de la séquence de motifs émergeant successivement d'une série de réseaux statiques successifs.

21. Concernant les communautés scientifiques, nous ne nous appuyons pas sur un jeu de données générique mais sur une série de bases dont nous décrirons les spécificités et la modélisation au fur et à mesure.

tel-00464316, version 1 - 16 Mar 2010

sidé au choix de l'un ou l'autre des types de communauté résultent notamment de considérations pratiques liées à la disponibilité et à la pertinence des données vis-à-vis des besoins de chaque étude. Nous tenterons, néanmoins, d'expliciter autant que possible les prolongements et perpectives que l'application de nos méthodes à d'autres communautés de savoirs appellent.

tel-00464316, version 1 - 16 Mar 2010

Résumé du chapitre:

Dans ce chapitre nous avons explicité, en premier lieu, la perspective théorique adoptée pour étudier les *communautés de savoirs* et notamment la nécessité de suivre simultanément leurs dynamiques aux niveaux micro et macro. Un des attraits de la modélisation sous forme de réseaux est justement de permettre d'appréhender dans un même cadre les dynamiques locales au niveau le plus atomique : l'individu ou le concept, et les motifs émergeant de ces dynamiques individuelles qui prennent la forme de structures mésoscopiques ou macroscopiques du réseau dans son entier.

Nous avons également souligné la nécessité d'une approche longitudinale permettant d'apprécier les évolutions des structures relationnelles à tous les niveaux. Un des objectifs de cette approche est la compréhension de la morphogenèse des communautés de savoirs qui se décline en deux points : régularités des dynamiques micros et émergence de structures de haut-niveau dont on puisse également évaluer la stabilité. L'accès à des données longitudinales à grande échelle offre également l'opportunité de suivre les processus de diffusion dont le réseau social est le support.

Enfin, nous avons précisé notre ambition épistémologique générale de reconstruction des dynamiques des communautés de savoirs à partir des traces digitales qu'elles produisent. L'opération de modélisation de ces traces textuelles sous la forme d'un réseau épistémique a été décrite, d'abord de façon générale, puis, plus pratiquement, lors de la définition de nos jeux de données liés aux blogosphères politiques française et américaine. Les hypothèses afférentes à cette modélisation ont été clairement délimitées et discutées.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Deuxième partie

Morphogenèse dans les réseaux de savoirs

DANS cette deuxième partie nous aborderons les modalités d'apparition et de stabilisation de certaines structures caractéristiques des communautés de savoirs. Nous souhaitons interroger les dynamiques de nos communautés de savoirs à différents niveaux ainsi que les couplages existant entre les dimensions sociale et sémantique. Quels sont les régularité observables dans la dynamique micro des entités (agents et concepts) ? Peut-on décrire les comportements individuels comme résultant de paramètres endogènes au système liés à l'état de la communauté en son entier ? Nous nous intéressons dans cette partie non seulement à ces déterminants structurels, mais aussi aux propriétés dynamiques de haut niveau aussi bien sociales, sémantiques que socio-sémantiques qui caractérisent nos communautés de savoirs. La dynamique de ces propriétés émergentes est également interrogée. Sont-elles stables ? Comment les représenter ? Peut-on mesurer l'influence qu'elles exercent en retour sur les dynamiques individuelles ?

L'analyse de la morphogenèse dans ces systèmes socio-sémantiques peut donc se décrire à deux niveaux différents.

Au *niveau micro, caractérisation des dynamiques locales, chapitre 3*. Il s'agit, d'une part, de caractériser les mécanismes individuels à l'origine de la morphogenèse. Les dynamiques locales seront appréhendées aussi bien dans leur dimension sociale (nouvelles relations/interactions), socio-sémantique (production de nouveaux contenus par les agents), que sémantique (évolution des contextes d'usage des concepts). Dans notre cadre, l'ensemble de ces questions revient à étudier le comportement de création de liens dans les trois réseaux : sociaux, socio-sémantiques et sémantiques. D'autre part, ces dynamiques locales sont intimement liées à l'émergence de propriétés de plus haut niveau structurant les trois réseaux qui composent notre réseau épistémique, certaines de ces propriétés simples comme la forme de la distributions de degré, ou le clustering seront mises en regard des régularités observées dans les dynamiques microscopiques ; nous tâcherons également d'évaluer la stabilité relative de ces propriétés de haut niveau. Outre le développement d'une méthodologie de caractérisation des communautés de savoirs, ouvrant la voie à leur comparaison ou à la modélisation de leur morphogenèse (que nous laisserons néanmoins de côté ici), ces méthodes d'analyse permettent d'apporter des éclaircissement sur les questions généralement posées

quant à leur "fonctionnement" ou à leur organisation. Comment l'autorité de certains agents est-elle construite et stabilisée au cours de la morphogenèse ? Peut-on évaluer les risques de balkanisation liés à l'agrégation des agents en fonction de leur profil sémantique ? Comment quantifier l'influence sociale exercée entre voisins ?

Au *niveau macro, caractérisation des motifs structurels globaux, chapitre 4.* Nous tâcherons d'identifier un certain nombre de structures émergentes liées à l'organisation de nos communautés en conservant à nouveau une perspective duale entre les dimensions sociales et sémantiques. Nous nous intéresserons particulièrement aux ensembles mésoscopiques que constituent les agrégats d'entités sociales et sémantique dotés d'une "cohésion interne", la structure et l'articulation de ces ensembles révélant l'organisation de nos communautés de savoirs. La dynamique de ces motifs structurels sera également examinée (*phylogenèse*). Nous nous concentrerons principalement sur la reconstruction multi-échelle des dynamiques de plusieurs communautés scientifiques en nous focalisant sur le réseau sémantique dynamique produit à partir d'un corpus de publications. Est-il possible à travers ce réseau, de reconstruire la phénoménologie de l'organisation des sciences sous la forme d'une *carte des sciences* pertinente et intelligible articulant un ensemble de *champs épistémiques* ? Dans une perspective dynamique, peut-on décrire l'évolution des sciences au niveau de ces champs épistémiques en retraçant la phylogénie des champs épistémiques ? Enfin, dans ce même chapitre, nous tâcherons de quantifier la façon dont la structure d'un paysage conceptuel peut contraindre les dynamiques individuelles des agents qui s'y trouvent immergés.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

CHAPITRE 3

Dynamiques locales

Sommaire

3.1 Dynamiques locales dans le réseau social 69
3.1.1 Attachements préférentiels 69
3.1.2 Attachement préférentiel aux degrés : capital social et capital sémantique 71
3.1.2.1 Définition 71
3.1.2.2 Résultats 73
3.1.2.3 Degré et activité 74
3.1.2.4 Propriétés macroscopiques émergentes associées . . 75
3.1.3 Attachement préférentiel à la distance sociale et sémantique . 76
3.1.3.1 Définitions 76
3.1.3.2 Résultats 77
3.1.4 Motifs cohésifs locaux 81
3.1.5 Capitaux, homophilie sociale et sémantique, découpler les effets 84
3.2 Dynamiques locales dans le réseau socio-sémantique 86
3.2.1 Similarité et interaction 86
3.2.2 Cohésion socio-sémantique locale 91
3.3 Dynamiques locales dans le réseau sémantique 93
3.3.1 Mesures d'occurrences 93
3.3.2 Mesures de co-occurrences 95

Nous tâcherons dans ce chapitre de décrire les dynamiques au sein des communautés de savoirs en nous plaçant dans le cadre général des réseaux socio-sémantiques. Les acteurs sont donc caractérisés à la fois par leur activité "sociale" de production de relations ou d'interactions avec d'autres agents et leur activité "cognitive" de production de contenus.

L'objectif est de comprendre comment ces deux dynamiques co-évoluent l'une avec l'autre, la production de contenus étant supposée couplée à la production d'interactions inter-individuelles. Ce couplage est perceptible au travers de l'influence sociale qu'un agent peut exercer sur un autre quant à son activité de production. Il peut encore se manifester par des processus de sélection qui peuvent présider à la création de nouvelles interactions, les effets d'homophilie entre agents augmentant *a priori* la probabilité que deux agents sémantiquement proches interagissent l'un avec l'autre. Les deux processus qui viennent d'être mentionnés -

influence sociale et sélection - ont des effets très différents sur la forme du système socio-sémantique résultant. *L'influence sociale* tend, à l'équilibre, à produire une homogénéité sémantique généralisée sur l'ensemble des agents, alors que *les effets de sélection* peuvent provoquer une "balkanisation" progressive des agents se regroupant dans des agrégats sémantiquement semblables. L'analyse du couplage de ces deux effets a été l'objet de nombreux modèles analytiques ou simulatoires (Axelrod, 1997b; Deffuant, 2006). Nous nous plaçons dans une perspective plus descriptive et empirique, et cherchons à *quantifier* ces effets à partir de l'analyse de bases de données longitudinales rendant compte de l'évolution de l'activité de communautés de savoirs. Nous appliquerons notre étude aux blogosphères politiques américaine et française présentées chapitre 2. Nous nous restreindrons donc à un seul type de communauté de savoirs et n'aborderons le cas des communautés scientifiques qu'incidemment en signalant, dans la section 3.1, les convergences et divergences existantes entre les résultats que nous présentons sur la blogosphère politique et ceux obtenus, par ailleurs, sur les communautés scientifiques. Si notre ambition première n'est donc pas de faire une analyse comparative exhaustive de nos deux cas d'étude, l'ensemble des analyses de ce chapitre forme néanmoins un grille générique de caractérisation de la morphogenèse d'une communauté de savoirs qui est parfaitement reproductible sur d'autres terrains.

Ce chapitre s'appuie (pour la majeure partie de la section 3.1) sur un travail réalisé en collaboration avec Camille Roth (Roth and Cointet, 2009). Le jeu de données employé est néanmoins légèrement différent, un léger décalage temporel (de l'ordre de deux semaines) ayant été rajouté.

La question du couplage entre les attributs cognitifs des agents et leur profil relationnel a été relativement négligée dans la modélisation des réseaux sociaux, les aspects cognitifs apparaissant généralement comme le parent pauvre de l'analyse traditionnelle des réseaux (voir section 1.2.1). Néanmoins, de nombreuses recherches ont tâché d'étudier les comportements individuels au sein des réseaux sociaux en attribuant aux agents un certain nombre de paramètres exogènes (entre autres âge, sexe, métier ou nationalité). Sans nécessairement introduire un cadre co-évolutionnaire, ces études se sont efforcées de mettre en évidence des comportements réguliers dans la dynamique du réseau en fonction de paramètres structurels endogènes (propres au réseau) mais aussi en fonction d'un certain nombre d'attributs qui caractérisent les individus (par exemple dans un contexte dyadique simple concernant l'évolution d'un réseau d'emails (Kossinets and Watts, 2006) ou dans un contexte hypergraphique sur la composition d'équipe d'entrepreneurs (Ruef et al., 2004)). Ces études considèrent généralement ces attributs comme extérieurs au réseau social, et les intègrent comme des variables purement exogènes et immuables.

Dans notre cadre, nous soutenons que les dimensions sociale et sémantique sont toutes deux dotées d'autonomie et co-évoluent l'une avec l'autre : si les attributs sémantiques sont susceptibles d'impacter la structure du réseau social, ce

tel-00464316, version 1 - 16 Mar 2010

dernier est également susceptible d'influencer la distribution des contenus sur le réseau. Seules des études très récentes, relevant plutôt de la "computational sociology" et s'intéressant principalement aux communautés en ligne, abordent la question de la co-évolution des mécanismes d'interactions sociales et la distribution des connaissances. Sans être exhaustif, la caractérisation des processus d'influence entre la distribution des contenus sur la structure des relations a notamment été abordée dans le cadre de la Wikipedia (Crandall et al., 2008a), de la blogosphère (Adar et al., 2004a), dans les "tagging system" (Cattuto, 2006) ou encore au sein de la plate-forme communautaire de partage de photos Flick'r (Cha et al., 2008).

Si l'on se réfère au chapitre 2, nous nous positionnons au niveau des dynamiques individuelles (évolution micro : plan inférieur du parallélogramme, voir figure 2.7). Nous étudierons les nouveaux événements qui animent nos réseaux (production de nouveaux contenus ou de nouvelles interactions entre deux périodes successives) afin d'en saisir les régularités par rapport à l'état de l'ensemble du réseau épistémique considéré au pas de temps précédent. Ainsi nous souhaitons par exemple pouvoir interroger les comportements individuels de création de nouveaux liens dans le réseau social en fonction de l'état du réseau socio-sémantique. Nous nous contenterons d'appréhender ces dynamiques individuelles en fonction d'observables locales du réseau épistémique associé. Ces propriétés peuvent aussi bien être des propriétés dites monadiques, relative à *ego* ou à *alter* comme le degré des noeuds, que des propriétés dyadiques, relatives à une couple de nœuds, comme la distance entre deux agents. Parallèlement, ces dynamiques locales sont intimement liées à l'émergence de propriétés de plus haut niveau structurant les trois réseaux qui composent notre réseau épistémique. Certaines de ces propriétés simples (distributions de degrés, clustering, etc.) seront placées en regard des régularités observées dans les dynamiques microscopiques. On tâchera enfin d'évaluer la stabilité relative de ces propriétés de haut niveau.

Comme on l'a déjà précisé dans l'introduction de ce chapitre, bien que nous distinguions clairement les trois types de réseaux servant à formaliser nos communautés de savoirs, nous les traitons de façon séparée et envisagerons successivement les dynamiques locales de nos trois réseaux : réseau social, réseau socio-sémantique, et réseau sémantique. Nous débutons notre analyse en tentant de caractériser la dynamique microscopique de notre réseau social $\mathcal{G}^{\mathrm{S}}$.

3.1 Dynamiques locales dans le réseau social

3.1.1 Attachements préférentiels

Différentes méthodes ont été introduites pour rendre compte des dynamiques de formation de liens dans les réseaux sociaux. Une famille importante de modèles appelés "p* models" vise, à travers un modèle markovien d'évolution du réseau, à estimer la part respective d'une grande gamme de phénomènes suscep-

tel-00464316, version 1 - 16 Mar 2010

tibles d'influencer la dynamique de production de liens. Cette méthode d'estimation statistique s'appuie sur une fonction décrivant la dynamique de formation de liens exprimée comme la somme d'effets supposés importants dans la morphogenèse du réseau (rôle de la réciprocité, homophilie, influence du degré, transitivité, etc.) (Snijders, 2001; Wasserman and Pattison, 1996). Ces méthodes d'estimation ont l'avantage de permettre de saisir, à travers un ensemble de coefficients, la part respective d'une grande gamme de phénomènes à l'œuvre dans la dynamique du réseau. Une des hypothèses liminaires posées par les modèles p* est que les agents agissent en maximisant une fonction d'utilité qui dépend des configurations structurelles de leur voisinage. Une large gamme de processus sociaux peuvent être incorporés au sein de cette fonction, au moyen d'un ensemble de fonctionnelles dont la forme doit être fixée *a priori*. Dès lors, l'analyse de données empiriques permet d'obtenir, grâce à différentes méthodes d'estimation de paramètres, le poids respectif de chacun des processus envisagés dans la dynamique de formation du réseau.

À cette méthode d'estimation statistique, nous préférons une approche moins holiste, puisque nous n'avons pas pour ambition de quantifier la part respective prises par chacun des processus susceptibles d'influencer la dynamique de production de nouveaux liens. Nous chercherons plutôt à caractériser de façon précise chacun de ces processus sous la forme d'une distribution de la propension de création d'un lien en fonction d'une variable explicative. Ainsi, nous n'avons pas besoin de postuler une forme fonctionnelle explicative *a priori*. Nous adopterons donc une approche fondée sur le calcul d'attachements préférentiels (Roth, 2005; Jeong et al., 2003a) qui permet de décrire fidèlement la phénoménologie de notre communauté de savoirs .

L'attachement préférentiel mesure simplement la propension d'apparition entre deux moments successifs (t et $t+dt$) d'un lien entre deux nœuds d'un réseau en fonction de propriétés monadiques (dépendant d'un seul des nœuds pouvant être *ego* ou *alter*) ou dyadiques (dépendant du couple de nœuds) propres au "réseau épistémique" en son entier considéré à l'instant t. Notre hypothèse est que l'état du réseau à un moment donné forme un ensemble de "contraintes" pour les interactions à venir. L'attachement préférentiel, ou la propension à une propriété m, permet de caculer les corrélations existantes entre les caractéristiques relationnelles ou cognitives des agents engagés dans l'interaction et la probabilité d'observer une telle interaction. Plus précisément, on cherche à estimer la probabilité conditionnelle qu'un nœud (respectivement une dyade soit un couple de nœuds) de propriété m reçoive un lien : $P(\mathrm{L}|m)$. Pratiquement on estime la fonction de propension $f(m)$ [1], proportionnelle à cette probabilité, qui permet de dire d'un agent (resp. d'une dyade) de propriété m qu'il (resp. elle) est $f(m)$ fois plus attractif (attractive) que ne le serait un nœud (resp. une dyade) quelconque si la distribution

1. Dans la littérature, la propension d'interaction au degré est fréquemment notée Π.

tel-00464316, version 1 - 16 Mar 2010

des nouveaux liens était aléatoire par rapport au paramètre m. On peut également dire qu'un agent (resp. une dyade) de type m sera plus attractif, toutes choses égales par ailleurs, d'un facteur $f(m)/f(m')$ qu'un agent (resp. qu'une dyade) de type m'. $f(m)$ est estimé en effectuant le quotient de la distribution d'une propriété m sur l'ensemble des nouveaux liens créés divisée par la distribution de m sur l'ensemble des couples de nœuds du réseau. La propension d'interaction f peut-donc être estimée en calculant $\hat{f}$ tel que

$$\hat{f}(m) = \frac{\nu(m)}{N(m)} \frac{N}{\nu}$$

où $\nu(m)$ est le nombre de nouveaux liens pointant vers un agent de type m (resp. le nombre de dyades de type m créées) pendant un intervalle de temps donné, et $N(m)$ mesure le nombre d'agents (resp. de dyades) de type m (ν et N désignent le nombre total de nouveaux liens créés sur la période considérée, et le nombre de nœuds (resp. de couples de nœuds) possibles).

Concrètement, nous estimons $f(m)$ en calculant, sur une série de périodes discrètes $[t+1,\, t+T]$, la propension par rapport à une propriété m observée à t de création d'un nouveau lien dans le réseau pendant cet intervalle. Nous avons fixé 8 périodes d'observation définies de la façon suivante : $\Big\{[t_k+1, t_k+T] \text{ tel que } t_k = 60+(k-1)T, T=7\Big\}_{k\in\{1,\ldots,8\}}$. Chaque nouvelle période d'une semaine permet de calculer un vecteur de propension pour les différentes valeurs possibles de la propriété m étudiée. Une période d'initialisation du réseau de 60 jours a été respectée avant toute mesure de propension ; la première mesure de propension se rapporte donc à une semaine d'activité du système mesurée par rapport à un réseau ancien de 2 mois. Dans son état final, le réseau agrège l'ensemble des interactions collectées pendant approximativement 4 mois. Les 8 mesures réalisées sont ensuite moyennées pour fournir un profil sur l'ensemble des deux derniers mois d'activité. Nous appliquerons un protocole similaire lorsque nous présenterons des données liées à la blogosphère politique française. Seul le découpage temporel diffère : la durée des périodes temporelles est fixée à 2 semaines, et le calcul des propensions s'étend sur 4 mois. .

3.1.2 Attachement préférentiel aux degrés : capital social et capital sémantique

3.1.2.1 Définition

Une première forme d'attachement préférentiel est l'attachement préférentiel au degré. Ce type de caractérisation a été popularisée par le modèle de morphogenèse de Barabási and Albert (1999) (dans ce modèle, un nouveau nœud entrant dans le réseau a une probabilité de se lier à un nœud de degré k proportionnelle à

tel-00464316, version 1 - 16 Mar 2010

$k : \Pi(k) \propto k$) même si de Solla Price (1976) avait déjà montré que ce type de processus (dit d'*avantage cumulé*) pouvait expliquer les distributions de degré en loi de puissance observées dans les réseaux de citation scientifiques (pour un éclaircissement historique sur les origines de la notion voir (Keller, 2005)). La très forte attractivité des nœuds de fort degré souvent résumée à l'adage "rich get richer" et plus précisément la dépendance (sub ou supra)-linéaire entre la propension d'attachement à un nœud et son degré observée dans une large gamme de réseaux réels a été largement commentée dans la littérature (Jeong et al., 2003a; Eisenberg and Levanon, 2003). Cette propriété a également alimenté nombre de modèles de morphogenèse de réseaux (Barabási et al., 1999; Almaas and Barabási, 2005), *l'attachement préférentiel au degré* étant, comme on l'a vu, considéré comme une explication plausible de la forme de la distribution de degré en loi de puissance des réseaux réels.

Dans le cadre qui nous occupe ici, celui des communautés de savoirs, nous définissons le degré d'un agent de deux façons. Le degré peut d'abord désigner la connectivité (Freeman, 1979) dans le réseau social, *i.e.* le nombre de voisins d'un noeud i donné, ou, plus précisément dans le cas orienté, le nombre de liens entrants [2] d'un nœud i. Il peut aussi désigner la connectivité dans le réseau socio-sémantique, soit le nombre de concepts mentionnés par un agent i.

Plus formellement, dans le réseau social $\mathcal{G}^{\mathbf{S}}$, qui est par définition croissant, on définit le voisinage d'un agent i comme $\mathcal{V}_t(i) = \{j \mid \exists t' \leq t,\ (j, i, t') \in \mathcal{R}^{\mathbf{S}}\}$, le degré social de i que l'on note $k(i, t)$ vaut $|\mathcal{V}_t(i)|$. Le degré social est donc simplement un décompte du nombre total d'agents ayant pointé vers i jusqu'à t. Cette définition de la connectivité d'un agent peut être rapprochée d'une mesure de position "dominante" dans le réseau, correspondant à un capital cumulé. En tant que telle, la connectivité peut être interprétée comme une mesure "d'autorité" (Coleman, 1988) [3]. Concernant notre cas d'étude, réduire l'autorité d'un blog à son degré entrant est d'autant plus justifié que la quasi-totalité des moteurs de recherche actuels s'appuient sur un "ranking" des sites essentiellement construit à partir du nombre de liens pointant vers un site donné. Compte tenu de l'importance des moteurs de recherche dans ces espaces virtuels à géographie variable, le degré entrant s'avère être une mesure approchée satisfaisante de la *visibilité* d'un site.

De façon équivalente, on définit dans le réseau socio-sémantique $\mathcal{G}^{\mathbf{SC}}$ le degré socio-sémantique d'un nœud i, ou son "capital sémantique" à un instant t, comme le nombre de concepts mobilisés par i à t (le réseau socio-sémantique étant également considéré comme croissant, il s'agit ici encore d'une mesure cumulée sur l'ensemble de la période d'observation jusqu'à t). On le définit formellement

2. Le nombre de liens sortants semble plus naturellement lié à l'activité de l'agent i qu'à une notion d'un capital agrégé, nous reviendrons plus tard sur cette notion d'activité.

3. Cette définition du capital social, même en supposant une approche purement structurale, est une notion largement débattue dans la littérature sur les réseaux sociaux (Burt, 1997, 2004). Nous aurons l'occasion de revenir sur cette définition plus longuement dans la partie III.

tel-00464316, version 1 - 16 Mar 2010

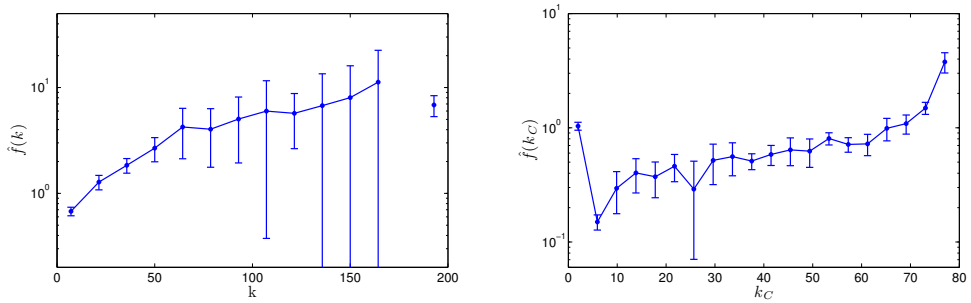


FIGURE 3.1: *À gauche :* propension d'interaction $\hat{f}$ au degré social k. *À droite :* propension d'interaction $\hat{f_{\mathbf{C}}}$ au degré sémantique $k_{\mathbf{C}}$. Moyenne sur 8 périodes. Afin de pallier le manque de données, celles-ci ont été regroupées dans des "bins" et tronquées dans le cas de l'attachement préférentiel au capital social($k < 200$).

ainsi : $k_{\mathbf{C}}(i,t) = |\mathcal{V}_t^{\mathbf{C}}(i)| = \{c \mid \exists t' \leq t,\ (i,c,t') \in \mathcal{R}^{\mathbf{C}}\}$.

3.1.2.2 Résultats

En utilisant la méthode et le protocole empirique présentés au paragraphe précédent (section 3.1.1), nous estimons les propensions d'interaction f et $f_{\mathbf{C}}$ sur notre réseau de blogs politiques américains par rapport aux capitaux sociaux et sémantiques k et $k_{\mathbf{C}}$, respectivement. Les fonctions de propension obtenues sont représentées figure 3.1 [4]. Nous observons que l'attachement préférentiel au degré est bien croissant avec le degré, c'est à dire que les nouvelles interactions pointent préférentiellement vers des agents ayant de plus grand capitaux aussi bien sociaux que sémantiques. La croissance de la probabilité d'attachement avec le capital social est cohérente avec les résultats classiques obtenus sur des réseaux de collaboration scientifique, et illustre un comportement partagé par un grand nombre d'autres réseaux (réseaux physiques de l'internet, réseaux de citation scientifique, réseau de co-appartenance à l'équipe d'un film pour ne citer que les principaux exemples traités dans (Jeong et al., 2003b)). Cette propriété est censée illustrer un processus d'accumulation du capital social dans la dynamique de la communauté. Il faut néanmoins noter que la propension d'interaction f observée semble plafonner après une première phase de croissance exponentielle. Ainsi sur notre réseau de blogs, l'attractivité exercée par les agents de forts degrés semble être limitée par un phénomène de saturation qui permet de "redistribuer les liens" de façon moins hétérogène sur l'ensemble des blogs. Notre analyse montre donc, *qu'en plus* d'une corrélation entre la création de nouvelles interactions et le capital social du blog cité, on observe également une corrélation liant la probabilité d'attachement au capital sémantique des blogs cités. Cette dépendance du comportement de citation au capital sémantique des agents a également été observé dans la dynamique d'une communauté scientifique (Roth and Cointet, 2009).

4. L'ensemble des intervalles de confiance est calculé pour une probabilité de 0.95.

tel-00464316, version 1 - 16 Mar 2010

3.1.2.3 Degré et activité

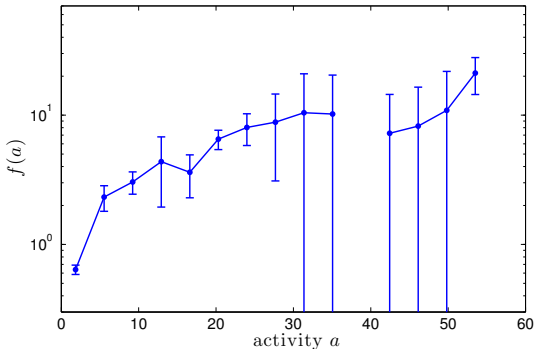


FIGURE 3.2: Attachement préférentiel en fonction de l'activité a, mesurée en nombre moyen de liens sortants créés hebdomadairement

Cette mesure de propension par rapport au capital social et sémantique doit être mise en regard de la propension d'un acteur à attirer des liens en fonction de son activité dans le réseau. Dans le cas des résaux symétriques, comme les réseaux de collaboration scientifique, le degré d'un nœud est, de façon mécanique, très fortement corrélé à son activité (estimée par exemple par le nombre d'articles publiés). Ainsi la mesure de propension au degré pourrait être la conséquence d'une hétérogénéité dans l'engagement des acteurs au sein de la communauté. Si on mesure l'attachement préférentiel à un nœud d'activité a estimé par exemple à travers la fréquence de production de liens sortants (l'activité est donc proportionnelle au degré sortant), on observe à nouveau (figure 3.2) un attachement préférentiel croissant avec ce même effet de saturation. Les blogs les plus actifs dans la communauté sont donc également ceux qui reçoivent le plus de liens (à titre d'illustration, un blog dont l'activité dépasse 30 liens hebdomadaires produits aura tendance a attirer 10 fois plus de liens que la moyenne). L'hétérogénéité des distributions de degré que nous observons dans ces réseaux sociaux pourraient donc simplement être la conséquence d'une hétérogénéité dans l'implication de ses membres.

Cette croissance de l'attractivité d'un blog en fonction de l'activité du blogueur au sein de la communauté est en fait très liée à la pratique même de construction d'une audience propre aux territoires virtuels. Comme l'analysait déjà Licoppe and Beaudoin (2002) à propos de sites personnels amateurs, la construction et la maintenance d'un site nécessitent "un investissement et un effort continuel" (au travers de mises à jour régulières et d'une activité de "gestion" des emails des visiteurs). Cardon and Delaunay-Teterel (2006) ont mené une série d'entretiens auprès de blogueurs et insistent sur les "tactiques" d'occupation de l'espace (par le biais de commentaires laissés sur des sites tiers notamment) mises en œuvre par les blogueurs pour attirer des commentateurs sur leur site et augmenter l'audience de

tel-00464316, version 1 - 16 Mar 2010

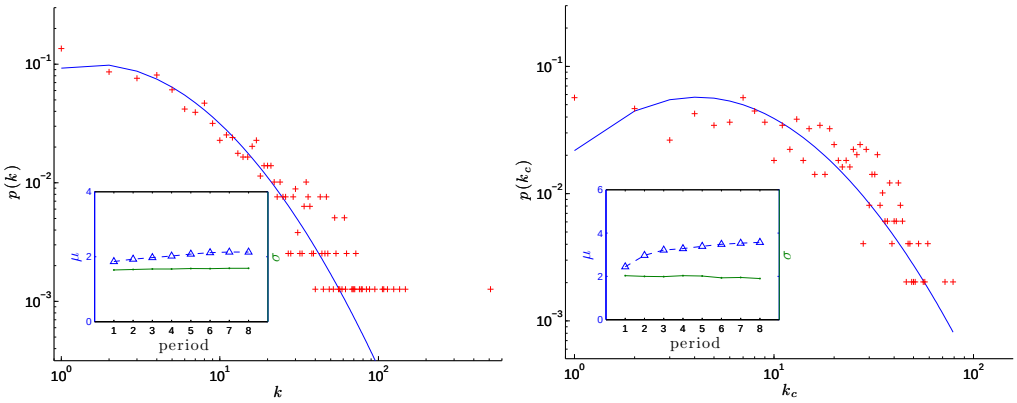


FIGURE 3.3: *À gauche :* distribution du " capital social", i.e. degré sortant dans le réseau social ; *à droite :* distribution du "capital semantique", i.e. nombre de concepts mobilisés dans le réseau socio-sémantique. *Croix rouges :* $p(k)$ (respectivement $p(k_{\mathbf{C}})$, proportion de blogs de degré k (resp., de degré $k_{\mathbf{C}}$), distribution calculées lors du dernier jour d'observation. *ligne continue :* meilleur fit log-normal. *encart :* évolution des paramètres du meilleur fit log-normal : μ (triangles bleus) et de σ (points rouges), sur l'ensemble des 8 périodes.

leur blog.

3.1.2.4 Propriétés macroscopiques émergentes associées

Ces résultats sur les dynamiques microscopiques peuvent être mis en relation avec les distributions macroscopiques de capitaux à travers le réseau et leur évolution. D'un point de vue macroscopique, la distribution du capital social nous renseigne ainsi sur la configuration du capital social des acteurs dans le réseau et la présence éventuelle d'une structure "hiérarchique" qu'une hétérogénéité de la distribution du capital social illustrerait.

Nous calculons pour notre réseau de blogs la distribution de connectivité dans le réseau social (k) et dans le réseau socio-sémantique ($k_{\mathbf{C}}$). Ce type de distribution ou du moins les queues de ces distributions sont généralement approchées par une loi puissance. Dans notre cas, le caractère relativement plat de la distribution pour les degrés faibles nous pousse à privilégier une approximation de ces distributions à l'aide d'une distribution de type log-normal [5].

Ces distributions se caractérisent d'abord par leur hétérogénéité, les valeurs sont distribuées sur plusieurs ordres de grandeur, et illustrent une très forte asymétrie entre agents. Un nombre faible mais non-négligeable d'agents reçoivent un nombre très conséquent de liens, alors, que près de la moitié des agents ont un

5. la forme analytique d'une densité log-normal est la suivant : $p(x) = \frac{e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}}$ (μ et σ^2 étant respectivement la moyenne et la variance de la distribution normale sous-jacente) ou sous une autre forme $p(x) = \lambda x^{\frac{-\ln x+2(\mu-\sigma^2)}{2\sigma^2}}$. σ est donc le paramètre prépondérant pour les x grands, c'est à dire pour apprécier la vitesse de décroissance de la queue de la distribution.

tel-00464316, version 1 - 16 Mar 2010

capital social ou sémantique inférieur à 10. Cette hétérogénéité reflète les hiérarchies inhérentes à cette communauté de blogueurs que l'on retrouve dans d'autres travaux sur les communautés de blogueurs (Adamic and Glance, 2005a; Shi et al., 2007) ou de façon plus générale sur le web (Hindman et al., 2003).

D'un point de vue dynamique (encart figure 3.3), on constate que les distributions sont relativement stables pour le capital sémantique et pour le capital social. σ, le paramètre principal caractérisant la vitesse de décroissance de la queue de la distribution, est extrêmement stable pour l'ensemble des 8 périodes temporelles envisagées. Cette stabilité dynamique peut s'expliquer en partie par la forme croissante des propensions au capital social tracées précédemment.

3.1.3 Attachement préférentiel à la distance sociale et sémantique

3.1.3.1 Définitions

Nous nous intéressons maintenant à la façon dont les dynamiques locales de formation de nouveaux liens dépendent de propriétés dyadiques du réseau. Nous nous concentrerons principalement sur une propriété simple : la distance entre deux nœuds. À nouveau, nous pouvons décliner la définition de cette distance sur les deux réseaux faisant intervenir des agents : le réseau social, et le réseau socio-sémantique. L'attachement préférentiel à la distance sociale permet d'apprécier les phénomènes transitifs au sein de notre communauté de savoirs tandis que la propension à la distance sémantique doit nous permettre de quantifier le caractère homophile ou hétérophile des interactions entre blogs.

La distance sociale entre deux nœuds i et j est notée $d(i,j)$; elle désigne la longueur du plus court chemin reliant ces deux nœuds dans le réseau social. Si aucun chemin ne permet de relier i à j (s'ils n'appartiennent pas à la même composante connexe), on considérera alors qu'ils sont à distance infinie l'un de l'autre.

La distance sémantique entre deux nœuds nécessite de faire appel aux profils sémantiques des blogs que l'on a défini dans le chapitre 2. Nous adoptons donc une distance standard de type *cosinus* pour comparer les profils sémantiques [6] de deux agents basée sur le traitement classique d'un corpus par le *tf.idf* (Salton et al., 1975)

Une fois les profils $\hat{\mathbf{w}}$ des agents calculés, la distance sémantique entre les agents i et j s'écrit alors simplement $\delta(i,j) = 1 - \frac{\hat{\mathbf{w}}_i \cdot \hat{\mathbf{w}}_j}{\|\hat{\mathbf{w}}_i\| \|\hat{\mathbf{w}}_j\|}$. Cette distance, comprise entre 0 et 1, permet de rendre compte des divergences "cognitives" existant entre agents. Dans le cas de notre réseau de blogs, cette distance permet de me-

6. Pour rappel, le profil d'un agent $\hat{\mathbf{w}}_i$ s'exprime donc sous la forme $\hat{\mathbf{w}}_{i,c} := \mathrm{tf}_{i,c} \cdot \mathrm{idf}_c = \frac{\mathbf{w}_{i,c}}{\sum_{c=1}^{|\mathbf{C}|} \mathbf{w}_{i,c}} \cdot \log \frac{B}{d_c}$ où $\mathbf{w}_{i,c}$ désigne le nombre d'occurrences d'un concept c dans l'ensemble des contenus produits par i, B est le nombre total d'agents, et d_c correspond au nombre d'agents mentionnant le terme c.

tel-00464316, version 1 - 16 Mar 2010

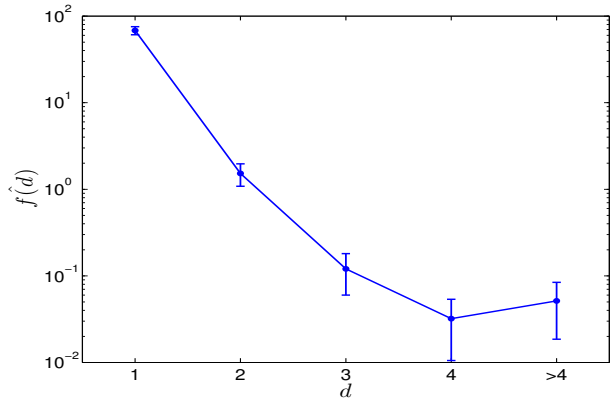


FIGURE 3.4: Propension $\hat{f}(d)$ de création d'un lien par rapport à la distance sociale d (les valeurs de propension pour des distances supérieures à 4 ont été agrégées en un seul point)

surer à quel point deux blogs traitent de sujets distincts (δ proche de 1) ou similaires (δ proche de 0). Par contre, elle ne permet en rien de signaler des lignes d'accord ou de désaccord entre agents, deux blogs spécialisés sur les questions de politique étrangère se retrouveront sans doute à proximité l'un de l'autre vis-à-vis des thématiques abordées (et auront donc un δ faible) mais peuvent très bien être en désaccord l'un avec l'autre.

3.1.3.2 Résultats

La fonction d'attachement préférentiel $\hat{f}(d)$ à la distance sociale d (figure 3.4) révèle une propension très forte à répéter les interactions avec ses propres voisins. Ainsi, la propension normalisée $\hat{f}(d)$ est très proche de 100 pour une distance égale à 1, tandis qu'elle est inférieure à $0,1$ pour deux agents à distance $d > 3$. Cela signifie, toutes choses égales par ailleurs, qu'un voisin auquel un blog est déjà lié sera préféré à un agent à distance 4 dans plus de 999 cas sur 1000. On peut également dire qu'il y a 100 fois plus de chances qu'un agent interagisse à nouveau avec un de ses voisins que ne le prédirait un modèle aléatoire de création de liens. Cette courbe permet également de rendre compte du phénomène de création de transitivité dans le réseau : chaque fois qu'un lien est produit entre deux nœuds à distance 2, un triangle est formé. La formation d'un lien à distance 2 reste bien plus probable, toutes choses égales par ailleurs, que la création d'une lien à une distance plus grande (la décroissance est quasi-exponentielle jusqu'à $d = 4$) indiquant une forte influence de l'environnement proche des blogueurs pour "sélectionner" ses futurs voisins.

On peut s'interroger légitimement sur la significativité de ces courbes de propensions par rapport à la distance sociale entre agents sachant que notre base constitue un échantillon (très réduit) du web. Ainsi, certains sites manquants pourraient fausser nos mesures ; leur absence entraînant une augmentation artificielle des distances entre blogs. Il y a deux réponses possibles à cette objection.

tel-00464316, version 1 - 16 Mar 2010

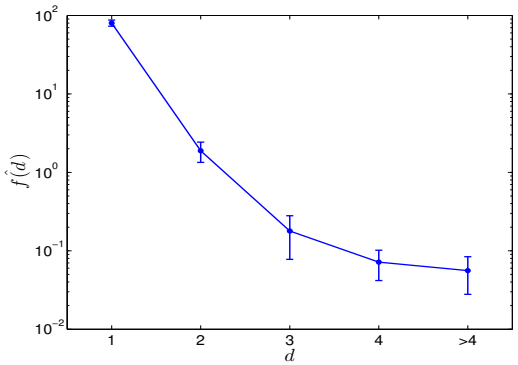


FIGURE 3.5: Propension $\hat{f}(d)$ par rapport à la distance sociale d calculée sur un échantillon de 500 blogs choisis aléatoirement à partir de notre base de données originale.

En premier lieu, notre sélection de sites a été réalisée à partir de critères d'appartenance à la communauté des blogueurs politiques ; ainsi, certains sites se situant à proximité (au sens où une courte séquence d'hyperliens permettrait de naviguer de l'un à l'autre) des blogs sélectionnés ont été exclus volontairement de l'échantillon. Si nous les avions inclus, nous aurions mesuré non pas une distance sociale dans un espace thématique et communautaire cohérent mais une distance dans le graphe du web qui serait moins pertinente compte tenu de notre objectif de caractérisation des communautés de savoirs [7].

En second lieu, nous avons effectué une mesure de contrôle de cette même propension à s'attacher à une distance sociale d donnée sur une sous-sélection de notre échantillon de blogs de départ, afin de contrôler la robustesse de nos résultats au processus de collecte (qui peut malgré tout être incomplet ou comporter des erreurs). En calculant cette propension sur une sélection aléatoire de 500 blogs (soit environ la moitié de notre échantillon de départ ; nous restreignons donc le réseau social et le calcul des distances afférentes aux seuls liens entre ces 500 blogs), nous observons, figure 3.5 un profil quasiment identique à celui obtenu sur la totalité des blogs. Cette stabilité s'explique en partie par le protocole de mesure employé (un attachement préférentiel est un ratio entre ce que laisse apparaître un comportement réel observé vs. un modèle de comportement aléatoire. En limitant notre jeu de données à 500 blogs, l'ensemble des distances entre blogs peut avoir été modifiée sans que ne soit bouleversé le rapport existant entre la distribution des distances entre blogs entrant en interaction et la distribution de distances entre l'ensemble des couples de blogs.).

Ce résultat nous conforte donc sur l'existence d'un processus "d'homophilie structurelle" au sein de notre communauté. On a observé la même tendance au

7. Ainsi, pour prendre un exemple extrême, nous ne souhaitons pas que la présence de sites tels qu'*adobe acrobat* (extrêmement fréquent sur le web) dans le voisinage de nos blogs rentre en compte dans les mesures que nous souhaitons développer au niveau de la communauté des blogueurs politiques américains.

tel-00464316, version 1 - 16 Mar 2010

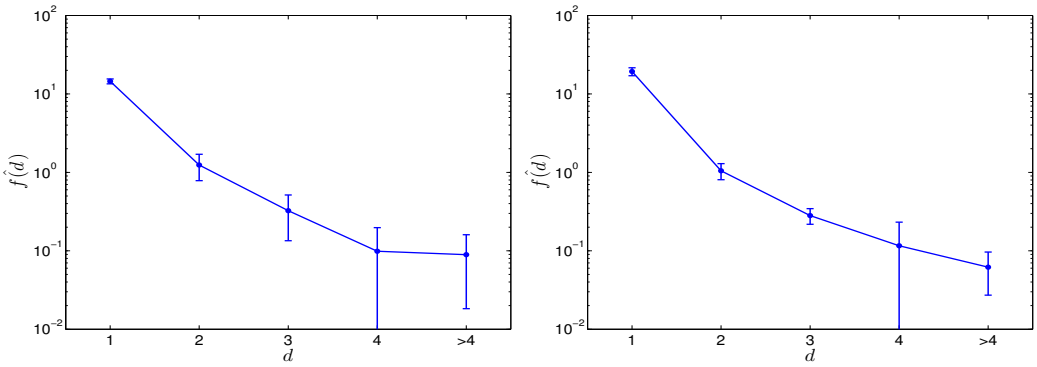


FIGURE 3.6: Propension $\hat{f}(d)$ par rapport à la distance sociale d (les valeurs de propension pour des distances supérieures à 4 ont été agrégées en un seul point) dans le réseau de citations (*à gauche*), et dans le réseau de commentaires (*à droite*)

sein de la communauté de blogs politiques français dont les courbes de propension pour les réseaux de citations et de commentaires sont représentés figure 3.6. Les dynamiques des réseaux de citation et de commentaire se caractérisent par des propension à la distance sociale très ressemblantes. On note néanmoins, que la tendance à la répétition des interactions est moins marquée dans ce jeu de données. Enfin, le même type de comportement a été observé sur un réseau de collaboration scientifique, les chercheurs ayant naturellement tendance à co-publier avec d'anciens collaborateurs ou avec des chercheurs à faible distance dans leur réseau social Roth (2006).

La figure 3.7 représente la propension de création d'un lien à une distance sémantique donnée. Cette courbe montre une très forte tendance des blogueurs à *l'homophilie* (Lazarsfeld and Merton, 1954; McPherson and Smith-Lovin, 2001; Ruef et al., 2004). Les blogs dont les profils sémantiques sont éloignés ont une propension systématiquement plus faible à se citer que des blogs dont les bagages sémantiques sont proches. Notre objectif étant d'éclairer les effets de sélection qui président à la création de nouvelles interactions, cette courbe est calculée en excluant les interactions répétées. Seuls sont considérés les couples de blogs n'étant jamais entrés en contact préalablement dans le réseau de citation, ainsi, nous éliminons le biais éventuellement induit par des interactions entre blogs déjà voisins susceptibles d'avoir une distance sémantique moindre à cause d'effets dûs à l'influence sociale que les blogs voisins exerceraient l'un sur l'autre.

Il est intéressant de comparer ce résultat à d'autres mesures d'attachement préférentiel à la distance sémantique obtenues sur d'autres types de réseaux. Ainsi Roth (2008a) a constaté dans un réseau de collaboration scientifique une forte tendance à l'homophilie comme on l'observe dans notre cas, mais doublée d'une légère hétérophilie vis-à-vis des distances faibles ce qui semble indiquer que les chercheurs tendent naturellement à collaborer avec des chercheurs sémantiquement proches, mais pas parfaitement identiques. Ainsi la propension d'interaction

tel-00464316, version 1 - 16 Mar 2010

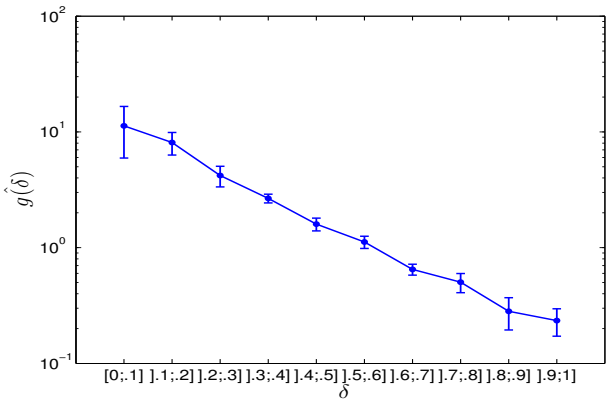


FIGURE 3.7: Propension $\hat{g}(\delta)$ de création d'un lien par rapport à la distance sémantique δ.

est maximale pour une distance sémantique faible mais non nulle avant de chuter brutalement pour des δ plus importants.

Un effet similaire a également été observé sur notre réseau de blogs politiques français. Si nous traçons l'attachement préférentiel à la distance sémantique dans les réseaux de citation (liens de post) et d'interaction (liens de commentaires) : figure 3.8, nous observons que le comportement de création de commentaires (courbe rouge) est nettement plus hétérophile que celui de création de liens de citation (courbe verte). Cet effet illustre combien ces mesures de comportements locaux varient selon le type de processus sociaux représenté par le réseau : un lien de citation implique généralement une forme de continuité entre les centres d'interêts de deux blogs qui ont donc, au moins partiellement, des thématiques recouvrantes, le commentaire peut avoir comme fonction de servir de lieu de débat où s'expriment les différences entre blogueurs. D'après les courbes de propension, l'activité associée au fait de commenter le billet d'un autre blog semble encourager la formation de liens entre des couples de blogs relativement distants sémantiquement, contrairement au comportement de citation qui a tendance à encourager la mise en rapport de blogs semblables, et donc à créer des agrégats sociaux très focalisés sémantiquement. L'activité de commentaire semble donc induire une plus grande hétérophilie que le comportement de citation qui tend vers une uniformisation des groupes (Davis, 1963). Ces comportements homophiles microscopiques peuvent notamment expliquer la division structurelle très marquée entre républicains et démocrates observée au sein de la blogosphère politique américaine par (Adamic and Glance, 2005b).

En suivant l'hypothèse d'une coévolution des contenus et des relations entre les agents (Crandall et al., 2008b; Roth, 2005), nous pouvons également nous interroger, cette fois-ci de façon statique, sur la façon dont la distance sémantique est distribuée sur l'ensemble des agents voisins dans le réseau de citation *vs.* sur l'ensemble des dyades possibles. Les distributions qui ont été représentées figure 3.9 révèlent une relative homogénéité de la distribution des distances sémantiques

tel-00464316, version 1 - 16 Mar 2010

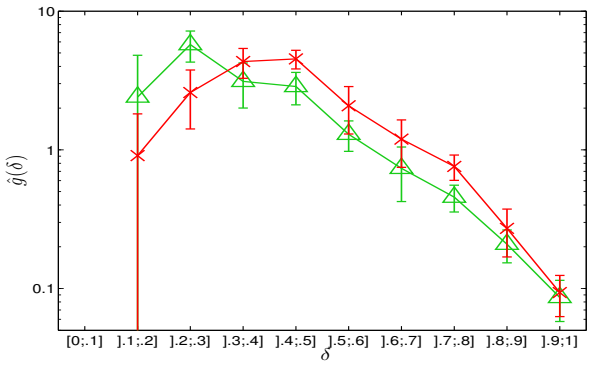


FIGURE 3.8: Propensions $\hat{g}$ de création d'un lien de citation (en vert)) et de création d'un lien de commentaire (en rouge), par rapport à la distance sémantique δ dans la blogosphère politique française.

entre les couples d'agents liés, alors, qu'en *moyenne*, la distribution sur l'ensemble des dyades potentielles est fortement portée sur les valeurs de δ importantes. Ainsi près de 30% des couples d'agents dans le réseau de citation sont à une distance sémantique inférieure à 0.3, ceux-ci représentent pourtant moins de 4% de l'ensemble des dyades possibles. Cette dernière courbe est cohérente avec les observations précédentes sur l'homophilie accrue des dyades étant déjà entrées en contact les unes avec les autres. Nous reviendrons ultérieurement (section 3.2.1) sur la caractérisation précise de l'influence sociale entre deux blogs entrant en interaction.

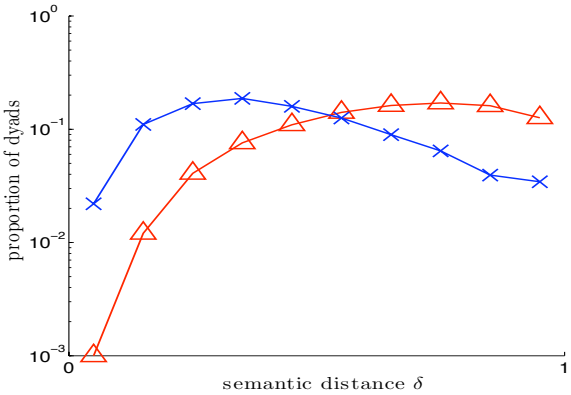


FIGURE 3.9: Distributions des distances sémantiques dans la dernière période dans le réseau de citation *en bleu (croix)* et sur l'ensemble des dyades possibles *en rouge (triangles)*.

3.1.4 Motifs cohésifs locaux

Nous tâchons maintenant de caractériser les comportements des agents en complétant notre analyse des dynamiques de création de liens par la caractéri-

tel-00464316, version 1 - 16 Mar 2010

sation d'agrégats locaux limités à trois nœuds. Cette caractérisation, très simple, consiste à faire le décompte des motifs triadiques observés dans nos réseaux.

La cohésion sociale locale peut en effet être appréhendée en énumérant les triades existantes dans le réseau social (Snijders and Stokman, 1987; Holland and Leinhardt, 1976; Milo et al., 2004). 16 configuration de sous-graphes de taille 3 distinctes peuvent être dénombrées dans le cas orienté. Cette approche démographique permet d'évaluer, en cas de sur-repéresentation ou de sous-représentation de certaines triades, la validité de certaines hypothèses quant aux processus structurels sous-tendant la formation d'un réseau (Davis and Leinhardt, 1972).

Par exemple, l'analyse de réseaux sociaux s'est largement penchée sur la caractérisation au travers des triades, des processus sociaux transitifs ou de réciprocité (Davis, 1963). Dans notre cas, nous souhaitons interroger la cohésion locale soit la tendance pour les voisins d'un agent d'être également voisins les uns des autres [8]. On peut quantifier la cohésion locale en nous concentrant sur deux types d'observation :

- le coefficient de clustering $c_3(i)$ d'un agent i qui s'exprime en calculant la proportion de liens existants entre l'ensemble des couples de voisins de i :
  $c_3(i) = \frac{|(j,j') \subset \mathcal{V}(i) \text{ tel que } j' \in \mathcal{V}(j)|}{k(i)\,(k(i)-1)}$
- le coefficient de transitivité t_3 d'un agent i défini comme la proportion, parmi les couples d'agents (j', i) reliés par un chemin de longueurs 2 pointant vers i (i.e. parmi les agents dont i est un voisin de voisin) de liens directs de j' vers i :
  $t_3(i) = \frac{|(j,j') \subset \mathcal{V}(i) \text{ tel que } j' \in \mathcal{V}(j)|}{|(j,j') \text{ tel que } j \in \mathcal{V}(i) \text{ et } j' \in \mathcal{V}(j)|}$

Ces deux quantités mesurent l'importance de deux processus distincts : respectivement, la tendance qu'ont deux voisins d'un même nœud à entrer en contact, et la tendance d'ego à être cité par des blogs appartenant au voisinage de ses voisins. Pour plus de clarté, ces motifs ont été schématisés figure 3.10 (haut). Dans notre entreprise générale d'appréhension des caractéristiques structurelles de notre réseau dans ses dimensions sociales et sémantiques, nous nous intéressons également à la corrélation existant entre le capital social et sémantique d'un nœud et la probabilité qu'il a de participer à une triade donnée. Nénamoins, contrairement aux mesures de formation de dyades, la distribution de ces motifs ne semble pas être affectée par le capital sémantique, et nous avons donc seulement représenter la courbe d'évolution de nos deux coefficients par rapport au capital social k. L'évolution de leur valeur moyenne ($c_3 = \langle c_3(i) \rangle$, $t_3 = \langle t_3(i) \rangle$) dans le temps figure également en encart.

Nous avons également vérifié en simulant une série de réseaux aléatoires uniformes (Erdös and Rényi, 1959) que les valeurs moyennes de clustering et de transitivité mesurées sont supérieures d'un ordre de grandeur au cas aléatoire. Cette

8. Pour rappel $j \in \mathcal{V}(i)$ signifie qu'il existe un lien pointant de j vers i.

tel-00464316, version 1 - 16 Mar 2010

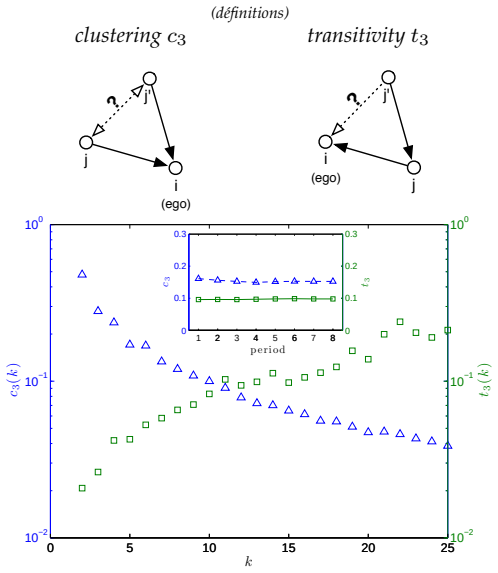


FIGURE 3.10: Coefficients de clustering c_3 (triangles bleus) et transitivité t_3 (carrés verts)en fonction du capital social k (*encart :* évolution des valeurs moyennes sur les 8 périodes).

sur-représentation de motifs triangulaires est exemplaire des réseaux de terrain et *a fortiori* des réseaux sociaux (Newman and Park, 2003a) et a donné lieu à de nombreux modèles de reconstruction de réseaux (Watts and Strogatz, 1998; Pattison et al., 2000).

Le coefficient de clustering c_3 qui mesure la densité dans le voisinage immédiat d'ego est décroissant avec le capital social ce qui indique que le voisinage des agents les plus connectés a tendance à être moins clusterisé. Cette tendance est corroborée par des études antérieures décrivant la décroissance du clustering avec le degré d'un nœud, mais cette décroissance est difficile à intepréter plus précisément, car elle résulte en partie d'un biais systématique dans toute mesure du clustering local (voir (Soffer and Vázquez, 2005)). Le coefficient de transitivité t_3 a un comportement contraire, il est croissant [9] avec k. Les agents dotés d'un fort capital social semblent attirer une proportion plus importante de liens provenant de leur voisins de voisins que des agents de capital social moindre. De plus, cette tendance ne semble pas être générique à l'ensemble des réseaux, on ne l'observe notamment pas dans le cas d'un réseau de collaboration scientifique (Roth and Cointet, 2009). Dans le cas des blogosphères politiques, toutes choses égales par ailleurs, le capital social d'un agent agit comme une propriété particulièrement favorable pour créer de la transitivité, *i.e.* se faire citer par des agents à distance 2.

Ces deux types de mesure, c_3 et t_3 bien que caractérisant des comportements *a priori* distincts, sont relativement stables dans le temps (cf. encart figure 3.10 qui mesure l'évolution des valeurs moyennes des deux coefficients sur les 8 périodes

9. t_3 ne souffre *a priori* pas du même biais de mesure que c_3.

tel-00464316, version 1 - 16 Mar 2010

étudiées). Cette stabilité est d'autant plus remarquable que le réseau se densifie largement durant cet intervalle comme l'attestent les courbes d'évolution du nombre de liens figure 2.6.

3.1.5 Capitaux, homophilie sociale et sémantique, découpler les effets

Nous avons rendu compte d'un certain nombre de régularités dans les comportements de création de liens au sein du réseau social. Ainsi un agent de capital social important, situé à une distance sociale et sémantique proche semble constituer un candidat idéal pour former une nouveau lien. Il est moins évident, compte tenu des corrélations très fortes qu'entretiennent les observables les unes avec les autres (distances sociales et distances sémantiques sont notamment clairement corrélées, des agents à distance 1 par exemple sont nettement plus proches sémantiquement que la moyenne (cf. figure 3.9)), d'apprécier, d'une part l'importance respective des effets les uns par rapport aux autres, d'autre part la façon dont ces propriétés peuvent interférer constructivement ou destructivement les unes avec les autres. Notre objectif est donc, dans une perspective plus holiste, d'illustrer ces couplages.

Pour ce faire nous restons dans le cadre de l'estimation de fonctions d'attachement préférentiel mais en l'étendant à des propriétés mixtes. Beaucoup de combinaisons entre propriétés sont envisageables, mais nous nous concentrerons sur la façon dont la propension par rapport à la distance sociale d, est couplée à une autre propriété m. Autrement dit, étant donné un agent i, nous souhaitons estimer la propension que cet agent se lie à un autre agent j connaissant la distance $d(i,j)$ qui les sépare dans le réseau social *et* une autre propriété m liée à l'un des nœuds ou à la dyade (i,j) : $P(\mathrm{L}|d,m)$. Nous limitons notre étude à l'examen de deux propriétés m : le capital social k_j et la distance sémantique $\delta(i,j)$, nous noterons $p(d,k)$ et $p(d,\delta)$ les deux fonctions de propensions associées.

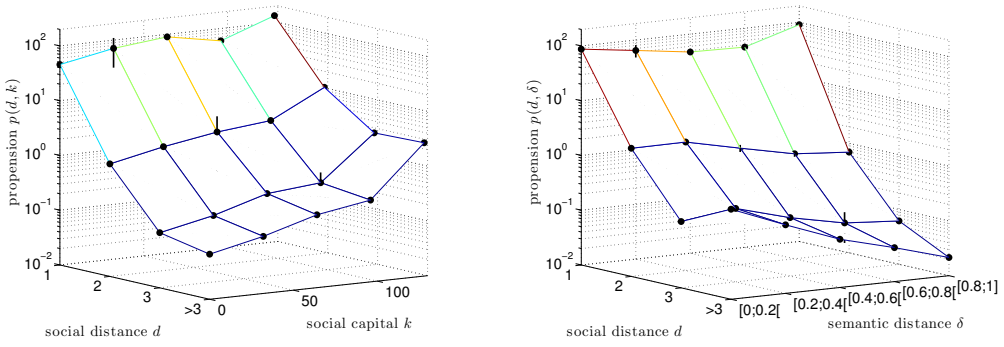


FIGURE 3.11: Propensions d'interaction par rapport à la distance sociale couplée au capital social :$p(d,k)$ (*gauche*), et à la distance sémantique $p(d,\delta)$ (*droite*)

Les courbes de propension calculées, qui prennent la forme de surfaces tridi-

tel-00464316, version 1 - 16 Mar 2010

mensionnelles (figure 3.11), illustrent une forme de prééminence de la distance sociale sur le capital social et la distance sémantique. Dans les deux cas, le paramètre qui semble influencer le plus fortement les valeurs d'attachement préférentiel est bien d soit la distance sociale. Le capital social k a un effet mineur sur les interactions répétées, et n'induit des inflexions de la courbe de propension que pour les couples de blogs à distance élevée l'un de l'autre. Il en va de même de la distance sémantique δ dont les effets sur la propension d'interaction n'apparaissent que pour une distance sociale suffisante. On note d'ailleurs que pour les blogs à distance 1, la propension d'interaction est décroissante puis croissante en fonction de la distance sémantique, ce qui traduit une tendance à répéter des interactions avec des blogs soit très semblables soit très différents [10].

Inversement, le rôle de d sur la propension est d'autant plus discriminant que le capital social est faible, et que la distance sémantique est importante. Ces observations tendent à montrer qu'au sein de ces communautés, la proximité dans le réseau social est déterminante pour le choix des nouvelles interactions. Hors de l'horizon des proches voisins d'autres critères de choix ont droit de cité tels que la similarité sémantique ou la popularité d'une source. Ces courbes pourraient résulter d'un compromis entre deux modes de navigation au sein de ces communautés en ligne. Dans un voisinage proche, les liens hypertextes entre blogueurs peuvent encore guider l'exploration. Au delà, les moteurs de recherche prennent le relais de cette exploration locale et induisent une plus grande dépendance au degré et à la similarité sémantique Cet effet est peut-être amplifié par la navigation sur internet fortement guidée par les moteurs de recherche, ces derniers privilégiant des critères de popularité (mesurés au travers du nombre des liens entrants) et de pertinence par rapport à une requête (on peut imaginer qu'un blogueur fera des recherches sur des thématiques proches de celles dont il alimente son blog).

Nous avons mis en évidence un certain nombre de régularités dans la dynamique du réseau social semblant fortement dépendre de la structure du réseau épistémique décrivant les dynamiques agrégées de création de liens dans la blogosphère politique américaine. L'espace virtuel décrit par ces blogs est loin d'être homogène et ne semble pas pouvoir être simplement réduit à un modèle hiérarchique. La structure du réseau social importe pour déterminer les interactions futures, de même que la distribution des contenus dans cet espace.

10. les profils d'attachement préférentiel à $d = 1$ sont bien différents de ceux calculés figure 3.4, en effet, nous calculons dans le cas présent la propension à créer un lien à une distance sémantique donnée δ_0 *et* à répéter une interaction qui est proportionnelle à $P(L|d = 1, \delta = \delta_0) = \frac{\nu(d=1,\delta=\delta_0)}{N(d=1,\delta=\delta_0)}$, précédemment, nous mesurions simplement *pour le sous-ensemble* des interactions non répétées, la propension pour un blog de se connecter à un autre blog à une distance sémantique δ_0 proportionnelle à $P(L, d > 1|\delta = \delta_0) = \frac{\nu(d>1,\delta=\delta_0)}{N(d>1,\delta=\delta_0)}$. La propension représentée figure 3.4 correspond donc à une agrégation de l'ensemble des valeurs observées pour $d > 1$.

tel-00464316, version 1 - 16 Mar 2010

3.2 Dynamiques locales dans le réseau socio-sémantique

Nous souhaitons maintenant appréhender les dynamiques à l'œuvre dans le résau socio-sémantique liant les acteurs aux concepts qu'ils emploient. Nous cherchons toujours à illustrer et caractériser, au niveau des comportements individuels, la co-évolution entre les dynamiques de production de contenus et les dynamiques sociales. Pour tenter d'évaluer l'influence potentielle du contexte social sur l'évolution de "l' état cognitif" d'un acteur, nous tenterons, à travers une approche purement dyadique, de quantifier l'évolution de la similarité sémantique entre deux agents amenés à interagir en amont et en aval du moment de l'interaction.

Dans un second temps nous nous intéresserons, à un niveau plus mésoscopique, à la présence de motifs socio-sémantiques caractéristiques de la tendance qu'ont certains concepts à être de façon plus ou moins systématique conjointement employés par des agents.

3.2.1 Similarité et interaction

Nous nous intéressons à nouveau à la co-évolution entre les attributs cognitifs des agents et la topologie du réseau social support des interactions entre ces agents en tâchant, cette fois ci, non pas d'apprécier des effets de "sélection" (à savoir la façon dont de nouvelles interactions sont produites dans le réseau social en fonction de propriétés dyadiques ou monadiques du réseau social et du réseau socio-sémantique) mais de caractériser l'influence exercée sur un agent par son voisinage en terme de modification de son profil sémantique. La très forte proximité sémantique des couples d'agents voisins dans notre réseau de citation (voir figure 3.9) peut en effet avoir deux causes : (i) l'homophilie sémantique mise en évidence figure 3.4 a tendance à créer des effets de sélection qui connectent préférentiellement des agents sémantiquement semblables, (ii) l'influence sociale exercée entre les voisins dans le réseau social est susceptible d'encourager certains acteurs à adopter des comportements similaires à ceux de leurs voisins. La question du découplage des effets de sélection et d'influence sociale a récemment bénéficié d'un certain nombre d'avancées méthodologiques s'appuyant sur l'analyse de données relationnelles longitudinales (Burk et al., 2008; Steglich et al., 2004).

Nous nous plaçons dans un cas simplifié et ne tâcherons de mettre en évidence que des influences de type dyadique ; la question que nous poserons est donc la suivante : comment les propriétés cognitives d'un couple d'agents et plus précisément leur distance sémantique est modifiée en amont et en aval d'un premier événement d'interaction ?

Contrairement à notre cas d'étude, dans lequel les contenus publiés par les blogueurs définissent leur profil cognitif, Crandall et al. (2008a) décrivent l'activité des agents à partir des pages qu'ils éditent dans la *Wikipedia*, la méthode de caractérisation des processus d'influence à laquelle nous nous livrons est néanmoins

tel-00464316, version 1 - 16 Mar 2010

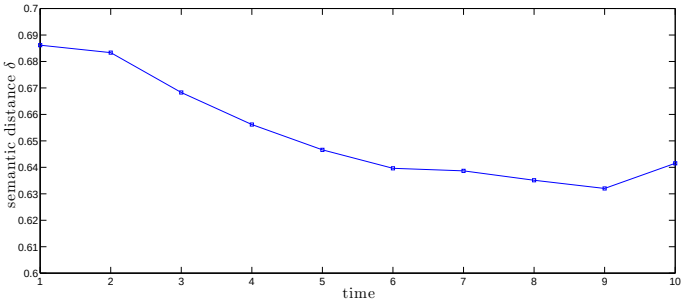


FIGURE 3.12: Evolution moyenne de la distance sémantique entre deux blogs semaines 1 à 10 ($T = 7$) .

très semblable à la leur.

Le protocole de mesure employé est le suivant : pour chaque couple d'agents (i, j) et pour une périodicité T donnée (dans la suite nous utilisons une granularité temporelle d'un jour ou d'une semaine), nous calculons le vecteur $\boldsymbol{\delta}_{ij} = (\delta_{ij}(t_k))_k$ qui prend comme valeurs la distance sémantique entre les agents i et j aux temps $\{t_k\}_k$ tel que $t_k = 46 + (k-1)T$ [11]. $\boldsymbol{\delta}_{ij}$ décrit l'évolution temporelle de la distance sémantique entre i et j. Doté de ces vecteurs, notre objectif est de comprendre comment $\boldsymbol{\delta}_{ij}$ est modifié lorsque les deux agents i et j entrent en interaction à un instant t.

Si l'on calcule l'évolution de la distance sémantique moyenne entre l'ensemble des couples de blogs au sein de notre système au cours du temps, on constate que cette distance est légèrement décroissante (voir figure 3.12). Cette décroissance pourrait signaler une polarisation d'ensemble de la communauté de blogueurs, mais, nous souhaitons bien ici examiner les effets locaux induits par les influences entre voisins deux à deux. Aussi pour pouvoir comparer de façon non biaisée les profils d'évolution de la distance sémantique entre couples de blogs, nous avons préféré normaliser l'ensemble des $\{\boldsymbol{\delta}_{ij}\}_{ij}$ mesurés par rapport aux valeurs moyennes prises au cours du temps par la distance sémantique : $\boldsymbol{\delta_0} = \frac{\sum_{i \neq j} \boldsymbol{\delta}_{ij}}{\sum_{i \neq j} 1}$. L'ensemble des vecteurs $\{\hat{\boldsymbol{\delta}}_{ij}\}_{ij}$ correspond donc à l'évolution temporelle d'un taux de dissimilarité sémantique par rapport à une évolution moyenne calculée sur l'ensemble des couples d'agents. Une valeur de $\hat{\delta}_{ij}(t)$ de 0.8 signifiera donc que les contenus produits par un couple d'agents considéré (i, j) sont moins dissemblables de 20% en moyenne au moment t qu'un couple d'agents choisi aléatoirement sur l'ensemble des dyades possibles.

Nous sélectionnons maintenant pour une série de temps d'observation $\{t_k\}_{k \geq 1}$

11. Nous avons étendu la période temporelle d'observation de deux semaines afin d'être à même de décrire des évolutions de distance sémantique sur une période plus longue, les résultats de la section précédente sont naturellement robustes à ce changement de référentiel.

tel-00464316, version 1 - 16 Mar 2010

les couples d'agents entrant en relation l'un avec l'autre (quel que soit le sens du lien [12]) *pour la première et unique fois au temps* t_k [13]. Nous calculons alors la moyenne des vecteurs de dissimilarité sur l'ensemble de ces couples [14]. Pour chaque période t_k, et pour l'ensemble des dyades associées à une première interaction à cette date, on obtient ainsi un profil d'évolution moyen de la distance sémantique sur toute la période d'observation noté $\hat{\boldsymbol{\delta}}^{t_k}$.

Ces profils d'évolution ont été tracés figure 3.13 pour quatre valeurs de t_k, avec une périodicité d'une semaine ($T = 7$). Cette figure appelle plusieurs commentaires. D'une part, on constate, que pour l'ensemble des temps d'interaction $\{t_k\}_{1\geq k\geq 4}$, et pour l'ensemble des moments de mesure de la distance sémantique normalisée, celle-ci est largement inférieure à 0.8. Même très en amont de leur première interaction, deux blogs amenés à interagir l'un avec l'autre dans un futur proche (ici de 1 à 4 semaines) ont déjà une distance sémantique sensiblement inférieure à la moyenne. D'autre part, on constate qu'en moyenne la distance sémantique normalisée décroît brusquement au moment de l'interaction (le moment de l'interaction a été matérialisé par des cercles rouges) ainsi qu'en amont et en aval de l'interaction. La portée de l'interaction sur le comportement de production de contenus des deux blogs ne semble donc pas se limiter au seul moment de création du lien. Naturellement, la création du lien n'exerce pas de causalité rétrospective sur l'activité des blogueurs, cette décroissance doit plutôt être interprétée comme une similarité sémantique minimale à franchir avant que ne s'établisse une relation (dans le cadre des paramètres fixés figure 3.13, la distance sémantique normalisée est inférieure à 0.7 au moment de l'interaction (il faut noter que ces profils d'évolution agrègent chacun plusieurs milliers de dyades, et c'est bien un comportement moyen qui est caractérisé ici)). La création effective d'un lien entre deux blogs pourrait être l'aboutissement d'une phase d'apprentissage des deux blogs éventuellement médiatisée par la lecture d'un blog tiers, ce qui expliquerait la présence d'une influence antérieure au moment de l'interaction.

Afin de mieux saisir l'influence de l'événement d'interaction sur la similarité de deux blogueurs en amont et en aval de celui-ci, nous avons réalisé le même calcul en décalant l'ensemble des séries temporelles $\hat{\boldsymbol{\delta}}^{t_k}$ correspondant à chaque t_k de façon à ce que le moment de l'interaction t_k de chaque série coïncide en une même origine temporelle t_0. Nous faisons l'hypothèse, compte tenu de la longue

12. en pratique plus de 90% de ces nouvelles interactions sont en réalité symétriques pour $T = 7$.

13. Ne considérer que les blogs ayant interagi pendant une seule période permet *a priori* de définir de façon univoque le moment de l'interaction. Néanmoins, affaiblir la contrainte et envisager les autres couples de blogs interagissant sur plusieurs périodes en choisissant la période de première interaction comme origine temporelle fournit des résultats qualitativement similaires. Rappelons également qu'un mois d'activité en amont a permis de contrôler que les blogs qui interagissent pour la première fois le jour 1 ou le jour 20, n'ont pas interagi durant le mois et demi qui avait précédé.

14. Plus formellement, pour chaque temps t_k, on calcule : $\hat{\boldsymbol{\delta}}^{t_k} = \sum_{(i,j)\in M(t_k)} \frac{\hat{\boldsymbol{\delta}}_{ij}}{|M(t_k)|}$ où l'ensemble des dyades $M(t_k)$ considérées sont définies telles que : $M(t_k) = \{(i,j), \mathbf{P}_\infty(i,j) = P_{t_k}(i,j) > 0\}$

tel-00464316, version 1 - 16 Mar 2010

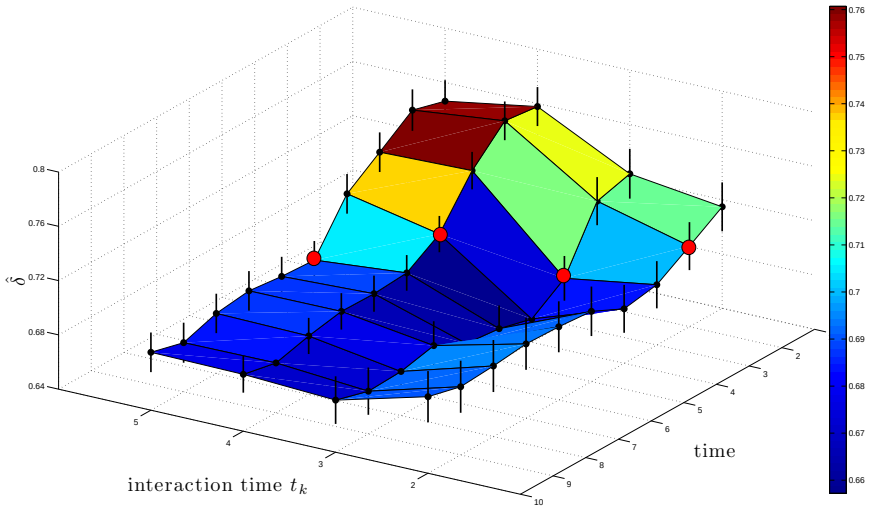


FIGURE 3.13: Evolution moyenne sur les 10 semaines de la distance sémantique normalisée $\hat{\boldsymbol{\delta}}^{t_k}$ entre deux blogs ayant interagi pour la première fois au temps t_k. Les temps d'interaction sont matérialisés par un point rouge pour chaque série. À chaque temps de première interaction t_k ($\{t_k\}_{k=[2...5]}$ tel que $t_k = 46 + (k-1)7$), on associe le vecteur d'évolution $\hat{\boldsymbol{\delta}}^{t_k}$ qui correspond donc à un nouvel ensemble de dyades.

période d'observation du réseau qui précède la mesure (un mois et demi), que le profil d'évolution de deux blogs amenés à interagir le $10^{ème}$ jour diffère peu de celui d'un couple interagissant le 1^{er} jour moyennant notre changement d'origine temporelle. Comme Crandall et al. (2008a), nous mesurons donc la moyenne des profils d'évolution de la distance sémantique normalisée en prenant comme origine des temps le jour de l'interaction t_k. Ce changement de référentiel permet de représenter en une seule courbe l'ensemble des profils d'évolution, c'est ce gain d'information qui permet également de réduire le grain d'observation à 1 jour tout en garantissant une significativité satisfaisante des résultats. On a représenté figure 3.14 l'évolution du vecteur d'évolution de la distance sémantique normalisée pour l'ensemble des dyades rentrant en interaction entre le jour 1 et le jour 13 en opérant ce changement de référentiel ($\hat{\boldsymbol{\delta}}^{t_k}(t) \rightarrow \hat{\boldsymbol{\delta}}^{t_k}(t - t_k)$). Cette courbe est décroissante dès $t_0 - 8$ et continue à décroître bien après l'interaction au moins jusqu'à $t_0 + 8$. À nouveau, on observe au temps t_0 de la création du lien une chute conséquente de la distance sémantique entre les deux blogs. En dehors des quelques jours qui encerclent le moment de l'interaction, les profils sémantiques semblent stables.

Nous avons reproduit le même protocole en choisissant une granularité de l'ordre de la semaine. Les deux profils d'évolution moyens sont représentés figure 3.14. Cette dernière courbe confirme bien la stabilité de la distance sémantique au

tel-00464316, version 1 - 16 Mar 2010

delà de deux semaines avant ou après le moment de l'interaction.

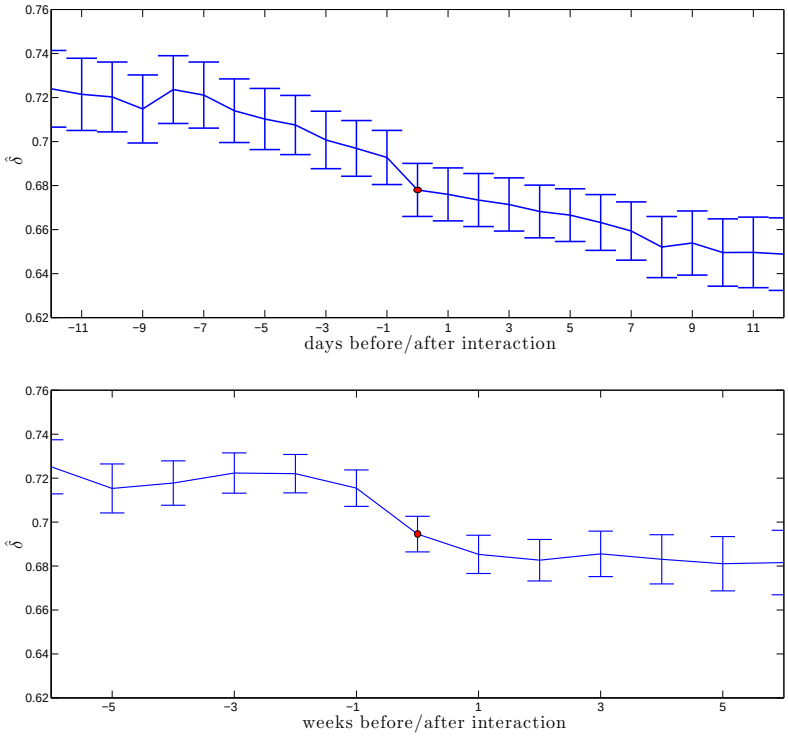


FIGURE 3.14: Evolution de la distance sémantique moyenne entre les couples de blogs ayant interagi au temps $\{t_k\}_k$ agrégés par le changement de référentiel $t \rightarrow t - t_k$: *en haut* : granularité : un jour, ensemble des blogs ayant interagi pour la première fois durant les 13 premiers jours ($\{t_k\}_{k=[1...13]}$ tel que $t_k = 46 + (k-1)$), *en bas* : granularité : une semaine, ensemble des blogs ayant interagi pour la première fois durant les 7 premières semaines ($\{t_k\}_{k=[1...7]}$ tel que $t_k = 46 + (k-1)7$)

Crandall et al. (2008a) ont observé des effets qualitativement similaires sur les dynamiques d'édition de contenus et d'interactions sur Wikipedia. Ils interprètent la diminution de la distance sémantique en amont de l'interaction entre deux contributeurs comme une forme de sélection et la décroissance observée après l'interaction comme une conséquence de l'influence sociale exercée entre les rédacteurs. Il semble néanmoins difficile d'être aussi catégorique sur l'interprétation de nos courbes. D'une part, l'influence sociale pourrait être un facteur explicatif pour l'ensemble de la courbe, y compris en amont de l'interaction, au moins sous l'effet d'une influence indirecte. En effet, nous avons vu combien la transitivité est un processus fréquent dans notre réseau, ainsi, sans avoir à entrer en interaction directe avec alter, ego peut très bien être déjà lié à un blog tiers qui, vient de se lier à alter. Une influence sociale indirecte transitant via un blog tiers (ou des blogs tiers) peut donc déjà être à l'œuvre en amont de l'interaction entre alter et ego. D'autre part, ego peut être familier en tant que simple lecteur (et donc potentiellement sous influence) de l'activité d'alter sans jamais l'avoir cité explicitement. Le début de la

tel-00464316, version 1 - 16 Mar 2010

décroissance de nos profils de distances sémantiques signalerait alors un temps caractéristique entre la découverte d'un nouveau blog et la citation effective.

Pour résumer, ces courbes révèlent un scénario moyen autour de l'événement d'interaction. Premièrement, la production d'un lien de citation entre deux blogs a tendance à rapprocher leur profil sémantique, et ceci aussi bien au moment de l'interaction, en amont de l'interaction et en aval de celle-ci. On peut décrire ce processus comme la succession de trois phases :

1. une première phase de rapprochement en amont de l'interaction qui semble suggérer un processus dynamique d'apprentissage des blogs vis-à-vis de leur environnement,
2. un climax au moment de la réalisation effective du lien qui fait décroître brutalement le taux de dissimilarité sémantique entre les deux blogs,
3. une phase ultérieure de mise sous influence plus directe qui aligne encore un peu plus les profils sémantiques des blogueurs.

3.2.2 Cohésion socio-sémantique locale

Nous avons mis en évidence la présence d'une structuration cohésive locale dans le réseau social au travers des coefficients de clustering, nous cherchons maintenant à identifier des motifs équivalents dans le réseau socio-sémantique. Le réseau socio-sémantique étant formalisé par un réseau biparti, la structure cohésive minimale est un cycle de longueur 4 mettant en jeu deux concepts reliés à une même paire d'agents. Le coefficient de clustering biparti c_4 permet de mesurer la tendance d'un agent, partageant déjà un concept avec un autre agent, à partager également un second concept. Le coefficient c_4 peut également être défini comme la proportion de motifs cycliques de longueur 4 autour d'un agent (ce qui est équivalent au calcul de la probabilité qu'une paire de concepts dans le voisinage d'un agent appartienne également au voisinage d'un autre agent) dans le réseau socio-sémantique biparti (Robins and Alexander, 2004). Formellement la formule de calcul du clustering d'ordre 4 est la suivante dans le cas biparti [15] :

$$c_4(i) = \frac{\sum_{\{c,c'\}\subseteq \mathcal{V}^{\mathbf{C}}(i)} [\kappa(c,c') - 1]}{\sum_{\{c,c'\}\subseteq \mathcal{V}^{\mathbf{C}}(i)} [k_{\mathbf{C}}(c) + k_{\mathbf{C}}(c') - \kappa(c,c') - 1]}$$

15. Des définitions alternatives du clustering biparti ont été proposées (Robins and Alexander, 2004; Lind et al., 2005) et peuvent diverger quantitativement de notre mesure, car elles décrivent le ratio des cycles fermés de longueur 4 sur tous les cycles ouverts de longueur 4 (Zhang et al., 2008), et calculent ce ratio directement sur l'ensemble du réseau. Nous avons contrôlé que nos résultats sont robustes à ces changements de définition, nos conclusions restent donc qualitativement identiques quelle que soit la formule employée.

tel-00464316, version 1 - 16 Mar 2010

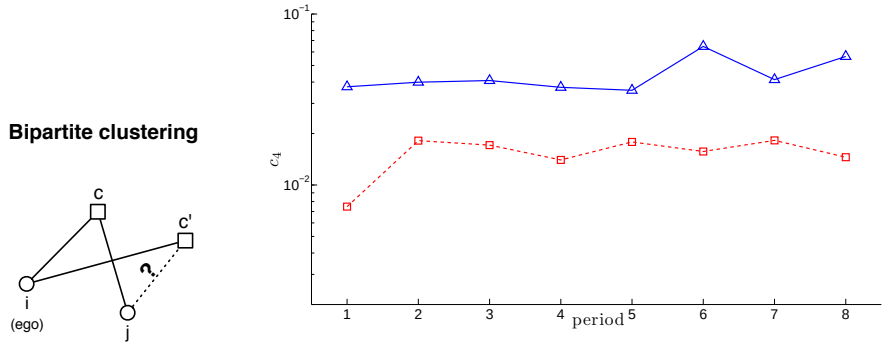


FIGURE 3.15: Clustering socio-sémantique c_4, en bleu (triangles) évolution du coefficient sur les 8 périodes, en rouge (carrés), valeurs obtenues pour un réseau biparti aléatoire respectant le modèle configurationnel.

où $k_{\mathbf{C}}(p)$ est le degré du concept $p \in \mathbf{C}$, $i \in \mathbf{S}$, $p \in \mathcal{V}^{\mathbf{SC}}(i) \Leftrightarrow \exists t, l = (i, p, t) \in \mathcal{R}^{\mathbf{SC}}$, $\kappa(c, c')$ désigne le nombre d'agents liés à la même paire de concepts (c, c'), i.e. $\kappa(c, c') = |j \in \mathbf{S}$ telle que $\{c, c'\} \subseteq \mathcal{V}^{\mathbf{C}}(j)|$.

Le réseau socio-sémantique étant très dense (certains agents sont quasiment liés à l'ensemble des concepts, et ceux-ci produisant mécaniquement des cycles de taille 4 avec l'ensemble des autres agents), nous appliquons d'abord une procédure de seuillage sur l'ensemble des liens bipartis. Nous partons des profils sémantiques des agents, et calculons pour chacun des concepts la valeur moyenne du *tf.idf* observé sur l'ensemble des blogs pendant toute la période d'observation. Nous ne conservons par la suite que les liens bipartis agent vers concept correspondant à des valeurs de *tf.idf* supérieures à 10 fois cette valeur moyenne. Cette procédure permet de réduire le degré moyen $k_{\mathbf{C}}$ en conservant *a priori* l'essentiel de la structure socio-sémantique du réseau. Cette procédure agit donc comme un filtre sur le réseau biparti initial, seuls les liens agents-concepts d'une "intensité" suffisante sont conservés. Afin de pouvoir apprécier les valeurs obtenues nous les comparons à celles obtenues pour une série de réseaux aléatoires respectant les distributions de degré des agents et des concepts (distribution parfois appelées *à gauche* et *à droite*) du réseau biparti préalablement seuillé. Dans notre modèle aléatoire, les concepts restent donc employés par le même nombre d'agents, et le nombre de concepts utilisés par les agents suit la même distribution que dans le cas réel. Les valeurs obtenues pour ces réseaux aléatoires sont représentées figure 3.15. Les coefficients de clustering socio-sémantiques mesurés sont au moins 3 fois plus importants dans le cas réel que dans le cas aléatoire. Cette différence indique que les paires de concept ne sont pas choisies aléatoirement. Certains concepts ont donc tendance à apparaître conjointement plus fréquemment dans les profils sémantiques des agents qu'un modèle aléatoire ne le prédirait [16].

16. Symétriquement on peut également interpréter la sur-représentation de ces motifs comme la

tel-00464316, version 1 - 16 Mar 2010

Les valeurs de clustering ont été calculées sur l'ensemble de la période d'observation (voir figure 3.15), et malgré les dynamiques couplées de production de nouveaux contenus, et de modification du réseau social, le clustering socio-sémantique reste stable dans le temps (courbe bleue), et significativement plus important que le cas aléatoire sur toute la période d'observation (courbe rouge). Une même stabilité du taux de clustering socio-sémantique dans le temps a été mise en évidence dans une communauté scientifique (Roth and Cointet, 2009). Là encore, le clustering d'ordre 4 obtenu est largement supérieur à celui attendu d'un réseau socio-sémantique aléatoire.

3.3 Dynamiques locales dans le réseau sémantique

3.3.1 Mesures d'occurrences

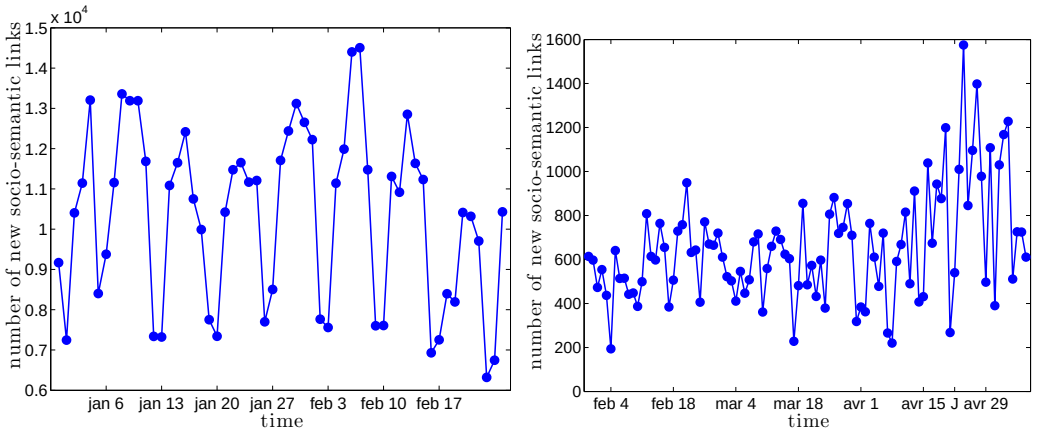


FIGURE 3.16: Evolution temporelle du nombre de concepts mobilisés quotidiennement, *à gauche* au sein de la blogosphère américaine, *à droite*, au sein de la blogosphère française, les dates figurant en abscisse correspondent à des dimanches (à un ryhtme hebdomadaire pour la blogosphère politique américaine, toutes les deux semaines pour la blogosphère française)

Nous nous intéressons dans un premier temps à une caractérisation de l'activité de mobilisation de concepts par les blogueurs dans nos deux jeux de données : blogosphère politique française et américaine. L'analyse de flux informationnel dans les media sociaux ouvre des perspectives inédites autant d'un point de vue théorique pour la fouille de données, qu'en termes applicatifs comme laboratoire d'observation des sources de concernenement à travers la planète pour les

tendance pour un couple d'agents à être associés aux mêmes concepts, cette interprétation dans la dimension sociale est une forme d'illustration de la présence d'agrégats socio-sémantiques, constitués d'agents liés densément liés les uns aux autres, et mobilisant des concepts identiques. (de façon similaire (Uchida et al., 2007) ou (Chi et al., 2007) et toujours Adamic and Glance (2005b) ont mis en évidence la forte polarisation sémantiques de territoires localement denses dans la blogosphère)

tel-00464316, version 1 - 16 Mar 2010

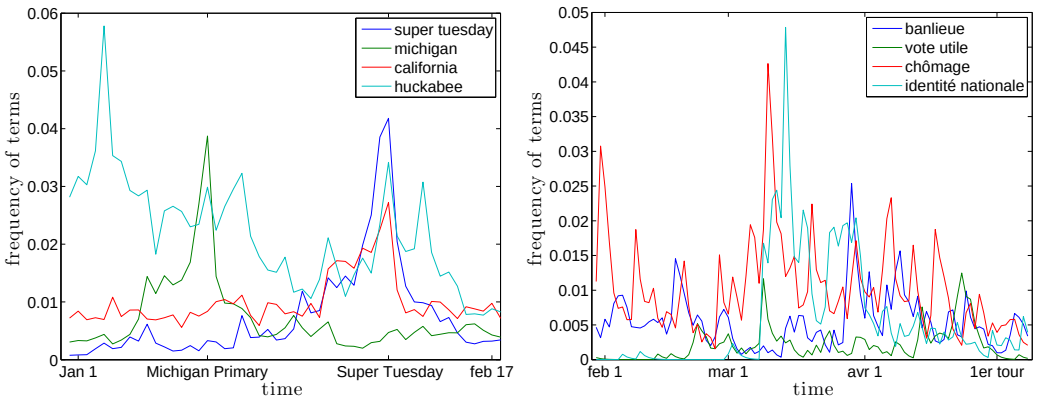


FIGURE 3.17: Evolution de la fréquence de 4 concepts, *à gauche* au sein de la blogosphère américaine, *à droite*, au sein de la blogosphère française

sciences sociales, d'aucuns voyant dans la détection automatique de "tendances" dans les blogs une opportunité pour sonder la "conscience collective" des internautes (Shalizi, 2007).

Au niveau global, si nous observons l'activité de la blogosphère dans son ensemble, sans nous soucier des contenus qui y circulent, à travers une statistique regroupant le nombre total d'occurrences de concepts publiés chaque jour, nous pouvons observer des variations volumétriques caractéristiques de l'activité humaine sous-jacente. La figure 3.16 illustre cette statistique dans nos deux jeux de données. Même si le signal est bruité dans le cas de la blogosphère française à cause de sa taille, nous observons une périodicité typique de l'ordre de la semaine dans ces séries temporelles, les blogueurs étant surtout actifs durant la semaine et semblant moins productifs durant les week-ends. De nombreux travaux basés sur l'analyse de grands flux de données textuelles sur internet (entre autres :(Leskovec et al., 2009; Balog et al., 2006; Thelwall, 2006; Lloyd et al., 2006)), ont déjà illustré ces effets et mis en évidence différents types de fréquence caractéristiques de l'activité des blogueurs. Certains jours particuliers, tels que le 1^{er} janvier, et le 31 décembre dans le cas de la blogosphère américaine ou le 1^{er} mai dans le cas français présentent des profils typiques d'une journée de week-end, tandis que les jours qui ont suivi le dimanche du 1^{er} tour des élections présidentielles françaises (signalé par un J sur la courbe) sont caractérisés par une recrudescence d'activité. Globalement le niveau d'activité est d'ailleurs accru dans le dernier mois précédent le vote dans le cas français, tandis que l'activité est légèrement déclinante dans le cas de la blogosphère américaine quelques jours après le pic d'activité associé au "Super Tuesday" (5 février 2008).

Mais cette activité agrégée mesurée sur l'ensemble des contenus (ou du moins l'ensemble des termes qui nous a permis de définir les bagages conceptuels des agents) cache en réalité une grande variabilité quant aux thèmes qui alimentent les discussion à un moment donné dans la blogosphère. Ainsi si on trace l'évolu-

tel-00464316, version 1 - 16 Mar 2010

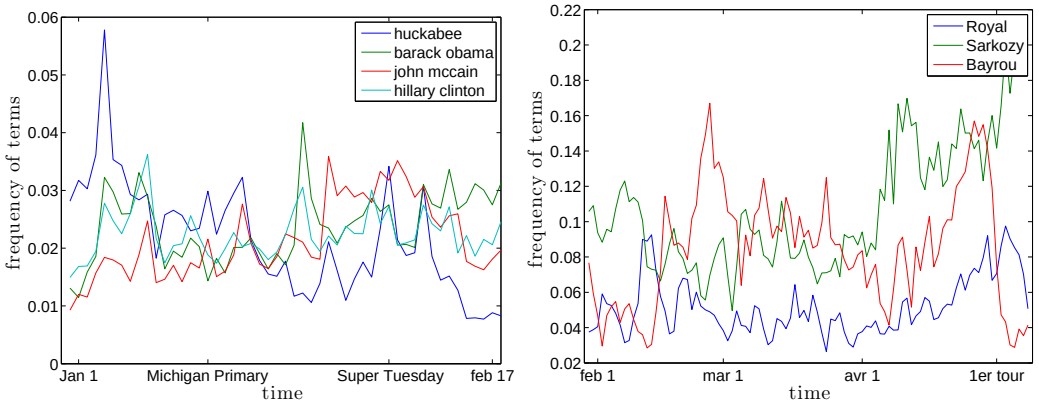


FIGURE 3.18: Evolution de la fréquence d'apparition de candidats aux élections présidentielles, *à gauche* au sein de la blogosphère américaine, *à droite*, au sein de la blogosphère française

tion de la fréquence d'apparition de quatre concepts clés dans le temps figure 3.17, on observe d'une part des périodes d'augmentation brutale dans l'usage de certains concepts liés à l'actualité politique (primaires du *Michigan* ou *Super Tuesday* dans le cas américain, pics d'activité autour de l'utilisation du concept *banlieue* ou *Identité nationale* lors de l'irruption de la thématique au cours de la campagne présidentielle française). Ces variations d'activité qui se caractérisent par une augmentation rapide de la fréquence d'usage d'un concept suivie d'une décroissance plus ou moins brutale constituent des *bursts* d'activité (Kleinberg, 2003; Barabási, 2005; Kleinberg, 2005; Bentley and Ormerod, 2009). Ces bursts montrent à quel point la nature de ce qui est discuté dans le système en son entier est soumise à des évolutions très rapides.

À titre d'anecdote, ces mesures simples de fréquence d'occurrences de concepts peuvent permettre de sonder en temps réel les sujets d'intérêt ou les personnalités politiques qui attirent l'attention des blogueurs (voir figure 3.18). Sans avoir valeur de "sondages d'opinion", ces mesures permettent de signaler les sujets qui (pré)occupent le plus la blogosphère. Des méthodes de traitement linguistique plus fines, telles qu'elles sont développées actuellement notamment sous l'impulsion de la recherche en marketing, réunies sous la catégorie de *sentiment analysis*, tentent de qualifier la façon (positive, négative, émotive, etc.) dont sont mobilisés les concepts (Pang and Lee, 2008; Mishne and de Rijke, 2006) afin d'affiner ces méthodes de fouille de données.

3.3.2 Mesures de co-occurrences

Les profils d'évolution temporels du nombre d'occurrences calculés au sein de nos deux blogosphères sont donc extrêmement irréguliers, ils se caractérisent par une périodicité hebdomadaire, une forte résonance aux événements du "monde réel" qui provoquent des "bursts" autour de certaines thématiques.

tel-00464316, version 1 - 16 Mar 2010

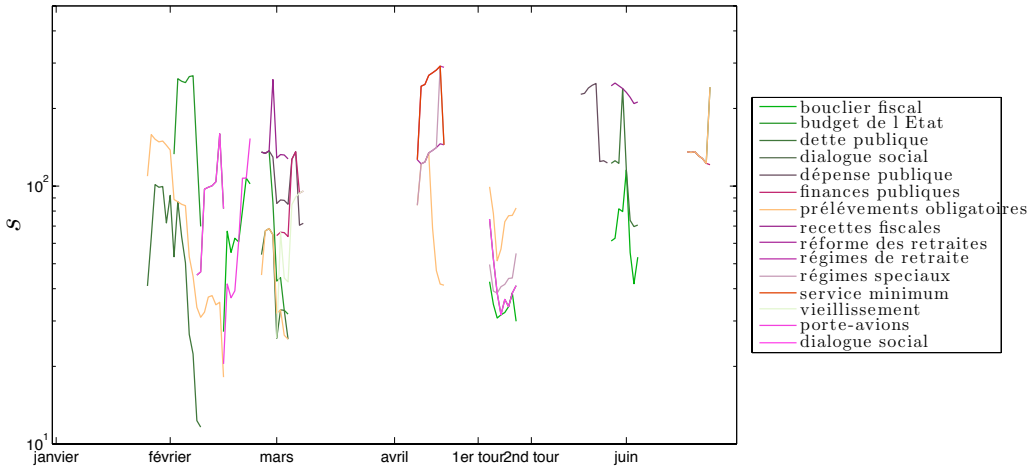


FIGURE 3.19: Evolution du log des similarités entre le concept "*baisse des prélévements*" et ses 15 plus proches voisins sur les 6 premiers mois de 2007

Malgré cette forte variabilité, la structure du réseau sémantique semble dotée d'une certaine forme de stabilité. Partant des statistiques fournissant pour chaque paire de concept et à chaque temps t le nombre de ses co-occurrences : n_{ij}^t (*i.e.* le nombre de billets publiés au temps t et mentionnant conjointement les concepts i et j) dans l'ensemble de la communauté de savoirs, on peut construire, une mesure de similarité entre termes inspirée des méthodes employées en scientométrie (on rediscutera longuement de ces méthodes dans le chapitre 4) : $s^t(i,j) = n_{ij}^t N^t / n_i^t n_j^t$ où n_i^t désigne le nombre total de billets mentionnant le concept i au temps t et N^t désigne le nombre total de billets publiés au temps t. Cette similarité mesure en fait la fréquence à laquelle les concepts i et j sont mentionnés dans un même billet : n_{ij}^t/N^t, divisée par la probabilité théorique de les retrouver ensembles si i et j étaient distribués aléatoirement que l'on estime en effectuant le produit des probabilités de mentionner chacun des termes : $n_i(t)/N(t) \times n_j(t)/N(t)$ [17]. Les deux termes i et j sont donc "indifférents" l'un à l'autre au temps t si $s^t(i,j)$ est proche de 1, "anti-corrélés" si $s^t(i,j) \ll 1$ et "corrélés" si $s^t(i,j) \gg 1$. Nous pouvons dès lors construire le réseau sémantique dynamique de similarité entre concepts, qui relie à un moment t l'ensemble des couples de concepts dont la similarité est supérieure à s_0 (pratiquement, nous choisirons comme valeur seuil $s_0 = 10$).

Si nous traçons maintenant figure 3.19 l'évolution de la similarité d'un concept : *baisse des prélévements* par rapport aux concepts auquel il est lié au cours de la période d'observation, nous observons d'une part que le terme *baisse des prélévements*

17. De nombreuses autres mesures de similarité ont été introduites : cosinus, coefficient de Jaccard, rapport de vraisemblance, indice de Dice, coefficient de colligation de Yule, coefficient de corrélation produit-moment de Pearson, mesure basée sur le test du χ^2 ou encore mesure d'information mutuelle ponctuelle, etc.

tel-00464316, version 1 - 16 Mar 2010

n'est employé que sporadiquement durant les 6 mois d'observation (les discontinuités dans les profils de similarité indiquent qu'au moins un des deux concepts est absent, nous avons vérifié que les zones vierges de tout concept sont bien associées à l'absence de notre terme cible). La question des *baisse des prélévements* a donc été l'objet de discussions épisodiques. De plus, chacune des périodes durant lesquelles le terme a été employé sont caractérisées par la co-appartition de concepts différents.

Pour mieux saisir le contenu de ces épisodes, et pour faire un premier pas vers la caractérisation de structures complexes de haut niveau, nous avons également calculé sur l'ensemble de la période d'observation les cliques (soient les sous-graphes complets maximaux) du réseau de similarité dans lesquelles le concept *baisse des prélévements* apparaît. Le résultat est représenté tableau 3.1. Une visualisation par cliques permet de saisir des ensembles de concepts qui agrègent des éléments qui co-occurrent tous fréquemment les uns avec les autres, fournissant ainsi des ensembles de termes dotés d'une forte contiguïté sémantique. On constate ainsi que le concept *baisse des prélévements*, a successivement été associé du $40^{ème}$ au $44^{ème}$ jour au sein d'une même clique aux concepts : *dialogue social* et *prélèvements obligatoires*, à laquelle viennent se rajouter le $45^{ème}$ jour les concepts *bouclier fiscal* et *porte-avions* qui continuent à former une clique avec *baisse des prélévements* jusqu'au $52^{ème}$ jour, etc. Au delà du récit détaillé du développement des débats autour de ce concept, la caractéristique la plus importante de la figure 3.19 tient au fait que sur les 15 concepts apparaissant dans le voisinage du concept *baisse des prélévements*, aucun de ceux-ci n'apparaît avec une similarité inférieure à 10, malgré une certaine variabilité dans les mesures de similarité. On observe donc, à une certaine échelle temporelle (de l'ordre de plusieurs jours et donc largement supérieure à celle caractéristique de l'activité micro qui anime l'ensemble des blogueurs), une stabilité remarquable des motifs dyadiques et n-adiques dans le réseau de similarité sémantique.

tel-00464316, version 1 - 16 Mar 2010

numéro	cliques	temps
1	dette publique ; prélévements obligatoires	$25 \rightarrow 31$
2	budget de l'Etat ; dette publique ; prélèvements obligatoires	$32 \rightarrow 38$
2	budget de l'Etat ; dette publique ; dialogue social ; prélèvements obligatoires ; dialogue social	39
3	dialogue social ; prélèvements obligatoires	$40 \rightarrow 44$
4	bouclier fiscal ; dialogue social ; prélèvements obligatoires ; porte-avions ;	45
4	bouclier fiscal ; porte-avions	$46 \rightarrow 52$
5	budget de l'Etat ; dette publique ; dépense publique ; prélèvements obligatoires ; recettes fiscales ;	$55 \rightarrow 58$
6	budget de l'Etat ; dette publique ; dépense publique ; finances publiques ; prélèvements obligatoires ; recettes fiscales ; vieillissement	$59 \rightarrow 62$
7	dépense publique ; finances publiques ; vieillissement	$63 \rightarrow 66$
8	prélèvements obligatoires ; réforme des retraites ; régimes de retraite ; régimes spéciaux ; service minimum	$96 \rightarrow 104$
9	bouclier fiscal ; dialogue social ; prélèvements obligatoires ; régimes spéciaux ; dialogue social	$115 \rightarrow 122$
10	depense publique	$140 \rightarrow 146$
11	bouclier fiscal ; dette publique ; recettes fiscales	$147 \rightarrow 154$

TABLE 3.1: Ensemble des cliques successives auxquelles le concept *baisse des prélévements* est associé dans le réseau de similarité.

tel-00464316, version 1 - 16 Mar 2010

Résumé du chapitre:

Dans ce premier chapitre de notre seconde partie consacrée à la morphogenèse, nous avons interrogé un certain nombre d'hypothèses de "fonctionnement" de nos communautés de savoirs en nous concentrant particulièrement sur les blogs politiques, en en découpant notre analyse entre dynamiques au sein du réseau social, du réseau biparti socio-sémantique, et du réseau sémantique. Un certain nombre de régularités ont été observées dans l'ensemble des réseaux notamment grâce au calcul de fonctions d'attachement préférentiel qui permettent de rendre compte de comportements systématiques des agents dans un environnement donné. On a également mis en évidence des corrélations très fortes entre les dimensions sociales et sémantiques accréditant l'hypothèse d'une co-évolution entre tissu relationnel inter-individuel et production de contenus.

Nos principaux résultats sont les suivants :

- importance du capital social et sémantique, mais également de l'activité des agents, dans leur attractivité vis-à-vis de la création de nouveaux liens de citations,
- homophilie sémantique très forte au sein de nos communautés de savoirs , renforçant l'homogénéité des agrégats socio-sémantiques,
- prépondérance des aspects sociaux vis-à-vis des aspects sémantiques dans le choix de nouvelles interactions, le réseau social définit une topologie qui contraint largement les interactions futures,
- mise en évidence d'une influence sociale entre agents voisins après leur première interaction mais également en amont de celle-ci,
- mise en évidence d'une forme de stabilité dynamique de certains motifs macroscopiques simples au sein des communautés de savoirs en dépit de la richesse des dynamiques individuelles (hiérarchisation du réseau social, stabilité de la cohésion socio-sémantique ou des motifs de co-occurrences de termes).

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

CHAPITRE 4

Structures émergentes

Sommaire

4.1 Communautés thématiques et communautés structurelles 104
4.1.1 Une portion du web social français 104
4.1.2 Détection des communautés structurelles 105
4.1.3 Hétérogénéité des topologies 107
4.1.4 Conclusion 107
4.2 De l'analyse de l'activité scientifique à la cartographie des sciences 108
4.2.1 Les mutations contemporaines de l'activité scientifique 108
4.2.2 Les bases de données de publications scientifiques, une opportunité pour la cartographie des sciences 110
4.2.3 Un modèle de l'activité scientifique 110
4.2.4 un modèle multi-échelle de la connaissance 112
4.2.5 Méthodes scientométriques de cartographie des sciences . . . 113
4.3 Cartographier les sciences 114
4.3.1 Jeux de données 114
4.3.2 Une mesure asymétrique de proximité entre termes 116
4.3.3 Construction du réseau lexical 122
4.3.4 Echelle microscopique : voisinages locaux 122
4.4 Echelle mésoscopique : la notion de champ épistémique 123
4.4.1 Définitions 123
4.4.2 Identifier les champs épistémiques 125
4.4.3 Plongement des clusters dans un espace bi-dimensionnel . . . 127
4.4.4 Qualifier les clusters 129
4.4.5 Représentation macroscopique 130
4.4.6 Reconstruction multi-échelle 132
4.4.7 Procédures de validation 136
4.5 Méthode de reconstruction dynamique 137
4.5.1 Dynamiques de voisinage 138
4.5.2 Dynamique d'un champ épistémique 138
4.5.3 Vers les dynamiques macroscopiques 141
4.5.4 Reconstruction de la phylogénie des sciences 144
4.5.4.1 Méthode de reconstruction des lignages entre champs épistémiques 145
4.5.4.2 Exemples de phylogénies 147

4.5.4.3 Motifs phylogénétiques 148
4.6 Trajectoires des individus au sein des paysages sémantiques. . . . 153
4.6.1 Opérateur de projection 153
4.6.2 Rétroaction macro-micro 156
4.6.3 Se déplacer dans un espace mouvant 160

Un ensemble de structures de haut-niveau non triviales résultant des interactions individuelles entre agents a été mis en évidence dans une grande variété de réseaux sociaux réels : réseaux de collaboration scientifique (Newman, 2001, 2004b; Garas and Argyrakis, 2008), d'amitié (Adamic and Adar, 2005), d'email (Kossinets and Watts, 2006; Zhou et al., 2005), de contact téléphonique Onnela et al. (2007), de contact sexuel (Bearman et al., 2004), d'interaction dans des communautés en ligne (Holme et al., 2004; Viegas et al., 2007a), etc. On rencontre également des structures de haut niveau similaires ou différentes (Newman and Park, 2003a) dans d'autres types de réseaux d'interaction (Strogatz, 2001; Girvan and Newman, 2002; Barrat et al., 2004) tels que les réseaux biologiques (réseaux métabolique, trophique, neuronal ou réseau de régulation), les réseaux d'infrastructure (comme Internet ou un réseau électrique), mais aussi les réseaux sémantiques (Steyvers and Tenenbaum, 2005).

Parmi ces structures, on peut signaler des motifs classiques déjà traités dans le chapitre 3 tels que les distributions de degré hétérogènes ou la présence d'une forte densité locale ou plus généralement la sur- ou la la sous-représentation de certains motifs triadiques (Newman, 2001; Milo et al., 2004), mais aussi l'organisation du réseau en petit-monde (Milgram, 1967; Watts and Strogatz, 1998), ou encore la structure dite modulaire des réseaux sociaux faite d'agrégats mésoscopiques dont les nœuds sont fortement connectés entre eux et peu liés vers l'extérieur (Girvan and Newman, 2002; Newman and Park, 2003a).

Ces propriétés structurelles sont remarquables au sens où elles sont caractéristiques des "réseaux de terrain" et complètement absentes du réseau aléatoire prototypique que constitue le graphe aléatoire de Erdős-Rényi (Erdös and Rényi, 1959) : $G(n, p)$, formé de n nœuds connectés deux à deux avec une probabilité p[1]. Comme nous l'avons déjà illustré pour certaines d'entre elles dans le chapitre précédent, bien qu'émergeant d'un grand nombre d'interactions locales, ces propriétés structurelles se caractérisent par une grande stabilité.

Au delà de la mise en évidence de ces motifs, la littérature récente sur les grands réseaux d'interaction s'est orientée, notamment sous l'impulsion des approches physiciennes, vers le développement de modèles de morphogenèse de

1. Ces réseaux aléatoires se différencient des réseaux réels vis-à-vis de l'ensemble des caractéristiques structurelles que nous venons de mentionner : leur distribution de degré est homogène (elle suit une loi de Poisson), leur clustering est faible, ils se caractérisent également par l'absence d'une structure modulaire ou hiérarchique, par contre, tout comme les réseaux réels leur diamètre (plus grande distance entre deux nœuds du réseau) est relativement faible (de l'ordre de $log(n)$)

tel-00464316, version 1 - 16 Mar 2010

réseaux à même de reconstruire ces propriétés, l'ambition étant de reproduire le plus grand nombre de ces faits stylisés de haut niveau à partir de modèles aussi parcimonieux que possible (Watts and Strogatz, 1998; Barabási and Albert, 1999; Holme and Kim, 2002; Klemm and Eguiluz, 2001).

Ces propriétés de haut-niveau ont également une influence déterminante sur les propriétés dynamiques de ces réseaux. Une distribution de degré en loi de puissance garantit une certaine robustesse du réseau et une tolérance aux attaques non ciblées (Albert et al., 2000; Callaway et al., 2000). Elle induit également nombre de conséquences vis-à-vis des processus de diffusion(Watts, 2002; Cowan and Jonard, 2004a) [2] et de synchronisabilité (Motter et al., 2005; Dorogovtsev et al., 2007). La corrélation entre les structures locales (liées au clustering du graphe) et les connections à longue distance (propriété de petit-monde) peuvent produire des propriétés remarquables vis-à-vis de la naviguabilité dans ces réseaux (capacité des agents à "router" efficacement une information vers un autre agent à partir d'informations purement locales) (Kleinberg, 2000; Watts et al., 2002).

Malgré la variété des motifs de haut niveau possibles, nous nous intéresserons particulièrement, dans ce chapitre, aux agrégats mésoscopiques d'entités formant des sous-ensembles cohésifs qui structurent nos communautés de savoirs. Compte tenu de l'approche duale entre les dimensions sociale et sémantique que nous adoptons, il s'agit d'interroger la présence de ces agrégats dans les communautés de savoirs en distinguant entre ceux rassemblant soit des individus (on parlera alors de groupes ou plus précisément de sous-communautés) soit des concepts (on parlera de champs, et plus précisément de champs épistémiques dans le cas de la Science) soit les deux. Ces agrégats apparaissent comme des structures émergentes de nos trois réseaux composant notre réseau épistémique. Nous discuterons également des couplages éventuels que ces structures entretiennent les unes avec les autres.

Dans ce chapitre, nous resterons succincts sur la dimension sociale en proposant une approche simple visant à saisir l'articulation entre *communautés structurelles* et *communautés thématiques* à travers une étude sur les communautés affinitaires au sein de la blogosphère française. Nous ne détaillerons pas non plus les motifs d'ordre socio-sémantique qui peuvent structurer les communautés de savoirs. On peut renvoyer à ce sujet à l'ensemble du champ de *l'analyse formelle de concepts* qui permet, notamment au travers du formalisme des treillis de Gallois, de cartographier dans un treillis de "concepts" des ensembles d'entités reliées à des attributs (Wille, 1992). Concernant les communautés de savoirs, cette approche a été notamment employée pour repérer les "communautés épistémiques" au sein d'un ensemble d'embryologistes (Roth and Bourgine, 2006). L'essentiel de nos efforts se sont concentrés sur la dimension sémantique à travers la mise en évidence

2. Nous aurons l'occasion dans la partie III de revenir sur d'autres propriétés des réseaux réels susceptibles de modifier la dynamique de diffusion.

tel-00464316, version 1 - 16 Mar 2010

et la cartographie de structures caractéristiques des réseaux lexicaux construits à partir de statistiques d'occurrences et de cooccurrences extraites de publications scientifiques. Cet exercice de cartographie des dynamiques scientifiques nous permettra également de mettre en évidence la rétroaction de structures de haut-niveau sur la dynamique des thématiques de recherche des scientifiques.

4.1 Communautés thématiques et communautés structurelles

Nous commençons notre exploration des agrégats d'entités structurant nos communautés de savoirs en nous concentrant en premier lieu sur l'analyse de la "structure de communautés" d'un réseau social construit à partir d'une partie du Web social français que suit RTGI. Ce travail a été réalisé en collaboration avec Camille Roth, Guilhem Fouetillou, Nils Grunwald et Camille Maussang.

4.1.1 Une portion du web social français

De nombreux travaux sur l'évolution du web postulent depuis le travail séminal de Kleinberg (1999) que les sites thématiquement proches les uns des autres ont tendance à être liés les uns aux autres, formant ainsi des structures socio-sémantiques cohérentes. Comme nous l'avons illustré dans le chapitre précédent, les individus se connectent préférentiellement les uns aux autres lorsqu'ils sont proches sémantiquement.

Le travail de RTGI s'appuie sur ce postulat pour définir les bases d'une *géographie* des communautés du web social basée sur l'exploration semi-automatisée du web par des documentalistes qui identifient, étiquettent et délimitent des territoires du web thématiquement cohérents. Ce travail a permis de classer $12,000$ sites actifs au sein de la blogosphère française dans plusieurs dizaines de communautés thématiques différentes (à chaque ensemble de blogs n'est assigné qu'une "étiquette", un blog ne peut donc pas appartenir à deux communautés différentes et les frontières entre communautés thématiques sont univoques).

Dans cette étude, nous nous sommes concentrés sur 18 de ces communautés thématiques sélectionnées aléatoirement. Cet ensemble couvre $4,980$ sites web. La liste de ces communautés est donnée dans la première colonne du tableau 4.1. Ces communautés couvrent des thématiques très variées : décoration d'intérieur, loisirs, ou encore politique. En plus de bénéficier du travail de catégorisation de RTGI, nous avons également accès à la structure hypertextuelle qui relie l'ensemble de ces sites - soit au réseau des sites Dans ce réseau, tous les liens sont incorporés indépendamment de leur nature, (liens de citation, de commentaire, etc.) grâce à un crawl en profondeur qui extrait l'ensemble des liens hypertextes des sites.

tel-00464316, version 1 - 16 Mar 2010

Nom	Taille	Comm. structurelles		Degré	Clustering	ratio de liens	
		Taille	Similarité(%)	moyen	(c_3)	Entrant	Sortant
deco design	111	57	39	3,27	0,19	0,42	0,29
cuisine	744	753	95	28,03	0,06	0,07	0,04
tech revolution	259	628	37	4,48	0,12	0,4	0,23
foyer	597	544	61	16,52	0,09	0,23	0,26
deco brico	227	227	57	12,33	0,15	0,33	0,36
freemen	123	157	74	26,96	0,05	0,07	0,16
carnet bd	267	249	86	6,98	0,12	0,22	0,13
gauche	563	377	59	4,54	0,22	0,2	0,2
musique	129	130	0	2,12	0,21	0,34	0,26
jardinage	130	79	59	3,56	0,16	0,43	0,6
droite	197	67	0	3,09	0,22	0,29	0,24
automobile	366	124	35	2,38	0,21	0,13	0,07
beaute	128	142	62	7,73	0,18	0,38	0,33
centre	160	160	78	5,68	0,27	0,22	0,4
cinema	248	95	37	3,88	0,2	0,15	0,08
mode lifestyle	451	363	64	7,49	0,11	0,43	0,43
liberaux	85	82	81	5,15	0,23	0,14	0,22
actu opinion	195	43	4	2,76	0,21	0,67	0,64

TABLE 4.1: **Communautés thématiques :** meilleur appariement possible avec les communautés structurelles détectées et propriétés topologiques.

4.1.2 Détection des communautés structurelles

Notre objectif est de comparer la catégorisation validée par des documentalistes à partir de critères essentiellement thématiques avec une catégorisation automatique obtenue sur la base d'hypothèses purement structurelles. On applique un algorithme de détection de communauté au réseau des liens hypertextes liant l'ensemble de ces sites web pour calculer cette catégorisation structurelle. La notion de communauté structurelle a donné lieu à une littérature pléthorique en physique et en informatique même si, en sociologie, l'analyse des réseaux sociaux avait déjà proposé de nombreuses méthodes pour détecter des ensembles localement denses. Nous avons utilisé l'algorithme de Blondel et al. (2008) qui propose une méthode innovante pour optimiser la modularité d'un réseau [3].

3. La modularité d'un réseau généralement notée Q a été introduite par Newman and Girvan (2004) et mesure, pour une partition des nœuds d'un réseau en communautés donnée, la fraction des liens du réseau qui relient des nœuds appartenant à la même communauté moins la valeur de cette fraction si on la calculait à partir d'une distribution aléatoire des liens appliquée à cette même partition.

tel-00464316, version 1 - 16 Mar 2010

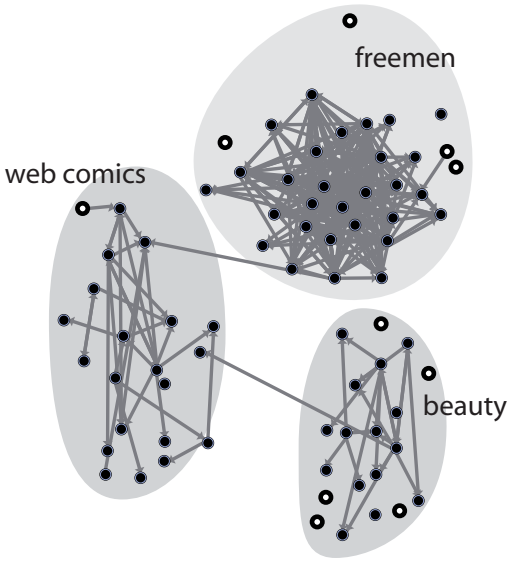


FIGURE 4.1: **Portion de notre jeu de données.**. On a sélectionné aléatoirement 10% des nœuds issus de trois communautés thématiques - *carnet bd*, *freemen*, *beauté* - ainsi que l'ensemble des liens entre eux. Nous avons matérialisé en noir les nœuds que l'algorithme de détection de communautés avait "correctement" catégorisé ; les autres, en blanc. Cette représentation illustre également les différences de topologies entre les différentes communautés.

À partir du réseau hypertexte symétrisé constitué de $4,980$ nœuds, nous avons obtenu 148 *communautés structurelles* distinctes, la plupart étant de taille très réduite (moins de 10 nœuds). Pour simplifier notre procédure de comparaison entre communautés structurelles (issues de l'algorithme de détection de communautés) et *thématiques* (issues du travail d'expert sur les sites), nous n'avons retenu que les 18 communautés issues de notre catégorisation automatique, qui couvrent $4,222$ blogs (soit 84.8% de l'ensemble). Nous avons ensuite effectué un appariement optimal entre nos 18 communautés structurelles et les 18 communautés thématiques originales, qui maximise le nombre de nœuds "correctement" identifiés.

Plus de 80% des sites inclus dans les 18 plus grandes communautés structurelles ont été correctement catégorisés, au sens où ils ont été associés au même ensemble de sites qu'attendu par la catégorisation thématique. Néanmoins, les résultats sont extrêmement variables en fonction des communautés - voir tableau 4.1. Ainsi, si l'on calcule pour chaque communauté un taux de similarité (calculé par une simple distance de Jaccard) entre communauté structurelle et thématique, les résultats sont très contrastés. Certaines communautés structurelles sont très semblables à leur alter-ego thématique comme la communauté "cuisine" ou "carnet bd" dont le taux de similarité avoisine 90% [4]. D'autres sont beaucoup moins

4. Il faut rappeler que plus de 15% des sites ont été catégorisés par l'algorithme dans des commu-

tel-00464316, version 1 - 16 Mar 2010

bien reconstruites comme les communautés "actu-opinion", "musique" ou "droite" dont les similarités sont nulles ou négligeables. Dans ces cas là, l'algorithme a sans doute opéré des partitions a un grain trop fin par rapport à la division thématique choisie par les documentalistes, les sites de ces communautés thématiques sont donc éparpillés au sein des communautés structurelles de petites tailles.

4.1.3 Hétérogénéité des topologies

Le tableau 4.1 nous informe également sur la très forte variabilité topologique des communautés thématiques. Nous avons simplement mesuré, pour chaque communauté thématique, le degré moyen des sites qui la composent, leur clustering, et le ratio de liens entrants ou sortants de la communauté. L'ensemble de ces mesures montre une très forte diversité des topologies rencontrées. Tandis que certaines communautés comme "cuisine" ou "news-opinion" semblent être très extraverties (*i.e.* fort ratio de liens sortants de la communauté), d'autres , comme "cuisine", semblent nettement plus introverties. D'autres paramètres comme le clustering ou le degré moyen sont soumis à une forte variabilité qui laisse à penser que la structure topologique des territoires thématiques peut être extrêmement différente en fonction des processus de régulation ou d'organisation qui les régissent (voir tableau 4.1 et figure 4.1 pour un aperçu visuel). Il semble donc que certaines communautés se construisent sur la base d'une structuration interne très forte (faible ratio de liens entrants ou sortants), ce qui les rend facilement détectable par les algorithmes de détection de communautés fonctionnant sur des bases purement structurelles. D'autres communautés, par nature plus ouvertes sur leur environnement, sont moins facilement détectables par ces méthodes. À ce titre, la figure 4.1 semble indiquer que les nœuds les moins connectés sont également les moins bien catégorisés.

La variabilité topologique que nous observons illustre le caractère mosaïque de la blogosphère, pourtant souvent traitée comme un seul bloc indifférencié. Ces premiers résultats appellent à un prolongement vers une analyse ethnographique systématique des territoires virtuels (Thelwall et al., 2005) qui explicite les comportements à un niveau mesoscopique — celui de la communauté — en insistant sur les différentes pratiques éditoriales et relationnelles au sein et entre ces différentes communautés.

4.1.4 Conclusion

Pour résumer, nous avons montré que : la structure relationnelle du jeu de données analysé se caractérise par la présence de communautés au sens structurel, dont les frontières sont relativement proches des frontières thématiques dessinées

nautés de petites tailles qui ont été éliminées, les similarités mesurées ne peuvent donc naturellement pas toutes avoisiner les 100%

tel-00464316, version 1 - 16 Mar 2010

par des experts. Certaines communautés thématiques ne semblent pas être suffisamment "structurées" pour être reconstruites à partir d'une simple analyse de la structure relationnelle. Aussi, leur détection automatique pourrait nécessiter de faire appel à d'autres critères s'attachant plus aux contenus qui y sont mobilisés. Néanmoins, chaque communauté thématique semble se caractériser par des motifs topologiques spécifiques, ouvrant la voie à une exploration comparative des modes d'organisation des communautés en ligne.

4.2 De l'analyse de l'activité scientifique à la cartographie des sciences

Dans la suite de ce chapitre nous nous concentrerons sur la reconstruction des dynamiques scientifiques en nous focalisant sur la dimension sémantique de notre schéma général 2.7 (plan arrière de notre parallélogramme). Nous souhaitons (i) mettre en évidence et cartographier les structures conceptuelles qui organisent notre réseau sémantique (ii) décrire leur dynamique et les représenter. Ces méthodes de reconstruction ont été conçues et développées avec David Chavalarias et ont donné lieu aux publications suivantes : (Chavalarias and Cointet, 2008; Cointet and Chavalarias, 2008; Chavalarias and Cointet, 2009; Cointet, 2008). Enfin, nous tâcherons, en cohérence avec notre programme de description des dynamiques multi-échelles des communautés de savoirs, (iii) d'évaluer les rétroactions que ces structures de haut-niveau exercent sur la dynamique des profils sémantiques des chercheurs (immergence du haut niveau sémantique sur les dynamique microscopiques du réseau socio-sémantique).

4.2.1 Les mutations contemporaines de l'activité scientifique

Certains soutiennent qu'un nouveau régime de production de connaissances aurait émergé consécutivement à la transformation de la nature même du processus de recherche. Selon Nowotny et al. (2001, 2003), la Science serait entrée dans un nouveau *mode* de production de connaissance donnant toute sa place à une *transdisciplinarité*, définie comme la circulation d'outils, de perpectives théoriques et de personnes. Selon ces auteurs, les réflexes habituels de classification de la connaissance, suivant des taxonomies bien codifiées et clairement délimitées héritées de la structure prévalant dans le seul monde académique, seraient caduques. L'ère de la trans-disciplinarité rend toujours plus floues les frontières entre communautés scientifiques : les *assemblages d'acteurs et de concepts* se multipliant, la mobilisation de connaissances techniques ou scientifiques n'est plus circonscrite à une seule communauté scientifique bien clôturée. Les acteurs engagés dans la chaîne — ou même dans le réseau — de production de connaissances proviennent d'horizons divers : ingénieurs, chercheurs, usagers, etc. Indépendamment des débats que cette

tel-00464316, version 1 - 16 Mar 2010

théorie a suscités, la sociologie des sciences et des techniques s'accorde aujourd'hui sur le fait que la production scientifique est une activité éminemment *socio-technique*, et que de nouveaux modes de production de la connaissance ont émergé ces dernières décennies, entraînant des changements aussi bien épistémiques, organisationnels, que politiques. Ainsi, la notion de "modernité réflexive" introduite par Beck (1992) illustre la façon dont le public s'est éveillé aux risques techno-scientifiques durant le XX^{ème} siècle, transformant de par la même le regard que l'on portait sur les sciences. Mais au delà de ces théories globales, des études plus micros montrent que la construction de polémiques techno-scientifiques se fonde également sur des dynamiques socio-cognitives complexes de co-constuction des concernements des acteurs et des programmes de recherche. Les frontières entre science et société sont sans cesse renégociées par des acteurs hétérogènes plongés dans de nombreuses arènes aux régimes de justifications variés (Bonneuil et al., 2008).

Ces transformations ont été accompagnées d'une mutation profonde des modes d'échange dans le monde de la recherche. Internet a impacté à plusieurs titres les dynamiques scientifiques : (i) la fluidification des échanges entraîne une plus grande liberté dans la construction des "équipes scientifiques" qui sont de plus en plus internationales et trans-institutionnelles (Jones et al., 2008), (ii) la multiplication des supports de publication ainsi que la mise à disposition d'archives en ligne de plus en plus souvent gratuites démocratisent en partie l'accès à la connaissance et créent des ponts inédits entre les travaux de différentes disciplines, (iii) ces bases de données permettent, dans le prolongement des premiers travaux de scientométrie, d'effectuer des mesures de l'impact des publications, des chercheurs ou même des institutions au sein du paysage de la recherche internationale. Cette nouvelle donne n'a pas fini de modifier en profondeur les pratiques de recherche.

De façon générale, le transfert des publications scientifiques vers des supports numériques tel que les journaux en ligne, ou les bases de données d'archives scientifiques, a complètement modifié la façon dont nous interagissons avec la production scientifique. Nous sommes passés d'un régime fortement hiérarchisé de circulation des contenus à un régime plus horizontal et plus résiliaire dans lequel la navigation dans les bases de publications scientifiques est susceptible de mettre côte à côte des productions provenant d'univers potentiellement très différents [5].

5. La principale forme de hiérarchie qui prévaut encore dans la navigation à travers ces espaces est issue d'une autorité hypertextuelle, directement liée à l'idée d'un "ranking" des publications en fonction du nombre de citations reçues, mais finalement *a priori* différente de la notion plus traditionnelle de réputation d'une revue.

tel-00464316, version 1 - 16 Mar 2010

4.2.2 Les bases de données de publications scientifiques, une opportunité pour la cartographie des sciences

L'accessibilité de ces bases de données de publications dans un format numérique nous semble être une opportunité réelle pour retracer l'évolution de la production scientifique. La contrepartie d'une telle méthode tient naturellement à la masse et à l'hétérogénéité de ces bases de données. L'idée de s'appuyer sur les marqueurs textuels de l'activité scientifique pour en retracer la dynamique n'est certainement pas neuve, l'histoire de la scientométrie est d'ailleurs étroitement liée à celle du développement des bases de données massives et des outils de traitement de ces bases. Un auteur aussi crucial dans le champ de la scientométrie et de la bibliométrie qu'Eugene Garfield, créateur de l'ISI [6], témoigne bien de la façon dont la scientométrie s'est développée de concert avec ces innovations technologiques.

Interroger l'organisation des sciences à travers ces bases de données — la façon dont disciplines et sous-disciplines se déploient et s'hybrident ou cartographier ces articulations — constitue naturellement un enjeu crucial pour les chercheurs qui doivent s'informer continûment sur les travaux plus ou moins connexes à leur spécialité. C'est également un besoin stratégique pour les gestionnaires de la recherche, telle ou telle institution devant comprendre l'organisation et la direction prise par les communautés scientifiques existantes afin de définir de façon optimale leur politique scientifique. C'est enfin une opportunité pour la sociologie ou l'histoire des sciences d'être à même de voir se déployer les dynamiques de production de connaissance à partir d'une observation *in-vivo* afin d'étayer les hypothèses et modèles que conçoivent les chercheurs.

4.2.3 Un modèle de l'activité scientifique

L'activité scientifique résulte de la combinaison de processus complexes (Hull, 1988) portés par de multiples réseaux d'interactions hétérogènes mêlant chercheurs, ingénieurs, objets d'expérimentation, outils, journaux, institutions (etc.) (Morris and Yen, 2004). Les voies de communication et d'interaction au sein de ce système sont multiples que ces liens soient formels (dynamiques de construction d'équipes de coauteurs pour publier un article, citations, etc.) ou informels (correspondance par mail entre chercheurs ou rencontres dans les congrès) (Mullins, 1972).

La publication scientifique, dont la validité est attestée à l'issue d'un processus d'évaluation par les pairs, est généralement considérée comme l'un des principaux produits de ces interactions multiples. Les centres d'intérêt, objets, concepts d'une

6. ISI : Institute of Science Information, dès 1961 cet institut créa le "Genetics Citation Index" sur la demande du *National Institute of Health* avant de l'étendre en 1963 à d'autres disciplines à travers le Science Citation Index (SCI). Ces indices mesurent le nombre de citations reçues par des articles publiés dans des revues académiques

tel-00464316, version 1 - 16 Mar 2010

communauté scientifique sont ainsi cristallisés au sein des publications qu'elle produit. Nous ferons donc l'hypothèse que les dynamiques épistémiques d'un domaine peuvent être appréhendées au travers des publications scientifiques qui y sont produites. S'intéresser exclusivement aux publications scientifiques pour décrire l'activité sceintifique peut paraître réducteur. L'importance de cette "inscription littéraire" dans l'activité de recherche (Latour and Woolgar, 1986) nous encourage néanmoins à faire cette hypothèse. Les publications scientifiques sont en effet omniprésentes dans la communication entre chercheurs de façon directe (projet d'écriture d'un article par exemple, conférences) ou indirecte (comme élément de référence privilégié au cours des discussions entre chercheurs).

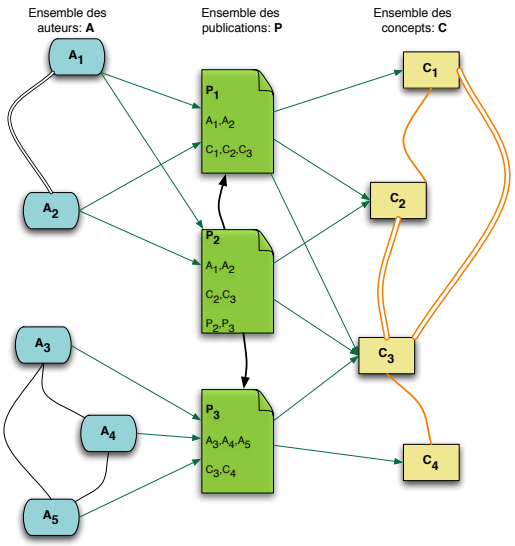


FIGURE 4.2: Les chercheurs interagissent au sein d'un réseau de collaborations scientifiques liés par un réseau de citation, tandis que la distribution des connaissances est ici formalisée par un réseau de co-apparition de concepts au sein de ces publications.

La figure 4.2 représente de façon schématique le processus de base de l'activité scientifique. Des auteurs (A) produisent des publications (P) qui mettent en relation des concepts (C). Ce schéma fait apparaître le réseau de co-publication (liens noirs entre auteurs), le réseau de citations (liens dirigés noirs entre publications) ainsi que le réseau de co-présence des concepts au sein des publications (liens rouges pondérés entre concepts). Il est extrêmement réducteur comparativement à la richesse des réseaux animant l'activité scientifique. Certains auteurs ont pris le parti d'une démarche exhaustive vis-à-vis de l'ensemble des entités (publications, journaux, auteurs, références, termes, etc.) "mises en réseau" au sein de l'activité scientifique (Chen, 2006; Morris and Yen, 2004). Nous prenons une direction différente en privilégiant une orientation cognitive s'attachant à reconstruire les dynamiques des communautés scientifiques uniquement au travers des agencements conceptuels produits au sein des publications. Nous nous concentrons

tel-00464316, version 1 - 16 Mar 2010

donc sur la reconstruction de la dynamique des "champs épistémiques" entendus comme l'ensemble des termes (qu'ils se rapportent à des outils, des objets, des méthodologies, des théories, etc.) qui sont fréquemment employés conjointement dans les corps des publications.

Notre objectif va donc consister à révéler les structures remarquables d'un réseau sémantique de proximité entre concepts (parfois également appelé réseau sémantique lexical) construit à partir des statistiques sur les occurrences et cooccurrences de termes au sein des articles afin de cartographier le paysage conceptuel construit par, et dans lequel se déploie, l'activité scientifique.

4.2.4 un modèle multi-échelle de la connaissance

Une autre contrainte de notre travail de reconstruction est de rendre compte du caractère naturellement multi-échelle de la structure de la connaissance scientifique. Par exemple, les universités opèrent classiquement une division des sciences en grands départements qui correspondent à autant de disciplines comme la biologie, l'économie, l'informatique, la physique, etc... Chacune de ces disciplines peut par la suite être elle-même morcelée en sous-champs : biologie végétale, animale, moléculaire, évolutive (voir figure 4.3 pour une illustration)... On peut certainement critiquer la pertinence de certaines de ces divisions. Certaines frontières anciennes peuvent être rendues caduques par l'évolution des sciences, ou sembler relever de critères non épistémiques. Ces frontières peuvent être de différentes natures et ne suivent pas nécessairement les même lignes de démarcation selon que l'on tâche de différencier un objet d'étude (biologie animale/végétale) ou un type d'approche (micro-biologie/physiologie/écologie/développement) par exemple. Dans la plupart des cas, un "sous-champ" est spécifique d'un seul champ qui l'englobe, mais, dans certains cas, un sous-champ est précisément défini comme l'intersection de plusieurs champs (la bio-physique par exemple). Les frontières entre champs ne sont pas parfaitement hermétiques et nombre de ces ensembles se recouvrent.

Néanmoins, on ne peut nier que la division en disciplines, champs, sous-champs (etc.) constituent une taxonomie efficace pour se donner une représentation mentale intuitive de l'organisation des sciences. Néanmoins, la structure générale que nous aimerions pouvoir mettre en évidence n'est certainement pas celle d'une structure hiérarchique *stricto sensu* prenant la forme d'un arbre, mais plutôt celle d'un treillis qui autorise des formes de ramifications plus variées entre entités. Le premier objectif que nous nous fixons sera donc de reconstruire cette "hiérarchie" propre à l'organisation des sciences en respectant la complexité des "motifs d'inclusion" et leurs articulations, et donc en autorisant un certain degré d'héterarchie. Cette reconstruction sera réalisée à l'aide d'outils d'analyse quantitative de nature *scientométrique* s'appuyant exclusivement sur la connaissance de statistiques de base sur les occurrences et cooccurrences d'un ensemble de termes

tel-00464316, version 1 - 16 Mar 2010

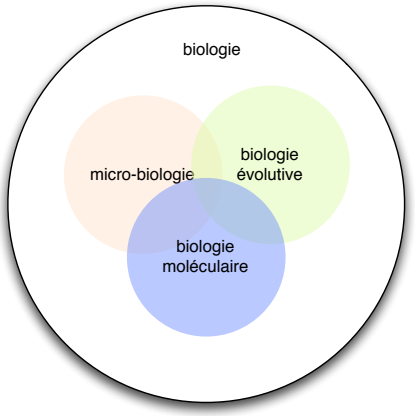


FIGURE 4.3: Exemple schématique de l'organisation d'un champ. Les trois sous-champs représentés relèvent tous de la biologie, mais leur intersection est non vide.

extraits d'un corpus de publications scientifiques.

4.2.5 Méthodes scientométriques de cartographie des sciences

La scientométrie est une science récente qui prit son envol à la fin des années 70 grâce au développement combiné des outils de traitement automatisé de larges bases de données et la mise à disposition de ces bases dans des formats numériques. Elle désigne de façon générique l'application de méthodes statistiques à des données quantitatives dans le but de caractériser un certain état de la science.

La cartographie des sciences est un des objectifs premiers de la scientométrie et figurait parmi les ambitions des pionniers de la discipline (de Solla Price, 1965). Les cartes des sciences sont généralement construites à partir de données de cooccurrences en suivant l'hypothèse qu'un couple d'entités qui apparaissent conjointement "fréquemment" (ou en tout cas plus fréquemment que ne le prédirait une distribution aléatoire) entretiennent l'une avec l'autre une certaine "proximité". Ces mesures de proximité permettent par la suite de construire des réseaux de similarité entre ces entités.

Ces données de co-occurrences peuvent aussi bien s'appliquer à des auteurs co-signant le même article (réseaux de co-publication ou de collaboration (Newman, 2004a; Palla et al., 2005a)), à des références étant citées dans un même article (réseaux de co-citation (Small, 1973)), ou à des termes figurant dans le même titre - abstract - ou texte d'un article (réseaux dits de "mots associés" (ou co-word) Callon et al. (1983)). Dans cette dernière catégorie, des structures de haut niveau sont déduites des textes en analysant les motifs récurrents qui y figurent (Callon et al., 1986). Le lien entre deux termes est pondéré par une intensité qui correspond à leur similarité supposée. Ces mesures de similarité permettent de construire des clusters de termes, puis des cartes des sciences censées refléter les domaines d'ac-

tel-00464316, version 1 - 16 Mar 2010

tivité qui organisent le domaine scientifique étudié.

Réseaux de citation et de mots associés constituent les deux méthodes privilégiés pour cartographier les sciences. Les deux techniques ont leurs inconvénients respectifs (Noyons, 2001) et ont donné lieu à un certain nombre de critiques (voir notamment la critique de Leydesdorff (1997) sur la pertinence de l'analyse lexicographique). Les études fondées sur les co-citations peuvent s'avérer biaisées par l'absence de certains papiers pertinents, ou au contraire par l'inclusion de publications non pertinentes. Le décalage temporel entre l'émergence de nouvelles spécialités scientifiques et leur détection sur une carte des sciences peut également s'avérer problématique. Les techniques à base de mots associés, peuvent également souffrir d'un choix inapproprié de l'ensemble initial de termes à cartographier. Mais l'objection principale qui est adressée à ce type d'analyse est que les mots peuvent être ambigus ou porteurs de plusieurs sens.

Dans la suite, nous introduirons une méthode de reconstruction des dynamiques scientifiques à partir d'une analyse de mots associés. Nous proposons un certain nombre de méthodes permettant de dépasser les limites classiques de la cartographie des sciences. Ces méthodes sont introduites à différents niveaux du travail de reconstruction : (i) introduction d'une mesure de proximité asymétrique entre termes qui tienne compte de l'hétérogénéité de leur distribution (section 4.3), (ii) utilisation d'une méthode de clusterisation des termes qui autorise les clusters recouvrants - et en corrollaire la polysémie de certains termes (section 4.4)(iii) définition d'une véritable représentation multi-échelle articulant champs et sous-champs dans une structure de treillis (et qui ne se limite donc pas à un arbre hiérarchique) (iv) reconstruction du réseau phylogénétique des champs scientifiques (section 4.5).

4.3 Cartographier les sciences

Nous allons décrire dans cette section la méthodologie que nous avons développée pour *cartographier les sciences* en nous contentant pour le moment d'une reconstruction statique de *l'organisation* d'un ensembles de termes $\mathcal{L}$ pertinents structurant un domaine d'étude donné. Nous présenterons les trois jeux de données qui serviront à illustrer notre méthodologie. La construction de ces cartes se déroule en trois étapes. Une fois le travail d'indexation réalisé sur un corpus de textes datés, il s'agit dans un premier temps de définir un réseau de proximité entre nos termes. Nous proposons d'introduire une mesure asymétrique de proximité entre deux termes qui rendent compte des relations "d'inclusion" entre termes.

4.3.1 Jeux de données

Nous décrivons dans cette partie les trois cas d'étude qui serviront à illustrer notre méthode par la suite. Chaque domaine d'étude est constitué d'un premier

tel-00464316, version 1 - 16 Mar 2010

corpus de termes ou d'expressions noté $\mathcal{L}$ que l'on cherche à cartographier, et d'un second corpus de publications à partir duquel sont calculées les statistiques d'occurrences et de cooccurrences de notre corpus de termes. L'analyse de mots associés peut dépendre de façon critique du corpus initial de termes. Le risque est de biaiser cet ensemble (on parle de l'"indexer effect" (Whittaker (1989), Callon et al. (1986), He (1999)) en omettant des termes capitaux ou en sélectionnant des termes trop généraux. Dans notre cas, nous tâchons d'éviter ces biais en proposant d'une part une méthode semi-automatique de sélection des mots-clés, et d'autre part en développant des outils d'analyse robustes par rapport au bruit présent dans la base de termes initiale.

Le premier cas d'étude a trait au champ des *systèmes complexes*, il est constitué d'un corpus de près de 450 termes (la liste des termes est disponible dans l'annexe B.1. Ce corpus de termes a été construit à partir d'une liste de mots-clés d'un appel à projet dédié à la science des systèmes complexes de l'Union Européenne dans le cadre du $6^{ème}$ programme cadre. On a par la suite extrait le nombre de cooccurrences (dans le texte intégral des articles) observées dans la base de données *Scirus* de 1975 à 2005. La base originale est composée de plus de 20 millions de publications couvrant un large éventail de plate-forme de publications scientifiques [7].

La seconde base de données traite de la biologie contemporaine exposée aux évolutions paradigmatiques introduites par l'introduction de la "métaphore réseau". Nous avons utilisé Pubmed-Medline comme source de publications. Cette plateforme couvre la plupart des publications en biologie (plus de $17M$ de références), dont les titres et les abstracts sont publiquement accessibles. Nous avons construit une première requête réunissant un certain nombre de termes caractéristiques de la pensée réseau en biologie ("network, evolvable, evolvability, hub, feedback") afin de sélectionner dans notre corpus de publications les articles mentionnant au moins l'un de ces termes dans leur titre ou leur abstract. Cette requête nous a permis de collecter près de $2,4$ millions d'articles s'étalant sur plus de 50 années. La sélection du corpus de termes a été réalisée en deux étapes. Dans un premier temps, un ensemble de publications comportant le terme "network" dans leur titre a été sélectionné grâce à une requête sur l'*ISI Web of Knowledge* limitée à un ensemble de journaux de premier rang [8]. Cette première extraction a permis de

7. ScienceDirect, Society for Ind. & App. Mathematics, BioMed Central, Crystallography Journals Online, Institute of Physics Publishing, MEDLINE/PubMed, Project Euclid, Scitation and Pubmed Central.

8. la requête précise est la suivante : (TS=Network*)AND (SO=("Science" OR "Nature" OR "Proceeding of the National Academy of Science" OR "Nature Genetics" OR "Annual Review of Genetics" OR "Annual Review of Biochemistry" OR "Annual Review of Cell and Developmental Biology" OR "Annual Review of Genomics and Human Genetics" OR "Journal of Theoretical Biology" OR "Biochimica et Biophysica Acta" OR "Nucleic Acids Research" OR "Journal of Molecular Biology" OR "Genetics" OR "Current Biology" OR "Genome Research" OR "Genome Biology" OR "Bioinformatics" OR "Biosystems" OR "BMC Systems Biology"))

tel-00464316, version 1 - 16 Mar 2010

construire une liste de termes caractéristiques extraits des abstracts de cette collection d'articles. Un ensemble de plus de 800 termes (liste en annexe B.2) a ainsi été sélectionné en privilégiant les termes les plus fréquents (après avoir éliminé les "stop-words"). Une matrice de cooccurrences des termes a été construite à partir de notre corpus de publications de *Pubmed*.

Notre dernier domaine d'étude concerne le champ du *développement durable*. Le corpus de publications a été construit par Marc Barbier et Andrei Mogoutov à partir de la base de données de publications CAB [9]. Cette plate-forme regroupe près de 5 millions de publications scientifiques de sciences naturelles appliquées couvrant les disciplines suivantes : sciences de l'agriculture, sciences de l'environnement, alimentation & santé humaine, microbiologie et parasitologie, sciences des plantes. Une requête spécifique que nous avons reproduit en annexe C a été construite afin d'extraire de façon aussi précise que possible l'ensemble des publications rattachées au domaine d'étude, le développement durable (Barbier et al., 2008). Ce sont finalement environ $70,000$ publications apparentées à la thématique qui ont été rassemblées. Nous avons par la suite extrait des mots-clés de ces publications (en nous référant aux mots-clés choisis par les revues) afin d'en extraire les quelques 650 termes qui apparaissent plus de 80 fois dans le corpus (cf. l'annexe B.3)

4.3.2 Une mesure asymétrique de proximité entre termes

Notre premier objectif est de définir une mesure de proximité entre termes en nous appuyant exclusivement sur les statistiques brutes du nombre d'occurrences et de cooccurrences calculées sur un ensemble de termes $\mathcal{L}$. Ainsi, étant donnés deux termes i et j, l'indexation du corpus de publications permet d'extraire les valeurs : n_i, n_j, et n_{ij}, qui correspondent respectivement au nombre d'articles dans lesquelles apparaissent i, j ou les deux termes i et j. Nos statistiques sont en réalité équivalentes au nombre de pages renvoyées par un moteur de recherche auquel on adresse une requête de type : "i", "j" ou "i AND j". Ces statistiques brutes sont, de plus, dynamiques. On peut les calculer chaque année afin d'obtenir leur profil d'évolution temporel comme l'illustre la figure 4.4 sur une sélection de termes.

De nombreuses mesures du degré de similarité ou de proximité entre deux termes à partir de leurs occurrences et co-occurrences ont été utilisées en scientométrie (voir He (1999) pour une revue). Nous pouvons citer entre autres l'indice d'inclusion qui s'exprime sous la forme : $\frac{n_{ij}}{min(n_i,n_j)}$ l'indice d'équivalence : $\frac{n_{ij}^2}{n_i.n_j}$ (Michelet, 1988; Callon et al., 1986), l'indice de Jaccard $\frac{n_{ij}}{n_i+n_j-n_{ij}}$(qui mesure un taux de recouvrement de i vis-à-vis de j), ou encore l'indice de proximité $\frac{n_{ij}N}{n_in_j}$. D'autres mesures ont été introduites par la suite. Néanmoins, la plupart synthétisent la relation entre deux termes sous la forme d'un simple scalaire. Or, une

9. http ://www.cabi.org/datapage.asp ?iDocID=228

tel-00464316, version 1 - 16 Mar 2010

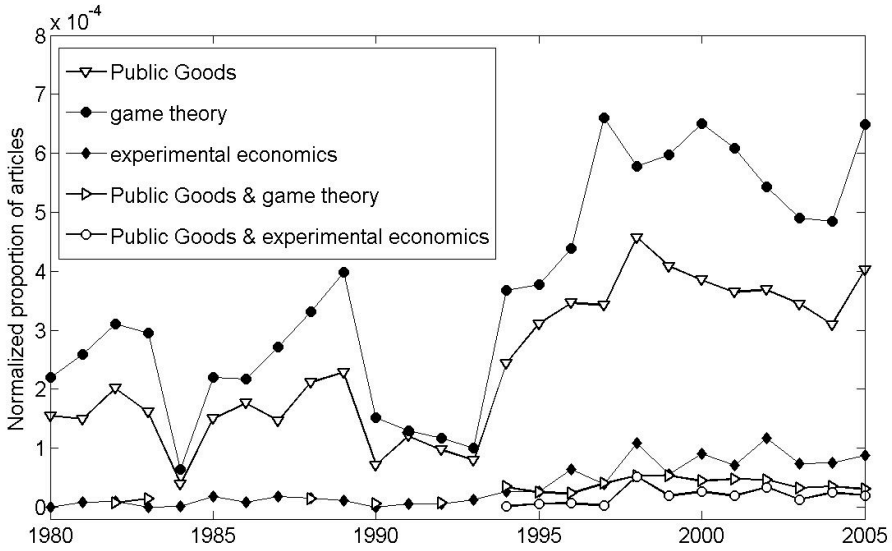


FIGURE 4.4: Dynamiques comparées des fréquences d'occurrence et de cooccurrence des termes *Public Goods*, *Game Theory* et *Experimental Economics* extraites de notre base de données Systèmes complexes.

mesure de proximité symétrique ne permet pas *a priori* de rendre compte de l'hétérogénéité des fréquences d'usage des termes. Ces mesures ne permettent pas, étant donnés deux termes dont les fréquences sont très différentes, de distinguer entre le terme le plus "générique" et le plus "spécifique".

Or la structure hiérarchique théorique des champs scientifiques que nous mettions en exergue dans la section précédente nous paraît capitale pour la bonne compréhension de l'organisation des champs scientifiques. Et cette hiérarchisation se retrouve naturellement dans la distribution du nombre d'occurrences des termes apparaissant dans nos corpus qui se caractérise par une forte hétérogénéité. On peut décrire ces distribution par une loi de Zipf, comme il est classique de l'observer sur les fréquences d'occurrences de termes tirés de corpus de langage naturel (voir (Steyvers and Tenenbaum, 2005) même si cette propriété n'est pas systématique (Lieberman et al., 2007)...). Nous avons représenté pour l'un de nos jeux de données la distribution de ces fréquences d'occurrences figure 4.5.

Cette hétérogénéité naturelle de la distribution des fréquences des concepts d'un corpus de texte peut avoir une conséquence primordiale sur les mesures de proximité que l'on réalise. À titre d'exemple, sur la figure 4.4, les occurrences et co-occurrences du terme *Public Goods* (noté PG par la suite) avec *Game theory* (GT) et *Experimental economics* (EE) ont été tracées. A priori, *Game theory* et *Experimental economics* sont deux termes pertinents par rapport aux études liées aux *Public Goods*. Néanmoins le terme *experimental economics* est par nature plus spécifique et par conséquent moins fréquent que *game theory* dans la littérature (ce dernier est

tel-00464316, version 1 - 16 Mar 2010

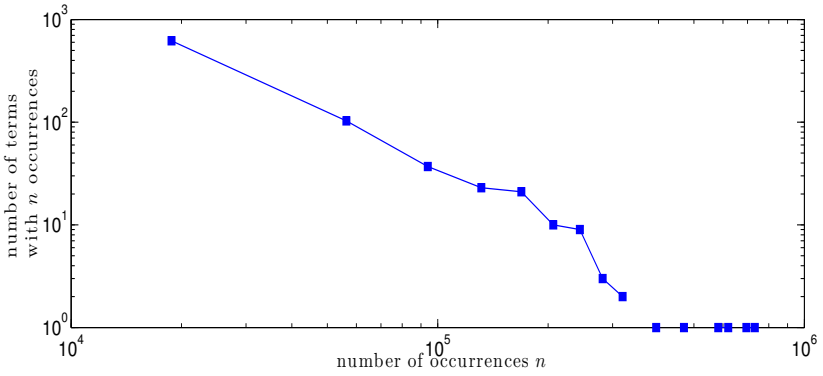


FIGURE 4.5: Distribution des occurrences des termes extraits de notre base de données biologie & réseau

au moins 5 fois plus fréquent que le premier).

Les probabilités conditionnelles $P(GT|PG)$ et $P(EE|PG)$ qu'un article incluant déjà *Public Goods* mentionne également *Game theory* ou *Experimental economics* sont comparables. Par contre, les probabilités conditionnelles inverses : $P(PG|GT)$ et $P(PG|EE)$ sont très différentes, la dernière étant beaucoup plus importante que la première. Cela est directement imputable aux différences de fréquences entre les termes [10]. Le terme *Public Goods* est donc très largement utilisé dans les études invoquant l'économie expérimentale alors que les études s'appuyant sur la théorie des jeux ne relèvent pas nécessairement de la question des *Public Goods*.

Les mesures de proximité utilisées classiquement ne permettent pas de repérer des relations proches de l'inclusion lorsque les deux termes ont des fréquences très différentes (c'est le cas de l'indice d'équivalence (d'ailleurs également appelé coeffcient d'inclusion mutuelle (Turner et al., 1988)) l'indice de Jaccard ou encore de l'indice de proximité) ni de distinguer entre un terme générique et un terme spécifique à partir de la simple observation de leur proximité (l'indice d'inclusion étant symétrique, même cette mesure ne permet en rien d'indiquer lequel des deux termes "est inclus" dans l'autre).

Nous pouvons, en complément de la figure 4.6 qui en donne une représentation ensembliste, résumer les différentes configurations possibles lorsqu'on entreprend de mesurer la proximité entre deux termes i et j :

1. $p(i|j)$ **haut,** $p(j|i)$ **haut :** i et j appartiennent au même champ et ont des fréquences similaires,
2. $p(i|j)$ **bas,** $p(j|i)$ **haut :** j est générique relativement au terme i (*e.g.* i = *Public Goods* et j = *Experimental economics*),
3. $p(i|j)$ **haut,** $p(j|i)$ **bas :** j appartient à un sous-domaine spécifique de i (*e.g.* i=

10. Rappelons la formule de Bayes : étant donné deux événements A et B, $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$

tel-00464316, version 1 - 16 Mar 2010

Experimental economics et j = *Public Goods*),

4. $p(i|j)$ **bas,** $p(j|i)$ **bas :** i et j sont peu connectés l'un avec l'autre, indépendamment de leur fréquence respective.

Seule la combinaison des indices d'équivalence et de l'indice d'inclusion permettrait *a priori* de couvrir l'ensemble des configurations possibles (recouvrement mutuel de termes de fréquences semblables pour le premier et configuration d'inclusion hiérarchique (direction exclue) pour le second). Afin de pouvoir discriminer ces configurations, nous proposons une nouvelle mesure de proximité $\mathcal{S}$ que nous appellerons également mesure de similarité. $\mathcal{S}$ est conçue comme une mesure *asymétrique* de façon à rendre compte de la directionnalité des relations d'inclusion entre termes (i spécifique/générique relativement à un terme j plus générique/spécifique).

Notre mesure de proximité $\mathcal{S}$ doit remplir les conditions suivantes :

1. $\mathcal{S}(i,j) = f(n_{ij}, n_i, n_j) \geq 0$, on souhaite exprimer la similarité entre deux termes i et j à partir des seules statistiques sur leurs nombres d'occurrences et de cooccurrences, cette mesure est toujours positive.
2. $\mathcal{S}(i,j) = 0$ ssi $n_{ij} = 0$, deux termes à recouvrement nul (aucun article ne les mentionne simultanément) ont une proximité nulle.
3. $\mathcal{S}(i,i) = 1$, la proximité maximale vaut 1, elle est obtenue ssi lorsque les ensembles d'articles mentionnant chaque terme sont parfaitement identiques.
4. $\mathcal{S}(i,j)$ croissant avec n_{ij}, un plus large recouvrement entre deux termes traduit une plus grande proximité entre termes. $\frac{\partial f}{\partial n_{ij}} > 0$ à nombres d'occurrences (n_i et n_j) constants.
5. $\mathcal{S}(i,j)$ est décroissant par rapport à n_i et n_j : toutes choses égales par ailleurs, l'augmentation de la fréquence du terme j l'éloigne de i vis à vis de $\mathcal{S}$. $f(n_{ij}, n_i, n_j)$ est donc une fonction croissante vis à vis de sa première coordonnée et décroissante vis à vis des deux autres, toutes choses égales par ailleurs.
6. La similarité doit être indépendante de la taille de l'échantillon d'articles sur lesquels sont calculées les statistiques, ainsi la fonction f doit être homogène, *i.e.* $f(\lambda x, \lambda y, \lambda z) = f(x,y,z)$. Nous en déduisons que f peut s'exprimer comme une fonction de deux paramètres : $\mathcal{S}(i,j) = f(1, n_{ij}/n_i, n_{ij}/n_j)$. On réécrit donc $\mathcal{S}(i,j) = f(n_{ij}/n_i, n_{ij}/n_j)$ qui s'exprime donc comme une fonction de deux variables : n_{ij}/n_i et n_{ij}/n_j vis à vis desquelles elle est croissante.
7. La fonction f doit être continue en $(0,0)$ avec $f(0,0) = 0$ d'après 2.

Si nous écrivons maintenant le développement de Taylor de $\mathcal{S}$ en 0, nous avons : $f(x,y) = \mu_0 + \mu_{1,0}x + \mu_{0,1}y + \mu_{2,0}x^2 + \mu_{0,2}y^2 + \mu_{1,1}xy + \mu_{3,0}x^3 + \mu_{3,0}y^3 + \mu_{1,2}xy^2 + ...$ D'après les conditions (2) et (7) on peut déduire que $\mu_0 = 0$, $\mu_{1,0} =$

tel-00464316, version 1 - 16 Mar 2010

$\mu_{0,1} = \mu_{2,0} = 0$ etc.... Ainsi f peut être écrit comme la somme des produits croisés :

$$f(\frac{n_{ij}}{n_i}, \frac{n_{ij}}{n_j}) = \sum_{k=1}^{\infty} \sum_{l=1}^{i-1} \mu_{k,l-k} (n_{ij}/n_i)^k (n_{ij}/n_j)^{l-k}$$

.

Les fonctions de type *Cobb-Douglas* $f_{\alpha,\beta}(x, y) = x^\alpha y^\beta$ constituent la classe de fonctions la plus simple répondant à l'ensemble de ces contraintes. f étant une fonction croissante en $\frac{n_{ij}}{ni}$ et $\frac{n_{ij}}{nj}$, $\alpha > 0$ et $\beta > 0$. Afin de réduire la gamme des paramètres et garantir certaines conditions de navigabilité, nous optons pour la mesure suivante :

$$\mathcal{S}^\alpha(i, j) = \left((n_{ij}/n_i)^\alpha (n_{ij}/n_j)^{\frac{1}{\alpha}}\right)^{min(\alpha, \frac{1}{\alpha})}$$

Nous appellerons *focus* le paramètre α. Cette mesure répond à toutes les contraintes listées et possède une propriété supplémentaire dont nous discuterons les conséquences dans la section suivante : $\mathcal{S}^\alpha(i, j) = \mathcal{S}^{\frac{1}{\alpha}}(j, i)$.

Notre proximité permet de définir le voisinage d'un terme cible i étant donné un seuil s et un focus α comme :

$$V_{s,\alpha}(i) = \{j | \mathcal{S}^\alpha(i, j) > s\}$$

Elle peut s'interpréter géométriquement comme une mesure de pseudo-inclusion (pour un focus suffisant). Asymptotiquement on obtient une mesure d'inclusion pure : si $\alpha \to 0$, $\mathcal{S}^\alpha(i, j) > 0 \iff n_{ij} = n_j$ et si $\alpha \to \infty$, $\mathcal{S}^\alpha(i, j) > 0 \iff n_{ij} = n_i$. De plus si on choisit $\alpha = 1$, notre proximité redevient symétrique, et on retrouve l'indice d'équivalence e (Callon et al., 1991). e aura tendance à rapprocher les couples de termes (i, j) dont la partie recouvrante (n_{ij}) est importante vis à vis des occurrences de chacun des termes (n_i et n_j), *i.e.* il y a inclusion mutuelle. Cette mesure est symétrique et privilégie à recouvrement égal les termes de fréquences semblables. La mesure d'inclusion pure ($\alpha = 0$), inscrira dans le voisinage immédiat d'un terme i l'ensemble des termes co-occurrant systématiquement avec i, indépendamment des valeurs respectives de n_i ou de n_j.

La mesure que nous adoptons se situe entre ces deux extrêmes (indice d'équivalence et inclusion pure), c'est pourquoi nous l'appelons également mesure de pseudo-inclusion. Dans ce cas, notre mesure de proximité permet vis-à-vis d'un terme i donné, de favoriser dans son voisinage les termes j qui sont fortement recouvrants avec i, en pénalisant ceux dont le ratio $n_{ij}/n_i = p(j|i)$ est faible (si $\alpha > 1$) ou dont le ratio $n_{ij}/n_j = p(i|j)$ est faible (si $\alpha < 1$). Ainsi notre mesure $\mathcal{S}^\alpha$ paramétrée par un focus $\alpha > 1$ aura tendance à privilégier (au sens où ($\mathcal{S}^\alpha(i, j)$ est important, et toujours à recouvrement égal) vis-à-vis d'un terme cible i des termes j qui sont des bons contextes vis à vis de i. Inversement, si $\alpha < 1$, les voisins les plus proches de i seront plutôt bien contextualisés par i.

Pour résumer, étant donné un terme i et cherchant ses plus proches voisins vis-à-vis de $\mathcal{S}$,

tel-00464316, version 1 - 16 Mar 2010

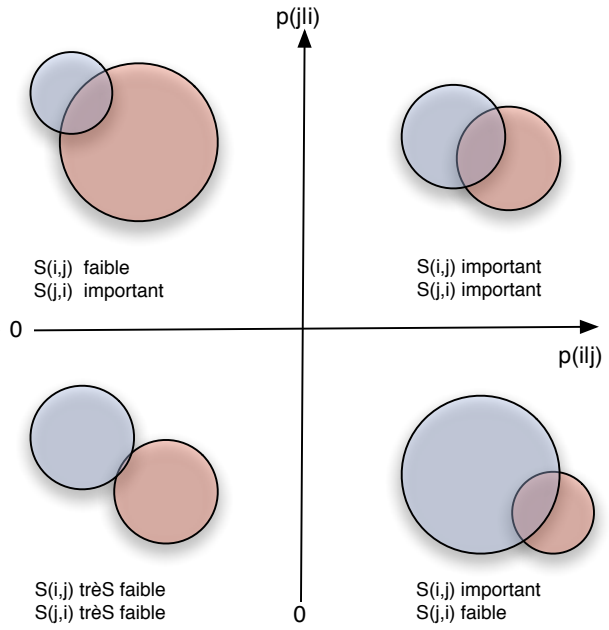


FIGURE 4.6: Représentation schématique des effets de la mesure de similarité $\mathcal{S}$ en fonction de différents agencement possibles entre deux termes, plus ou moins fréquents, et plus ou moins recouvrants, ($\alpha = 0.1$)

- pour un focus $\alpha < 1$, notre mesure $\mathcal{S}$ sera importante vis-à-vis de termes qui sont plutôt spécifiques dans le contexte de i,
- pour $\alpha > 1$, $\mathcal{S}$ sera importante pour des termes j qui constituent des bons contextes pour i.

Dans la suite le paramètre de focus est fixé et vaut $\alpha = 0.1$ [11]. La figure 4.6 permet de se représenter les valeurs respectives prises par $\mathcal{S}^{\frac{1}{10}}(i,j)$ en fonction des différentes configurations attendues. Dans notre exemple le terme i est proche de j (au sens où $\mathcal{S}(i,j)$ est important) lorsque i constitue un bon contexte pour j ou lorsque i et j ont des fréquences similaires tout en ayant un grand nombre de cooccurrences, mais pas lorsque j est un bon contexte pour i. L'identité $\mathcal{S}^{\alpha}(i,j) = \mathcal{S}^{1/\alpha}(j,i)$ est illustrée dans la même figure par la symétrie que l'on observe en comparant les configurations mettant en jeu des termes de fréquences différentes (au haut à gauche et en bas à droite). Intervertir i et j dans la formule revient alors à inverser la valeur de α, ce qui a pour conséquence de rapprocher du terme cible les termes qui constituent un bon contexte à son égard. Comme nous le préciserons dans la

11. NB : la valeur de α retenue est relativement petite et nous sommes en réalité dans une situation quasi équivalente à une expression de la proximité beaucoup plus simple de la forme : $\mathcal{S}(i,j) = n_{ij}/n_i$, néanmoins, notre expression étendue permet qualitativement de privilégier des couples de termes dont les termes se situent dans une gamme de fréquences "raisonnable". Dans la pratique, les opérations de cartographie à venir sont relativement peu affectées par la valeur de α.

tel-00464316, version 1 - 16 Mar 2010

section suivante, le paramètre de focus permet "d'orienter" la recherche de termes voisins soit vers des termes de même importance ($\alpha = 1$) soit vers des termes plus spécifiques que le terme cible original ($\alpha < 1$), soit vers des termes plus génériques ($\alpha > 1$).

4.3.3 Construction du réseau lexical

Etant donné un ensemble $\mathcal{L}$ de termes et un corpus de textes dont on peut extraire les statistiques brutes d'occurrences et de cooccurrences des termes de $\mathcal{L}$. On définit alors $G_s = (\mathcal{L}, E_s)$, le réseau lexical dirigé dont l'ensemble des liens E_s correspond à la matrice d'adjacence $\mathbf{S}_s$ telle que $\mathbf{S}_s(i,j) = H_s(\mathcal{S}^\alpha(i,j))$ où $H_s(x)$ est une fonction seuil valant 1 si $x \geq s$, 0 sinon. Plus simplement, on peut définir G_s comme le réseau dont les liens relient les termes i aux termes dans le voisinage de i : $V_{s,\alpha}(i)$.

Naturellement, l'ensemble des mesures et notations que nous avons introduit jusque-là, dépend d'un jeu de données couvrant une période bien définie. Ainsi, même si nous n'abordons pas les aspects dynamiques pour le moment et ne précisons pas cette dépendance dans les notations, notre mesure de distance entre termes, le voisinage d'un terme, ainsi que le réseau lexcial G_s qui en découle dépendent naturellement de la période d'observation retenue et sont susceptibles d'évoluer lorsqu'on les applique à un corpus dynamique. Nous ne préciserons cette dépendance dans nos notations que lorsqu'il y a ambiguïté. Nous simplifierons également $\mathcal{S}^\alpha$ en $\mathcal{S}$ lorsque le focus prendra sa valeur de référence : 0.1.

4.3.4 Echelle microscopique : voisinages locaux

Le paramètre de focus α offre un moyen élégant de naviguer avec un point de vue local à travers notre corpus grâce à la notion de voisinage. Etant donné un terme cible, nous cherchons à identifier ses plus proches voisins. Le paramètre α permet d'accéder à deux types de voisinage. Pour de faibles valeurs de α ($\alpha < 1$), les plus proches voisins ont tendance à avoir un caractère plus générique que le terme cible. Si le paramètre α est plus important ($\alpha > 1$), on retrouvera préférentiellement des termes plus spécifiques que i. La figure 4.7, extraite de notre cas d'étude sur un corpus de termes portant sur les *systèmes complexes*, illustre cette propriété. Nous avons tracé le voisinage $V_{s,\alpha}$ du terme *knowledge discovery* pour $\alpha = 0.1$ et $\alpha = 10$ et une valeur de seuil s fixée. Pour $\alpha = 0.1$, les termes les plus proches de *Public Goods* le spécifient *via* les termes utilisés dans les sous-spécialités du domaine (dans l'exemple figure 4.7 "collective action", "consumer sovereignty", etc...). *A contrario*, un paramètre de focalisation supérieur à 1, ici $\alpha = 10$ a tendance à assigner au voisinage du terme cible l'ensemble de ses contextes (dans notre exemple : "policy", "development", "environment", etc...).

tel-00464316, version 1 - 16 Mar 2010

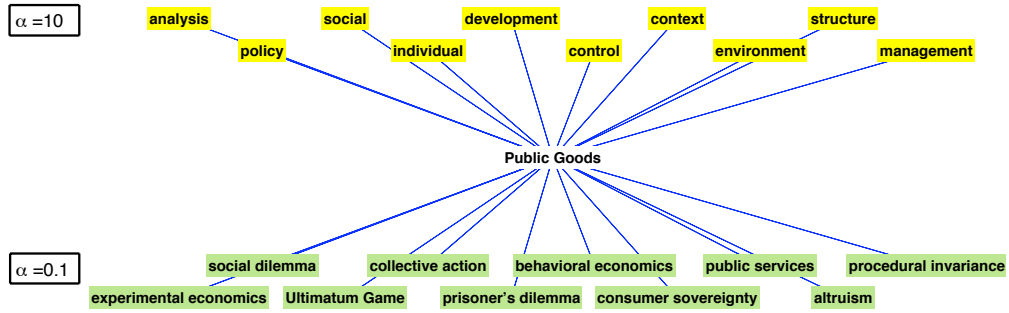


FIGURE 4.7: Voisinages en spécificité et en généricité du terme *Public Goods*. Selon la valeur de α, on obtient parmi les plus proches voisins de notre terme cible, l'ensemble des termes qui ont tendance à le spécifier ($\alpha = \frac{1}{10}$ en vert), ou l'ensemble des termes qui le contextualisent ($\alpha = 10$, en jaune)

4.4 Echelle mésoscopique : la notion de champ épistémique

Dans cette section, nous abordons la seconde partie de notre travail de reconstruction en proposant une méthode de détection et de représentation des agrégats de termes qui structurent notre réseau lexical G_s. Nous proposons une méthode pour détecter ces ensembles de termes fortement interconnectés les uns aux autres, appelés *champs épistémiques*. En nous appuyant sur la mesure de proximité asymétrique $\mathcal{S}$, nous introduisons également une représentation bi-dimensionnelle du contenu de ces champs ainsi que deux indices permettant de quantifier la structuration de ces champs. Enfin, nous définirons un carte conceptuelle à partir de l'articulation entre ces champs et tenterons d'en fournir une représentation intelligible.

4.4.1 Définitions

L'opération de cartographie nécessite souvent, eu égard au grand nombre d'entités à représenter, une première étape de *réduction* du réseau lexical à travers des méthodes de *catégorisation* qui réunissent au sein de clusters des ensembles de termes densément interconnectés. Mais cette opération n'est pas uniquement guidée par des besoins techniques, elle vise également à identifier des *domaines de spécialité* ("research specialities") pour reprendre le terme employé par Chubin (1976). Morris and der Veer Martens (2008) définissent de la façon suivante ces assemblages hybrides de textes et de chercheurs :

> "the research specialty is the largest homogenous unit in the self-organizing systems of science, in that each specialty tends to have its

tel-00464316, version 1 - 16 Mar 2010

> own set of problems, a cohesive core of researchers, shared knowledge, vocabulary, and archival literature." [12]

Cette définition peut être directement rapprochée de la (ou de l'une des) définition(s) que Kuhn (1970b) donne d'un paradigme :

> "a paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men who share a paradigm

Même si l'essentiel des études sur ces " spécialités de recherche" s'est concentrée en sociologie des sciences sur la structure sociale qui les anime (espaces de communication, système d'évaluation, processus d'accumulation de ressources et de capital, diffusion de connaissances au sein de collèges invisible (Crane, 1972) etc.), ces champs sont également déterminés et peuvent être détectés grâce aux propriétés cognitives qui les structurent (Chubin, 1976; Chen et al., 2002). Dans notre analyse nous nous concentrerons sur la seule dimension cognitive de la production scientifique tracée à travers le recueil des publications. Cela ne signifie pas que les structures que nous cherchons à mettre en évidence sont des constructions exclusivement cognitives, mais que nous en cherchons des traces dans le seul réseau lexical G_s.

La notion de "spécialité de recherche" a été largement discutée dans la littérature. Knorr-Cetina (1982), notamment, critique sévèrement le concept "quasi-économique" et fonctionnaliste de spécialité scientifique comme unité d'étude pertinente pour comprendre l'organisation technique et sociale de la science. Elle lui dénie toute forme opératoire vis-à-vis de l'activité scientifique dans les laboratoires ou même vis-à-vis des représentations mentales des chercheurs. Elle lui oppose la notion d'arènes de recherche trans-épistémiques qui mêlent problématiques techniques et non-techniques, spécialisées et non-spécialisées. La critique avancée vise plutôt à dénoncer le caractère fermé et la croyance en une dynamique endogène des processus de production de connaissance au sein de ces espaces :

> "The point here is that if we cannot assume that the 'cognitive' or 'technical' selections of scientific work are exclusively determined by a scientist's specialty membership groups, it makes no sense to search for a 'specialty community' as the relevant setting for knowledge production."

L'approche "micro", suivant l'activité quotidienne du chercheur développée par Knorr-Cetina (1982) montre bien la multiplicité des "transactions" négociées entre spécialistes et non-spécialistes dans le processus d'élaboration de la connaissance.

12. "la spécialité de recherche est la plus grande unité homogène dans le système auto-organisé que forme la science, au sens où chaque spécialité de recherche tend à avoir ses propres problématiques, un cœur cohesif de chercheurs, une connaissance partagée, un vocabulaire spécifique, et un ensemble de références communes."

tel-00464316, version 1 - 16 Mar 2010

Néanmoins ces éléments d'analyse n'interdisent pas, selon nous, l'existence de "champs épistémiques" qui malgré l'hétérogénéité intrinsèque de leur constituants (qu'on parle ici d'éléments purement cognitifs de tout ordre (des outils de recherche comme un programme informatique, à un animal modèle en passant par des artefacts argumentatifs mobilisés dans un article), ou d'intervenants humains (du chercheur au manager de la science en passant par le technicien)) ne forment pas moins des ensembles cohérents et signifiants pour l'ensemble des acteurs engagés. Il ne s'agit certes pas d'assigner à un ensemble de scientifiques et de concepts une spécialité et d'en fermer la porte à double tour, mais de repérer des structures émergentes signalant à un moment donné "la cristallisation" (comme l'appelle Chubin) d'une singularité remarquable au sein du réseau d'interaction complexe mettant en jeu un ensemble d'acteurs et d'artefacts cognitifs. Ces structures ne sont pas des constructions sociales fantasmées, leurs formes institutionnelles dont on ne peut nier le caractère performatif en témoignent (conférences, organisation par départements des organismes de recherche, organisation thématique des appels à projet, etc.). Elles ne sont pas non plus des structures "en vase clos", l'activité d'un chercheur n'est pas nécessairement restreinte aux limites d'un seul champ. La circulation des personnes et des concepts est sans doute fondamentale à la viabilité d'un champ et la caractérisation des champs épistémiques est indissociable de l'identification des ponts qui les relient. La notion de multiplicité des appartenances est ici fondamentale. Et cette multiplicité se joue à nouveau aussi bien du point de vue des acteurs qui animent ces communautés que des concepts qui y circulent.

C'est pourquoi un champ épistémique ne saurait être défini de façon univoque comme la monade au sein de laquelle un certain type de connaissance est produite par un certain nombre de personnes bien identifiées, mais comme un lieu temporaire (mais suffisamment pérenne pour être observable) de cristallisation de certaines questions et de certains enjeux travaillés par un certain nombre d'individus potentiellement engagés en parallèle dans d'autres activités.

4.4.2 Identifier les champs épistémiques

Plusieurs méthodes de clusterisation ont été proposées et testées dans la littérature en scientométrie : on peut citer par exemple la méthode des "k-means" (Zitt and Bassecoulard, 2006; Boyack et al., 2005), les *Self-Organized Maps* (Skupin, 2004), ou une méthode récente basée sur les flux d'information développée par Rosvall and Bergstrom (2008a), etc. Malgré la diversité de ces méthodes, la majorité d'entre elles opère, en guise de clusterisation une partition du réseau (quelle que soit la nature du réseau : citations, co-publications, ou mots associés) qui cantonne par définition un nœud à un seul et unique cluster.

Or, dans la cas qui nous occupe, celui de la catégorisation d'un ensemble de termes saisis au travers du réseau de proximité G_s, on conçoit aisément que

tel-00464316, version 1 - 16 Mar 2010

certains termes puissent être mobilisés dans différents champs, ou même qu'ils possèdent différentes significations, ou que leur sens soit modifié selon les communautés dans lesquelles il est employé. C'est d'ailleurs la critique principale qu'adresse Leydesdorff (1997) à l'analyse des mots associés :

> "The subsumption of phenomenologically similar words or other textual signals under keywords or other concept symbols assumes stability in the meanings of the indicated concepts."

Nous affirmons au contraire que c'est précisément l'instabilité et la volatilité des sens qui nous intéressent ici car elles permettent de définir nos champs comme des agencements *plastiques*, susceptibles d'autoriser des appartenances multiples. Notre objectif est donc d'intégrer cette variabilité intrinsèque à notre entreprise de modélisation mais cette exigence implique de faire appel à des méthodes de catégorisation permettant la détection de clusters recouvrants.

Pour détecter les clusters au sein de notre réseau dirigé de proximité entre termes, nous faisons appel à l'algorithme de détection de percolation de cliques développé par Palla et al. (2005b). Cet algorithme fait partie de la famille récente des méthodes de détection de clusters recouvrants (comprenant entre autres les approches de Zhang et al. (2008) ou de Lancichinetti et al. (2009)). Ainsi les méthodes de détection de communautés classiques (Danon et al., 2005) visent à trouver la meilleure partition d'un graphe possible (comme le font les méthodes classiques d'optimisation de la modularité (Girvan and Newman, 2002; Blondel et al., 2008) ou d'autres méthodes fondées sur la construction de partitions à partir de l'analyse spectrale des graphes (Capocci et al., 2005; Newman, 2006) ou à partir de la dynamique de marches aléatoires (Latapy et al., 2008; Rosvall and Bergstrom, 2008b)).

La méthode que nous employons ici, *a contrario* de ces dernières méthodes algorithmiques, est une méthode purement algébrique et déterministe. Elle se déroule en deux étapes. En premier lieu, l'ensemble des cliques (dirigées, dans le cas d'un graphe orienté (Palla et al., 2007b)) du graphe considéré sont détectées. Puis l'algorithme construit les communautés de k-cliques pour toutes les tailles de cliques k possibles en opérant une percolation de k-cliques. Plus précisément, une communauté de k-cliques est définie comme un ensemble de k-cliques (sous-graphes de taille k), qui partagent la propriété suivante : il est possible, depuis n'importe quelle k-clique d'une communauté de k-clique donnée, d'atteindre n'importe quelle autre k-clique de cette communauté en suivant une série de k-cliques adjacentes (deux k-cliques étant adjacentes si elles partagent $k-1$ nœuds). La méthode employée permet de réunir en une même "communauté" un ensemble de termes fortement inter-connectés les uns aux autres, ce qui nous semble être un critère pertinent pour repérer les structures cognitives régulières dans l'activité scientifique reflétant l'usage d'un vocabulaire, d'outils techniques ou conceptuels communs à un champ épistémique donné.

tel-00464316, version 1 - 16 Mar 2010

Nous appliquons cet algorithme dans sa version orientée [13] (Palla et al., 2007b) à notre réseau lexical G_s [14]. L'algorithme [15] permet de détecter un ensemble de clusters $\mathcal{C} = \{C_i\}_{i \in I}$ que nous appellerons *champs épistémiques* [16]. Un champ $C_i \subset \mathcal{L}$ correspond donc à l'ensemble des termes appartenant à la même communauté de k-cliques.

Le choix d'un seuil s pertinent, en-deça duquel les liens sont considérés comme négligeables, est directement lié à l'algorithme de percolation de cliques. Si l'on se concentre sur l'ensemble des communautés de k-cliques obtenues pour différentes valeurs de seuil s, on constate pour un seuil $s_0(k)$ donné un phénomène de percolation qui produit une "communauté géante" agrégeant une grande partie des nœuds du réseau. En diminuant le seuil s légèrement en-dessous de $s_0(k)$, la structure de communautés obtenue est la plus informative possible. Plutôt que de nous cantonner à une taille de clique k fixée pour définir nos champs épistémiques, nous faisons l'inventaire de l'ensemble des communautés de k-cliques pour $k \geq 3$, en choisissant un seuil s_1 proche mais inférieur à $s_0(3)$. L'indice k d'une communauté de k-clique C_i donnée fournit alors un premier indice de cohésion de la dite communauté. Dans la suite on considérera ce seuil comme fixé, et on notera notre réseau lexical $G = G_{s_1}$.

Nous avons représenté figure 4.8 un exemple de catégorisation obtenue en appliquant l'algorithme de percolation de cliques qui illustre la possibilité de multi-appartenance d'un terme à un champ épistémique. Nous avons représenté deux champs auxquels le terme "Public Goods" appartient dans la base de données sur les systèmes complexes : un premier cluster est orienté *théorie des jeux* ; le second est plus proche des *sciences politiques*.

4.4.3 Plongement des clusters dans un espace bi-dimensionnel

Une fois l'ensemble des champs épistémiques $\mathcal{C}$ construit, nous proposons de plonger chacun des champs dans un espace bidimensionnel. Etant donné un champ C et un terme w, on définit l'indice de généricité I_g et l'indice de spécificité I_s comme suit :

indice de spécificité Il fournit une mesure du positionnement du terme w dans

13. Les améliorations les plus récentes de cet algorithme de percolation de cliques (Farkas et al., 2007) permettent théoriquement d'étendre la procédure à des réseaux dirigés *et* pondérés. Cela nous permettrait de travailler directement sur la matrice de proximité entre termes et donc sur un réseau lexical pondéré sans avoir à définir un seuil s. Néanmoins, la version simplement dirigée fournit d'ores et déjà des résultats convaincants.

14. Nous avons donc effectué une opération de réduction sur notre graphe lexical G_s en retirant les poids portés par l'ensemble de ses liens.

15. dont l'implémentation des auteurs librement accessible a été utilisée : http ://www.cfinder.org/

16. Il faut noter que l'algorithme ne catégorise pas nécessairement l'ensemble des termes $\mathcal{L}$, certains pouvant se trouver trop distants de l'ensemble de leurs voisins pour figurer dans quelque clique que ce soit.

tel-00464316, version 1 - 16 Mar 2010

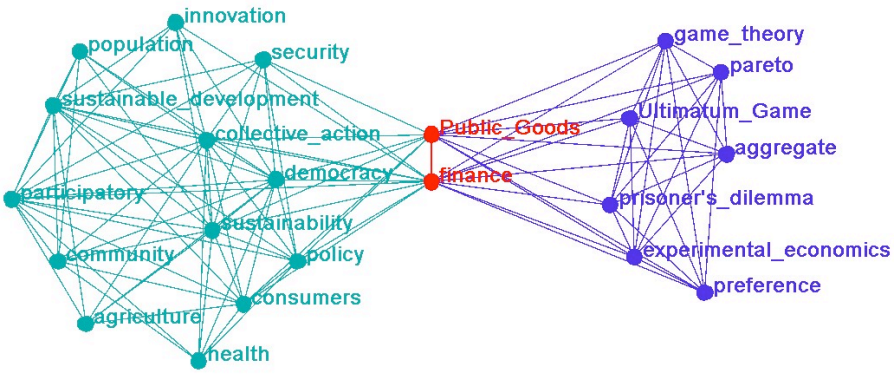


FIGURE 4.8: Deux clusters comprenant le terme *Public Goods*, en vert à gauche, un premier champ orienté *sciences politiques*, en mauve à droite, un champ orienté *théorie des jeux*. Les deux termes (*Public Goods* et *Finance*) partagés par les deux champs épistémiques sont en rouge (image extraite du logiciel CFinder).

un de ses champs d'appartenance C en tant que contexte vis-à vis de l'ensemble des termes du champ. Il est défini comme la somme des distances entrantes des termes de C vers w

$$I_s^\alpha(w) = \frac{1}{card(C)} \sum_{w' \in C} \mathcal{S}^{max(\alpha, \frac{1}{\alpha})}(w, w')$$

indice de généricité Il définit dans quelle mesure les éléments du champ C sont bien contextualisés par le terme w. On le définit comme la moyenne des distances sortantes de w à l'ensemble de ses voisins dans C :

$$I_g^\alpha(w) = \frac{1}{card(C)} \sum_{w' \in C} \mathcal{S}^{min(\alpha, \frac{1}{\alpha})}(w, w')$$

Ces deux indices permettent de représenter de façon intuitive les champs dans un espace à deux dimensions. À chaque terme, on attribue une coordonnée $(I_s^\alpha(w), I_g^\alpha(w))$ et une taille proportionnelle à son importance dans le champ (calculée comme la somme du nombre de ses co-occurrences avec les autres termes du champ). La couleur de chaque terme traduit le taux de croissance de son importance dans le champ entre deux périodes consécutives (du bleu pour les croissances négatives au rouge foncé, pour les croissances supérieures à 50% en passant par le blanc signalant une croissance nulle).

Pour l'illustrer, nous présentons figure 4.9 deux domaines qui partagent les termes "knowledge discovery". Tout comme le terme *Public Goods*, ce terme peut relever de plusieurs domaines distincts : un premier orienté *systèmes d'apprentissage automatique*, le second plus focalisé sur les enjeux propres à la *catégorisation* (cf figure 4.9). La représentation dans le référentiel des indices de spécificité et

tel-00464316, version 1 - 16 Mar 2010

de généricité permet d'organiser l'ensemble des éléments d'un champ selon une hiérarchie intuitive. Ce plongement dans un espace bi-dimensionnel correspond à une mesure au niveau mésoscopique tenant compte de l'ensemble des relations entre termes du champ. Elle est donc complémentaire mais différente de la représentation figure 4.7 qui offrait un point de vue purement local sur les voisinages de termes.

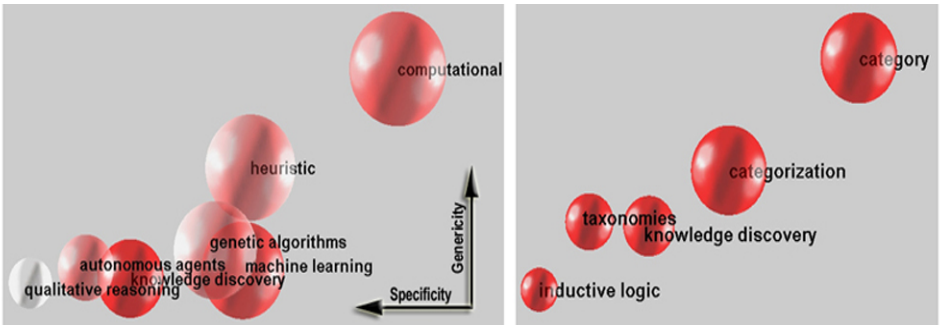


FIGURE 4.9: Deux champs épistémiques mentionnant le terme *Knowledge Discovery* sur la période 2002-2005 dans la base de données portant sur les systèmes complexes. *Knowledge Discovery* appartient à deux sphères de production de connaissance distinctes : à gauche un champ orienté vers le *machine learning*, à droite l'accent est mis sur les questions de *"categorisation"*. Dans cette représentation, I_s croît de droite à gauche, et I_g croît de bas en haut.

4.4.4 Qualifier les clusters

On peut associer un certain nombre de mesures pour qualifier les clusters obtenus. On s'orientera vers un type de caractérisation ou un autre en fonction de la façon dont on souhaite interroger ces cartes : quels sont les grands champs structurant un domaine scientifique, quels sont les champs qui émergent à un moment donné, quels sont les clusters interdisciplinaires, etc. ?

Nous introduisons deux types d'indices permettant de caractériser les champs épistémiques détectés : la *densité* et l'*indice de pseudo-inclusion*. La densité a été introduite par Callon et al. (1991) :

> "It characterizes the strength of the links that tie the words making up the cluster together. The stronger these links are, the more the research problems corresponding to the cluster constitute a coherent and integrated whole. It could be said that density provides a good representation of the cluster's capacity to maintain itself and to develop over the course of time in the field under consideration."

Formellement, la densité d'un champ C, notée $D(C)$, s'écrit avec nos notations : $D(C) = \frac{1}{Card(C)} \sum_{(w,w') \in C^2, w \neq w'} \mathcal{S}^1(w, w')$. D'autre part, on définit un autre indice de la cohésion d'un cluster, notre objectif étant de détecter des clusters dont

tel-00464316, version 1 - 16 Mar 2010

les termes satisfassent soit la contrainte d'être de bons contextes pour leurs voisins, soit la contrainte de bien spécifier leurs voisins. On définit donc l'*indice de pseudo-inclusion* d'un cluster : $I^{\alpha}_{\subset}(C) = \min_{w \in C} \frac{1}{2}\big(I^{\alpha}_{s}(w) + I^{\alpha}_{g}(w)\big)$. Cet quantité indique le degré de structuration de C. Les clusters avec un index de pseudo-inclusion peu élevé ont au moins un terme qui n'est ni bien contextualisé ni un bon contexte pour l'ensemble des autres termes (il n'est ni spécifique ni générique par rapport aux autres).

4.4.5 Représentation macroscopique

Nous avons défini les champs épistémiques comme des ensembles de termes qui cooccurrent préférentiellement les uns avec les autres, ces termes pouvant appartenir à plusieurs champs épistémiques différents. L'étape suivante consiste à donner un aperçu de l'articulation des différents champs épistémiques les uns avec les autres afin de fournir une vision globale et structurée du paysage scientifique formé par notre ensemble de termes au sein du corpus de publications. Cette représentation macroscopique de l'activité scientifique prend la forme d'une carte (Buter and Noyons, 2002; Shaikevich, 2005) qui reflète le paysage conceptuel.

Une procédure possible pour représenter la façon dont ces champs s'articulent les uns avec les autres est de définir un réseau dont chaque nœud correspond à un champ, et dont les liens sont pondérés par les valeurs de proximité entre champs. Afin de définir une mesure de similarité au niveau des champs épistémiques, nous étendons donc notre mesure de pseudo-inclusion entre deux termes en calculant cette fois la moyenne des similarités entre les termes respectifs de chaque cluster. On exprime ainsi la proximité $\hat{\mathcal{S}}$ entre deux champs C_a et C_b sous la forme suivante :

$$\hat{\mathcal{S}}(C_a, C_b) = \frac{1}{\mid C_a \mid} \sum_{i \in C_a} \left(\frac{1}{\mid C_b \mid} \sum_{j \in C_b} \mathcal{S}(i,j)\right)$$

Cette mesure permet de définir le réseau orienté entre champs épistémiques $\hat{G}_s = (\mathcal{C}, E_{\hat{s}})$ dont l'ensemble des liens $E_{\hat{s}}$ pondérés correspond à la matrice d'adjacence pondérée $\hat{\mathbf{S}}_s$ telle que $\hat{\mathbf{S}}_s(i,j) = \hat{\mathcal{S}}^{\alpha}(i,j) H_s(\hat{\mathcal{S}}^{\alpha}(i,j))$ où $H_s(x)$ désigne à nouveau une fonction seuil valant 1 si $x \geq s$, 0 sinon. À nouveau le seuil $s = s_2$ est choisi de façon à obtenir la structure la plus informative possible (en choisissant s_2 de façon à obtenir un réseau ni trop dense, ni trop déconnecté) [17]). Dans la suite on considérera ce seuil s_2 comme fixé, et on notera le réseau des champs épistémiques $\hat{G} = \hat{G}_{s_2}$.

Mais la définition du réseau des champs épistémiques $\hat{G}$ ne suffit pas nécessairement à fournir une description satisfaisante de l'activité scientifique d'un

17. NB : lorsqu'il s'agit de comparer deux cartes, calculées à deux moments distincts par exemple, il faut naturellement veiller à ce que les seuils retenus pour la construction des cartes soient identiques.

tel-00464316, version 1 - 16 Mar 2010

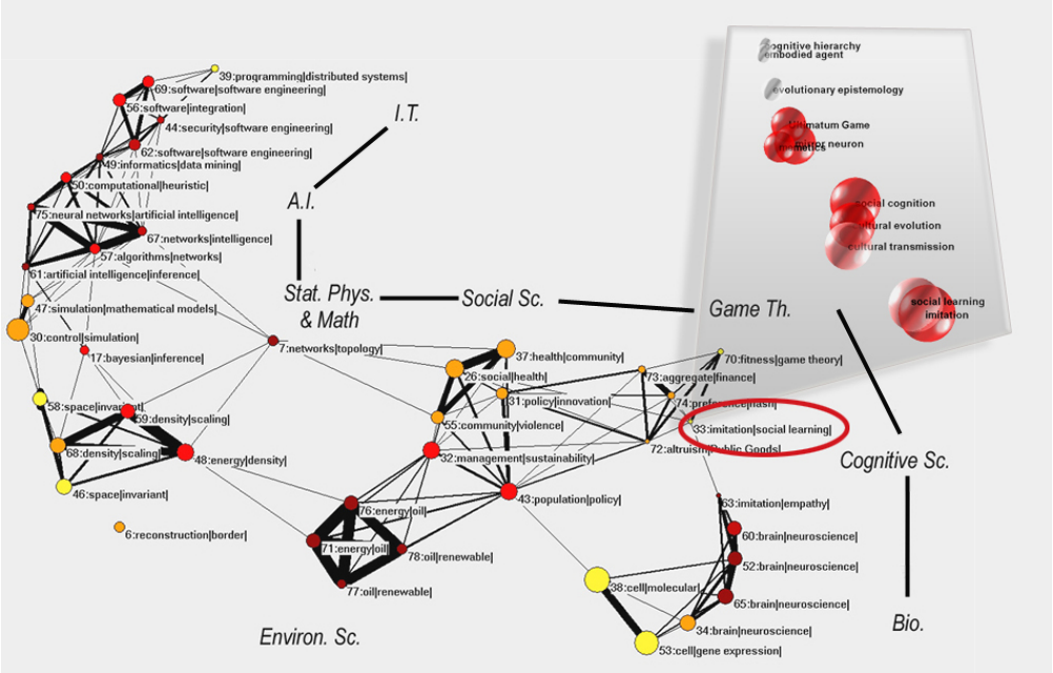


FIGURE 4.10: Carte macroscopique du champ des *systèmes complexes*. La taille des champs correspondent à l'activité des champs (échelle logarithmique), la couleur, du jaune le plus clair au rouge le plus foncé représente le taux de croissance de cette activité entre la période courante (années 2002-2005) et la période précédente (1998-2001). Chaque champ peut être décomposé en l'ensemble des termes qui le constitue et représenté dans notre référentiel bidimensionnel (voir encart, l'axe I_g a été inversé par rapport à la figure précédente)

domaine. En effet les champs épistémiques peuvent être composés d'un grand nombre de termes, et les opérations précédentes paraîtraient superflues s'il s'agissait *in-fine* de faire figurer sur nos cartes l'ensemble des termes en guise de légendes des champs détectés, aussi ingénieux leur agencement spatial soit-il. Pour simplifier notre représentation il nous faut donc étiqueter de la manière la plus pertinente et économique possible chaque champ.

Plusieurs stratégies sont envisageables pour donner à chaque champ la légende qui le représente le plus fidèlement. Une première possibilité est d'employer une méthode basée sur l'analyse des fréquences des termes au sein de l'ensemble des champs afin d'extraire les termes les plus prototypiques de chaque champ. Nous exposerons précisément cette méthode ultérieurement dans un contexte plus favorable qui est celui des représentations de plus haut niveau mettant en jeu des clusters de champs épistémiques (cf section 4.4.6). À ce stade, nous nous appuierons directement sur la décomposition des champs dans notre espace d'indices de spécificité/généricité pour réaliser notre opération de labellisation. En fonction de ce qui paraîtra le plus pertinent au destinataire de la carte, on peut privilégier différentes stratégies pour représenter le contenu d'un champ :

tel-00464316, version 1 - 16 Mar 2010

- sélectionner les termes les plus spécifiques vis-à-vis du champ, *i.e.* ceux dont les indices de spécificité I_s sont les plus importants,
- sélectionner les termes les plus générique vis-à-vis du champ, *i.e.* ceux dont les indices de généricité I_g sont les plus importants,
- sélectionner les termes "médians" qui sont les plus centraux vis-à-vis de la dimension dont la variance est la plus grande,
- ou encore adopter des solutions mixtes entre ces trois premières méthodes.

Nous avons représenté une carte du domaine des *systèmes complexes* pour la période 2002-2005 figure 4.10. La représentation de cette carte (ainsi que les suivantes) a été réalisée grâce au logiciel d'analyse et de représentation de graphes Pajek (Batagelj and Mrvar, 1998) qui s'appuie sur des algorithmes classiques de spacialisation (Fruchterman–Reingold ou Kamada-Kawai). Les champs ont été étiquetés par leurs deux termes les plus génériques. La description des champs est alors suffisamment condensée pour fournir une représentation macroscopique du domaine, permettant de repérer très facilement les grands sous-domaines qui le compose (ceux-ci ont été rajoutés à la main sur la carte (*I.T.*, *A.I.*, *Physique statistique*, etc.)). La taille d'un nœud est proportionnelle à l'activité a du champ définie comme la moyenne des occurrences normalisées (p_i^T) $= \frac{n_i^T}{\sum_{j\in\mathcal{L}} n_j^T}$) des termes exprimés au sein d'un champ (l'activité d'un champ C à une période T s'exprime donc sous la forme : $a_C^T = \frac{1}{card(C)}\sum_{i\in C} p_i^T$). Nous pouvons également visualiser sur cette carte la croissance A_C^T de l'activité de chaque champ C à T, simplement définie comme la croissance moyenne des occurrences normalisées des termes d'un champ donné entre la période précédente T^- et la période actuelle T : $A_C^T = \frac{1}{card(C)} \sum_{i\in C} \frac{p_i^T}{p_i^{T-}}$.

4.4.6 Reconstruction multi-échelle

Pour illustrer cette partie nous nous appuyons sur une base de données de publications décrivant les activités de recherche autour du thème du *développement durable* (voir description du jeu de données section 4.3.1). Nous avons appliqué nos méthodes de reconstruction statique sur ce jeu de données en nous concentrant sur les 10 dernières années du corpus, soit tous les articles parus publiés entre 1998 et 2007. La carte correspondante est présentée figure 4.11.

Le réseau $\hat{G}$ articulant les champs scientifiques les uns avec les autres a été construit grâce à la distance inter-cluster $\hat{\mathbf{S}}$ précédemment définie. Notre objectif est maintenant de soumettre $\hat{G}$ à une nouvelle opération de réduction en catégorisant les champs épistémiques au sein de classes de champs. Nous appliquons à nouveau la méthode de détection de communautés de cliques afin de construire les ensembles de *communautés de champs épistémiques* $\{\mathbf{C}_i\}_i$ que nous appellerons par la suite meta-communautés (nous utilisons la même notation que les champs épistémiques mais en gras), qui sont formés d'ensembles de champs : $\mathbf{C}_i \subset \mathcal{C}$

tel-00464316, version 1 - 16 Mar 2010

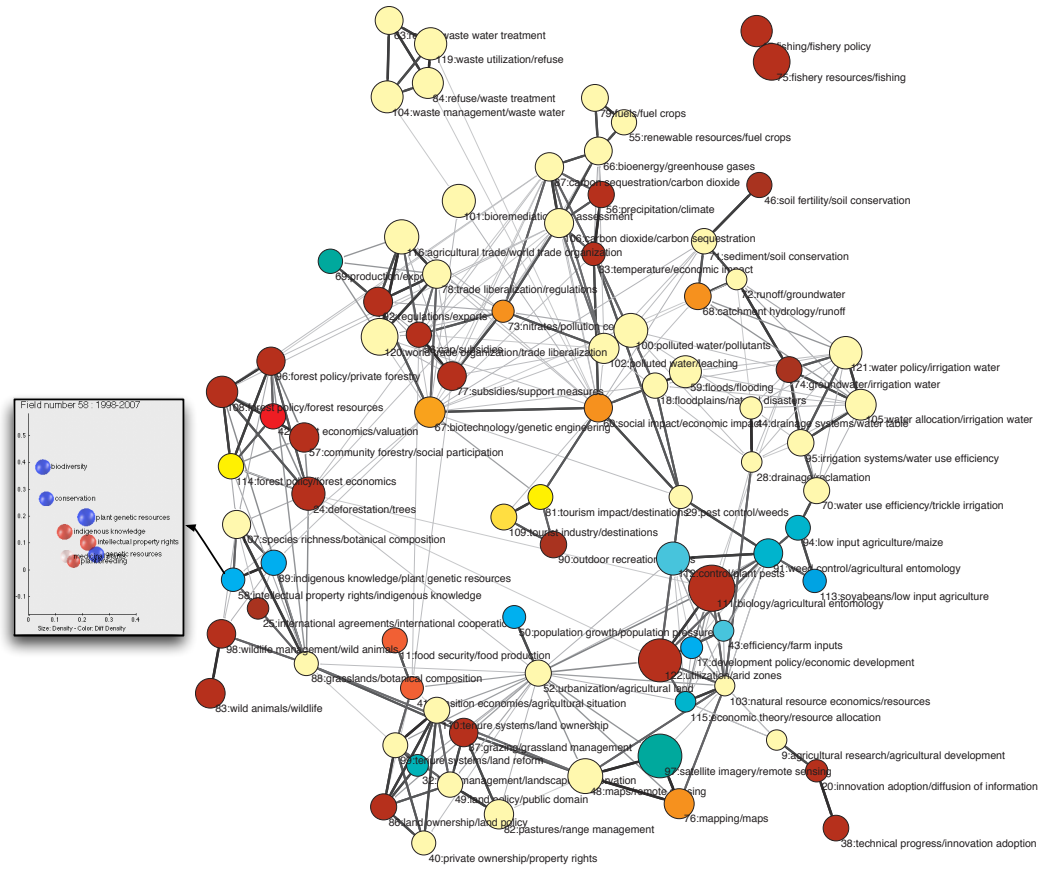


FIGURE 4.11: Carte du "domaine développement durable" sur la période 1998-2007. Le détail de la constitution du champ 58 (*intellectual property rights* & *indigeneous knowledge*) est représenté en encart. Cette fois-ci nous avons représenté la taille des champs en fonction de leur densité tandis que leur couleur traduit la différentiel de densité par rapport à la période précédente (ce taux de croissance augmente des couleurs froides (bleu) aux couleurs chaudes (rouge)). Les champs ont été labellisés par les deux termes les plus "médians".

(pour rappel $\mathcal{C} = \{C_j\}_j$ désigne l'ensemble des champs épistémiques détectés). Une représentation de l'ensemble des communautés obtenues par une percolation de k-cliques ($k \geq 5$) est représentée figure 4.12. Elle présente à nouveau un fort taux de recouvrement entre les communautés de champs détectées et n'est pas directement intelligible en l'état. Au niveau précédent, nous avions adopté une stratégie de labellisation des champs en fonction des indices de spécificité et de généricité des termes présents dans chaque champ. Dans le cas des agrégats de champs, une telle méthode n'est pas envisageable car elle consisterait à labelliser les communautés de champs avec des étiquettes déjà complexes et éventuellement variables (car dépendantes de la stratégie de labellisation retenue pour définir les champs épistémiques au premier niveau). Nous préférons employer une méthode plus classique qui permet de reconstruire les étiquettes de haut-niveau à partir de

tel-00464316, version 1 - 16 Mar 2010

la distribution des termes se retrouvant dans la même meta-communauté.

Chaque méta-communauté peut être décrite par l'ensemble des termes qui composent les champs épistémiques qu'elle recouvre. Certains termes pouvant être communs à plusieurs champs, on associe donc à chaque communauté de champs $\mathbf{C}_i$ un vecteur d'occurrences $W_{\mathbf{C}_i}$ qui dénombre pour chaque terme j le nombre de champs épistémiques dans $\mathbf{C}_i$ comprenant le terme j. Ainsi, pour un terme j, la coordonnée $W_{\mathbf{C}_i}(j)$ vaut : $|\{C_k \in \mathbf{C}_i, \text{ tel que } j \in C_k\}|$

Ces vecteurs "d'occurrences" de termes dans les communautés de champs permettent d'appliquer une procédure de type *tf.idf* afin d'extraire, pour chacune, les termes les plus significatifs à même de qualifier de façon pertinente leur contenu. On choisit ainsi d'étiqueter chaque communauté par les cinq termes dont les *tf.idf* sont les plus importants. [18]



FIGURE 4.12: Méta-communautés du champs "développement durable" sur la période 1998-2007. Les communautés de champs épistémiques sont représentées par des ensembles colorés. Les nœuds du réseau sont des champs épistémiques.

Une fois définies les étiquettes de chaque communauté de champs, on peut représenter le réseau de proximité entre les méta-communautés. Les liens entre méta-communautés sont construits comme une agrégation des liens entre champs au niveau inférieur. Nous appliquons la même procédure de passage au niveau supérieur que celle employée précédemment [19]. La figure 4.13 représente une

18. Pour rappel et dans notre contexte, le *tf.idf* d'un terme j pour une meta-communauté i vaut : $\text{tf.idf}_i(j) = \frac{W_{\mathbf{C}_i}(j)}{\sum_{k=1}^{n} W_{\mathbf{C}_i}(k)} log\left(\frac{|\{\mathbf{C}_k\}_k|}{|\{\mathbf{C}_k, \text{ tel que } W_{\mathbf{C}_i}(k) \geq 1\}_k|}\right)$

19. Plus précisément on définit la distance entre deux méta-communatés $\mathbf{C}_a$ et $\mathbf{C}_b$ sous la forme suivante : $\hat{\hat{\mathcal{S}}}(\mathbf{C}_a, \mathbf{C}_b) = \frac{1}{|\ \mathbf{C}_a\ |} \sum_{i \in \mathbf{C}_a} (\frac{1}{|\ \mathbf{C}_b\ |} \sum_{j \in \mathbf{C}_b} \hat{\mathcal{S}}(i,j))$

tel-00464316, version 1 - 16 Mar 2010

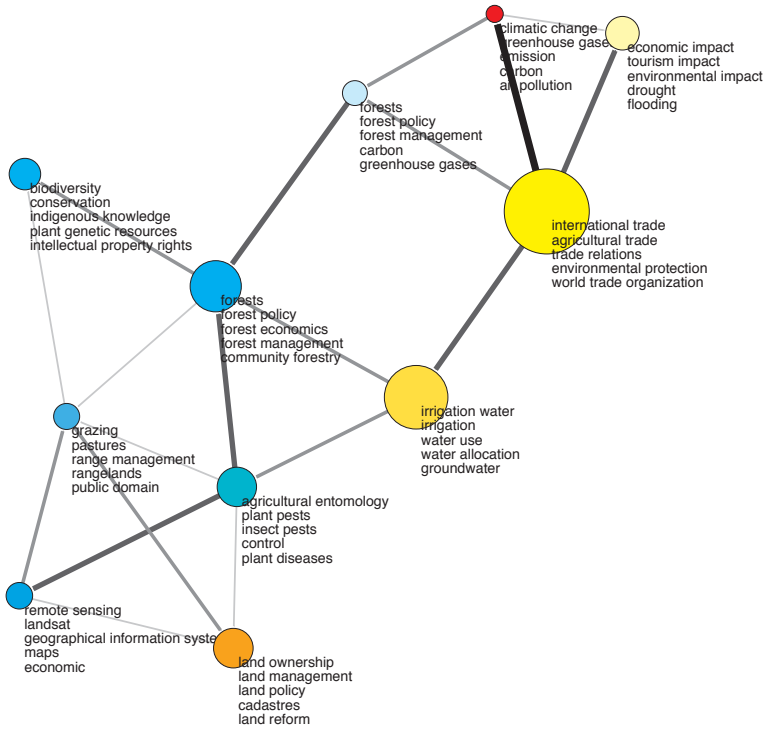


FIGURE 4.13: Méta-communautés du champs "développement durable" sur la période 1998-2007. Les communautés de champs épistémiques sont étiquetées par leurs cinq termes les plus cararctéristiques (plus haut tf.idf). La taille correspond à la somme des fréquence des mots-clés, la couleur, à la différence par rapport à la période précédente. Les nœuds du réseau sont des communautés de champs épistémiques.

carte de haut niveau du champ du développement durable tel qu'il se structure dans notre jeu de données. 11 communautés de champs distinctes ont été identifiées, leurs légendes forment des ensembles semblant conceptuellement pertinents. Ainsi de haut en bas et de droite à gauche, on reconnaît les sous-domaines suivants :

- économie de l'environnement et catastrophes naturelles,
- changement climatique (activité faible mais croissance très rapide),
- gestion forestière et développement durable,
- commerce international (très forte activité, croissance moyenne),
- biodiversité et propriété intellectuelle indigène,
- gestion forestière (aspects économiques),
- gestion de l'eau,
- gestion des zones agricoles,
- agronomie et nuisibles,
- aménagement du territoire, et occupation des sols,
- SIG.

tel-00464316, version 1 - 16 Mar 2010

4.4.7 Procédures de validation

Nous définissons l'indice de *qualité empirique* sur les clusters, noté Q_e, pour nous donner un critère de validation des champs reconstruits. La validation empirique est liée à l'adéquation des champs scientifiques reconstruits avec l'activité de publication à proprement parler. Pour représenter l'activité des communautés scientifiques, les champs scientifiques que nous mettons en exergue se doivent de rendre compte d'une littérature correspondant aux assemblées de termes repérées. Le principe de validation empirique que nous proposons est donc le suivant : pour chaque champ, nous contrôlons simplement, à l'aide d'une requête dans notre base de données de publications, qu'un nombre significatif d'articles réunit l'ensemble des mots-clés. Compte tenu de l'hétérogénéité des fréquences d'usage des concepts, nous proposons d'utiliser la notion de *self-information* (Shannon and Weaver, 2002) qui permet de mesurer la quantité d'information associée à un tirage d'une variable aléatoire X, dont les probabilités d'apparition valent p_i. À l'observation d'un événement X_i, on peut alors associer $I(X = X_i) = -log_2(p_i)$ qui est une mesure en bits de la quantité d'information. La quantité d'information correspond donc à une mesure du degré de "surprise" associé à un événement plus ou moins improbable.

Nous souhaitons évaluer la quantité d'information associée à la présence d'un nombre k d'articles mentionnant l'ensemble des termes d'un champ C. Nous pouvons mesurer à l'aide d'une requête dans notre base de données le nombre réel d'articles contenant l'ensemble des termes de C, on note n_C ce nombre. Pour calculer la quantité d'information associée à cet événement (n_C articles regroupent l'ensemble des termes de C), on fait l'hypothèse d'un modèle nul dans lequel les probabilités d'occurrence des termes sont indépendantes les unes des autres[20]. Ainsi la probabilité théorique qu'un des N articles emploie l'ensemble des termes vaut $\prod_{i \in C} p_i$. On en déduit que la probabilité théorique d'observer l'ensemble des termes de C conjointement mobilisés dans k articles vaut $\binom{N}{k}(\prod_{i \in C} p_i)^k(1 - \prod_{i \in C} p_i)^{N-k}$. L'indice de *qualité empirique* d'un cluster C est donc défini comme la quantité d'information, notée $Q_e(C)$, associée à l'événement "n_C articles regroupent l'ensemble des termes de C", soit

$$Q_e(C) = -log_2\Big[\binom{N}{n_C}(\prod_{i \in C} p_i)^{n_C}(1 - \prod_{i \in C} p_i)^{N-n_C}\Big]$$

.

20. les probabilités d'occurrences de nos termes ne sont en réalité naturellement pas indépendantes les unes des autres, nous nous servons néanmoins de la formule de la quantité d'information en traitant les fréquences d'usages des termes "comme si" elles correspondaient à des variables indépendantes dans le but de fournir une hypothèse nulle simplifiée par rapport à laquelle apprécier combien ces agencements de termes sont non aléatoires.

tel-00464316, version 1 - 16 Mar 2010

4.5 Méthode de reconstruction dynamique

Nous avons essayé de montrer dans la section précédente la manière dont la cartographie des sciences pouvait bénéficier d'une mesure asymétrique de proximité entre termes qui, associée à une méthode de catégorisation avec recouvrement, nous a permis de reconstruire une structure multi-échelle "hiérarchisée" des sciences robuste à la polysémie des termes et aux enchâssements complexes des communautés scientifiques. Ces méthodes de reconstruction ouvrent la voie à de nouveaux modes de navigation et d'interrogation des corpus de publications à travers des interfaces proposant de parcourir des paysages conceptuels multi-échelles qui pourraient s'avérer utiles pour les chercheurs, les théoriciens des sciences ou encore les gestionnaires.

Mais l'analyse statique de la structure des champs paradigmatiques n'est qu'une première étape vers la caractérisation et la représentation de l'activité scientifique. Nous poursuivons notre étude par la reconstruction des dynamiques des champs épistémiques. L'ensemble des méthodes décrites dans la section précédente permet de reconstruire la structure multi-échelle d'un domaine scientifique à n'importe quelle période. On peut ainsi aisément imaginer rajouter une dimension temporelle à notre analyse visant à la caractérisation des dynamiques scientifiques.

La reconstruction des dynamiques des communautés scientifiques présente des enjeux théoriques forts. Appréhender l'évolution des sciences à partir de données réelles présente un intérêt particulier en épistémologie, ou en histoire des sciences. L'objectif est double : d'une part, caractériser finement les évolutions des champs épistémiques, d'autre part, à une échelle de temps plus grande, observer les mutations opérées dans la dynamique d'évolution des sciences afin de caractériser des transitions des régime de régulation des communautés scientifiques elles-mêmes. Est-il possible de repérer une tendance à la balkanisation des sciences associée à une spécialisation croissante des communautés ? Peut-on mesurer l'impact des nouveaux outils de communication introduits par Internet sur l'organisation des communautés ? Est-il possible de reconstruire l'évolution des changements paradigmatiques majeurs ? peut-on identifier de façon automatique les champs émergents ? Peut-on également retracer les grandes mutations dans les régimes de régulation et d'organisation des communautés scientifiques ? Sans avoir prétention à répondre à toutes ces questions, des outils de suivi des dynamiques scientifiques observées in-vivo pourraient nous informer sur les conséquences de ces mutations.

La question de la dynamique des sciences a largement animé le champ de la scientométrie. On peut se référer, en ce qui concerne l'analyse par réseau de co-citation, aux historiographes de (Garfield, 2004) ou à l'analyse de la continuité des "bases intellectuelles" propres à une spécialité (Braam et al., 1991). L'analyse de mots associés a également été employée pour cartographier le développement

tel-00464316, version 1 - 16 Mar 2010

d'un champ sur une longue période Cambrosio et al. (1993) ou pour étudier les relations au cours du temps (influence, circulation) qu'entretiennent recherche académiques et recherche appliquée autour d'un même domaine (Callon et al., 1991). Des approches mixtes ont proposé de représenter les tendances émergentes et les motifs transitoires dans la littérature scientifique en s'appuyant aussi bien sur les "fronts de recherche" (analyse lexicale) que sur les "bases intellectuelles" des spécialités (Chen, 2004, 2006). Enfin, on peut mentionner les réseaux de citation animés entre journaux de Leydesdorff and Schank (2008b) qui illustrent les dynamiques émergentes de domaines inter-disciplinaires.

Naturellement, les processus d'évolution des sciences ont également été interrogés par d'autres disciplines (Börner and Scharnhost, 2009) telles que la philosophie des sciences qui a fourni nombre de descriptions et d'explications plus ou moins compatibles les unes avec les autres sur les dynamiques de changements et de révisions scientifiques (Kuhn, 1970a; Mulkay, 1976), l'ethnographie à travers des études réalisées *in-situ* dans l'espace du laboratoire (Knorr-Cetina, 1995; Latour and Woolgar, 1988), les sciences de gestion (Bonaccorsi, 2008) ou/et par la sociologie des sciences qui a vu dans l'analyse des "polémiques " une méthodologie privilégiée pour comprendre les dynamiques socio-techniques complexes qui se déploient dans les communautés scientifiques exposées à un changement (Pestre, 2007).

Nous découpons notre analyse des dynamiques scientifiques selon les différents niveaux, microscopique, mésoscopique et macroscopique, auxquels nous appréhendons nos données.

4.5.1 Dynamiques de voisinage

Au niveau local, on peut s'interroger sur l'évolution des voisinages associés à un terme. Etant donné un seuil s fixé et un terme i, on peut représenter l'ensemble des termes qui appartiennent au voisinage de i à différentes périodes. Cette représentation offre un premier mode d'observation du glissement de sens d'un terme au cours du temps comme illustré figure 4.14. On observe sur cet exemple que les études sur les biens publics ont été appréhendées récemment à travers des approches de type théorie des jeux. Parmi les termes émergeant dans le voisinage de "Public Good", on trouve notamment "heterogeneous agents" ou "procedural rationality". Cette dynamique correspond bien aux transformations actuelles du domaine.

4.5.2 Dynamique d'un champ épistémique

Nous interrogeons maintenant la dynamique d'un champ épistémique en le considérant en isolation complète du reste du réseau. Etant donné un ensemble de termes participant à un champ épistémique à une période donnée, nous souhai-

tel-00464316, version 1 - 16 Mar 2010

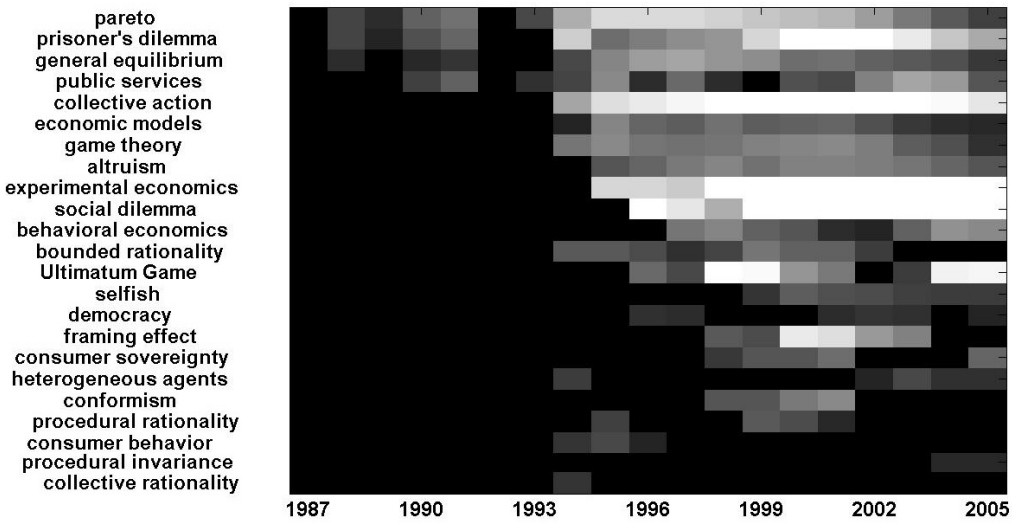


FIGURE 4.14: Représentation dynamique de l'évolution du voisinage du terme "Public Good" de 1987 à 2005 pour un focus $\alpha = 1$ (base de données : *systèmes complexes*). Une zone noire signifie que le terme associé n'est pas dans le voisinage de "Public Good" durant l'année considérée. Les cases les plus claires correspondent par contre aux voisins les plus proches.

tons retracer les conditions d'émergence de ce champ en visualisant l'évolution de la structure de ses termes dans la représentation bidimensionnelle que nous avons introduite précédemment (section 4.4.3) [21]. Un exemple d'une telle représentation est donné figure 4.15. Il représente un de nos champs épistémiques détectés dans notre base de données dédiée à la métaphore réseau en biologie sur la période 2003-2007. Ce champ est lié à la morphogenèse et au rôle des réseaux de régulation au cours de l'embryogenèse.

Cette évolution, qui se représente plus naturellement comme un film, permet d'apprécier la structuration progressive du champ avec l'aide des mesures *d'importance* et de croissance de l'importance [22] associées à chaque terme dans le cluster couplées à la "forme" prise par les termes de notre champ dans notre référentiel. En effet, un champ bien structuré se caractérise par une certaine linéarité des termes dans cet espace, un terme bien ancré au sein de son champ étant vis-à-vis de ses voisins soit bien contextualisé par certains soit un bon contexte pour les autres. La combinaison de ces deux voisinages garantit que la somme de ses indices de spécificité et de généricité est élevée. De plus cette somme est relativement stable en fonction des termes. Empiriquement, on a ainsi observé que les clusters détectés prenaient une forme caractéristique : celle de termes alignés le

21. Il est également possible de s'intéresser au devenir d'un champ si l'on souhaite tenter d'apprécier ses transformations ultérieures.

22. L'importance mesure la somme des cooccurrences qu'un terme a avec les autres termes du champ, sa croissance correspond simplement au taux d'accroissement de cette quantité.

tel-00464316, version 1 - 16 Mar 2010

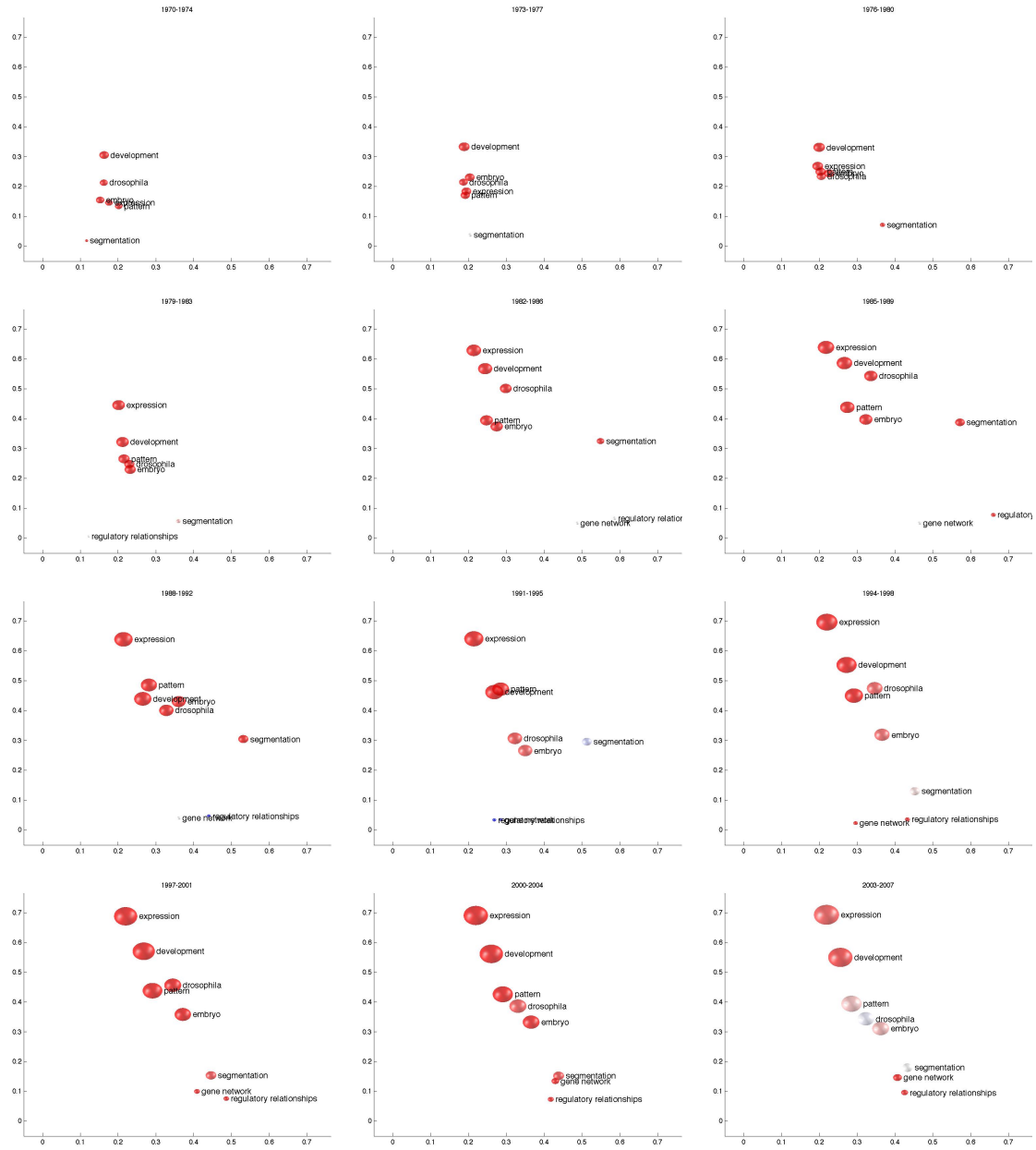


FIGURE 4.15: Evolution de la strucutre d'un champ épistémique détecté durant la période 2003-2007. Les années croient de de gauche à droite et de haut en bas (depuis la période 1970-1974) par fenêres glissante se décalant de 3 ans à chaque représentation. L'indice de spécificité I_s est en abscisse, l'indice de généricité en ordonnées I_g. La taille d'un terme correspond à son importance dans le champ, la couleur représente son taux de croissance

long d'une diagonale qui rend constant la somme des indices de spécificité et de généricité ; c'est d'ailleurs cette propriété de bon alignement que tente de traduire notre indice de pseudo-inclusion. Nous observons donc figure 4.15 que l'ensemble

tel-00464316, version 1 - 16 Mar 2010

des termes s'aligne progressivement alors que l'importance de chaque terme au sein du champ croît dans les périodes qui précèdent la détection effective du cluster par nos algorithmes.

Cette représentation dynamique permet également d'observer la dynamique de termes au sein de cet espace. Par exemple, le terme "segmentation" paraît très distant de l'ensemble des autres termes (I_s et I_g faibles) durant les premières périodes, il a de plus une très petite taille et donc une importance mineure dans la structure du champ. On peut observer son déplacement progressif et la croissance de son importance dans le cluster au fil des ans. Nous en restons au stade d'une simple observation, mais il semble possible de développer une cinématique des termes au sein de ces espaces, qui permette éventuellement d'extraire un certain nombre de régularités voire de prédiction vis-à-vis de l'évolution d'un terme. On observe la même dynamique "d'alignement" avec le reste du cluster pour les termes "gene network" et "regulatory relationships". Le terme "pattern" quant à lui, bien que relativement bien situé au sein du cluster dès les premières périodes, voit son importance et sa généricité augmenter durant les années 60 et 70, ce qui traduit un usage accru du terme en biologie du développement durant cette période.

4.5.3 Vers les dynamiques macroscopiques

Une première approche "naïve" du suivi des dynamiques scientifiques consisterait à mettre bout à bout l'ensemble des cartes macroscopiques obtenues à des périodes successives. Une analyse de ce type (schématisée figure 4.16) n'est informative que du point de vue de la structuration globale du domaine. Observe-t-on une augmentation ou une diminution du nombre total de champs ? Est-ce que la cohésion de l'ensemble des champs a tendance à augmenter ou à diminuer ? La comparaison des cartes obtenues à différentes périodes nécessite donc un travail d'interprétation de la part de "l'utilisateur". Dans le cas présent, ces cartes ont été annotées et interprétées par deux sociologues et historiens des sciences (Jean-Paul Gaudillière et Christophe Bonneuil), sous la forme d'ensembles de champs épistémiques regroupés au sein de domaines plus larges, se répétant ou non d'une période à l'autre. La délimitation et la labellisation de ces domaines est le résultat d'une analyse plus fine et "manuelle" des champs épistémiques qui les constituent.

L'analyse des cartes réalisées sur 4 périodes courant de 1976 à 2007 à ainsi permis de révéler les traits saillants suivants dans l'évolution des travaux autour de la notion de réseau et de complexité en biologie. On voit notamment émerger de façon très nette à compter de la période 1999-2003 la confirmation d'un discours réseau, centré sur les aspects instrumentaux de la génomique qui se traduit par l'apparition de clusters autour des termes *micro-array & expression*, ou *protein-protein interactions & data bank*. Ces clusters remplacent les champs caractéristiques de l'ère

tel-00464316, version 1 - 16 Mar 2010

pré-génomique dont les rapports aux réseaux sont plus incidents. Avant cette période, on constate également une forme de continuité du vocabulaire réseau sans réelle intégration des approches contemporaines de l'analyse des grands réseaux d'interaction dans les clusters liés à l'immunologie, ou aux réseaux neuronaux ; et des intégrations mais avec changement de sens dans les clusters liés à *expression regulation*, *feedback* ou *kinase & phosphorylation*.

La percée des approches réseau en biologie semble donc reposer fondamentalement sur des bases matérielles (expérimentale et bio-informatique) issues notamment de la biologie à haut débit. Ces cartes révèlent également que les aspects plus théoriques liés aux réseaux en terme d'architecture du vivant ne commencent à se constituer en un champ de recherche à part entière que plus tardivement, essentiellement après 2002 (vocabulaire autour de l'architecture, la modélisation, les propriétés de connectivité ou de robustesse des réseaux) - à la suite de la multiplication des études sur les grands réseaux d'interaction en physique (popularisation des notions de "scale-free", "small-world", etc. à partir de 1999-2000). Le discours réseau, se généralisant et s'affirmant autour de 2000, on assiste maintenant à sa diffusion vers d'autres champs de la biologie auxquels il n'était pas traditionnellement lié. Par exemple, un *effet de dissémination* est visible lorsqu'on examine l'amas de clusters portant sur le cancer apparaissant sur la période 2004-2007 qui est fortement lié aux approches réseaux d'interaction et données d'expression génétique à haut débit.

Cette méthode d'analyse requière néanmoins un investissement coûteux puisqu'il est nécessaire d'explorer en profondeur le contenu détaillé des champs détectés à chaque période. Les nœuds figurant les champs dans chaque carte sont *a priori* différents entre deux périodes successives, et il semble délicat de juger de la transformation de la structure globale d'un réseau dont les nœuds sont eux-mêmes modifiés à chaque pas de temps. Notre objectif est donc d'accompagner le travail d'interprétation des cartes en fournissant une procédure de détermination des dynamiques scientifiques au niveau mésoscopique, c'est à dire directement liées aux mutations subies par les champs épistémiques. L'observation des changements dans les associations de mots-clés qui modifient la nature de nos champs épistémiques constitue sans doute un niveau d'analyse pertinent pour observer les dynamiques de l'activité scientifique. L'évolution de ces ensembles de termes permettent en effet de retracer les processus, parfois discontinus (Kuhn, 1970a; Mulkay, 1976), de fertilisation croisée entre champs, la circulation de concepts à travers les domaines, ou encore les augmentations et diminutions d'activité propres à un champ. Nous appelons *méso-dynamiques* les transformations qui affectent les champs épistémiques. La reconstruction de ces méso-dynamiques revient en fait à spécifier la fonction déterminant les "relations de parenté intellectuelle" entre les champs épistémiques obtenus entre deux périodes successives.

Callon (1994) oppose deux dynamiques extrêmes des dynamiques des collectifs "locaux" qui animent l'évolution des sciences. Il distingue d'un côté des dy-

tel-00464316, version 1 - 16 Mar 2010

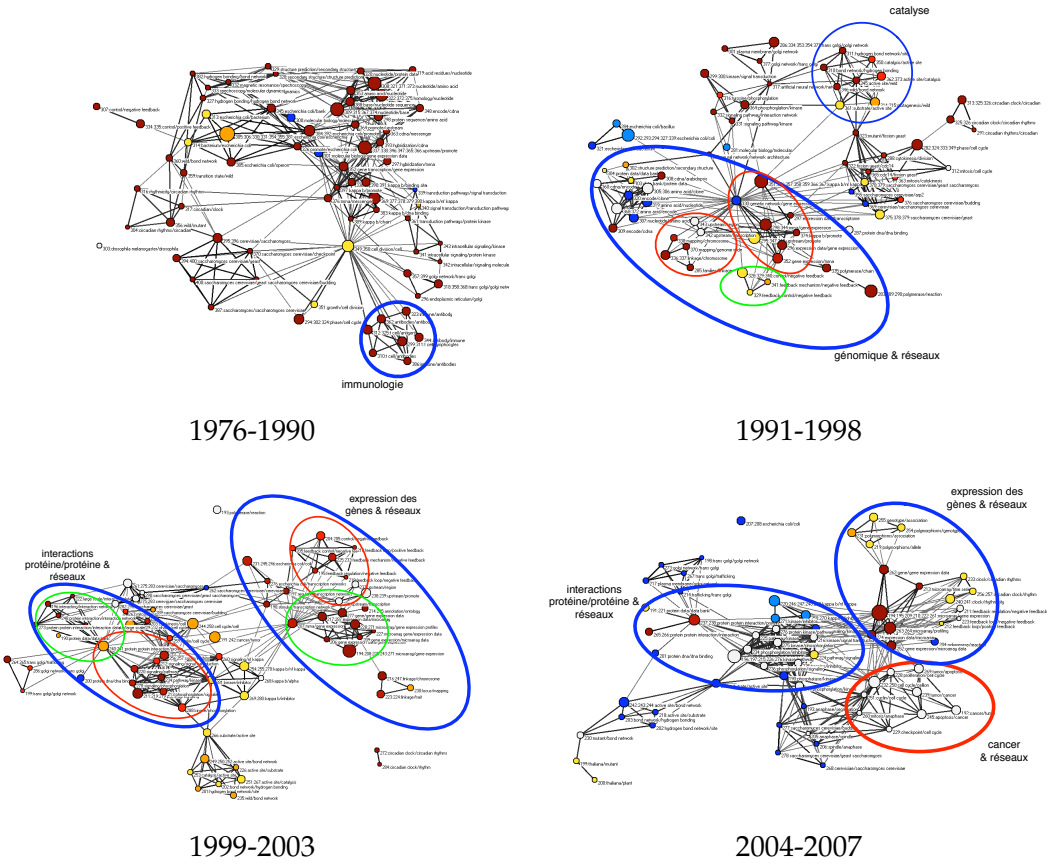


FIGURE 4.16: Evolution de la structuration des champs de notre jeu de données *biologie & réseaux* sur quatre périodes : 1976-1990 ; 1991-1998 ; 1999-2002 ; 2003-2007. (la couleur des champs correspond à leur croissance d'activité, leur taille à leur activité (échelle logarithmique))

namiques *conservatives* qui "consolident" un état des choses" formées par "des collectifs locaux qui opèrent des reconfigurations mineures sans bouleverser les connections existantes" [23], à des dynamiques de "prolifération de nouveaux états du monde", qui se caractérisent par l'émergence de "collectifs locaux à même de proposer des reconfigurations très originales et innovantes de réseaux non connectés jusqu'à lors" [24].

Dans notre perspective, nous n'envisageons les dynamiques des champs que comme la transformation des ensembles de "concepts", censés refléter les "connaissances de bases" (Morris and der Veer Martens, 2008) partagées par les

23. "The local collective does some minor reconfiguration work, but this will not shake up existing connections." (Callon, 1994)

24. "the local collective is in a position to propose some very original, innovative reconfigurations linking together networks that had been separate. This leads to the proliferation of new states of the world. " *ibid*, p. 415

tel-00464316, version 1 - 16 Mar 2010

participants d'un champ épistémique. Nous n'intégrons pas dans notre analyse la même richesse descriptive que celle que (Callon, 1994) accorde aux collectifs locaux ; nous pouvons néanmoins nous en inspirer pour tâcher de retracer les conséquences de ces transformations en termes de dynamiques de nos champs.

La composition d'un champ scientifique peut en effet être soumise à un certain nombre de transformations qui en modifieront les frontières et la forme. Le répertoire de ces dynamiques comprend les événements potentiels suivants. Un champ peut *croître* en acquérant de nouveau concepts, ou bien *décroître* s'il en perd. Les deux versions extrêmes de ces cas de figure correspondent à une *naissance ex-nihilo* ou à la *disparition* du champ. Un champ peut également *fusionner* avec d'autre champs pour former un un nouveau champ épistémique ou encore se *diviser* en plusieurs champs.

4.5.4 Reconstruction de la phylogénie des sciences

En nous adonnant à la métaphore biologique, nous pouvons faire un parallèle entre la dynamique des champs épistémiques et l'évolution des espèces. En biologie, un arbre phylogénétique représente l'histoire évolutive des espèces et des organismes. Une large gamme de méthodes est utilisée pour reconstruire les relations phylogénétiques entre entités (Nei, 1996), ces méthodes s'appuient, généralement, sur la comparaison de séquences génétiques. Une méthode classique consiste à faire appel à des algorithmes de clustering basés sur des mesures de distances génétiques entre espèces (*Fitch-Margoliash* par exemple). D'autres méthodes comme le *maximum de vraisemblance* ou *l'inférence bayésienne* s'appuient sur des hypothèses évolutives spécifiques (principe de parcimonie par exemple) afin d'inférer l'arbre le plus vraisemblable (Huelsenbeck and Rannala, 1997).

Ces méthodes de reconstruction d'arbres ont été critiquées comme limitées voire inadaptées lorsqu'on souhaite intégrer la possibilité de dynamiques plus réticulées comme des recombinaisons génétiques, ou des transferts de gènes horizontaux. Confronté à ces épisodes d'hybridation, il est nécessaire de passer d'une modélisation des liens de parenté par des *arbres phylogénétiques* à une modélisation par des *réseaux phylogénétiques*.

Contrairement aux biologistes qui peuvent avoir accès à un certain nombre de connaissances sur les mécanismes microscopiques qui sous-tendent le processus évolutif (taux de mutation de certains gènes, mesures de pression de sélection, etc.) nous devons adopter une perspective agnostique et ignorer les mécanismes susceptibles de guider l'évolution des sciences [25]. Nous pouvons néanmoins supposer que l'évolution des sciences est animée "d'événements d'hybridation". Les lignages conceptuels ne progressent sans doute pas linéairement, la fertilisation

25. Même si certains auteurs ont soutenu que les changements conceptuels en science pouvaient être guidés par des principes similaires à ceux qui s'exercent sur les systèmes biologiques (Hull, 2001).

tel-00464316, version 1 - 16 Mar 2010

croisées de champs parfois distants est fréquente, et il serait naïf de supposer que l'évolution des sciences puisse être représentée comme un arbre dont les branches s'étendraient infiniment sans jamais se croiser.

En l'absence d'un principe de parcimonie, qui permettrait par exemple de fournir un critère global à minimiser (tel que le nombre de mutations nécessaires) pour reconstruire les relations de parenté, nous adopterons une approche purement locale. La reconstruction du réseau phylogénétique des sciences revient alors à répondre à la question suivante : étant donné à un moment t un champ épistémique C^t, de quels champs à la période précédente, C^t hérite-t-il conceptuellement ?

4.5.4.1 Méthode de reconstruction des lignages entre champs épistémiques

Pour réussir cet appariement inter-temporel entre champs, nous devons trouver pour chaque champ C^t à t, le ou les champs dont il hérite. Nous faisons l'hypothèse que l'échelle de temps de transformation des champs est suffisamment lente pour nous permettre d'en suivre la dynamique à l'aide d'une simple mesure de similarité entre champs calculée entre deux périodes successives (l'échelle de temps étant l'année). Nous nous appuyons donc sur un principe de continuité de la composition des champs qui n'est pas sans rappeler celui proposé par Simmel (1898) pour suivre la "persistence" des groupes sociaux. Nous cherchons ainsi à identifier le champ ou la combinaison de champs à $t-1$ (puisque nous souhaitons intégrer la possibilité d'événements de fusion) qui sont les plus semblables et donc le plus probablement parents de C^t. Plus formellement, étant donné $C_i^t \in \mathcal{C}^t$, nous créons un lien de parenté dans le réseau phylogénétique entre C_i^t et le sous-ensemble des champs de la période précédente $\Phi^t(C_i^t) \in \mathcal{P}(\mathcal{C}^{t-1})$ (pour rappel $\mathcal{C}^{t-1} = \{C_k^{t-1}\}_{k \in K^{t-1}}$ où K^{t-1} désigne l'index de l'ensemble des champs à $t-1$) vérifiant :

$$\Phi^t(C_i^t) = \begin{cases} \{C_j^{t-1}\}_{j \in \kappa_i^t}, \kappa_i^t = \underset{\kappa \subset K^{t-1}}{\arg\min}\, d\big(C_i^t, \bigcup_{k \in \kappa} C_k^{t-1}\big) & \text{si } d\big(C_i^t, \bigcup_{k \in \kappa_i^t} C_k^{t-1}\big) < d_0; \\ \emptyset & sinon. \end{cases}$$

Il semblerait abusif de faire se correspondre deux ensembles de champs trop distants l'un de l'autre, même en l'absence d'un meilleur appariement. Nous définissons donc un seuil d_0 de distance entre clusters inter-temporels au-dessous duquel nous considérerons l'appariement comme satisfaisant. Comme nous le verrons il existe un large intervalle de valeurs pour lesquelles ce seuil ne modifie pas de façon critique les motifs apparaissant au sein du lignage.

Une façon simple de choisir une distance d est d'opter pour la distance de Jaccard qui mesure le ratio : nombre de termes non-recouvrants entre deux champs divisé par le cardinal de l'union des champs. Cette mesure est également l'inverse

tel-00464316, version 1 - 16 Mar 2010

de "l'indice de transformation" introduit par Callon et al. (1991) dans une perspective identique. Palla et al. (2007a) ont proposé une méthode de reconstruction de l'évolution des groupes sociaux (basée sur l'analyse d'un réseau de contact téléphonique) en s'appuyant sur une même mesure de distance entre clusters intertemporels ; néanmoins leur méthode est distincte de la nôtre au sens où les lignages qu'elle construit sont limités à des chaînes purement linéaires (*i.e.* le degré entrant et sortant d'un cluster à un moment donné vaut au plus 1). Notre méthode en autorisant une plus grande variété de dynamiques (et notamment un degré sortant et entrant dans le réseau phylogénétique non limité) permet la construction de structures plus riches. D'autres auteurs ont proposé des indices de stabilité des domaines de spécialité, obtenus avec des analyse de type co-citation, en s'appuyant sur une mesure de similarité entre clusters de type cosinus (Braam et al., 1991), ou sur d'autres types de mesure similaires (Hopcroft et al., 2004a). Dans le cadre de la reconstruction des filiations entre champs épistémiques, nous considérons que la stabilité des ensembles de termes régulièrement appariés constitue un bon critère de continuité entre champs. Avec une telle mesure, on obtient alors la définition suivante (en posant $\delta_0 = 1 - d_0$) :

$$\Phi^t(C_i^t) = \begin{cases} \{C_j^{t-1}\}_{j \in \kappa_i^t}, \kappa_i^t = \underset{\kappa \subset K^{t-1}}{\operatorname{argmax}} \dfrac{|C_i^t \bigcap (\bigcup\limits_{k \in \kappa} C_k^{t-1})|}{|C_i^t \bigcup (\bigcup\limits_{k \in \kappa} C_k^{t-1})|} & \text{si } \dfrac{|C_i^t \bigcap (\bigcup\limits_{k \in \kappa_i^t} C_k^{t-1})|}{|C_i^t \bigcup (\bigcup\limits_{k \in \kappa_i^t} C_k^{t-1})|} > \delta_0; \\ \emptyset & sinon. \end{cases}$$

La procédure d'identification est illustrée figure 4.17. Etant donnés deux champs A et B à une période t, on détecte à un moment ultérieur $t+1$, deux autres champs C et D (qui partagent un terme en commun). La question est donc de savoir de quels champs descendent le plus vraisemblablement les champs C et D.

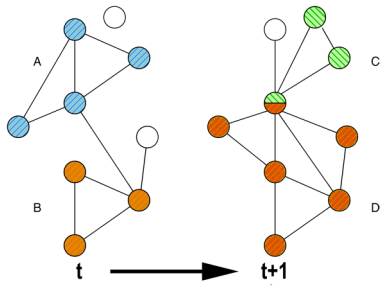


FIGURE 4.17: Exemple d'évolution de deux champs entre deux pas de temps successifs basé sur (Palla et al., 2007a). On distingue à la période t, les champs A (en bleu) et B (en orange) puis à la période suivante : les champs C (en vert) et D (en rouge)

Compte tenu de notre contrainte de continuité, on voit immédiatement que la

tel-00464316, version 1 - 16 Mar 2010

communauté C descend directement de la communauté A, même si deux noeuds ont disparu, tandis qu'un autre a été ajouté, la distance entre les champs A et C vaut : $d(A,C) = 1 - \frac{2}{5}$ et constitue la meilleure correspondance possible. En ce qui concerne D, l'appariement optimal est un peu moins évident. Si nous calculons pour chacun des cas possibles (le champ D peut descendre de A, B ou de $A \cup B$) les ratios correspondants : $d(A,D) = 1 - \frac{2}{8}$, $d(B,D) = 1 - \frac{3}{6}$ et enfin $d(A \cup B, D) = 1 - \frac{5}{7}$, on constate que la meilleure correspondance possible pour le champ D est donc offerte par $A \cup B$. Si $d_0 > \frac{5}{7}$ (ou de façon équivalente si $\delta_0 < 1 - \frac{5}{7}$), on considérera alors que D est un produit de la la fusion des champs A et B.

4.5.4.2 Exemples de phylogénies

Pour reconstruire la phylogénie d'un domaine, nous calculons pour l'ensemble des champs épistémiques à une période donnée, l'ensemble de leurs antécédents grâce à la formule décrite ci-dessus. L'ensemble des relations de filiation peut être réuni au sein d'un graphe dirigé acyclique formant ce que nous appelons la phylogénie du domaine. Ce graphe peut alors être représenté à l'aide d'un logiciel de représentation de réseau (nous avons choisi Graphviz [26], logiciel spécialisé dans la représentation d'arbre, qui minimise le nombre de croisements entre branches). La figure 4.18 fournit un exemple de phylogénie calculée sur notre base de données *biologie & réseau* retraçant la dynamique des champs entre 1994 et 2007. Il nous a été impossible de représenter l'ensemble de la phylogénie tout en rendant son contenu lisible, on peut néanmoins observer, sur cette représentation, la diversité des motifs de parenté existants. Certaines branches paraissent très linéaires, d'autres donnent lieu à un certain foisonnement avec de nombreux événements d'hybridation (un champ épistémique descendant de plusieurs communautés) et de diversification (une branche est à l'origine de plusieurs sous-branches).

Afin d'avoir une représentation plus resserrée, nous avons tracé une autre partie de la phylogénie sur une période de temps plus réduite (2001-2007) et en choisissant une fenêtre temporelle limitée à une année. On a ensuite sélectionné uniquement le sous-réseau constitué par les champs mentionnant les termes "cancer" ou "tumor". Le résultat de cette extraction est présenté figure 4.19 Sur ce détail de la phylogénie, on a annoté les "branches" en fonction de trois grandes familles de contextes dans lesquelles les problématiques liées au cancer sont étudiées.

On observe clairement 3 ensembles distincts dans cette phylogénie. Deux ensembles donnent lieu à des dynamiques complexes entre champs et traitent, d'une part, des relations entre le *cancer* et l' *ADN*, d'autre part, des problématiques liées aux termes *cancer, tumor* et *proliferation*. Ces sous-domaines semblent avoir multiplié leur interactions ces dernières années autour des concepts : *apoptosis*, *suppressor* et *cell cycle*. Le troisième ensemble, qui se distingue par un lignage conceptuel très linéaire, se rapporte aux relations entre *tumor* et *immune system*. Ces deux catégo-

26. http ://www.graphviz.org/

tel-00464316, version 1 - 16 Mar 2010

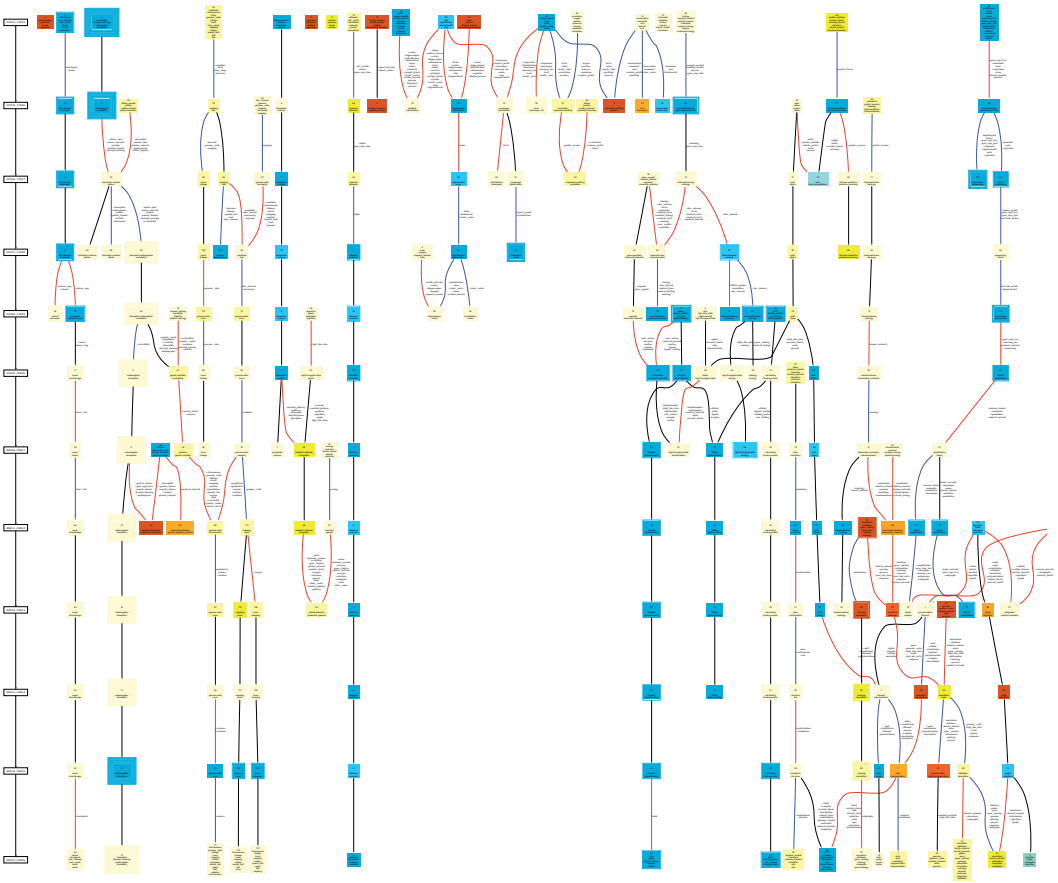


FIGURE 4.18: Extrait de la phylogénie des champs épistémiques liés aux approches *biologie & réseau* de 1994 à 2007. Nous avons uniquement sélectionné les champs constitués de quatre termes ou plus. Chaque ligne de la phylogénie correspond aux champs détectées sur une fenêtre de deux ans. On fait ensuite glisser cette fenêtre d'année en année (de ligne en ligne) pour remonter la phylogénie vers le passé (de bas en haut). Cette phylogénie est tronquée, pour des raisons de présentation. Les liens de parenté entre champs sont colorés soit en rouge en cas de croissance (gain net de concepts) soit en bleu en cas de décroissance (perte nette de concepts) en noir sinon. La taille (nombre de rectangles concentriques) des champs représente leur densité (D), leur couleur leur croissance d'activité.

ries de dynamiques ont également des indices de structuration (indice de pseudo-inclusion et densité, non représentés ici) très différents et qui semblent corrélées aux configurations locales prises par le lignage conceptuel de la phylogénie.

4.5.4.3 Motifs phylogénétiques

Afin d'étudier les couplages éventuels entre les indices de structuration des champs et les motifs de la phylogénie, nous proposons dans une première approche exploratoire, d'établir certaines statistiques corrélant la densité et l'indice de pseudo-inclusion aux "formes" prises par la phylogénie.

tel-00464316, version 1 - 16 Mar 2010

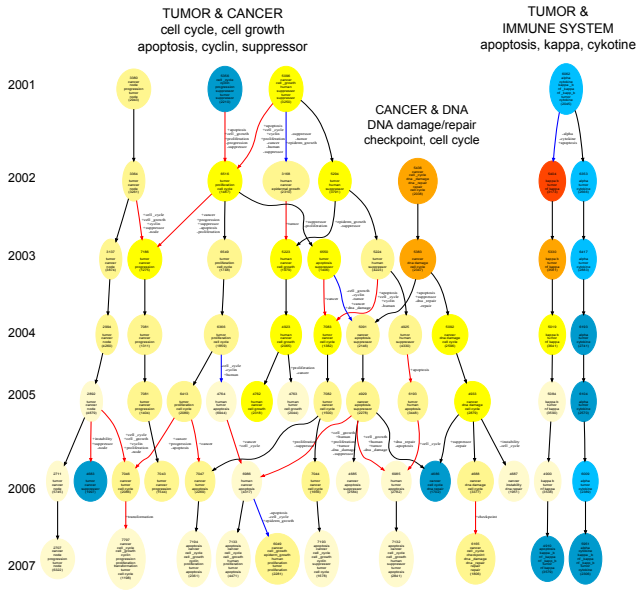


FIGURE 4.19: Détail du sous-réseau phylogénétique lié aux études sur le cancer. Les couleurs des cercles figurent la croissance de l'indice de pseudo inclusion $I_{\subset}$ (de faible à forte du bleu au rouge). Les champs sont étiquetés avec leurs termes les plus génériques (à part les débuts et les fins de branches dont le contenu intégral des champs est rappelé). Les flêches reliant les champs sont annotés des termes gagnés ou perdus par un champ entre deux deux périodes successives. Dans chaque cluster, on fait également figurer entre parenthèses le nombre d'articles mentionnant tous les termes du cluster.

Nous avons calculé une phylogénie [27] sur la même base de données (*réseaux & biologie*) entre 1990 et 2007 en choisissant une fenêtre temporelle d'une année. La qualité empirique (cf section 4.4.7) de l'ensemble des champs a été calculée, et nous n'avons sélectionné que les champs dont la qualité empirique était satisfaisante. Le réseau phylogénétique ainsi obtenu est composé de $7,758$ nœuds distribués sur les 18 années. Cette représentation permet de retracer finement l'histoire des champs en identifiant les influences croisées entre sous-domaines, et les périodes charnières d'émergence ou de disparition de champs par exemple. Une portion de cette phylogénie est représentée figure 4.20.

Une première façon d'apprécier les motifs phylogénétiques consiste simplement à faire le décompte pour chaque champ du nombre de "fils" (ou nombre d'enfants) qui en descendent. Le nombre de fils correspond dans notre réseau phylogénétique au degré sortant d'un champ. Alors que la plupart des champs

27. Sur cette carte, nous avons directement défini les champs épistémiques comme des k-cliques de G sans utiliser l'opération de percolation de cliques, cette méthode simplifiée garantit toujours la possibilité de clusters recouvrants. Elle consiste en réalité à reconstruire les noyaux ou les cœurs des champs épistémiques. C'est la raison pour laquelle nous avons obtenu un très grand nombre de champs.

tel-00464316, version 1 - 16 Mar 2010

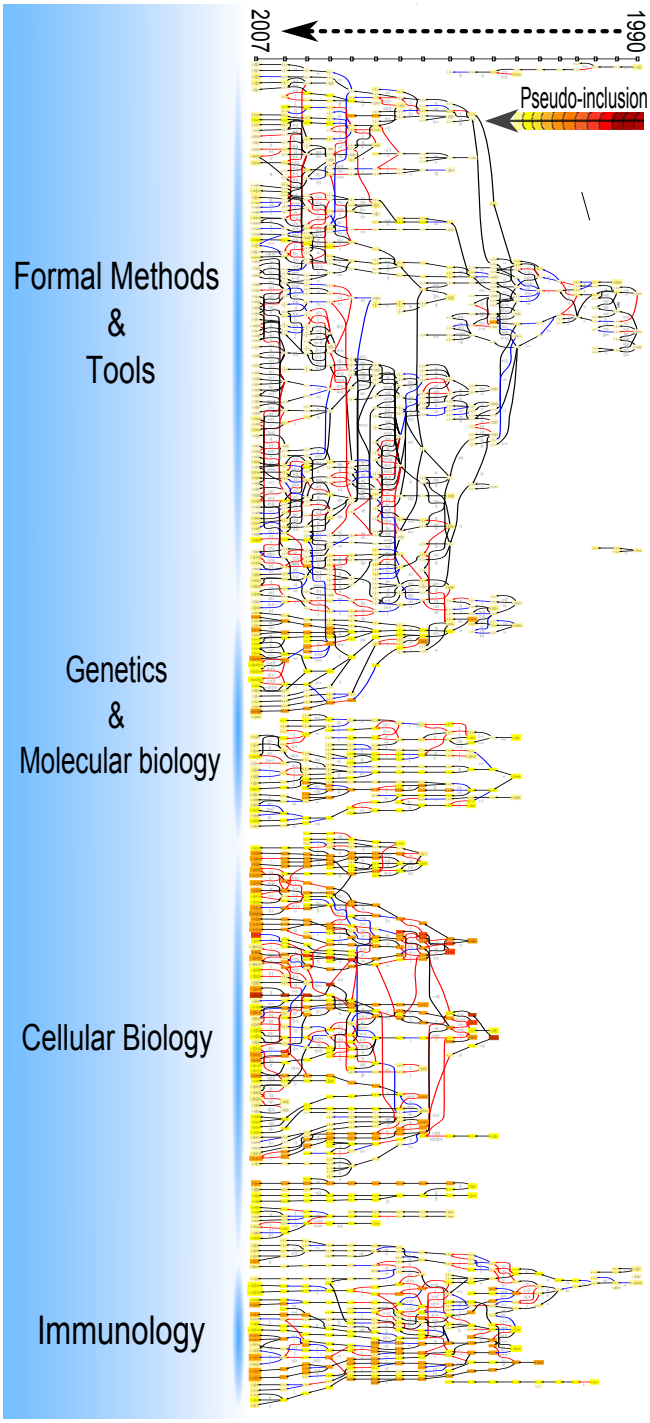


FIGURE 4.20: Extrait de la phylogénie des domaines liés à l'approche réseau en biologie (cette portion regroupe environ 1400 clusters) de 1990 à 2007. Nous avons uniquement sélectionné les champs constitués de quatres termes ou plus, et dont la qualité empirique était au-dessus d'un seuil fixé. Cette phylogénie est tronquée, pour des raisons de présentation. Les couleurs des champs correspondent à leur indice de pseudo-inclusion. Les liens de filiation entre champs sont colorés en rouge en cas de croissance et en bleu en cas de décroissance. Les grands domaines annotés dans la bande bleue ont été ajoutés manuellement.

tel-00464316, version 1 - 16 Mar 2010

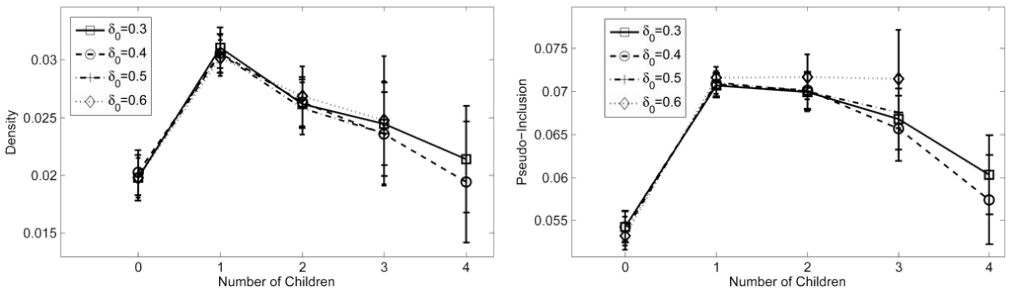


FIGURE 4.21: Densité moyenne (à gauche) et indice de pseudo-inclusion (à droite) en fonction du nombre de descendants d'un champ épistémique au sein de la phylogénie pour quatre valeurs d_0 distinctes : $0.3; 0.4; 0.5; 0.6$.

ont moins de deux fils, 44% d'entre eux n'en ayant qu'un, environ 14% en ont au moins trois. Nous avons calculé pour l'ensemble des champs ayant k fils leur densité et leur indice de pseudo-inclusion moyen. Pour interroger la façon dont nos indices de structuration se distribuent par rapport à cette observable, nous avons également souhaité intégrer le paramètre δ_0 (similarité minimale d'un lien de filiation) comme paramètre de notre analyse afin d'évaluer la robustesse de notre reconstruction par rapport à ce seuil. Les distributions correspondantes ont donc été calculées pour différentes valeurs de δ_0 qui correspondent à autant de réseaux phylogénétiques différents. La figure 4.21 regroupe l'ensemble de ces courbes.

On observe en premier lieu une grande robustesse de la distribution des indices de structuration par rapport au nombre d'enfants. Seul un paramètre très élevé ($\delta_0 = 0.6$) semble modifier sensiblement la structure du réseau phylogénétiques. Pour $0.3 \leq \delta_0 \leq 0.5$, on constate que les deux indices de structuration sont maximaux pour un seul enfant, la valeur des indices étant minimales soit lorsque le champ ne donne lieu à aucune filiation soit lorsqu'il est très fertile (nombre important d'enfants).

Nous avons également classé les champs en fonction de leur position dans le réseau phylogénétique, autant du point de vue de leur descendance que de leur ascendance. On distingue ainsi les champs : *isolés* (ni père ni fils), *émergents* (pas de père mais un ou plusieurs fils), *adultes* (présence d'un ou de plusieurs pères et d'un ou de plusieurs fils), et *déclinant* (pas de fils mais un ou plusieurs pères).

La distribution des champs scientifiques par rapport à ces catégories est particulièrement informative. On a représenté figure 4.22 la pseudo-inclusion moyenne ou la densité moyenne d'un ensemble de champs appartenant à une catégorie donnée, on observe à nouveau des régularités très nettes dans la plage de valeurs $0.4 \leq \delta_0 \leq 0.6$ sur laquelle les distributions sont semblables. Les quatre courbes calculées (figures 4.22 et 4.21) indiquent une forme de robustesse de ces motifs phylogénétiques vis-à-vis des deux indices de structuration à l'intérieur de la plage de valeurs de seuil : $0.4 \leq \delta_0 \leq 0.5$, ce qui constitue une forme de validation théo-

tel-00464316, version 1 - 16 Mar 2010

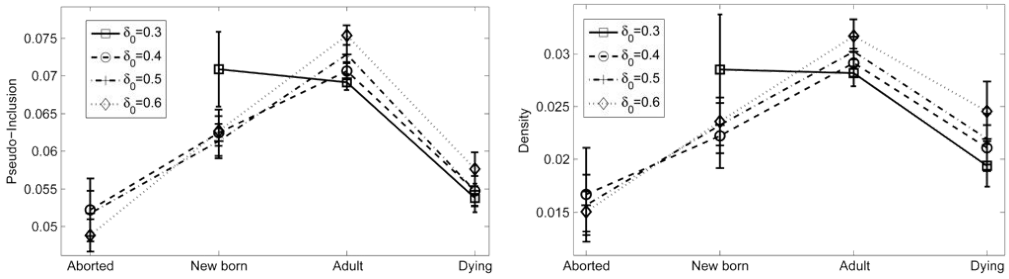


FIGURE 4.22: Densité moyenne (à gauche) et indice de pseudo-inclusion (à droite) en fonction du nombre de la nature (isolé, émergent, adulte, déclinant) ou des champs épistémiques au sein de la phylogénie pour quatre valeurs d_0 distinctes : $0.3; 0.4; 0.5; 0.6$.

rique indirecte de la méthode (Hopcroft et al., 2004b).

Les champs adultes ont des valeurs d'indice les plus fortes. Les champs en voie de disparition ou émergents ont des indices de structuration plus faibles tandis que ce sont les champs isolés qui, en moyenne, présentent les valeurs de structuration les plus faibles. La similitude de nos distributions pour nos deux indices de structuration ne signifie pas qu'ils sont équivalents. Ainsi on a observé empiriquement que certains champs pouvaient être caractérisés par un indice de pseudo-inclusion important et un indice de densité faible. C'est notamment le cas des champs dont le degré sortant (nombre important de fils) est important sur la figure 4.19 situés dans les zones dans lesquelles les fertilisations croisées sont fréquentes. Ceux-ci sont sans doute des champs émergents très cohérents, mais qui n'ont pas atteint une maturité suffisante. La faible densité de ces champs trahit alors leur jeunesse.

Ces analyses statistiques mériteraient d'incorporer d'autres paramètres (tel que le paramètre temporel par exemple) ainsi que d'être étendues aux dérivées de nos indices de structuration. On peut néanmoins conjecturer d'après ces premières observations l'existence d'une forme de cycle de vie des champs scientifiques, dont les indices de structuration augmentent après leur émergence jusqu'à l'état "adulte" avant de s'effondrer lorsque la communauté s'en désintéresse, donnant alors lieu à une extinction, ou à un morcellement en de multiples champs. Au contraire l'état de maturité qui se caractérise par un fort indice de structuration est *a priori* plus stable dynamiquement ; le faible renouvellement conceptuel induit alors un nombre limité mais non nul de descendants.

D'autres études s'appuyant sur des bases de données d'autres domaines confirmeront ou infirmeront ces hypothèses. Ces méthodes ouvrent en tout cas de nombreuses perspectives pour l'exploration comparative des motifs dynamiques mésoscopiques observés dans différents domaines.

tel-00464316, version 1 - 16 Mar 2010

4.6 Trajectoires des individus au sein des paysages sémantiques.

Nous avons décrit une méthode de reconstruction multi-échelle d'un domaine scientifique en une structure multi-échelle composée de *champs épistémiques* : ensembles de termes fortement connectés les uns aux autres et liés au sein d'un réseau $\hat{G}$ que l'on peut cartographier. Ces représentations de la connaissance sont obtenues de façon entièrement *bottom-up* (contrairement à des approches plus top-down qui proposent une labellisation des cartes des sciences à l'aide de catégories pré-existantes telles que la classification des journaux de l'ISI, *e.g.* (Moya-Anegón et al., 2004; Boyack et al., 2005; Leydesdorff and Schank, 2008a; Leydesdorff and Rafols, 2009)). On a enfin proposé une méthode de reconstruction de la dynamique de ces champs qui permet de retracer les filiations entre champs sous la forme d'un réseau phylogénétique.

Notre objectif est maintenant d'opérer un retour vers les scientifiques qui sont les véritables paysagistes de ces territoires conceptuels. En effet, les chercheurs, à travers leur production, modifient l'état du paysage scientifique, mais l'espace qu'ils contribuent à créer contraint également leur activité en retour, l'hypothèse que nous posons est que les champs épistémiques détectés réunissent l'ensemble des termes, concepts ou outils propres à une communauté scientifique. À chaque champ devrait donc correspondre un certain nombre de chercheurs qui échangent les uns avec les autres, se réunissent régulièrement au sein de congrès, ou publient dans les mêmes journaux. Nous illustrerons cette dernière partie à partir de notre base de données sur le *développement durable* décrite section 4.3.1.

4.6.1 Opérateur de projection

Maintenant que nous avons défini une méthode pour représenter les paysages scientifiques en spacialisant des réseaux de proximité entre champs $\hat{G}$, notre objectif est d'y "situer" également des chercheurs ou d'autres types d'entités, en projetant leur *bagage conceptuel* sur les cartes produites. L'opérateur de projection que nous concevons est générique au sens où il doit permettre de projeter aussi bien des chercheurs, des institutions, des journaux, ou des conférences. En toute généralité, il suffit d'attribuer à une entité un corpus défini par exemple par l'ensemble des articles auxquels elle est liée (les publications d'un chercheur, les articles publiés dans une conférence ou dans un journal donné, etc.) et d'en extraire un *bagage conceptuel* qui servira de signature de l'activité de cette entité au sein du domaine. Dans la suite, même si la méthode s'applique plus largement sans difficulté, nous ne traiterons dans nos exemples que des corpus de publications signés par *un chercheur*.

On définit donc le *bagage conceptuel* $B_i(T) \in \mathbb{N}^l$ ($l = |\mathcal{L}|$ désigne le nombre de concepts) d'un chercheur i à une période T comme le vecteur dénombrant le

tel-00464316, version 1 - 16 Mar 2010

nombre d'occurrences de chaque concept de $\mathcal{L}$ que i a mobilisé dans l'ensemble des publications dont il est l'auteur pendant la période T. Contrairement aux vecteurs C_j qui définissant les champs, le vecteur $B_i(T)$ peut prendre des valeurs supérieures à 1 si l'auteur i a publié plusieurs articles avec le même concept. Ainsi la j^{me} coordonnées de $B_i(T)$ est égale au nombre d'articles signés par i pendant la période T mentionnant le concept j.

On souhaite maintenant définir l'opérateur de projection $h : \mathbb{N}^l \rightarrow \mathbb{R}^n$ (n désignant le nombre de champs épistémiques reconstruits), qui, à un bagage sémantique donné, fait correspondre un vecteur de probabilités d'appartenance à l'ensemble des champs $\{C_j\}_{1\leq j\leq n}$, la probabilité $p_i(C_j)$ que le chercheur i appartienne à C_j peut s'écrire à partir de la distance inter-cluster déjà définie section 4.4.5. Nous proposons une définition plus générale de la distance inter-cluster précédemment définie de façon à ce qu'elle permette de prendre en compte des clusters définis comme des vecteurs dans $\mathbb{N}^l$ prenant des valeurs entières pouvant être supérieures à 1.Ainsi, la distance inter-cluster généralisée $\hat{\mathcal{S}}$ s'écrit sous la forme :

$$\hat{\mathcal{S}}(C_a, C_b) = \frac{1}{\sum_{k=1}^{n} C_a(k) \sum_{k=1}^{n} C_b(k)} \sum_{i,j=1}^{n} C_a(i) C_b(j) \mathcal{S}(i,j)$$

Cette nouvelle définition laisse inchangée la mesure entre deux clusters et permet de définir la proximité $\hat{\mathcal{S}}(B_i(T), C_j)$ entre le bagage sémantique de l'agent i et un champ C_j au temps T. Le degré d'appartenance d'un agent i à un champ j vaut alors : $p_i^T(C_j) = \hat{\mathcal{S}}(B_i(T), C_j)$

On peut également normaliser cette dernière quantité afin d'obtenir un vecteur des probabilités de présence de i sur l'ensemble des champs :

$$\hat{p}_i^T(C_j) = \frac{p_i^T(C_j)}{\sum_{k=1}^{l} p_i^T(C_k)} = \frac{\hat{\mathcal{S}}^T(B_i(T), C_j)}{\sum_{k=1}^{l} \hat{\mathcal{S}}(B_i(T), C_k)}$$

Cette projection définit un vecteur de probabilité de présence $\hat{p}_i$ qui peut s'interpréter comme une densité de présence d'un auteur sur l'ensemble des champs scientifiques [28]. Nous avons représenté figure 4.23 l'évolution de la densité de présence d'un auteur choisi au hasard dans notre base de données. Philip Lowe est professeur d'économie rurale et directeur du programme d'économie rurale et d'aménagement du territoire du Centre de l'université de Newcastle [29]. Cette représentation a comme simple ambition d'illuster la façon dont notre opérateur permet de situer un auteur au sein d'une "géographie" de la connaissance. Plusieurs observations peuvent être faites :

28. Cette normalisation n'est pas nécessairement souhaitable lorsqu'on compare la projection de différentes entités les unes avec les autres. Les distribution des degrés d'appartenance d'une entité aux champs d'un domaine peuvent être très différents d'une entité à l'autre, et il peut être souhaitable de conserver ces différences en abandonnant l'opération de normalisation.

29. Rural Economy School of Agriculture, Food and Rural Development University of Newcastle.

tel-00464316, version 1 - 16 Mar 2010

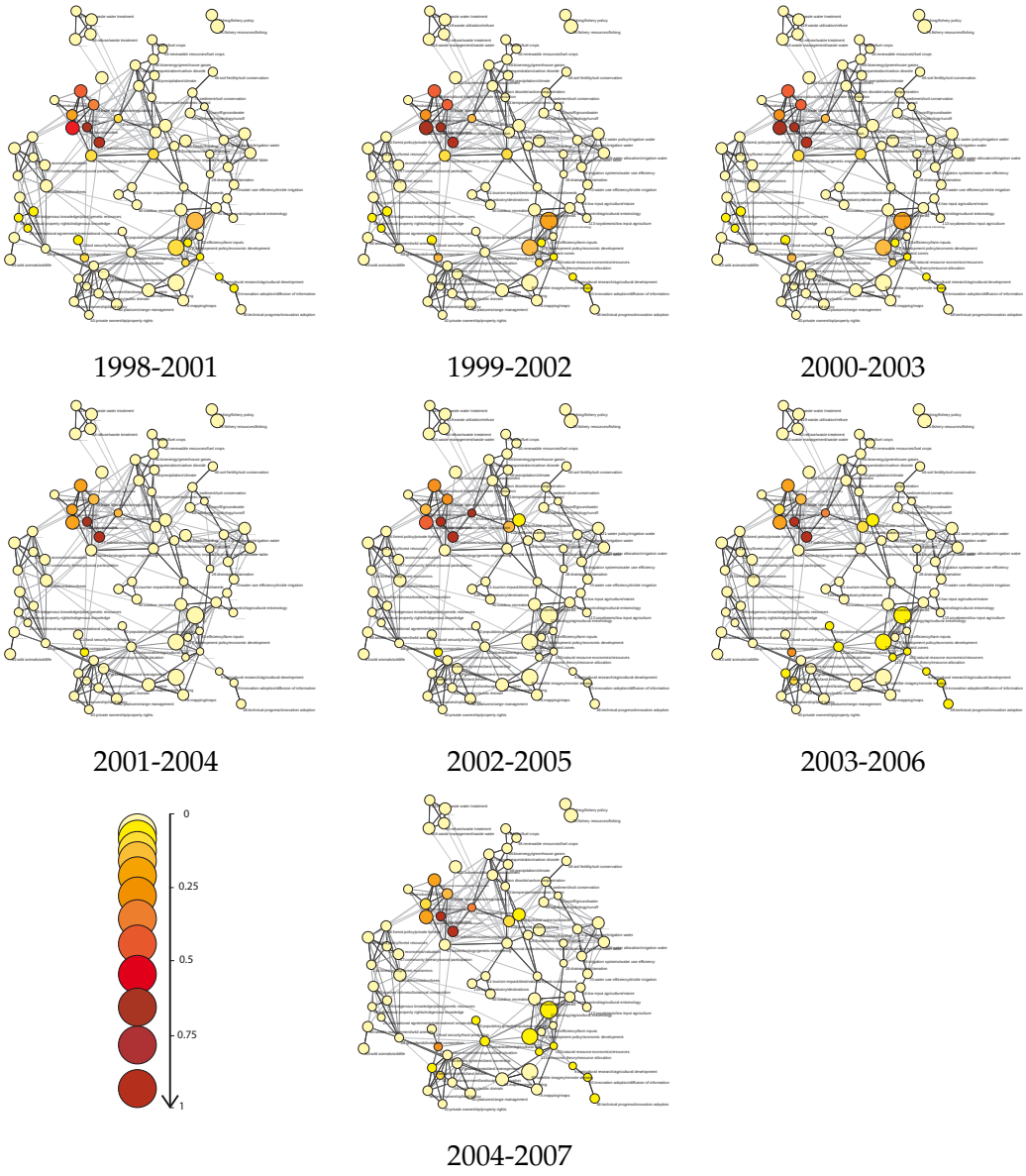


FIGURE 4.23: Evolution de la projection de l'auteur 478 (Philip Lowe) sur les 10 dernières années (fenêtres glissantes de 4 ans)

1. les champs associés à une forte densité de présence de l'auteur (zones rouges) sont contigus sur notre carte, ce qui tend à valider la reconstruction statique,
2. la densité de présence de l'auteur dans cet espace est centrée sur un ensemble de champs semblant correspondre aux domaines d'expertise du chercheur (principalement les aspects économiques et légaux des échanges commerciaux internationaux, mais également des domaines connexes)

tel-00464316, version 1 - 16 Mar 2010

3. la densité de présence de l'auteur évolue très peu dans le temps, laissant penser que le chercheur représenté ici est resté "fidèle" aux communautés épistémiques auxquelles il participe,
4. lorsqu'il y a dispersion vers d'autres champs — c'est à dire activation de nouveaux champs — les champs touchés semblent être proches des champs activés précédemment, *i.e.* les champs nouvellement peuplés se situent dans le voisinage immédiat des champs précédemment occupés [30].

Ces observations semblent donc valider nos reconstructions, et suggèrent une certaine forme de stabilité dans l'engagement des chercheurs auprès d'une communauté épistémique donnée. Nous cherchons maintenant à rendre compte de façon plus rigoureuse de l'attraction exercée sur les scientifiques par leurs champs épistémiques d'appartenance à travers le calcul de mesures d'attachement préférentiel liant la structure des champs à l'évolution de l'activité des chercheurs.

4.6.2 Rétroaction macro-micro

On s'attache maintenant à vérifier l'hypothèse d'une *stabilité dynamique* des champs d'appartenance des chercheurs en étudiant les motifs d'évolution d'un ensemble I de plus de 800 scientifiques du domaine ayant publié au moins 7 articles parmi notre base de publications de départ durant la période 1998-2007 [31]. Pour apprécier la vitesse de dispersion du vecteur de densité des auteurs au sein de notre carte, nous mettons en place deux mesures.

En premier lieu nous introduisons une mesure entre deux densités de présence normalisées $\hat{p}_a$ et $\hat{p}_b$ à l'aide de la version symétrisée de la divergence de KullBach Leibler D_{KL} (Kullback and Leibler, 1951). Etant donnée deux distributions de probabilités à valeurs non nulles [32], la divergence entre les deux distributions de probabilité P et Q est définie par : $D_{KL}(P,Q) = \sum_i P(i)log(P(i)/Q(i))$; cette distance est asymétrique mais est classiquement symétrisée en effectuant une moyenne. Dans sa version symétrique on définit alors la distance entre deux distributions comme $1/2[D_{KL}(P,Q) + D_{KL}(Q,P)]$ (Johnson and Sinanovic, 2001).

On définit donc δ comme la distance de Kull-Back Leibler symétrisée [33] entre deux vecteurs de probabilités de présence $\hat{p}_a$ et $\hat{p}_b$. Cette distance s'exprime donc

30. Cette observation reste valable en considérant une séquence de cartes sans recouvrement temporel.

31. Cette borne permet de réunir un nombre suffisant de chercheurs ayant *a priori* publié durant plusieurs années.

32. Dans notre cas, ll paraît improbable de trouver au sein des vecteurs de densité individuels des valeurs parfaitement nulle car cela supposerait que les termes mobilisés par un auteur ne co-occurrent pas une seule fois avec aucun des termes d'un champ donné.

33. Même si la divergence de KullBack Leibler, même symétrisée ne vérifie pas les conditions d'une distance (l'inégalité triangulaire n'est pas respectée) nous l'appellerons néanmoins distance par commodité.

tel-00464316, version 1 - 16 Mar 2010

sous la forme :

$$\delta(\hat{p}_a, \hat{p}_b) = 1/2(D_{KL}(\hat{p}_a, \hat{p}_b) + D_{KL}(\hat{p}_b, \hat{p}_a))$$

Cette mesure permet notamment d'estimer le déplacement qu'a effectué un auteur entre deux périodes successives. On calcule la propension moyenne de déplacement d'un chercheur à une distance donnée en contrastant la distribution des distances $\{\delta(\hat{p}_i^T, \hat{p}_i^{T^-})\}_{i \in I}$ [34] observées sur l'ensemble des chercheurs entre deux périodes successives T^- et T avec l'ensemble des distances $\{\delta(\hat{p}_i^T, \hat{p}_j^{T^-})\}_{(i,j) \in I^2}$ calculée sur la totalité des paires de chercheurs (i, j) entre les deux mêmes périodes. Ce mode de calcul revient à faire l'hypothèse d'un modèle nul construit en nous appuyant sur la distribution des mots-clés sur les agents à un temps donné comme une distribution typique d'un agent actif dans la communauté indépendamment de son activité antérieure. Les périodes successives choisies pour le calcul de la propension sont des fenêtres de trois ans non recouvrantes, soit l'ensemble des couples $\{(T_k; T_k^-)\}_{1 \leq k \leq 8}$ où $T_k^- = [1995 + k - 1, 1995 + k + 1]$ et $T_k = [1998 + k - 1, 1998 + k + 1]$. La propension représentée est une moyenne sur l'ensemble des couples de périodes $\{(T_k; T_k^-)\}_{1 \leq k \leq 8}$ accompagnée de l'intervalle de confiance associé.

La propension de déplacement en fonction de notre distance δ est représentée figure 4.24. Celle-ci est fortement décroissante, ce qui indique que la densité de présence d'un chercheur a tendance à être très stable d'une période à l'autre. Il est ainsi 10 fois plus probable pour un chercheur de limiter son déplacement conceptuel à $\delta < 0.1$ que ne le laisserait supposer un modèle aléatoire, *a contrario*, les déplacements importants ($\delta > 1.5$) sont 10 fois moins probable qu'attendu.

Mais cette mesure qui s'appuie sur le vecteur de densité de présence des agents, ne permet pas de rendre compte de la structure sous-jacente de l'organisation des champs. Afin d'illustrer quantitativement notre intuition sur la "diffusion" des champs d'appartenance des auteurs via les liens de notre réseau, nous proposons d'introduire une autre distance à même de rendre compte des déplacements des auteurs en fonction de la topologie du réseau des champs.

On introduit dans un premier temps un seuil θ (en pratique, $\theta = 0.15$, on a vérifié que les résultats restent extrêmement robustes aux modifications de ce seuil) qui permet d'attribuer à un auteur i l' ensemble des champs C_j, qui vérifient $\hat{p}_i(C_j) \geq \theta$ *i.e.* leur probabilité de présence dans un champ doit être supérieure à un seuil pour que ce champ soit retenu.

L'état d'un agent i au temps T est donc défini par un ensemble de champs d'appartenance $A_i^t = \{C_k\}_{\hat{p}_i^T(C_k) > \theta}$. La distance Δ entre deux états successifs d'un agent peut alors être définie comme la moyenne de la distance minimale (au sens du plus court chemin dans un graphe) pour se déplacer depuis $A_i^{T^-}$ l'ensemble

34. Dans le cas où un auteur n'aurait pas publié dans une des deux périodes considérée, cette mesure est simplement ignorée, la densité de probabilité de présence du chercheur en question n'étant pas définie.

tel-00464316, version 1 - 16 Mar 2010

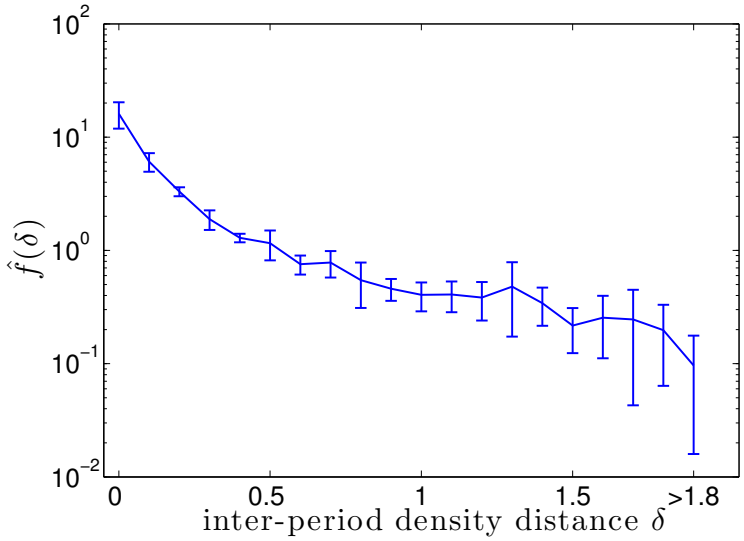


FIGURE 4.24: Propension à occuper un nouveau champ épistémique en fonction de la distance de déplacement des auteurs δ.

des champs d'appartenance de i à T^- vers chacun des champs de A_i^T auxquels i appartient à t. Plus formellement

$$\Delta(A_i^{T^-}, A_i^T) = \frac{1}{|A_i^T|} \sum_{j \in A_i^T} \left[\min_{C_k \in A_i^{T^-}} d(C_j, C_k) \right]$$

où d représente la distance dans le graphe (longueur du plus court chemin dans $\hat{G}$ permettant de naviguer d'un nœud à un autre) calculée avec l'algorithme de Dijkstra (Dijkstra, 1959) [35] [36].

La distance définie ci-dessus est une moyenne sur l'ensemble des déplacements opérés par un acteur. Ainsi, afin de rendre compte de l'hétérogénéité des déplacements et obtenir une mesure moins agrégée plus à même de rendre compte de la continuité ou de la discontinuité thématique dont font preuve les chercheurs dans leur déplacement, on peut également associer à chaque agent se déplaçant dans le paysage épistemique et pour chaque champ d'appartenance $C_j \in A_i^T$ de l'agent i à la période T, l'ensemble des distances $\{\Delta(A_i^{T^-}, C_j)\}_{C_j \in A_i^T}$. Ces déplacements sont comparés aux déplacements que l'on obtiendrait en appliquant la même hypothèse nulle que celle décrite précédemment afin de calculer la propension à "adopter" un champ situé à une distance Δ donnée.

35. Nous avons utilisé la version non pondérée du réseau même si les résultats sont sans doute semblables en conservant ces poids et en étendant la définition de la distance à un coût de circulation dans le réseau pondéré.

36. Dans le cas où un auteur n'aurait pas publié dans une des deux périodes considérée, cette mesure est ignorée comme précédemment.

tel-00464316, version 1 - 16 Mar 2010

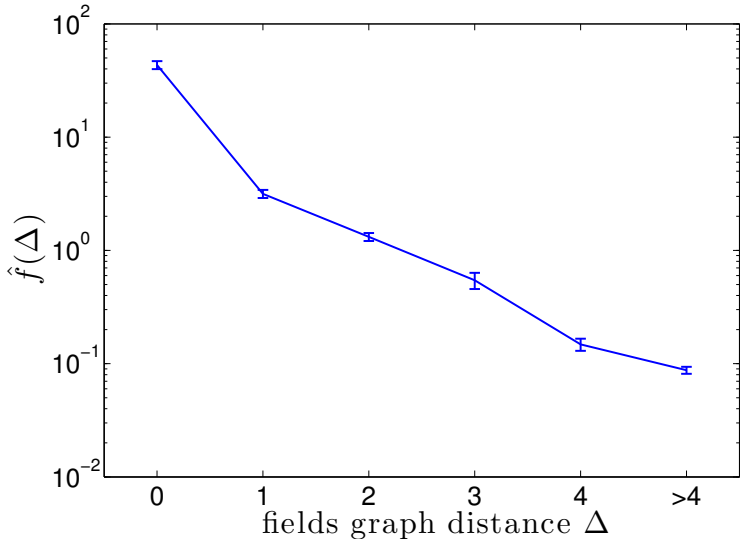


FIGURE 4.25: Propension à occuper un nouveau champ épistémique en fonction du déplacement Δ opéré dans le graphe des champs — périodes de trois ans par rapport aux trois années précédentes de publication.

La propension moyenne de déplacement dans le réseau des champs sur l'ensemble des auteurs et des périodes d'observation (en suivant les mêmes hypothèses que précédemment) est représentée figure 4.25. On observe que cette courbe est à nouveau fortement décroissante, indiquant que les chercheurs ont tendance à "adopter des champs" avec une propension d'autant plus faible que ces champs sont éloignés des champs auxquels ils participaient précédemment. Par contre, on constate une très forte tendance à la "répétition", la propension à rester dans un champ précédemment occupé ($\Delta = 0$) est proche de 80. La valeur obtenue signifie donc qu'un auteur a près de 80 fois plus de chances de continuer à occuper le même champ épistémique le pas de temps suivant que ne le laisserait supposer un modèle aléatoire. La propension chute ensuite d'un ordre de grandeur dès qu'on envisage des champs à distance 1. *A contrario* un déplacement à distance 4 ou supérieure est 10 fois moins probable que ne le prévoierait notre hypothèse nulle. Ce calcul confirme l'intuition que nous avions d'une grande stabilité de la dynamique des auteurs au sein de ce paysage. Il nous conforte également quant à la qualité du réseau de champ reconstruit dont la topologie semble exercer une influence capitale vis à vis de l'activité des auteurs et de leur évolution.

Les déplacements conceptuels des chercheurs semblent donc épouser les chemins formés par les relations de proximité entre champs. De façon symétrique au réseau social qui supporte la diffusion de concepts, on peut considérer de façon duale, que les cartes des sciences constituent le medium sur lequel les scientifiques "diffusent".

Cette dernière courbe peut également être interprétée comme la mise en évi-

tel-00464316, version 1 - 16 Mar 2010

dence d'une rétroaction du niveau macro sur le niveau micro, au sens où, la structure des champs qui émergent des statistiques brutes extraites de l'ensemble des publications "contraint" en retour la dynamique des scientifiques qui se retrouvent plus ou moins "emprisonnés" dans leurs champs d'appartenance ou dans leur voisinage proche. Ce résultat fournit donc l'illustration quantitative d'un effet d'immergence qu'exercent des structures de haut niveau (qui émergent pourtant directement de l'activité des chercheurs) sur les dynamiques individuelles.

Les mesures que nous avons introduites permettent également de définir des indices mesurant l'activité individuelle des agents dans ces paysages conceptuels ; on peut ainsi aisément déduire des distances précédemment introduites (δ et Δ), un indice global de déplacement général d'un agent agrégeant l'ensemble de ses déplacements dans le temps afin d'apprécier sa mobilité. De la même façon un indice de diversité (lié à la pluridisciplinarité d'un auteur) peut également aisément être construit à partir des vecteurs A_i^t en calculant par exemple la distance moyenne entre termes dans le réseau des champs.

4.6.3 Se déplacer dans un espace mouvant

L'analyse des dynamiques des scientifiques dans ces espaces conceptuels devrait également tenir compte des évolutions propres de l'espace. Si nous suivons Sewell (1992) :

> "...Of course, if cultural and societal (network) structures shape actors, then it is equally true that actors shape these structures in turn. Cultural an social structures do not, in other words, by themselves bring about or somehow "cause" historical change. Rather, it is the actions of historical subjects that actually "reconfigure" (given historically conducive circumstances) existing, long-term structures of action, both cultural and societal"

Les structures dans lesquelles se déplacent les agents sont également susceptibles d'évoluer sous l'effet de l'activité de ces mêmes agents. Or, nous avons fait l'hypothèse que la structuration du domaine restait relativement uniforme durant les 10 ans de notre analyse et avons donc calculé les propensions d'aborder un nouveau champ en fonction d'une cartographie tenant compte de l'activité scientifique sur l'ensemble de la période. Ainsi, certains déplacement d'auteurs observés sont certainement simplement dûs à une modification de l'espace sur lequel ils sont projetés. Les effets d'attachement à une communauté scientifique donnée seraient alors sans doute encore plus forts si nous avions envisagé des communautés dont le contenu évolue continuellement, les scientifiques présents dans une communauté ayant tendance à la fois à en modifier les frontières et à en suivre les glissements.

À titre d'illustration, nous avons tracé figure 4.26 la projection d'un auteur (P. Lowe à nouveau) sur une phylogénie. La méthodologie est identique à celle qui

tel-00464316, version 1 - 16 Mar 2010

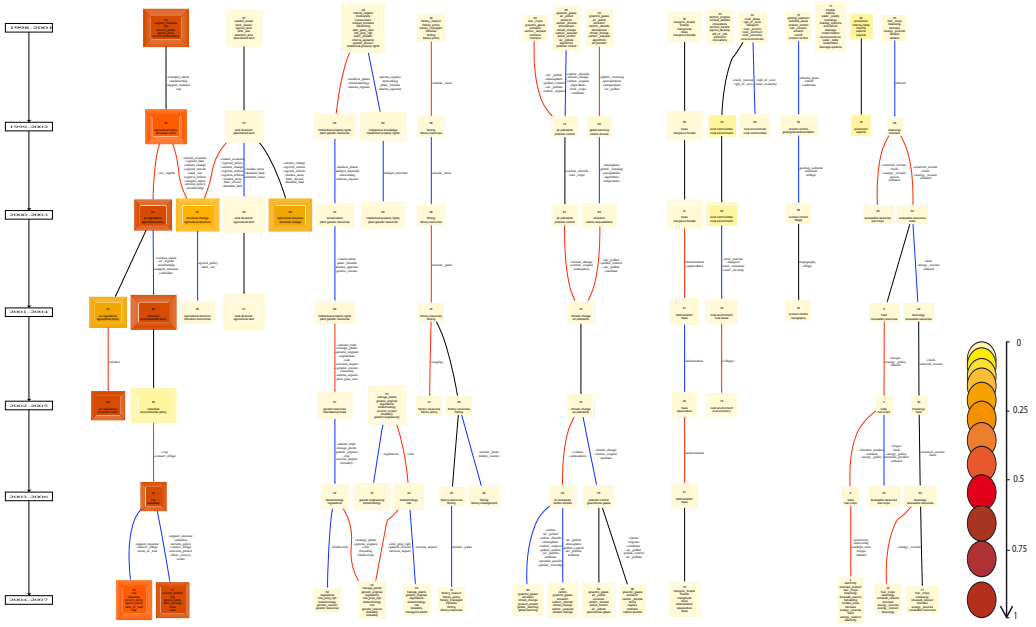


FIGURE 4.26: Extrait de la phylogénie du domaine *développement durable* sur les 10 dernières années (fenêtres glissantes de 4 ans), la couleur des champs correspond à la densité de présence d'un auteur (P. Lowe) sur les champs (du plus clair (blanc) au plus foncé (rouge)) ; la phylogénie complète est beaucoup plus étendue mais ne peut pas être aisément représenté. On remarque la présence privilégiée de l'auteur sur une seule branche de notre réseau.

nous avait précédemment permis de créer la projection des publications d'un chercheur sur une carte pendant une période donnée. Nous appliquons pour chaque période temporelle T notre opérateur de projection h sur l'ensemble des champs épistémiques détectés durant cette période. Cette opération permet de fournir une densité de présence associée à cet auteur pour chaque champ de chaque période. Nous représentons ensuite le résultat de cette projection sur la phylogénie calculée aux périodes successives en attribuant à un champ une couleur d'autant plus rouge que l'auteur a une probabilité de présence élevée dans ce champ. Le résultat, bien que nous n'ayons pas pu reproduire la totalité de la phylogénie pour des raisons de place, montre, pour l'auteur considéré (qui est le même que celui représenté figure 4.23), qu'il suit une trajectoire très linéaire au sein de la phylogénie. Une des branches de la phylogénie présente des taux d'occupation particulièrement importants et ce sur l'ensemble des périodes. Cette branche a subi des mutations importantes qui ont transformé le champ de départ (période 1998-2001) composé des termes : *support measure*, *CAP*, *environmental policy*, *agricultural policy* en un champ final (période 2004-2007) composé des mots-clés suivants : *environment protection*, *CAP*, *agricultural policy*, *roles* et *farm surveys* en passant par des champs ayant emprunté puis délaissé les termes *subsidies*, *support measure*, *EU regulation*,

tel-00464316, version 1 - 16 Mar 2010

ou *conservation tillage*. Malgré, ces transformations touchant à la définition même des champs, on observe une grande stabilité du chercheur considéré sur les 10 ans que traverse cette branche du réseau phylogénétique.

Perspectives

Nous avons développé un ensemble de méthodes de cartographie des sciences à partir de corpus électroniques de publications scientifiques en nous appuyant sur des outils de fouille de données et d'analyse de réseaux. Ces cartes permettent de représenter un domaine scientifique sous la forme de champs épistémiques qui s'articulent les uns avec les autres. L'enjeu est d'être capable de développer des outils permettant de développer une véritable épistémologie quantitative exclusivement fondée sur les traces de l'activité scientifique. Mais les perspectives ouvertes sont aussi bien théoriques qu'applicatives, ces méthodes de reconstruction permettent également de définir de nouvelles modalités d'interaction avec la science, notamment en ce qui concerne la navigation à travers de grandes bases de donneés. On peut même songer à des application du type reconstruction de connaissance à partir d'informations manquantes (données de micropuces, ou de réseaux de régulation en biologie).

Une partie de l'intérêt de ces cartes pour les théoriciens (sociologues des sciences et techniques, historiens ou philosophes des sciences), pour leurs chercheurs ou pour les gestionnaires de la science, est de les équiper de représentations et d'interfaces de manipulation de ces représentations afin de mieux saisir un paysage scientifique extrêmement mouvant, et donc de mieux anticiper, voir et comprendre les dynamiques qui l'animent.

Un des défis qu'il s'agit de relever tient précisément à la question de la représentation et de la communication de ces résultats. Les structures que nous avons construites ne sont pas toutes aisément manipulables. Ainsi au delà d'une interrogation d'ordre globale sur les motifs qui les composent, les phylogénies paraissent difficilement utilisables en l'état. Il nous faut donc inventer des représentations plus pertinentes ou des modalités de navigation plus locales qui interrogent l'usage que l'on souhaite faire de ces cartes et qui se construisent *de facto* en interaction très forte avec les destinataires des cartes.

Les méthodologies que nous avons décrites peuvent en grande partie être étendues à d'autres type de corpus. On peut envisager les appliquer à des domaines connexes aux sciences (base de brevets par exemple) ou plus éloignés : données de cooccurrences extraites de l'activité des communauté en ligne (nous pensons évidemment au texte brute des billets de blogs ou plus largement aux contenus véhiculés au sein de communautés de savoirs, mais également aux tagging system (folksonomy)) ou encore données extraites des requêtes des moteurs de recherche.

Dans une perspective plus directement liée à notre exploration des communau-

tel-00464316, version 1 - 16 Mar 2010

tés de savoirs, ces méthodes offrent une véritable opportunité pour améliorer leur modélisation. En effet, la modélisation des réseaux socio-sémantique et sémantique que nous avions introduit dans le chapitre précédent peut maintenant être étendue grâce à nos méthodes de reconstruction des dynamiques scientifiques (ou, dans une perspective plus large, sémantiques). Ainsi, nous pouvons utiliser notre notion de champ épistémique (ou simplement de champ) pour remplacer les entités linguistiques que nous avions privilégiées et qui étaient susceptibles d'induire un certain nombre de problèmes (liés notamment à la polysémie ou à la synonymie des termes par exemple). Le réseau socio-sémantique serait ainsi transformé en un réseau biparti liant les agents à leurs champs d'appartenance, tandis que le réseau sémantique est directement identifiable à notre réseau de champs $\hat{G}$. Nous n'avons pas pu reprendre l'ensemble des analyses du chapitre précédente à travers cette nouvelle modélisation, mais conjecturons que ces méthodes ouvrent la voie vers une prise en compte encore plus fidèle de la phénoménologie des communautés de savoirs.

tel-00464316, version 1 - 16 Mar 2010

Résumé du chapitre:

Après une rapide analyse des relations entre des communautés structurelles et communautés thématiques sur le web social français, nous avons proposé dans ce chapitre, un ensemble de méthodes de reconstruction des dynamiques des communautés scientifiques à différentes échelles. La cartographie des sciences se situe à la croisée de nombreux enjeux méthodologiques, politiques ou de gestion. Notre apport a consisté à proposer une série de méthodes de reconstruction entièrement *bottom-up* qui rende compte de la nature hiérarchique de l'organisation des sciences mais aussi de la polysémie des concepts.

Une mesure asymétrique de proximité entre termes adaptée à l'hétérogénéité de la distribution des occurrences d'apparition des termes dans les publications scientifiques a été introduite. Une méthode de catégorisation multi-échelle d'un ensemble de termes a ensuite été proposée. Les clusters ainsi produits ont été qualifiés au moyen d'indices permettant d'en apprécier la cohésion, l'importance, ou l'évolution. Nous avons également introduit un certain nombre de mesures et de méthodes à même de rendre compte de la dynamique de ces clusters.

La dynamique mésoscopique des champs épistémiques a été reconstruite sous la forme d'un réseau phylogénétique regroupant les motifs de filiation entre champs. Ces structures semblent dotées d'un certain nombre de propriétés remarquables vis-à-vis des indices de structuration de nos champs, ce qui ouvre la voie vers une véritable épistémologie quantitative. Enfin, les publications des chercheurs ont été reprojetées sur les champs épistémiques reconstruits. L'évolution de la distribution de ces chercheurs dans ces paysages conceptuels montre une forte stabilité de leurs déplacements dans ces espaces dont la topologie rétroagit sur le comportement des agents.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Troisième partie

Diffusion dans les réseaux sociaux

La partie précédente nous a permis de mettre en évidence un certain nombre de motifs à la fois sociaux, sémantiques, ou socio-sémantiques émergents qui structurent nos communautés de savoirs. On a également montré comment les dynamiques individuelles étaient pourvues d'un certain nombre de régularités liées directement à ces structures. Nous nous focaliserons, dans cette dernière partie, sur les processus de diffusion qui animent nos communautés de savoirs et qui peuvent également être interprétés comme des processus socio-sémantiques émergents.

Les phénomènes de diffusion sont par nature de type socio-sémantique. Non seulement ils décrivent la succession d' "états" pris par les agents d'un système par rapport à une innovation ou à une opinion, mais ils traduisent également la façon dont ces modifications sont médiatisées par les influences locales que les agents exercent sur leur environnement social. À ce titre, les processus de diffusion constituent une illustration du couplage entre le réseau social et le bagage cognitif des individus. Les épisodes de diffusion qui traversent les communautés de savoirs apparaissent également comme une propriété dynamique émergente des évolutions micros du système. Un processus de diffusion est par essence construit par la juxtaposition d'événements de transmission d'information locaux, pourtant, on peut définir, au niveau de l'ensemble du système des propriétés globales de ce processus telles que sa vitesse, ou sa longévité.

Nous adopterons deux stratégies pour comprendre les phénomènes de diffusion dans les communautés de savoirs. Nous aborderons dans un premier temps (chapitre 5) la question de l'influence entre différentes sources de contenus avec un point de vue exclusivement sémantique et en nous attachant à détecter des motifs inter-temporels dans l'activité de production de contenus d'un ensemble de groupes de blogs marqués politiquement et de la presse. Notre objectif est donc d'extraire des comportements dynamiques systématiques de reprise de tel ou tel contenu à partir de l'analyse longitudinale des contenus produits par une communauté de blogueurs politiques. Peut-on reconstruire un *diagramme d'influence* entre ces sources qui nous renseigne sur les motifs intertemporels qui corrèlent, les unes avec les autres, les activités de production de contenus de certains sous-ensembles de sources ? Peut-on, par exemple, décrire l'influence systématique de la presse sur les blogs de droite, ou, par exemple, de façon plus complexe, l'influence couplée

de la presse et des blogs centriste sur les thématiques qu'aborderont les blogs de droite ?

Dans les chapitres suivants, nous envisageons la question de la diffusion sur un réseau social à part entière en nous interrogeant sur l'influence des structures (locales et globales) du réseau d'interaction inter-individuel vis-à-vis des processus de diffusion. Dans une perspective macroscopique, nous proposons un protocole simulatoire (chapitre 6) afin de comprendre en quoi la topologie d'un réseau, support d'un épisode de diffusion, affecte la dynamique globale de cette diffusion. Plus précisément, et dans le prolongement de notre parti pris empirique, nous nous interrogeons sur les propriétés structurelles de réseaux sociaux réels susceptibles de modifier de façon sensible la vitesse de diffusion vis-à-vis d'un modèle de transmission inter-individuelle donné. Pour ce faire nous proposons de comparer de façon systématique les dynamiques de diffusion observées sur nos réseaux réels avec celles observées sur différents réseaux stylisés. Le réalisme des modèles de transmission choisis sera également interrogé.

Enfin, nous présenterons, dans le chapitre 7, une analyse des processus de diffusion à un niveau plus local, fondée sur le suivi *in-vivo* d'épisodes de diffusion d'URLs observés au sein des blogosphères politiques française et américaine. Cette analyse nous permet d'identifier, à un niveau égocentré, les paramètres structurels susceptibles d'être corrélés à l'influence d'un blog, mesurée comme le nombre de transmissions de ressources qu'il génère au sein du système. Quelles propriétés structurelles du réseau social sont-elles corrélées à l'influence d'un blog ? La longévité d'un épisode de diffusion peut-elle dépendre du chemin emprunté par l'information ?

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

CHAPITRE 5

Corrélations intertemporelles entre sources

Sommaire

5.1 Création des catégories de blogs 169
5.1.1 Définition des profils sémantiques instantanés des blogs . . . 169
5.1.2 Catégorisation des blogs selon leur sensibilité politique 170
5.2 Diagramme de corrélations intertemporelles 172
5.2.1 Contexte 172
5.2.2 Machines à états causaux 173
5.2.3 Alphabet des concepts 174
5.2.4 Définition d'une dynamique symbolique 175
5.2.5 Resultats 176
5.2.6 Perspectives 178

Avant d'explorer les mécanismes liés à la diffusion au sein d'un réseau social, nous abordons les dynamiques d'influence à un niveau d'agrégation supérieur en examinant la façon dont un ensemble de *catégories* de sources de contenu s'influencent mutuellement au sein d'une écosphère informationnelle. Notre approche sera à nouveau fondée sur l'observation *in-vivo* des dynamiques d'une communauté de savoirs mais l'apport essentiel sera néanmoins d'ordre méthodologique en proposant l'application d'une méthode générique de *mécanique computationnelle* Shalizi (2001a) à la reconstruction de corrélations intertemporelles entre des sources de production de contenus. Ce travail a fait l'objet d'une première analyse en collaboration avec Camille Roth et Emmanuel Faure. Ce chapitre étend donc la méthodologie déjà introduite dans (Cointet et al., 2007) tout en l'appliquant à un jeu de données plus large.

Les individus soulèvent des questions, échangent des points de vue, font part de leurs concernements au sein du système de production de contenus et d'interaction distribué que forment les communautés de savoirs. Bien que purement locales les modalités d'action des individus peuvent suivre certaines structures régulières. Notre objectif est donc d'extraire des comportements dynamiques systématiques de reprise de tel ou tel contenu à partir de l'analyse longitudinale des contenus produits par une communauté de blogueurs politiques.

Comme on l'a vu, la blogosphère peut être étudiée comme un système complexe à part entière doté d'une dynamique de production de contenus distribués sur l'ensemble des blogs ; les discussions ou conversations se déployant le long d'un réseau d'interaction lui même dynamique. Néanmoins, ce système de sources de contenus interconnectées bien que doté d'une dynamique propre, n'est pas isolé du "monde extérieur" ; on peut raisonnablement attendre d'un espace de débat public, fût-il médiatisé par un substrat purement numérique, qu'il entre en interaction avec un environnement plus large.

On a également déjà insisté dans le chapitre 4 sur le profil particulièrement "accidenté" de l'activité de production de contenus dans les blogs. Celui-ci se caractérise par la présence de "pics" (spikes) d'activité autour de certaines thématiques qui accaparent l'attention d'une grande partie des blogueurs pendant une courte durée et par des épisodes se développant sur des plages temporelles plus longues de "conversation" entre sources Gruhl et al. (2005, 2004). Dans tous les cas, il existe de fortes corrélations dynamiques voire des phénomènes de résonance entre l'ensemble des auteurs, *i.e.* des agents discutant d'un même sujet au même moment ou avec un certain décalage temporel. Les blogueurs s'influencent les uns les autres (le réseau des liens hypertextes peut être un moyen de mettre en évidence ces influences qu'elles soient directes Herring et al. (2005) ou indirectes Adar et al. (2004b)) mais ils sont également soumis à des influences extérieures à la blogosphère, notamment les media Lloyd et al. (2006), connaissances dans le "monde réel" (ou plutôt actuel par opposition à un monde virtuel), etc. Ces influences sont tout à fait primordiales voire constitutives dans le cas qui nous occupe ici : un échantillon de la blogosphère politique intervenant dans un espace public de débat plus large.

La question de l'indépendance relative ou de l'influence exercée par le monde extérieur est tout à fait prégnante lorsqu'on envisage les dynamiques des communautés dites citoyennes (réunies sous la quatrième catégorie de blogs dans la typologie des modes d'énonciation de la blogosphère de Cardon and Delaunay-Teterel (2006)) liées à l'actualité sociale, économique ou politique. En effet la dépendance de ces communautés vis-à-vis des arènes plus classiques telles que celle des media est une question largement débattue dans la littérature depuis la question de l'antériorité d'une source de contenus sur une autre (blogs *versus* media Leskovec et al. (2009))[1] jusqu'à la façon dont la blogosphère politique est questionnée quant à son autonomie réelle vis-à-vis d'arènes traditionnelles (Wallsten, 2005; Lloyd et al., 2006) - la question de fond étant de déterminer si les espaces

1. Ces problématiques s'inscrivent également dans un contexte de crise du journalisme, mis en péril selon certains par l'immédiateté des outils d'édition et de communication sur le web (qu'on songe simplement aux récents événements qui ont animé la plateforme de microblogging *Twitter* (Huberman et al., 2008; Java et al., 2007; O'Connor, 2009)), qui a donné lieu a la création de nouveaux modèles économiques de publication, ou à des formes émergentes de journalisme partcipatif et citoyen Pledel (2008)

tel-00464316, version 1 - 16 Mar 2010

de discussion ouverts par les technologies du Web 2.0 produisent réellement des espaces publiques à même de générer des nouveaux modes d'action et de participation à la vie politique (Goodman, 1964; Flichy, 2008; Thelwall and Price, 2006).

Nous appliquons la question de la corrélation des contenus publiés par différentes sources en observant les profils d'activité de notre échantillon de blogs politiques français précédemment introduit (voir section 2.3.6) que l'on catégorise selon leur inclination politique (que l'on résumera grossièrement aux blogs de sensibilité de droite, de gauche, ou du centre), et auxquels on adjoint un ensemble de sites web des trois grands quotidiens nationaux (*Le Figaro*, *Le Monde* et *Libération*). Notre objectif est de questionner la façon dont les comportements d'édition de ces 4 grandes catégories de sources peuvent être corrélés les uns aux autres.

Nous nous appuyons sur un formalisme basé sur des chaînes de Markov cachées pour extraire les motifs dynamiques de corrélation entre les occurrences de certains sujets entre ces groupes de sources distincts. La distribution des sujets au sein des groupes à un moment donné est donc modélisée comme un état du système, dont nous cherchons à reconstruire la dynamique. Pour détecter les motifs de corrélation intertemporelle entre sources dans la blogosphère et dans la presse, nous proposons de faire appel à un formalisme basé sur les machines à états causaux (causal states machines) Crutchfield and Young (1989); Shalizi and Shalizi (2004) issu de la mécanique computationnelle.

Au delà de la mise en évidence de corrélations entre les contenus discutés dans la blogosphère politique française et ceux issus de l'actualité politique publiés dans la presse - à même de mettre en évidence la "subordination" de l'agenda d'un territoire à un autre, nous visons également à identifier des motifs de corrélation inter-groupes plus riches et plus variés au sein même des différentes catégories de blogueurs. L'objectif est d'obtenir une description aussi exhaustive et concise que possible de la façon dont le comportement de production de contenus d'un groupe ou d'un ensemble de groupes est affecté par les contenus publiés par les autres groupes. À ce titre, nous étendons la notion de corrélation comparant deux variables, à une corrélation de nature n-adique permettant de révéler des relations de corrélation plus riches entre $n > 2$ variables (Funabashi et al., 2009).

5.1 Création des catégories de blogs

5.1.1 Définition des profils sémantiques instantanés des blogs

Nous nous appuyons sur l'ensemble des 120 blogs politiques français auquel on a ajouté 3 sources "institutionnelles" (éditions web des quotidiens nationaux) dont nous avons collecté l'activité pendant les 6 premiers mois de l'année 2007. Les profils de production de contenus sont estimés en suivant l'ensemble des 190 syntagmes déjà décrit section 2.3.6. Ils incluent un large ensemble de sujets et de noms de personnalités politiques ayant alimenté les débats durant la campagne

tel-00464316, version 1 - 16 Mar 2010

présidentielle française de 2007.

Les statistiques de base à partir desquelles nous travaillerons consistent donc pour une source i en un vecteur sémantique instantané $\hat{w}_t(i)$. Cette fois ci, et contrairement au chapitre 3, nous caractériserons le profil d'une source comme un vecteur dépendant uniquement des contenus publiés à t et non comme la résultante de l'agrégation de l'ensemble des contenus produits jusque là. La procédure de pondération des occurrences des syntagmes par le *tf.idf* reste semblable à celle employée dans la partie précédente (partie II), on définit donc le vecteur sémantique d'un blog $\hat{w}_t(i)$ selon la formule :

$$\hat{w}_t(i,c) = \frac{W_t(i,c)}{\sum_{c=1}^{|\mathcal{W}|} W_t(i,c)} \cdot \log \frac{|\mathcal{B}|}{|\{j, \mathbf{W}(j,c) > 0\}|}$$

où $W_t(i,c)$ désigne le nombre d'occurrences du concept c dans les contenus publiés au temps t (soit le jour t dans la suite) par le blog i, et $|\mathcal{B}|$ désigne le nombre de sources (soit 123)[2].

On emploiera à nouveau une mesure de similarité de type cosinus (mesure de corrélation) pour mesurer la proximité sémantique entre deux blogs i et j en fonction de leur profil sémantique $\hat{w}(i)$[3] et $\hat{w}(j)$ calculés à partir de l'ensemble des contenus qu'ils ont publié sur l'ensemble de la période d'observation : $s(i,j) = \frac{\hat{w}(i) \cdot \hat{w}(j)}{\|\hat{w}(i)\|\|\hat{w}(j)\|}$. Cette distance doit nous permettre de catégoriser les blogs selon des ensembles thématiquement cohérents, censés refléter la sensibilité politique des blogueurs.

5.1.2 Catégorisation des blogs selon leur sensibilité politique

On a montré que les blogueurs reproduisent en partie des motifs relationnels similaires à leur alter-ego "réel" : par exemple Adamic and Glance (2005b) ont montré que les regroupements (basés sur des critères structurels) de blogs au sein de la blogosphère politique américaine suivaient les frontières partisanes traditionnelles (blogs démocrates ou républicains). Cette similarité avec le monde réel nous encourage à proposer une catégorisation des blogs en fonction de leur sensibilité politique, dont nous ferons l'hypothèse qu'elle permet de définir des catégories de sources cohérentes et pertinentes vis-à-vis de notre problématique.

2. Comme nous souhaitons par la suite repérer les pics d'activité associés à un concept donné, nous avons fixé le terme de pondération – l'*idf* – comme constant dans le temps ; il est donc calculé pour l'ensemble des contenus agrégés dans le temps. Par la suite, nous définissons des seuils dépendant des concepts pour décrire la dynamique symbolique de nos sources, ainsi, ce terme de pondération est naturellement sans conséquence pour déterminer si une catégorie de blogs mobilise de façon particulière un concept à un moment donné, par contre, il est crucial au moment de l'étape de catégorisation des blogs que nous décrivons dans la section suivante.

3. $\hat{w}(i) = \frac{\sum_t W_t(i,c)}{\sum_t \sum_{c=1}^{|\mathcal{W}|} W_t(i,c)} \cdot \log \frac{|\mathcal{B}|}{|\{j, \mathbf{W}(j,c) > 0\}|}$

tel-00464316, version 1 - 16 Mar 2010

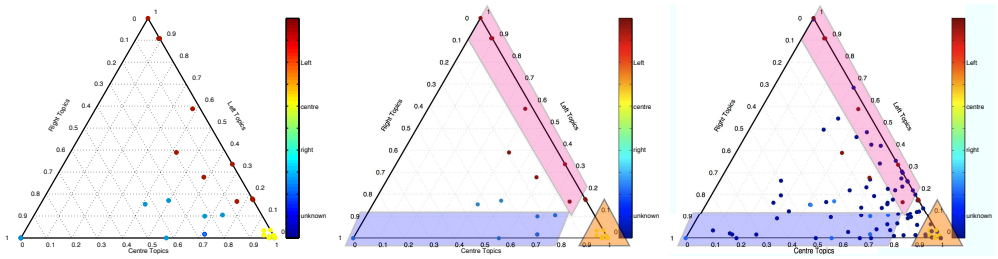


FIGURE 5.1: Diagramme ternaire représentant le vecteur des poids relatifs des concepts de la droite, du centre, et de la gauche, *à gauche*, projection des blogs pré-catégorisés (leur inclination politique est matérialisée par leur couleur), *au milieu*, définition des zones d'appartenance à chaque parti, *à droite*, projection de l'ensemble des blogs dans cet espace, les blogs appartenant aux zones pré-définies, sont catégorisés selon chacune des couleurs politiques.

Pour réaliser cette catégorisation, nous pouvons, compte tenu de la taille limitée du jeu de données, consulter chaque site et établir une première catégorisation "manuelle". Généralement, les blogueurs prenant ouvertement parti pour un camp politique placent des liens dans leur blogroll, vers des sites de campagne, ou les sites de soutien au candidat qu'ils supportent. Ces marqueurs nous ont permis d'attribuer sans ambiguïté une couleur politique : *droite*, *gauche*, *centre*, à près d'une trentaine de sites.

Une fois cette pré-catégorisation établie, nous définissons ensuite une série de concepts nous semblant caractéristiques des thématiques mises en avant par chaque parti au cours de la campagne[4]. Ces trois classes de concepts permettent d'attribuer à chaque source un vecteur tridimensionnel, dont les coordonnées correspondent à la moyenne de leur profil sémantique $\hat{w}_t(i)$ sur chaque classe de concepts. Ces vecteurs sont ensuite normalisés de manière à pouvoir définir chaque source comme un triplet censé indiquer la proportion de concepts de droite, du centre ou de gauche au sein de leurs publications.

Nous avons représenté l'ensemble des blogs catégorisés "manuellement" sur la figure 5.1. Le positionnement des blogs au sein du diagramme ternaire indique une focalisation des thématiques abordées en fonction de la couleur politique du blogueur, ce qui valide en partie notre hypothèse selon laquelle les classes de blogs que nous construisons ont un comportement d'édition relativement homogène. Ainsi, les blogs centristes (en jaune sur la figure) semblent aborder uniquement des thématiques centristes. Le profil des blogs de gauche et de droite est plus dispersé sur la dimension des thématiques centristes, on peut néanmoins aisément les caractériser par les thématiques qu'ils abordent peu, ainsi, d'après nos blogs pré-étiquettés, un blog de gauche emploiera systématiquement moins de 20% de

4. quelques exemples de concepts sélectionnés, pour le centre : *Ruralité, UDF, Francois Bayrou, budget de la recherche, dette publique...*, pour la gauche : *encadrement militaire, salaire minimum, capitalisme financier, Lionel Jospin, pacte présidentiel...*, et pour la droite : *réforme des retraites, service minimum, identité française, régimes spéciaux...*

tel-00464316, version 1 - 16 Mar 2010

thème de droite, et vice-versa pour les blogs de droite. Nous créons ainsi (diagramme central figure 5.1) trois zones censées caractériser l'inclination politique d'un blog : une première, matérialisée par un triangle orange pour les blogs centristes, correspond à des usages de concepts centristes à plus de 90%, le second, en bleu, correspond à l'ensemble des sources, dont les contenus publiés mobilisent moins de 10% de concepts de gauche et au moins 10% de concepts de droite. Le dernier ensemble, en rose, est le pendant du précédent en inversant gauche et droite.

Une fois ces "espaces de sensibilité politique" définis, nous projetons l'ensemble des blogs dans cet espace, et nous servons de leur distribution spatiale dans le diagramme ternaire pour les catégoriser. Un peu plus de 50 blogs sont ainsi étiquetés comme appartenant aux classes : gauche, droite ou centre, en fonction de leur appartenance aux trois zones pré-définies. Les trois classes de blogs sont de tailles équivalentes. Les blogs non catégorisés ne sont pas pris en compte par la suite, leur profil sémantique ne permettant pas de leur attribuer de façon claire une couleur politique.

Cette méthode de catégorisation peut être critiquée à cause des interventions manuelles qu'elle a requises, néanmoins, la contiguïté apparente des blogs qui affichent la même sensibilité politique au sein de cet espace offre quelque garantie vis-à-vis du résultat final. Nous avons maintenant à notre disposition 4 classes de sources notées $\mathcal{C}$: trois classes réunissent les blogs dont le profil d'activité est caractéristique d'un blog de droite, de gauche ou du centre, et une dernière classe de sources réunissant les 3 grands quotidiens nationaux [5].

5.2 Diagramme de corrélations intertemporelles

5.2.1 Contexte

Etant données ces différentes classes de sources, nous cherchons maintenant à savoir s'il existe certains groupes dont l'activité de publication est influencée par d'autres groupes. Le terme influence est ici employé dans une acception très large puisque nous cherchons simplement à détecter les motifs temporels systématiques du type : l'usage d'un concept par certains groupes induit ultérieurement l'usage de ce même concept par d'autre groupes.

Les corrélations que nous cherchons à exhiber peuvent aussi bien être la conséquence d'une relation causale directe (on peut penser à un "scoop" révélé dans la presse et qui donne lieu à des commentaires en écho dans la blogosphère par exemple) ou indirecte (un thème de campagne devenant particulièrement populaire dans un camp politique, est fréquemment mobilisé dans la communauté

5. Il est intéressant de noter que la projection de ces trois sources dans notre diagramme ternaire révèle qu'elles sont toutes trois très proches, et extérieures aux zones définies précédemment.

tel-00464316, version 1 - 16 Mar 2010

de blogueurs associée, les deux espaces (réels et virtuels) étant fortement recouvrants), ou simplement signaler un délai de réaction différent par rapport à une même cause extérieure.

5.2.2 Machines à états causaux

Nous nous appuyons sur la méthode des machines à états causaux introduite par Crutchfield and Young (1989). Avant d'appliquer cette méthode à nos données, nous allons en résumer brièvement le principe.

La méthode de Crutchfield et Young considère une dynamique symbolique discrète et vise à identifier des *classes d'équivalence* des états du système qui ont la même *probabilité de futur en probabilité* : on appelle ces classes d'équivalence des *états causaux*. Les états d'une même classe d'équivalence au temps t ont statistiquement les même futurs en probabilité, *i.e.* la même distribution de probabilités d'induire un futur état y au pas de temps suivant $t + 1$. Shalizi and Shalizi (2004) proposent un algorithme pour reconstruire l'ensemble des états causaux de la dynamique du système et simultanément trouver les probabilités de transition entre états causaux successifs. Cette reconstruction prend la forme d'une chaîne de Markov cachée (Shalizi, 2001b), le futur ne dépendant que de l'état présent du système.

Les états causaux représentent donc des ensembles d'états équivalents, au sens où ils ont le même futur en probabilité, tandis que les probabilités de transition entre ces états définissent des relations d'inférence systématiques (ou de "causalité") entre états.

	symboles émis														
signal	A	H	H	a	a	a	h	H	H	a	a	a	A	H	H...
								$\longrightarrow$*temps* $\longrightarrow$							

TABLE 5.1: Exemple de signal produit par une dynamique discrète symbolique sur un alphabet composé de quatre lettres : a, h, A et H)

De façon plus formelle l'algorithme `CSSR` reconstruit les états causaux d'une dynamique symbolique ainsi que l'ensemble des probabilités de transition associées. Les états réunis au sein d'une même classe peuvent être de longueurs variables allant jusqu'à $l_{\max}$ (c'est à dire qu'on peut considérer que le système peut être décrit par une succession d'"histoires" de longueurs variables). Par exemple si l'on considère le signal représenté tableau 5.1 qui définit une dynamique sur quatre symboles a, h, A et H, alors l'algorithme `CSSR`[6] paramétré pour intégrer des histoires de taille 1 ($l_{\max} = 1$) permet d'extraire les états causaux suivants : $S1 = \{a\}$, $S2 = \{H\}$, $S3 = \{A, h\}$, dont la séquence temporelle $\{S_t\}$ suit un

6. `CSSR` est librement disponible dans son implémentation originale http ://bactra.org/CSSR ainsi qu'à l'adresse http ://www.lsi.upc.es/%7Empadro/cssr.html dans une autre version.

tel-00464316, version 1 - 16 Mar 2010

processus markovien. Les états A et h ont le même futur en probabilité (émission d'un symbole H) et sont donc regroupés au sein du même état causal. Nous pouvons donc représenter la dynamique du signal comme une instance de la chaîne de Markov cachée sur les états causaux $S1$, $S2$ et $S3$ du système, comme représenté figure 5.2.

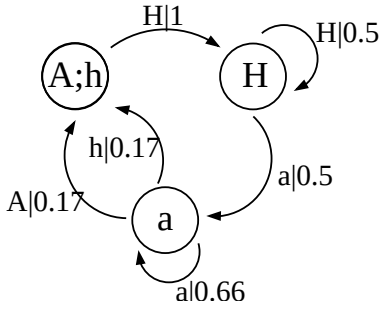


FIGURE 5.2: Machine à états causaux correspondant au signal présenté tableau 5.1 pour $l_{\max} = 1$. Les transitions entre états causaux sont accompagnées des symboles émis et leur probabilité. Ainsi, si le système est à un moment donné dans l'état causal S_1, à savoir (correspondant sans ambiguïté dans ce cas à l'état a), alors au pas de temps suivant, le système passera dans l'état causal S_3 avec une probabilité d'un tiers, en émettant avec la même probabilité le symbole A ou h.

5.2.3 Alphabet des concepts

Pour chaque concept c on décrit l'état de la blogosphère à un moment t, comme un vecteur binaire de dimension 4, qui représente l'ensemble des configurations d'apparition du terme c sur l'ensemble des classes de sources. Il existe ainsi 2^4 combinaisons possibles, qui forment l'*alphabet* (voir tableau 5.2) de notre dynamique symbolique. Les états en majuscule correspondent à la présence du terme dans la presse, les états signalés par une minuscule, signalent au contraire l'absence d'activité du concept dans la presse.

alphabet	a	b	c	d	e	f	g	h	A	B	C	D	E	F	G	H
$gauche$	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
$centre$	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
$droite$	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
$presse$	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
état causal	S_5	S_6	S_7	S_3	S_8	S_1	S_2	S_2	S_0	S_1	S_2	S_3	S_4	S_1	S_2	S_3

TABLE 5.2: Alphabet choisi, et états causaux associés ; par exemple, si le système se retrouve dans l'état f, cela signifie que seuls les groupes de droite et de gauche sont actifs vis-à-vis du concept envisagé.

tel-00464316, version 1 - 16 Mar 2010

5.2.4 Définition d'une dynamique symbolique

Nous appliquons la méthode de reconstruction des états causaux à la dynamique de production de contenus sur nos différentes catégories de sources. Nous considérons qu'il existe un "tissu d'influence" entre ces sources indépendant des concepts mobilisés, ainsi, nous faisons l'hypothèse que la dynamique observée sur chaque concept est une instance parmi d'autres du comportement global du système.

Il nous faut encore réduire l'activité de publication de l'ensemble des sources composant un même groupe en une seule série temporelle qui s'exprime dans l'alphabet que nous venons de décrire. Etant donnée une classe $C_j \in \mathcal{C}$ ($1 \geq j \geq 4$), on définit le profil sémantique de la classe comme la moyenne des profils individuels la composant, *i.e.* $\hat{w}_t(C_j) = \sum_{i \in C_j} \hat{w}_t(i)/|C_j|$[7]. Cette opération nous permet de construire la matrice temporelle tridimensionnelle M définissant la valeur du profil sémantique de la catégorie j pour un concept c à t : $M(j,c,t) = \hat{w}_t(C_j, c)$.

Parallèlement, nous construisons pour chaque concept c, un vecteur d'activité moyen à partir de l'ensemble des groupes de sources qui s'exprime comme la moyenne des profils sur chaque classe : $\hat{w}_t(c) = \frac{1}{4} \sum_{C_j \in \mathcal{C}} \hat{w}_t(C_j, c)$. La série temporelle $\hat{w}_t(c)$ permet donc de décrire l'évolution de l'usage moyen du concept c sur l'ensemble des classes de sources.

Cette série va nous permettre de définir un seuil au-dessus duquel l'activité d'une classe vis-à-vis d'un concept est "anormalement" élevée. Les vecteurs d'évolution propres à chaque concept étant relativement hétérogènes en dépit de la pondération que nous avons appliquée, nous définissons un seuil différent $\mu(c)$ pour chaque concept. On définit $\mu(c) = \langle \hat{w}(c) \rangle + 2\sigma(\hat{w}(c))$, le seuil est égal à la moyenne du profil sémantique du concept sur l'ensemble de la période d'observation à laquelle on rajoute deux fois l'écart type. Ce seuil permet de garantir qu'une classe de sources est "exceptionnellement" active vis-à-vis d'un concept, lorsque son usage dépasse largement l'usage moyen qui en est fait habituellement.

Ces seuils permettent de transformer la matrice M décrivant l'état du système sur l'ensemble des groupes sous une forme binaire M_0 en suivant la règle suivante pour chaque groupe j, concept c et temps t : $M_0(j,c,t) = 1$ ssi $M(j,c,t) \geq \mu(c)$, 0 sinon. Nous pouvons finalement construire les séries temporelles constituées à partir de notre alphabet en créant la matrice dont les lignes correspondent à nos 190 concepts, et les colonnes correspondent aux 181 jours de suivi de la blogosphère. Les éléments de cette matrice sont constitués par les lettres de notre alpha-

7. Les "normes" des vecteurs $\hat{w}_t(C_j)$ restent sensiblement comparables à travers les catégories et dans le temps, ce qui permet de comparer ces vecteurs par la suite. Une autre solution possible pour calculer le profil sémantique d'une catégorie de sources aurait consisté à calculer directement le tf·idf de l'ensemble des contenus agrégés produit par l'ensemble des sources d'un groupe (en concaténant simplement l'ensemble des textes). La solution que nous avons adoptée (qui consiste à établir une moyenne sur l'ensemble des profils sémantiques des sources) donne le même poids à chaque source, tandis que la seconde privilégierait les blogs produisant des billets plus longs.

tel-00464316, version 1 - 16 Mar 2010

bet qui permettent de décrire la distribution de l'activité autour d'un concept à un jour donné. Ainsi si le système est dans l'état "f" à t cela signifie que le concept c est fortement mobilisé par les blogs de la catégorie *Gauche* et *Droite* à t.

5.2.5 Resultats

Le paramètre l_{max} résulte d'un compromis entre la taille des données disponibles N et la taille de l'alphabet décrivant la dynamique symbolique k. Selon (Shalizi, 2001b), l_{max} doit rester inférieur au ratio $log(N)/log(k)$ pour que l'algorithme reste statistiquement fiable, aussi, non ne pouvons pas, compte tenu de nos données, espérer des résultats fiables pour des histoires de taille supérieure à 3. [8] Pratiquement, nous avons fixé l_{max} à 1. Nous avons privilégié une histoire courte essentiellement pour des raisons de clarté de présentation, notre objectif premier n'étant pas d'interpréter le résultat obtenu sur notre communauté de savoirs mais de présenter la méthodologie d'analyse [9].

L'algorithme `CSSR` fournit les états causaux, tels qu'indiqués dans la dernière ligne du tableau 5.2, ainsi que la chaîne de Markov entre états causaux dont on a représenté les principales transitions figure 5.3 (la composition des états causaux est également représentée sur la figure par les différentes combinaisons de cercles colorés). Neuf états causaux différents ont été construits par l'algorithme. Six d'entre eux correspondent à un seul état du système : S_5 : *aucune activation*, S_0, S_6, S_7 et S_8 : *activation d'une seule catégorie* (respectivement, Presse, Gauche, Centre et Droite) et enfin, S_4 : activation simultanée des catégories Presse et Droite. Les autres états causaux sont des assemblages de plusieurs états. S_1 par exemple regroupe l'ensemble des activations composées d'une combinaison des blogs de sensibilité de gauche avec au moins l'une des catégories Droite ou Presse. Cet état causal peut être interprété de la façon suivante : pourvu que les blogs de sensibilité centriste soient inactifs, si les blogs de sensibilité de gauche sont actifs ainsi qu'au moins une autre classe de source, alors, quel que soit le type de combinaison observée (états f, B, F), la même dynamique en probabilité sera observée le lendemain. Ces états sont causalement équivalents. Les probabilités de transition se lisent de la façon suivante : une fois le système dans l'état S_1, *i.e.* après l'émission d'un symbole f, B ou F, on observe de façon systématique l'émission le jour suivant d'un symbole b (seuls les blogs de la catégorie Gauche continuent d'utiliser le concept) dans 19% des cas, ou l'émission d'un symbole e (seuls les blogs de la catégorie Droite continuent d'utiliser le concept) dans 13 % des cas ou encore, l'émission d'un symbole f (blogs des catégories Gauche et Droite qui s'activent

8. dans notre cas, $N = 190 \times 181$ et $k = 16$

9. Théoriquement, augmenter la taille de l'histoire permet parfois de diminuer le nombre d'états causaux, dans notre cas, la dynamique restant très bruitée, le nombre d'états causaux obtenus augmente assez sensiblement avec la taille de l'histoire si bien que le diagramme d'influence obtenu devient rapidement trop compliqué pour être représenté graphiquement.

tel-00464316, version 1 - 16 Mar 2010

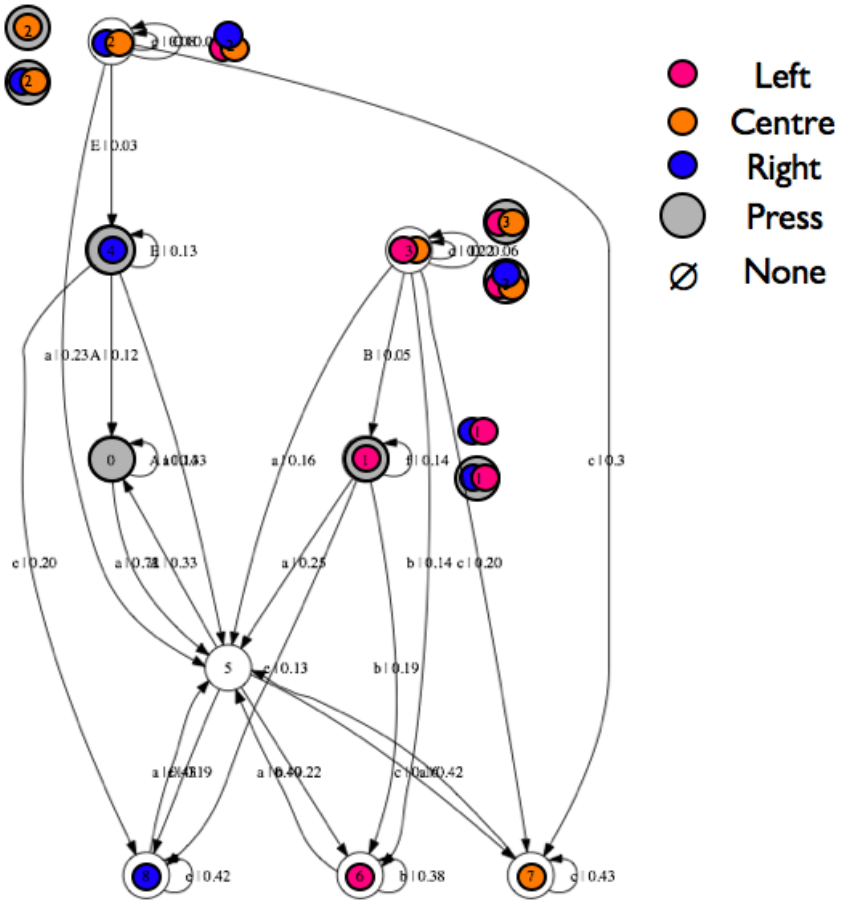


FIGURE 5.3: Machine à états causaux obtenue à partir de notre dynamique empirique. Seules les transitions les plus pertinentes ont été représentées (*i.e.* au moins les deux transitons les plus probables sortant de chaque état et toutes celles dont la probabilité est supérieure à 10%) ainsi que les symboles émis correspondants.

simultanément) dans 14% des cas.

Les deux autres états causaux composés de plusieurs histoires sont S_2 et S_3, ils réunissent des états du système pour lesquelles les blogs centristes sont toujours actifs, et accompagnés d'au moins une classe de source. Concernant S_2, presse mise à part, on observe que les états le composant sont essentiellement des états pour lesquelles la classe des blogs de droite est active. Concernant S_3, on observe essentiellement des états incluant une activations de la classe des blogs de gauche. Ces deux états causaux ont des futurs en probabilité extrêmement différents : ils donnent lieu à l'émission d'un symbole c (activation de la classe des blogs centristes uniquement) dans respectivement 30% et 20% des cas, à l'émission d'un symbole a (aucune classe de source n'est active) dans moins d'un quart des cas, pour le reste, les transitions représentées sur notre machine à états causaux montrent bien combien la dynamique à venir du système diffère selon l'état causal dans lequel il se trouve. S_5 et S_7 mis à part, l'état S_2 induira majoritairement des

tel-00464316, version 1 - 16 Mar 2010

transitions vers lui-même ou vers S_4, tandis que S_3 induira des transitions vers lui-même ou S_1.

Pour évaluer la significativité de notre reconstruction, nous avons appliqué `CSSR` à une sous-partie de notre jeu de données. En utilisant uniquement la moitié de nos concepts (choisis aléatoirement), et donc en divisant par deux la quantité de données initiale, nous obtenons les mêmes états causaux que ci-dessus. Les probabilités de transition entre états sont légérement modifiées mais, à ces quantités près, la représentation finale de la figure 5.3 reste la même. Cette stabilité illustre la robustesse de notre reconstruction.

À nouveau, notre objectif n'est pas de donner une interprétation complète de ce diagramme d'influence, mais de montrer la façon dont une information très riche (profils d'activité d'édition quotidienne de 120 blogs et de 3 journaux suivis pendant 6 mois) peut être résumée, réduite, en un diagramme synthétique qui représente les états du système causalement équivalents et les effets d'influence systématiques entre catégories de sources.

5.2.6 Perspectives

Cette méthode de reconstruction des diagrammes d'influence entre catégories de sources pourrait profiter de nombreux développements. Une première piste consisterait, en augmentant sensiblement le nombre de concepts, à en proposer une catégorisation susceptible de donner lieu à des motifs d'influence différents en fonction des catégories de concepts. Ainsi, on peut s'interroger sur la stabilité de notre diagramme selon le type de concepts employés. Est-ce que la dynamique d'influence des concepts ayant trait à des thématiques particulières (par exemple, des concepts exclusivement liés aux questions économiques ou écologiques) serait susceptible d'être reconstruite par un diagramme différent ? Certaines sources deviennent-elles plus "influentes", ou moins "influençables" selon le type de concept mobilisé ? Nous avons, dans ce travail, construit un diagramme qui agrège l'ensemble des influences observées sur l'ensemble de nos concepts. En ce sens, c'est une caractérisation des influences moyennes entre nos catégories de sources qui n'interdit pas à certaines catégories d'être dotées d'un pouvoir de prescription supérieur lorsque le débat se déplace dans leur domaine de spécialité.

Une autre amélioration consisterait à augmenter l'histoire des états possibles. Il existe sans doute des corrélations intertemporelles qui dépendent d'histoires de tailles supérieures à un jour ; on peut songer par exemple à une classe de sources devant systématiquement persister à discuter un concept pendant plusieurs jours d'affilée avant que d'autres ne lui "emboîtent le pas". La difficulté que nous risquons de rencontrer dans ce cas tient à la multiplication des états causaux construits par l'algorithme. Effectuer une seconde opération de catégorisation pourrait résoudre cette difficulté. Rassembler les états causaux les plus semblables (ayant des futurs "proches" en probabilité à défaut d'être stochastiquement iden-

tel-00464316, version 1 - 16 Mar 2010

tiques) pourrait permettre de conserver une représentation synthétique de notre diagramme d'influence tout en intégrant des histoires plus longues.

Une dernière piste d'approfondissement consisterait à tenter de corréler les diagrammes d'influence propres aux blogs avec les réseaux sociaux reliant les sources de contenus sous-jacentes. Notre méthode permet de retracer des influences systématiques d'un groupe de sources sur un autre. Le réseau social (comme le réseau de citations entre blogs) est-il un bon prédicteur des transitions entre états que nous avons exhibées ?

Enfin, dans une perspective plus large, nous pouvons nous interroger sur l'application de ce type de méthode à l'activité scientifique. Le même cadre pourrait être employé pour suivre les influences croisées induites par tel ou tel groupe de communautés scientifiques sur tel ou tel autre. Cette méthode permettrait ainsi de repérer les flux systématiques de concepts transitant entre communautés.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Résumé du chapitre:

Nous avons présenté une méthodologie générique de reconstruction des corrélations intertemporelles apparaissant entre les contenus produits par différentes catégories de sources. Nous nous sommes appuyés sur l'algorithme `CSSR` qui permet, à partir d'une dynamique discrète symbolique, de reconstruire l'ensemble des états causaux, définis comme des classes d'équivalence d'états du système ayant le même futur en probabilité. Cette méthode permet également de décrire la dynamique du système comme une chaîne de Markov cachée dont les éléments sont des états causaux, accompagnée de l'ensemble des probabilités de transitions entre états causaux et des symboles émis (les états du système) à chaque transition. La dynamique du système est ainsi reconstruite de façon statistiquement optimale.

Nous avons appliqué ce formalisme à la dynamique de production de contenus d'un ensemble de sources au sein de la blogosphère politique française ainsi que dans la presse. Notre objectif était de montrer qu'il est possible par la seule observation des profils sémantiques d'un ensemble de sources, d'exhiber les motifs d'influence systématiques entre ces sources afin de tenter de répondre aux questions concernant notamment la subordination d'une classe de sources à une autre (une forte activité observée autour d'un concept au sein des blogs est-elle susceptible d'être "reprise" le lendemain dans la presse ?) et plus largement d'exhiber des motifs d'activations de sources qui induisent de façon systématique un certain comportement du système.

Plus précisément, nous avons construit trois classes de blogs partageant les mêmes inclinations politiques (droite, gauche, centre), ainsi qu'une dernière classe regroupant un ensemble de quotidiens représentant l'activité des media durant les élections présidentielles françaises de 2007. L'examen des profils d'activité de chacune de ces classes nous a permis de définir une dynamique symbolique discrète sur les quatre classes déjà définies, *i.e.* la dynamique de l'ensemble du système peut être décrite comme une série temporelle discrète sous la forme d'un vecteur binaire dont les 4 éléments (correspondant aux 4 classes de source) valent 1 ou 0 selon que les différentes sources sont actives ou non vis-à-vis d'un concept à un moment donné. L'alphabet des états décrits par le système comprend 16 éléments. La totalité des transitions entre états est *a priori* possible, et la dynamique symbolique empirique observée en comprend un grand nombre. La reconstruction que nous en proposons permet de réduire cette dynamique sur l'ensemble de nos concepts à une chaîne de markov cachée que nous appelons *diagramme d'influence*, révélant les corrélations intertemporelles systématiques existant entre des profils d'activité de sources

réunies au sein d'états causaux. Le diagramme d'influence ainsi construit permet de faire différentes observations : d'une part repérer les états du système équivalents, *i.e.* appartenant à un même état causal - ces états sont équivalents d'un point de vue dynamique, ils induisent le même futur en probabilité, d'autre part observer à un niveau synthétique les dynamiques à l'œuvre dans le système de façon à représenter et quantifier les influences existantes entre groupes de sources.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

CHAPITRE 6

Du rôle de la topologie des réseaux sur la diffusion

Sommaire

6.1 Protocole de simulation 186
6.1.1 Protocole de simulation 186
6.1.2 Topologies de réseaux 188
6.2 Dynamiques de diffusion 191
6.2.1 Résultat des simulations 191
6.2.2 Interprétation 193
6.3 Rôle des règles de transmission 196
6.3.1 Directionalité de la transmission 196
6.3.2 Hypothèses de transmission réalistes 199
6.3.3 Modèles de transmission stylisés 201
6.3.4 Résultats des simulations 202

Les processus de diffusion de connaissance sont intimement conditionnés par la combinaison des comportements des agents (en situation d'incertitude, on peut s'attendre à différents comportements vis-à-vis d'une innovation (voir (Granovetter, 1978a) par exemple) et d'effets de structure inhérents au réseau social support des transmissions entre individus. Ce chapitre, qui s'appuie en grande partie sur un article (Cointet and Roth, 2007) publié en collaboration avec Camille Roth, vise, à travers un protocole simulatoire, à caractériser les paramètres topologiques susceptibles d'influencer la vitesse d'un processus de diffusion pour différentes hypothèses de transmission inter-individuelle.

Les modèles de diffusion d'innovation, de maladies ou de connaissance dans les réseaux sociaux ont suscité un intérêt accru ces dernières années. L'analyse des phénomènes de diffusion de connaissance remonte au milieu du XX^{ème} siècle et a initialement été abordée en sociologie, en économie ou en gestion (Coleman et al., 1957a; Rogers, 2003; Robertson, 1967; Rogers, 1976; Granovetter, 1978a; Burt, 1987; Valente, 1995).

Dès les premières études empiriques des processus de diffusion (Ryan and Gross, 1943; Menzel and Katz, 1955; Coleman et al., 1957b), une attention particulière a été portée à certaines propriétés du réseau social sous-jacent semblant liées à

la dynamique de diffusion (centralité des premiers innovateurs par exemple) tandis que Rogers (1976) insistait sur la nécessité de mettre en place des protocoles expérimentaux d'observation longitudinale des phénomènes de diffusion :

> "For network analysis to fulfill its potential, however, I feel we must improve the methods of data gathering and measurement (...). Longitudinal panel designs for networks analysis of diffusion process are also needed ; along with field experiments, they help secure the necessary data to illuminate the over-time process of diffusion." Rogers (1976)

L'analyse des grands réseaux d'interaction par des approches de type "physique statistique" a également insufflé un courant formalisateur dans l'appréhension des questions liées à la diffusion, dans un premier temps en s'appuyant sur la littérature en épidémiologie (Pastor-Satorras and Vespignani, 2001; Lloyd and May, 2001), avant de se pencher plus directement sur des processus plus spécifiques aux sciences sociales tels que la diffusion des rumeurs (Newman, 2002; Kempe et al., 2003), ou les dynamiques d'opinions (Axelrod, 1997b; Deffuant et al., 2002).

Néanmoins, même si certains auteurs ont insisté sur la nécessité de prendre en compte des topologies de réseau et des mécanismes de transmission réalistes à l'aide de mesures empiriques (Valente, 1996; Wu et al., 2004; Leskovec et al., 2007b), on peut s'interroger sur le degré d'adéquation des résultats analytiques ou simulatoires obtenus à partir des modèles de diffusion actuels par rapport aux phénomènes de diffusion "réels". Nous adresserons la question du réalisme de ces modèles en envisageant successivement les deux dimensions : topologie du réseau sous-jacent et mécanismes de transmission inter-individuelle.

Premièrement, la topologie de réseau retenue dans les études sur la diffusion est souvent basée sur des modèles classiques de morphogenèse de réseaux. Ainsi, les réseaux aléatoire dits à la Erdös-Renyi (que nous noterons ER par la suite) (Erdös and Rényi, 1959) ont été massivement employés (Barbour and Mollison, 1990; Wasserman and Faust, 1994; Zegura et al., 1996), tandis que d'autres ont privilégié des modèles plus simples ou plus géométriques (notamment fondés sur des grilles) (Ellison and Fudenberg, 1995; Deroian, 2002). Les modèles de type small-world (Watts and Strogatz, 1998) ont également suscité récemment un intérêt particulier (Cowan and Jonard, 2004b; Kuperman and Abramson, 2001), ainsi que d'autres modèles moins "classiques" (Bala and Goyal, 1998; Morris, 2000).

Mais le modèle de topologie qui a récemment attiré le plus d'attention, notamment dans le cadre de l'analyse des dynamiques de diffusion, est sans doute le réseau "sans échelle" ("scale-free") dont la distribution de degré suit une loi de puissance, caractéristique topologique dont les anciens modèles ne rendaient pas compte. Il existe différentes méthodes pour construire un réseau sans échelle. La plus populaire d'entre elles, introduite par Barabási and Albert (1999), s'appuie sur un processus de morphogenèse dans lequel de nouveaux nœuds sont ajou-

tel-00464316, version 1 - 16 Mar 2010

tés au réseau et sont connectés préférentiellement aux nœuds de fort degré. Un résultat en particulier a reçu un large écho dans les études ultérieures sur la diffusion : Pastor-Satorras and Vespignani (2001) ont montré que les réseaux dont la distribution de degré suit une loi de puissance ont un comportement radicalement différent d'un réseau aléatoire (ER) vis-à-vis d'un processus de diffusion. Plus précisément, ce travail prouve que le seuil épidémique [1] est nul sur un réseau sans échelle de taille infinie [2] alors qu'il est toujours positif dans le cas d'un réseau aléatoire de typer ER. Ainsi nombre d'études récentes sur la diffusion s'appuient sur des réseaux de type sans-échelle (Amblard and Deffuant, 2004; Ganesh et al., 2005; Crépey et al., 2006).

Au delà du choix d'une typologie ou d'un modèle de morphogenèse, il est important de noter que l'approche classique des processus de diffusion supportés par des réseaux, autant du point de vue des études simulatoires qu'analytiques, consiste à travailler à partir de *réseaux stylisés*. Les modèles de diffusion de connaissance ont rarement été simulés sur la base de réseaux réels [3].

Deuxièmement, les hypothèses employées quant au mécanisme de transmission même si elles paraissent plausibles n'ont que très rarement donné lieu à un contrôle empirique. Comme le mentionnent Leskovec et al. (2007b)

> "[while former] models address the question of maximizing the spread of influence in a network, they are based on assumed rather than measured influence effects."

Généralement, on postule un modèle de comportement individuel stylisé à partir de modèles "psychologiques" (Granovetter, 1978b; Goldenberg et al., 2001)), de modèles économiques (Ellison and Fudenberg, 1995; Morris, 2000), ou de modèles de connaissance visant à suivre l'évolution de profils d'opinions continus ou discrets, prenant la forme de vecteurs unidimensionnels (Axelrod, 1997a; Deroian, 2002; Deffuant et al., 2002) ou multidimensionnels (Gilbert et al., 2001; Cowan and Jonard, 2004b; Klemm et al., 2005).

Notre objectif est donc de caractériser la façon dont la dynamique de diffusion sur un réseau est modifiée en fonction du type de modèle stylisé retenu, autant au niveau de la topologie du réseau sous-jacent que des hypothèses de transmission.

1. Le seuil épidémique désigne le ratio d'agents infectés en-dessous duquel une épidémie suivant le modèle SIS (les agents du système peuvent se être dans trois états : Susceptible, Infecté ou Sain) s'interrompt

2. ce résultat est néanmoins limité à des réseaux de taille infinie, et ne tient plus pour des modèles de type SIR (Susceptible, Infected, Recovered).(May and Lloyd, 2001; Eguiluz and Klemm, 2002)

3. Wang et al. (2003) ont comparé les prédictions de leur modèle de diffusion à celui de Pastor-Satorras and Vespignani (2001) sur différentes topologies, dont un réseau informatique réel, mais sans s'interroger sur la façon dont leur modèle appliqué à différentes topologies pouvait modifier la dynamique de diffusion. De la même façon, Wu et al. (2004) ont simulé un processus de diffusion sur un réseau d'e-mail réel ; mais à nouveau sans chercher à estimer la façon dont leurs résultats seraient modifiés avec d'autres hypothèses de topologie.

tel-00464316, version 1 - 16 Mar 2010

Nous souhaitons donc comparer les résultats obtenus sur des modèles stylisés avec les résultats observés sur des réseaux ou des mécanismes de transmission réels. Malheureusement, nous manquons de données permettant de mesurer simultanément les comportements de transmission et la topologie du réseau sous-jacent[4]. Nous découplerons donc notre analyse en deux parties : d'une part, nous examinerons la façon dont un réseau réel et des réseaux stylisés dont la topologie est dépréciée à partir de ce dernier se comportent vis-à-vis d'un processus de diffusion, d'autre part, nous comparons un comportement de transmission réel et ses modèles stylisés vis-à-vis d'une dynamique de diffusion.

6.1 Protocole de simulation

Notre objectif étant de comparer les différents modèles de réseau et modèles de transmission à leur instance réelle, nous cherchons à définir un protocole aussi basique que possible.

6.1.1 Protocole de simulation

Nous considérons un ensemble de N agents et une entité d'information atomique binaire : l'état du système au temps t est décrit à l'aide du vecteur $c(t) \in \{0,1\}^N$, tel que $c_i(t) = 1$ si l'agent i a connaissance de cette entité à t — on dira alors qu'il est informé — $c_i(t) = 0$ sinon. Le processus est supposé strictement croissant ; une fois informés, les agents ne peuvent donc pas "oublier".

En terme épidémiologique, notre protocole est donc très proche d'un modèle de type "SI" (Sain, Infecté) (Hethcote, 2000). Il se distingue néanmoins des modèles classiques pour deux raisons. Premièrement nous envisageons un processus asynchrone de sélection des nœuds, de sorte qu'un seul nœud peut voir son état changer à chaque pas de temps, tandis que la plupart des modèles existants "mettent à jour" simultanément l'ensemble des états de nœuds du réseau en fonction des interactions qu'ils ont eu avec leur environnement de façon synchrone. Deuxièmement, on envisage un processus également asyncrhone d'interactions entre les nœuds : à chaque pas de temps l'agent susceptible de changer d'état interagit avec *un seul* de ses voisins. Chaque interaction met donc en jeu uniquement un couple de nœuds alors que les modèles classiques supposent généralement que l'état d'un nœud est modifié en fonction de l'état de l'ensemble de son voisinage. Nous décomposons en quelque sorte le processus de diffusion pour ne considérer à chaque pas de temps qu'un seul événement primitif : une interaction entre deux agents susceptible de modifier l'état d'un des agents. Les conditions initiales sont fixées de telle façon qu'une proportion λ d'agents est initialement "informée" (i.e. λN agents).

4. Le chapitre 7 propose une première avancée dans cette direction.

tel-00464316, version 1 - 16 Mar 2010

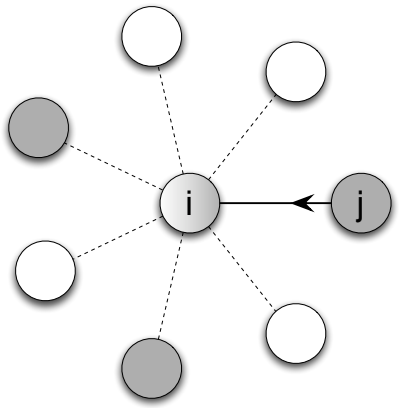


FIGURE 6.1: À chaque pas de temps, un nœud i est choisi aléatoirement, il interagit avec un de ses voisins j. Il y a alors transmission si j est informé, et que i ne l'est pas. Dans ce schéma les nœuds informés sont grisés.

Nous pouvons décrire notre processus de diffusion de la façon suivante. À chaque pas de temps, une interaction entre un agent i, choisi aléatoirement, et un de ses voisins (choisi aléatoirement parmi les voisins de i) induit une transmission d'information de j vers i si et seulement si j est informé. La séquence d'événements (représentée schématiquement figure 6.1) se déroulant à chaque pas de temps est donc la suivante :

1. un agent i est choisi aléatoirement,
2. un agent j est choisi aléatoirement parmi les voisins de i ($j \in \mathcal{V}_i$),
3. $c_j(t) = 1 \Rightarrow c_i(t) = 1$.

Notre modèle de transmission diffère de la plupart des modèles classiques car les agents entrent *a priori* le même nombre de fois en interaction avec leur voisins, en tant que récepteur potentiel d'une transmission, alors que la plupart des modèles existants confèrent une importance particulière aux nœuds de fort degré dont "l'activité" vis-à-vis de la diffusion est proportionnelle à leur degré.

Notre processus de transmission est bien asymétrique, nous reviendrons sur les conséquences de cette asymétrie ultérieurement. Ce type de processus de transmission appartient à la famille des modèles de diffusion de rumeurs (appelés "gossip-based models" (Kempe and Kleinberg, 2002)).

Nous procédons à une série de simulations dans des contextes expérimentaux distincts et mesurons l'évolution du ratio d'agents "informés" ρ au cours du temps où $\rho(t) = \frac{1}{N}\sum_{i=1}^{N} c_i(t)$, en utilisant pour chaque simulation un ensemble aléatoire d'agents initialement informés ($\rho(0) = \lambda$).

tel-00464316, version 1 - 16 Mar 2010

6.1.2 Topologies de réseaux

On considère deux réseaux sociaux réels. Notre premier réseau, appelé `Medline` par la suite, est un réseau constitué de collaborations entre des embryologistes travaillant sur le poisson zébré (zebrafish) extrait à partir d'une grande base de données de publications scientifiques [5]. Nous considérons uniquement la plus grande composante connexe composée de $6\ 453$ agents, liés par $67\ 392$ liens, issus de $2\ 476$ publications [6]. Sa distribution de degré (voir figure 6.2) est hétérogène comme il est classique de l'observer dans les réseaux sociaux (Barabási and Albert, 1999). Le second réseau a été collecté sur le site de "They rule" [7], ce réseau est un réseau *d'interlock*, il est composé de directeurs siégeant à des conseils d'administration de grandes firmes ou de grandes institutions américaines. Deux directeurs sont liés s'ils siègent dans un même conseil d'administration. La plus grande composante connexe de ce réseau est composée de $4\ 656$ nœuds pour $76\ 600$ liens extraits d'une base de données formée de 516 conseils d'administration.

Notre stratégie consiste à déprécier progressivement les caractéristiques topologiques de nos réseaux originaux afin de produire une série de réseaux stylisés. Nous comparerons ensuite les profils des dynamiques de diffusion sur ces différents réseaux afin de tenter d'apprécier le rôle des caractéristiques topologiques du réseau réel sur la dynamique de diffusion. Nous commençons par distinguer les 4 types de réseaux suivants :

- *Réseau réel (RN - Real Network)* — Le réseau réel non déprécié sera noté respectivement RN1 et RN2 selon qu'on se réfère au réseau de collaboration scientifique ou au réseau d'interlock.
- *Sans-échelle (SF - Scale-Free)* — Les réseaux SF (SF1 et SF2) sont construits à partir des réseaux réels (RN1 et RN2) en appliquant le modèle configurationnel (Molloy and Reed, 1995) qui consiste à connecter aléatoirement les demi-liens du réseau de façon à obtenir un réseau aléatoire préservant la distribution des degrés originale.
- *Erdös-Rènyi (ER)* — Les réseaux aléatoires (Erdös and Rényi, 1959) ER1 et ER2 conservent uniquement la densité des réseaux réels de départ, ainsi que le nombre d'agents N. Contrairement à SF et RN, les distributions de degré peuvent être approchées par une loi de Poisson (Bollobás, 1985), voir figure 6.2. Le nombre d'agents N et de liens M sont identiques au réseau réel.
- *Réseau complet (CN - Complete Network)* — Les réseaux complets CN1 et CN2 partagent uniquement le nombre d'agents avec leur alter-ego réel. Dans cette topologie chaque agent est connecté à l'ensemble des autres agents. La den-

5. la platforme `Medline` http ://www.ncbi.nlm.nih.gov/pubmed/ est spécialisée dans la littérature biomédicale

6. nous n'avons considéré que les articles dont l'abstract ou le titre mentionnait le terme "zebrafish" sur la période 2000–2004

7. http ://www.theyrule.net

tel-00464316, version 1 - 16 Mar 2010

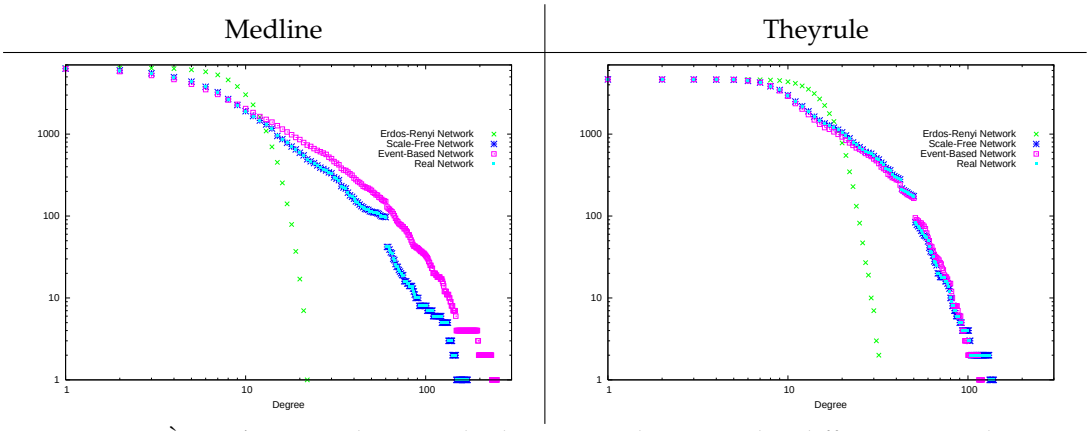


FIGURE 6.2: *À gauche :* Distributions de degré cumulées pour les différentes topologies envisagées (*abscisses* : degré k, *ordonnées* : $\mathcal{N}(k) = \sum_{k'=k}^{\infty} N(k')$), *à droite* réseau `Medline`, *à droite* réseau Theyrule).

sité n'est naturellement pas préservée contrairement aux réseaux précédents, et le nombre de liens dans le réseau vaut $N(N-1)/2$.

La structure de clustering de nos réseaux peut s'avérer un paramètre important pour étudier le processus de diffusion d'information (Bala and Goyal, 1998; Morris, 2000). La définition classique du clustering (voir chapitre 3) est une mesure liée au ratio du nombre de triangles sur le nombre de fourches. Dans sa version locale, on définit pour chaque agent son clustering comme le nombre moyen de ses voisins qui sont également voisins l'un de l'autre : $c_3(i) = \frac{[\text{nombre de paires de voisins de } i \text{ connectés }]}{k_i \cdot (k_i - 1)/2}$ où k_i désigne le nombre de voisins du nœud i. Le coefficient de clustering $\langle c_3 \rangle$ du réseau est alors une moyenne des clusterings calculés sur l'ensemble des agent. Comme on l'a déjà constaté dans la partie précédente, les réseaux réels sont généralement dotés de coefficients de clustering $\langle c_3 \rangle$ très grands comparés aux valeurs attendues sur un réseau aléatoire (et ce quelle que soit sa distribution de degré : de type SF ou ER (Boguna and Pastor-Satorras, 2002; Newman and Park, 2003b)). Dans notre cas, nous avons mesuré les valeurs de clustering suivantes : pour le réseau de collaborations, $\langle c_3 \rangle$ vaut .827 tandis que son réseau déprécié de type SF a un coefficient de clustering de l'ordre de .00539. Une même disparité entre réseau réel et sans-échelle est observée dans le réseau d'interlock dont le coefficient de clustering du réseau réel ($\langle c_3 \rangle = .889$) est supérieur de deux ordres de grandeur à celui du réseau SF ($\langle c_3 \rangle = .00395$).

Nous souhaiterions également intégrer dans notre panel de réseaux stylisés, un réseau capable de préserver ce fort taux de clustering. Pour ce faire, nous utilisons la structure événementielle sous-jacente de nos deux réseaux. En effet, les deux réseaux réels considérés sont en fait des réseaux d'affiliation de type biparti (les scientifiques sont liés aux articles qu'ils publient, les directeurs, aux conseils d'administration auxquels ils siègent) dont la projection sur la dimension sociale produit les réseaux réels pertinents pour notre étude. Nous proposons donc d'uti-

tel-00464316, version 1 - 16 Mar 2010

liser un modèle de morphogenèse "à base d'événements" qui s'appuie sur cette description bipartie sous-jacente des réseaux. À partir de la description originale des données de nos deux réseaux, qui est constituée d'événements regroupant des agents participant à une même activité, nous calculons d'une part la distribution des tailles des événements (distribution du nombre d'agents engagés dans chaque événement) et d'autre part la distribution du nombre d'événements auxquels chaque agent prend part. Ces deux distributions sont parfois appelées distributions à droite et à gauche du réseau biparti considéré. Le modèle de reconstruction est très simple et revient à un modèle configurationnel sur le graphe biparti original. Chaque agent conserve un nombre de liens sortants fidèle à la distribution de départ du nombre d'événements par agent. Ces liens sont reliés de façon aléatoire à l'ensemble des événements dont la taille respecte les distributions originales (distributions du nombre d'auteurs par article ou de directeurs par conseil d'administration). Une fois distribué l'ensemble des liens sortants des agents vers les événements, nous projetons le réseau biparti ainsi formé (deux agents se retrouvent liés s'ils participent au même événement), afin de construire un réseau de type EB (Event-Based).

Le réseau EB est encore plus proche de RN que SF au sens où il conserve un plus grand nombre de propriétés topologiques. Nous avons représenté figure 6.2 les distributions de degré pour l'ensemble de nos topologies de réseau. On observe que le réseau EB permet de reconstruire relativement fidèlement la queue de la distribution de degré. Cette propriété est une conséquence directe du processus de construction adopté (Guillaume and Latapy, 2004; Newman et al., 2001). La structure de clustering (voir tableau 6.1) est également reconstruite de façon satisfaisante. En effet, le coefficient de clustering est fortement influencé par l'opération de projection qui aboutit en l'addition d'autant de cliques [8] que d'événements (Newman et al., 2001). On s'attend donc à observer un large nombre de structures triangulaires dans ces réseaux.

Une mesure de clustering à plus longue distance, appelée clustering d'ordre 4 a récemment été introduite (Lind et al., 2005). Elle consiste à calculer la proportion moyenne de voisins communs parmi les voisins d'un nœud i :

$$c_4(i) = \frac{\sum_{i_1=1}^{k_i} \sum_{i_2=i_1+1}^{k_i} \kappa_{i_1,i_2}}{\sum_{i_1=1}^{k_i} \sum_{i_2=i_1+1}^{k_i} [(k_{i_1} - \eta i_1, i_2)(k_{i_2} - \eta_{i_1,i_2}) + \kappa_{i_1,i_2}]} \tag{6.1}$$

où κ_{j_1,j_2} désigne le nombre de nœuds que les voisins j_1 & j_2 de i ont en commun (i exclu). $\eta_{j_1,j_2} = 1 + \kappa_{j_1,j_2} + \theta_{j_1,j_2}$ où θ_{j_1,j_2} vaut 1 si j_1 et j_2 sont connectés, 0 sinon.

EB1 permet de reconstruire approximativement le clustering du réseau réel (RN1) mais semble mis en défaut pour reconstruire efficacement $\langle c_4 \rangle$ (un ordre de grandeur sépare les valeurs de EB1 et de RN1). Par contre, EB2 semble plus

8. une clique est un sous-graphe complet, *i.e.* un sous-graphe dont les nœuds sont tous connectés les uns aux autres.

tel-00464316, version 1 - 16 Mar 2010

performant par rapport à cet indice : $\langle c_4 \rangle$ vaut .280 ce qui reste relativement proche de la valeur originale de $.415$ pour le réseau réel : RN2. Les valeurs de l'ensemble des caractérisitiques topologiques de nos réseaux sont réunies tableau 6.1.

Pour résumer, nous considérerons 5 topologies de réseau différentes : (i) le réseau réel RN, (ii) un réseau à structure d'événements sous-jacente EB, (iii) un réseau sans-échelle SF, (iv) un réseau aléatoire de type Erdös-Rényi ER, (v) un réseau complet CN.

	RN1	**SF1**	**ER1**	**CN1**	**EB1**
N	6453				
M	$6.74 \cdot 10^4$			$2.08 \cdot 10^7$	$7.62 \cdot 10^4$
d	.00162			1	.00183
dist. degré	power-law tail		Poisson	—	power-law tail
$\langle c_3 \rangle$	.827	.00539	.00199	1	.753
$\langle c_4 \rangle$	.400	.000260	.000158	1	.0694

	RN2	**SF2**	**ER2**	**CN2**	**EB2**
N	4656				
M	$7.66 \cdot 10^4$			$2.17 \cdot 10^7$	$7.68 \cdot 10^4$
d	.0035			1	.0035
dist. degré	power-law tail		Poisson	—	power-law tail
$\langle c_3 \rangle$	.889	.00395	.00403	1	.897
$\langle c_4 \rangle$	.398	.000261	.000217	1	.268

TABLE 6.1: Principales caractéristiques topologiques des différents réseaux stylisés dérivés des réseaux réels RN1 et RN2 en termes de : nombre d'agents N, nombre de liens M, densité d, formes des distributions de degré et coefficients de clustering $\langle c_3 \rangle$ & $\langle c_4 \rangle$ (valeurs moyennées sur 1000 réseaux pour les modèles de type SF, ER & EB).

6.2 Dynamiques de diffusion

6.2.1 Résultat des simulations

Pour chaque type de topologie nous avons lancé $1\,000$ simulations en suivant le protocole décrit section 6.1.1. Pour chacune des instances de simulation, nous sélectionnons la plus grande composante connexe, qui dans tous les cas, couvre au moins 99.9% de l'ensemble des nœuds. Ainsi l'état final de la simulation tend naturellement vers $\rho(\infty) = 1$, *i.e.* l'ensemble des nœuds est informé *in fine*. La

tel-00464316, version 1 - 16 Mar 2010

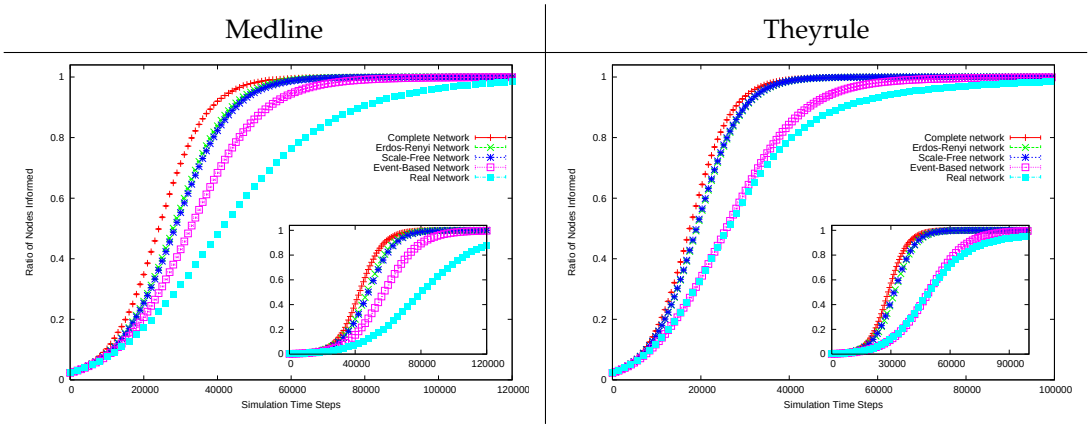


FIGURE 6.3: $\rho(t)$ simulé pour les réseaux complets (CN), Erdös-Rényi (ER), sans-échelle (SF), à base d'événements (EB) et réels (RN), $\lambda = 0.02$ et $\lambda = 0.002$ (en encart), les barres d'erreurs correspondent aux intervalles de confiance à 99% . Topologies extraites des réseaux de collaboration (*à gauche*) et du réseau d'interlock (*à droite*).

simulation est initialisée avec une fraction λ de nœuds informés.

La figure 6.3 récapitule l'ensemble des profils d'évolution temporelle de ρ sur l'ensemble des topologies de réseau pour $\lambda = 0.02$ et $\lambda = 0.002$. On observe que plus les réseaux ont une topologie "semblable" à celle du réseau réel, et plus la dynamique est lente. Le réseau complet est celui sur lequel la diffusion est la plus rapide. De façon plus surprenante, les réseaux ER et SF semblent se comporter de façon identique vis-à-vis de $\rho(t)$. Le comportement de EB est plus lent que les autres topologies stylisées et offre la meilleure approximation de la dynamique de diffusion sur RN. Néanmoins, les résultats sont contrastés selon que l'on s'intéresse au réseau de collaboration ou d'interlock : EB2 semble reconstruire de façon satisfaisante la dynamique de diffusion de RN2, tandis que EB1 diverge encore de façon significative de RN1.

Nos résultats ne semblent pas affectés par le paramètre λ fixant la proportion initiale d'agents informés (cf. encart figure 6.3), qualitativement nous observons simplement un ralentissement général de la dynamique de diffusion, sans que "l'ordre" entre les différentes topologies ne soit affecté.

Il semble donc que la focalisation sur les réseaux stylisés de type sans-échelle (Eguiluz and Klemm, 2002; Boguna and Pastor-Satorras, 2002; May and Lloyd, 2001) soit pertinente, dans les limites du processus de transmission mis en œuvre. Dans notre cas, le réseau SF a une dynamique de diffusion similaire à celle du réseau ER, beaucoup plus rapide que la dynamique que l'on observe sur le réseau réel.

Par contre, EB suggère une importance non négligeable de la structure de clustering dans le processus de diffusion. La lenteur du processus de diffusion sur RN pourrait être due à sa structure de communautés (ou encore à sa structure modulaire) sous-jacente (Girvan and Newman, 2002; Clauset et al., 2004; Blon-

tel-00464316, version 1 - 16 Mar 2010

del et al., 2008). De nombreuses études sur la diffusion des innovations ont ainsi souligné qu'une structure de réseau favorisant les interactions dans un cercle social proche plutôt qu'avec des agents distants est susceptible de ralentir la dynamique de diffusion (Granovetter, 1973; Bala and Goyal, 1998) — les clusters d'individus densément inter-connectés produisant mécaniquement de la redondance dans la distribution de la connaissance (i.e. l'ensemble des agents appartenant à un même groupe densément connecté sont rapidement alignés dans le même état, mais risquent également d'être dans un état d'isolement par rapport à d'autres groupes plus distants). Comme l'écrit Granovetter (1973),

> "if one tells a rumor to all his close friends, and they do likewise, many will hear the rumor a second and third time, since those linked by strong ties tend to share friends."

Ainsi, nombres des liens présents dans le réseau réel sont redondants vis-à-vis du processus de diffusion, ces redondances étant dommageables pour la diffusion de l'information sur l'ensemble des agents du réseau.

Une étude antérieure de Bala and Goyal (1998) avait montré que des voisinages recouvrants pouvaient ralentir le phénomène de diffusion, Eguiluz and Klemm (2002) ont également remarqué que le seuil épidémique était diminué dans les réseaux fortement clusterisés comparés à des réseaux aléatoires tandis qu'Onnela et al. (2007) ont observé l'alternance de plateaux et d'augmentations rapides du nombre d'agents infectés en simulant un processus de diffusion sur un réseau pondéré dont les liens représentent des contacts téléphoniques réels. Les auteurs interprètent ces plateaux comme des épisodes durant lesquels l'élément diffusant est "emprisonné" par des communautés locales. Enfin de façon connexe, Gallos et al. (2007) ont observé analytiquement et à l'aide de simulations une relation linéaire entre l'exposant fractal de modularité d'un réseau biologique et l'exposant caractéristique d'une marche aléatoire dans le réseau (qui traduit typiquement le temps caractéristique d'un processus de diffusion). Cette étude montre donc que selon l'exposant de modularité (déterminé pour un réseau fractal), on observe une "sub-diffusion" pour des réseaux très modulaires, et une "super-diffusion" pour des réseaux très peu modulaires.

6.2.2 Interprétation

Afin de vérifier notre hypothèse nous avons calculé un indice de corrélation entre agents voisins dans le réseau :

$$\nu(t) = \frac{1}{n}\Big[\sum_{i=1}^{n} c_i(t) \sum_{j \in \mathcal{V}_i} \frac{c_j(t)}{k_i}\Big]$$

ν mesure la somme des proportions de nœuds déjà informés dans le voisinage des nœuds informés. Pour un même taux de diffusion $\rho(t)$, un indice de corrélation $\nu(t)$ plus élevé indique que la diffusion risque d'être ralentie à cause d'un

tel-00464316, version 1 - 16 Mar 2010

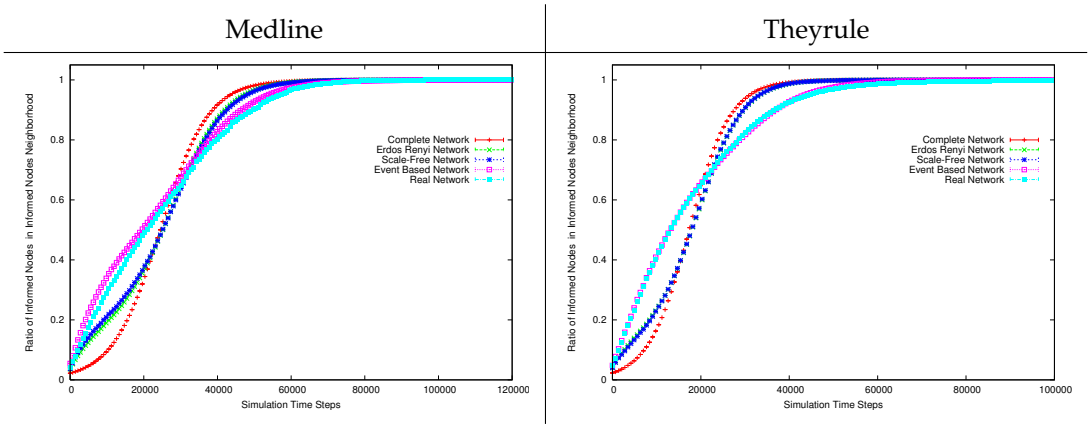


FIGURE 6.4: Evolution de $\nu(t)$ pour les réseaux complets (CN), Erdös-Rényi (ER), sans-échelle (SF), à base d'événements (EB) et réels (RN), $\lambda = 0.02$. Topologies extraites des réseaux de collaboration (*à gauche*) et du réseau d'interlock (*à droite*).

faible ratio d'agents non informés susceptible de propager l'information dans le voisinage des nœuds informés. Nous avons tracé l'évolution de ν durant le phénomène de diffusion sur nos deux réseaux et pour l'ensemble de nos topologies figure 6.4. L'indice de corrélation est croissant tendant vers 1 lorsque l'ensemble des nœuds est informé. Le profil d'évolution de l'indice de corrélation est extrêmement différent selon la topologie considérée. Pour les deux réseaux étudiés, la proportion d'agents susceptibles d'être infectés croît rapidement dans le cas des réseaux de type EB et RN, contrairement aux topologies de type SF et ER et a fortiori CN pour lesquels la croissance est beaucoup plus lente. Ces courbes doivent théoriquement être comparées pour un même taux de diffusion ρ. Les différences de comportement observées entre les différentes topologies vis-à-vis de ν seraient même encore plus flagrantes si les délais dûs aux écarts de vitesse de diffusion entre topologies étaient pris en compte.

L'ensemble des ces observations nous permet de mieux comprendre pourquoi EB est plus lent que SF, et pourquoi RN, dont la structure complexe de communautés n'est pas parfaitement reproduite (cf stastiques de c_4) par EB, est encore plus lent. En effet, les valeurs de $\langle c_4 \rangle$ mesurées semblent donner une indication sur la qualité de la reconstruction de la structure de communautés sous-jacente : lorsque $\langle c_4 \rangle$ est inférieur d'un ordre de grandeur dans EB1 par rapport à RN1, les vitesses de diffusion différent significativement ; alors que pour des valeurs de $\langle c_4 \rangle$ comparables, comme dans le cas de EB2 et RN2, les vitesses de diffusion sont comparables.

Il est probable que "la nature sans-échelle [des réseaux réels] ne doive pas être négligée dans l'estimation des seuils d'immunisation et épidémiques des réseaux réels" [9] (Pastor-Satorras and Vespignani, 2001), mais la seule prise en compte de

9. "the SF nature cannot be neglected in the practical estimates of epidemic and immunization

tel-00464316, version 1 - 16 Mar 2010

cette propriété pourrait s'avérer insuffisante : tous les réseaux sans-échelles ne sont pas équivalents (May and Lloyd, 2001; Boguna and Pastor-Satorras, 2002; Eguiluz and Klemm, 2002) et dans notre cas, la réseau sans-échelle le plus simple exhibe en réalité la même dynamique que celle observée sur une topologie ER, même si d'autres hypothèses de transmissions peuvent induire une différence de comportement entre les réseaux ER et SF (Dorogovtsev and Mendes, 2003; Barthélémy et al., 2005). Dans notre cas, ces différences sont minimes, tandis que les résultats constatés sur EB penchent plutôt pour un rôle déterminant de la structure de communautés.

On peut également s'interroger légitimement sur les différences observées entre nos deux réseaux vis-à-vis de la qualité de la reconstruction de type EB. Non seulement, EB1 a un coefficient de clustering d'ordre 4 largement inférieur à celui de RN1, mais on constate également figure 6.2 que la distribution de degré d'EB1 est légèrement décalée vers la droite par rapport à celle de RN1. Dans le cas du réseau d'interlock, on ne constate pas une telle divergence de EB2 par rapport à RN2, les distributions de degré de ces derniers étant très proches, et le coefficient de clustering d'ordre 4 étant relativement identique.

Cette observation est en fait cohérente avec les résultats obtenus par Newman et al. (2001, 2002a) sur l'évaluation des paramètres topologiques théoriques de réseaux aléatoires à structure bipartie sous-jacente respectant les distributions de degré du réseau biparti original. Plus précisément, les valeurs théoriques de clustering et de degré moyen d'un réseau aléatoire sont estimées analytiquement à partir des seules distributions de degré du réseau biparti sous-jacent (liant les agents aux événements auxquels ils prennent part). Ces valeurs théoriques sont ensuite comparées aux valeurs observées dans les réseaux réels ; le clustering est systématiquement sous-estimé tandis que le degré moyen est systématiquement surestimé dans un réseau d'acteurs (deux acteurs sont liés s'ils ont joué dans un même film), et deux réseaux de collaboration scientifique, tandis que le réseau d'interlock (Fortune 1000 - du même type que notre réseau Theyrule mais de taille plus importante) a des valeurs théoriques et réelles quasiment parfaitement concordantes. Comme le notent Newman et al. (2002b), " the discrepancy between theory and experiment may be highlighting real sociological phenomena in the networks studied ". Mais si l'on conçoit aisément que le réseau de collaboration scientifique puisse être fortement structuré par les institutions sous-jacentes, ou par des processus transitifs locaux [10], on comprend *a priori* moins bien pourquoi les réseaux d'interlock sont si fidèlement reconstruits par un modèle structurel (Robins and

thresholds in real networks"

10. Ces processus transitifs locaux sont générateurs de liens *répétés* lors de la production de nouveaux articles auxquels participent d'anciens collaborateurs. On s'attend également, dans un réseau de collaboration scientifique, à retrouver des ensembles d'individus reliés les uns aux autres en fonctions d'intérêts communs et formant par exemple des communautés d'experts, ces regroupements donnent lieu à une structuration particulière du réseau de collaboration.

tel-00464316, version 1 - 16 Mar 2010

Alexander, 2004; Uzzi et al., 2005) aussi simple fondé sur les seules distributions de degré du réseau biparti. Une explication possible (différente de celle avancée par Newman et al. (2001) qui insistent sur le fait que les directeurs siégeant dans de nombreux conseils d'administration ont également tendance à co-siéger avec des directeurs également impliqués dans plusieurs conseils d'administration) consiste à expliquer le caractère "typiquement aléatoire" des réseaux d'interlocks, par les règles entourant la composition des conseils d'administration. Ainsi, aux Etats-Unis (mais d'autre procédures du même type sont appliquées dans d'autres pays), le Clayton act (1914) interdit notamment à un individu de siéger dans les conseils d'administration de deux compagnies en situation de concurrence. Ce mode de régulation s'inscrit dans le dispositif légal anti-monopolistique américain et vise à empêcher les pratiques anti-compétitives. Il est possible que ce type de régulation induise un mélange de la composition des conseils d'administration qui rende "plus aléatoire" leur composition.

Pour résumer, nous avons constaté que, même avec un protocole de diffusion très basique, aucune des topologies stylisées n'est satisfaisante pour reconstruire la dynamique de diffusion que nous avons observé sur l'ensemble de nos réseaux réels. Le réseau SF se comporte comme le réseau ER vis-à-vis de la vitesse de diffusion mais très différemment du réseau réel. Cette observation amoindrit, pour le protocole de transmission retenu, l'importance qu'on porte généralement au rôle de la distribution de degré dans les processus de diffusion. La structure locale de communauté semble en revanche, comme le suggèrent partiellement les résultats du réseau EB, être un paramètre crucial pour modéliser fidèlement la dynamique de diffusion.

6.3 Rôle des règles de transmission

6.3.1 Directionalité de la transmission

Dans la partie précédente, nous avions fixé arbitrairement une direction dans la transmission d'information au moment de l'interaction entre deux agents. Un agent i est choisi aléatoirement, et "est informé" par l'un de ses voisins j (choisi aléatoirement) si celui-ci est lui-même informé : $c_j(t) = 1 \Rightarrow c_i(t) = 1$. Le flux d'information étant dirigé vers l'agent i, nous appellerons ce modèle, le modèle centripète (1) - il correspond à un phénomène de diffusion dans lequel les agents cherchent activement à acquérir une information (on pourrait également l'appeler le modèle de type *reporter*). Le modèle opposé est également envisageable, il correspond à un processus dans lequel les agents tentent activement d'informer un de leurs voisins, formellement il correspond à la même procédure de sélection aléatoire d'un agent i et à une règle de transmission du type : $c_i(t) = 1 \Rightarrow c_j(t) = 1$, on appellera ce modèle, le modèle centrifuge (2) (on pourrait également l'appeler le modèle de type *gossip* mais c'est également le type de transmission qui s'opère

tel-00464316, version 1 - 16 Mar 2010

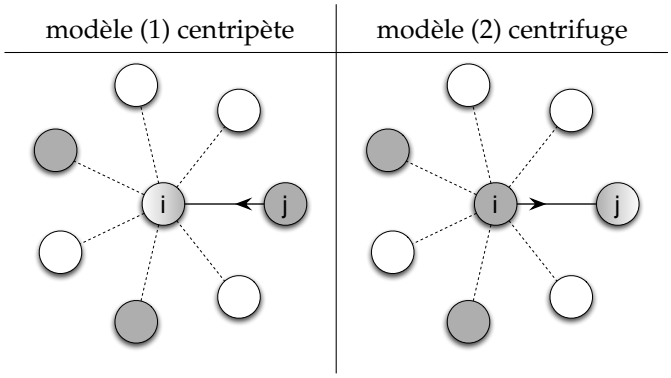


FIGURE 6.5: Représentation schématique du processus de transmission selon la direction retenue *À gauche :* modèle de transmission (1) dit centripète, *à droite :* modèle de transmission (2) dit centrifuge.

entre un professeur et son élève). Nous avons représenté figure 6.5 une représentation schématique du processus de transmission adopté pour chaque direction.

Nous avons évalué le rôle de ce changement de direction de la règle de transmission sur la dynamique de diffusion. Le modèle centrifuge (voir courbes d'indice *2* figure 6.6) ne modifie pas qualitativement les résultats obtenus sur le modèle centripète (voir courbes d'indice *1*). Pour simplifier nous n'avons représenté les dynamiques de diffusion pour les deux types d'hypothèses de transmission que sur les topologies CN, ER, SF et RN en nous basant sur le réseau de collaboration. Les vitesses de diffusion sont systématiquement plus lentes sous l'hypothèse (2) à l'exception du réseau complet, pour lequel le sens de la transmission n'a naturellement aucun effet : l'ensemble des nœuds étant équivalents, les agents i et j sélectionnés pour un événement d'interaction sont naturellement parfaitement interchangeables.

Le ralentissement de la diffusion dans l'hypothèse (2) peut s'expliquer à travers le rôle joué par les nœuds de plus fort degré. Après qu'un agent i a été choisi de façon aléatoire sur l'ensemble des agents, un voisin de i est choisi aléatoirement. Dans un premier temps, chaque agent a donc la même probabilité $1/N$ d'être sélectionné, mais les nœuds de fort degré sont favorisés lors du choix du second agent j parmi les voisins de i. Ainsi un agent aura une probabilité $k_j / \sum_j k_j$ de participer à une interaction comme second agent. Dès lors les nœuds de fort degré sont plus souvent mobilisés dans l'hypothèse (1) comme source de la transmission, et dans l'hypothèse (2) comme destinataire de la transmission. On s'attend donc dans l'hypothèse (2) à une acquisition rapide de l'information chez les agents de fort degré, et à un ralentissement de la vitesse d'acquisition chez les agents de faible degré beaucoup moins fréquemment sollicités. Par contre, dans l'hypothèse (1), la distribution de degré n'affecte pas directement la probabilité de participer à une interaction en tant que destinataire d'une information. Ainsi, en l'absence d'effets de corrélation de degrés, on s'attend, dans l'hypothèse (1) à ce que les agents

tel-00464316, version 1 - 16 Mar 2010

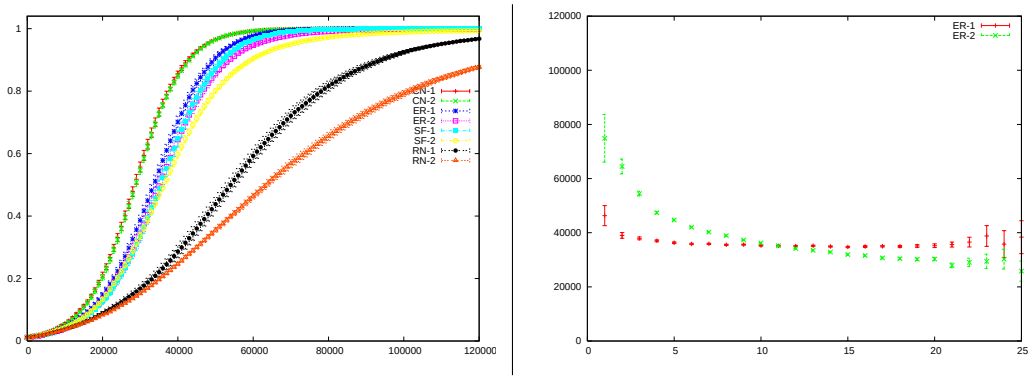


FIGURE 6.6: *À gauche :* évolution du taux de diffusion total ρ sur une série de 5 simulations pour différentes topologies de réseau : CN, ER, SF et RN, construites sur le réseau de collaboration Medline pour les deux hypothèses de transmission : (1) centripète (2) centrifuge (le temps est en abscisses, et $\rho(t)$ en ordonnées). *À droite :* évolution du temps de transmission moyen (en ordonnées) en fonction du degré des nœuds (en abscisses) sur une série de 50 simulations et une topologie de type ER construite à partir du réseau de collaboration Medline pour les deux hypothèses de transmission : (1) centripète (ER1) (2) centrifuge (ER2).

soient informés à la même vitesse indépendamment de leur degré.

Cette simple observation de l'hétérogénéité inhérente au nombre de participations à une interaction en tant que destinataire dans l'hypothèse (2) permet d'expliquer le ralentissement de la vitesse de diffusion par rapport à l'hypothèse (1) qui assure à l'ensemble des nœuds le même *taux d'exposition à un événement de transmission*. La fréquence d'exposition plus élevée prévisible pour les nœuds de haut degré implique un plus grand nombre d'interactions sans effet vis-à-vis de ces nœuds et d'autant moins de tentatives d'infection des nœuds de faible degré. Cette hétérogénéité est mise en évidence sur la figure 6.6 (droite) qui représente pour une série de simulations et pour chaque classe de nœuds de degré k donné le temps moyen écoulé avant que la moitié des nœuds d'une classe de nœuds donnée soit informée. On constate que le profil est quasiment plat dans le cas de l'hypothèse de transmission (1), *i.e.* les nœuds sont informés à la même vitesse quel que soit leur degré, tandis que dans le cas d'une hypothèse de transmission (2) la vitesse augmente fortement avec le degré. Des motifs similaires ont été observés sur les autres types de topologie. De façon générale, et sur l'ensemble des topologies, l'hypothèse centrifuge (2) induit une forte dépendance des vitesses de diffusion au degré des agents, qui est absent dans le cas de l'hypothèse centripète, la vitesse réduite d'acquisition de l'information des nœuds de faible degré est donc une explication plausible du ralentissement de la diffusion sur l'ensemble des topologies induit par le passage d'une transmission centripète à une transmission centrifuge.

tel-00464316, version 1 - 16 Mar 2010

6.3.2 Hypothèses de transmission réalistes

Indépendamment du sens de la transmission, le protocole de transmission utilisé dans nos simulations peut également être interrogé puisque celui-ci est extrêmement basique. Plusieurs modèles de transmission ont été proposés dans la littérature, notamment le "modèle à seuil", stipulant que les agents adoptent si une fraction de leurs voisins ont déjà adopté (Granovetter, 1978a; Valente, 1995, 1996; Abrahamson and Rosenkopf, 1997; Lew, 2000; Gruhl et al., 2004) et le "modèle à cascade", qui fixe une probabilité d'adoption à chaque nouvelle interaction (Goldenberg et al., 2001; Kempe et al., 2003).

De la même manière que sur les différentes topologies de réseaux, nous souhaitons comparer ces modèles stylisés à un comportement réel issu d'une mesure empirique. Cette fois-ci la topologie du réseau qui supporte la diffusion est fixée : nous choisissons le réseau réel (RN) comme topologie de référence.

Pour définir notre mécanisme de transmission "réel", nous nous basons sur des données empiriques extraites d'une analyse de Leskovec et al. (2007b) dans laquelle est évaluée la probabilité d'un comportement d'achat en fonction du nombre de recommandations reçues par email pour un produit donné. Nous nous servirons de "l'allure" de cette courbe de probabilité d'adoption comme d'un exemple de processus de transmission réel (voir figure 6.8).

Ces données empiriques permettent d'estimer une fonction de probabilité $P(n)$ d'avoir acheté un bien (en l'occurrence un DVD) en fonction du nombre n de recommandations reçues par un individu. Sachant que le protocole de mesure est le résultat d'une observation finale a posteriori, et que l'adoption a pu avoir lieu à la suite de chacun des événements de recommandation, nous déduisons la probabilité d'adoption pour chaque recommandation en appliquant la procédure suivante. Si on note $p(n)$, la probabilité d'adopter immédiatement après la $n^{\text{ème}}$ recommandation, nous pouvons écrire :

$$P(n) = 1 - \prod_{i=1}^{n} (1 - p(i)) \tag{6.2}$$

et nous en déduisons donc que :

$$p(n) = \frac{P(n-1) - P(n)}{P(n-1) - 1} \tag{6.3}$$

en faisant l'hypothèse que les probabilité individuelles d'adoption $p(i)$ sont indépendantes les uns des autres (*i.e.* la probabilité d'adopter après une tentative de transmission ne dépend pas du nombre d'interactions ayant précédées). Ainsi, la probabilité d'adoption pour une $n^{\text{ème}}$ recommandation (qui correspond donc dans notre cadre à une interaction avec un voisin informé) vaut $p(n)$.

Les données de ce type sont encore rares, Backstrom et al. (2006a) ont également mesuré la probabilité de participer à une communauté de "Livejournal" en

tel-00464316, version 1 - 16 Mar 2010

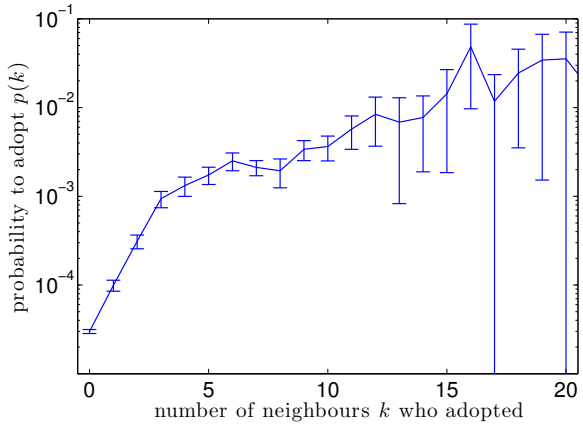


FIGURE 6.7: Probabilité de mentionner une URL en fonction du nombre de voisins dans le voisinage d'un blog l'ayant déjà mentionné.

fonction du nombre d'amis de l'agent considéré déjà engagés dans celle-ci, ou la probabilité de participer à une conférence en fonction du nombre de co-auteurs y participant et ont observé des courbes de probabilité d'adoption similaire.

Nous avons essayé de calculer le même type de courbe de probabilité d'adoption à partir de notre base de données sur les blogs politiques américains. Le protocole qui s'inspire du travail de Backstrom et al. (2006a) est le suivant : nous sélectionnons dans un premier temps un ensemble d'URLs $\mathcal{U}$ qui correspondent à des ressources mobilisées dans la blogosphère américaine (le protocole précis est décrit chapitre 7). Pour chacune de ces URLs $u \in \mathcal{U}$ nous calculons l'état des blogs (dont l'ensemble est désigné par $\mathcal{B}$) par rapport à cette URL. On considère qu'un blog a adopté u à t s'il l'a mentionnée dans un de ses billets à $t' \leq t$. L'état d'un blog i à un moment t donné s'exprime donc comme un vecteur d'état sur l'ensemble des ressources : $U_t(i)$ tel que $U_t(i,u) = 1$ ssi i a déjà adopté u à t, $U_t(i,u) = 0$ dans le cas contraire. À chaque pas de temps, on mesure également pour chaque blog i et chaque ressource u, le nombre de ses voisins $\kappa_t(i,u)$ ayant déjà adopté u à t. L'évolution temporelle des matrices U_t et κ_t nous permet de mesurer pour chaque moment t la probabilité moyenne pour un blog d'adopter une ressource u en fonction du nombre κ de voisins la partageant antérieurement [11].

11. Cette probabilité est simplement estimée en calculant la proportion sur l'ensemble des blogs n'étant pas informé à $t-1$ (la granularité est le jour) et ayant exactement κ_0 voisins ayant déjà mentionné u à $t-1$ du nombre de blogs adoptant u à t :

$$p_t(u,\kappa_0) = \frac{|i \in \mathcal{B}, U(i,t-1) = 0, \kappa_{t-1}(i,u) = \kappa_0, U(i,t) = 1|}{|i \in \mathcal{B}, U(i,t-1) = 0, \kappa_{t-1}(i,u) = \kappa_0|}$$

Ces probabilités sont ensuite moyennées sur l'ensemble des ressources (on a appliqué le calcul à environ 10 000 ressources différentes) pour construire des probabilités d'adoption p_t indépendantes des ressources. La figure 6.7 représente la moyenne de ces probabilités d'adoption dans le temps (avec les intervalles de confiance associés). NB : Cette courbe agrège des ressources très partagées

tel-00464316, version 1 - 16 Mar 2010

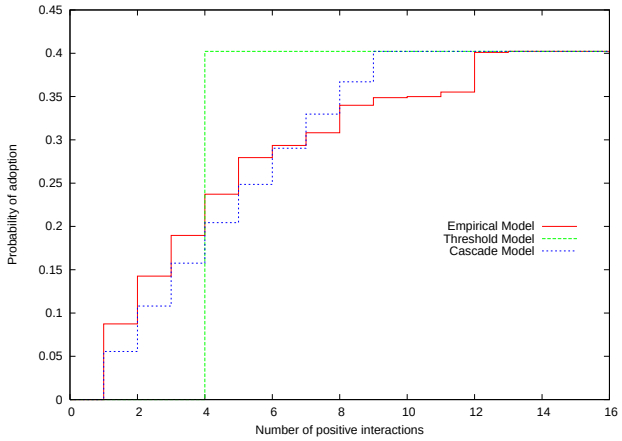


FIGURE 6.8: Probabilité d'adoption $P(n)$ après n événements de recommandations ($P_{max} = 0.4$) pour nos trois modèles de transmission (empirique, à seuil ou à cascade). Données empiriques adaptées de (Leskovec et al., 2007b), avec $P_{\max} = 0.04$.

Nous avons représenté la probabilité d'adoption en fonction de ce paramètre κ figure 6.7. Contrairement aux résultats déjà mentionnés et malgré le caractère assez bruité de la courbe obtenue, nous observons une croissance exponentielle (la croissance de la probabilité semblant linéaire malgré l'échelle logarithmique de l'axe des ordonnées) - avant un aplanissement de la courbe. Malgré l'allure plutôt logarithmique des probabilités d'adoption observées par Backstrom et al. (2006b), les auteurs notent que celles-ci ont néanmoins des allures légèrement sigmoïdales ("S-shaped"), avec une dérivée seconde positive pour les premiers κ ($\kappa = 0$ ou 1) avant de s'inverser. Le jeu de données et le type de données que nous observons permet peut-être d'observer, sur une plus grande plage de valeurs et avec une meilleure résolution, cet effet d'accélération supralinéaire de la probabilité d'adoption qui précède la phase de saturation au delà de laquelle l'augmentation du nombre de voisins ayant déjà adopté ne modifie quasiment plus la probabilité d'adoption.

6.3.3 Modèles de transmission stylisés

Nous nous concentrons sur les modèles classiques à seuil ou à cascade. Nous choisissons les paramètres de ces modèles de sorte que leur probabilité d'adoption $P(n)$ soit la plus proche possible du profil empirique.

Dans cette partie, nous envisageons également des épisodes de diffusion pour lesquels les agents ne sont pas nécessairement tous informés. Nous bornons donc les probabilités d'adoption à une valeur $P_{\max}$ telle qu'observée empiriquement

et d'autres qui n'ont été adoptées que par peu de blogs, néanmoins, elle semble stable lorsqu'on ne s'intéresse qu'à une sous-classe de ressources étant *in-fine* partagée par un nombre K de blogs pourvu que K soit suffisamment grand.

tel-00464316, version 1 - 16 Mar 2010

$\forall n, P(n) \leq P_{\max}$ et $\exists n_0, P(n_0) = P_{\max}$ — après un certain nombre d'interactions, toute forme d'adoption ultérieure est considérée comme impossible, dans l'étude originale de Leskovec et al. (2007b) on obtient $P_{\max} \approx 0.06$.

Trois types de comportement de transmission sont donc envisagés : le modèle empirique, et les modèles à seuil et à cascade les plus proches pour différentes valeurs de P_{max}. Pour résumer, nous allons comparer la dynamique de diffusion en fonction des hypothèses de transmission suivantes :

- *Modèle réaliste (RM)* — : $P_{\text{RM}}(n) = P_{\text{empirical}}(n)$ correspond aux observations empiriques. $p_{\text{RM}}(n)$ est calculé à l'aide de l'équation 6.3.
- *Modèle à seuil (TM)* — : pour ce type de modèle, les agents ne peuvent pas adopter avant un certain nombre d'interactions ν (appelé le seuil), les agents adoptent avec une probabilité $P_{\max}$ au moment de la ν^{me} interaction : pour n interactions avec $n \neq \nu$, $p(n) = 0$, if $n = \nu$, $p(n) = P_{\max}$. La probabilité finale d'adoption est donc de la forme :

$$P_{TM}(n) = P_{\max} \cdot H_\nu(n)$$

  où H_ν est définie telle que : $H_\nu(n) = 1$ si $n \geq \nu$, 0 sinon .
- *Modèle à cascade (CM)* — : nous introduisons un modèle de cascade "borné", dont la probabilité d'adoption finale est limitée à $P_{\max}$; de façon à modéliser un facteur de saturation (de façon similaire aux modèles de cascades "décroissantes" introduites dans la littérature (Kempe et al., 2005)). Ainsi, $p(n) = p$ est la probabilité fixe d'adoption à chaque interaction, mais après un certain nombre d'interactions ν, $p(n) = 0$. La probabilité finale d'adoption est donc de la forme : $P_{CM}(n) = 1 - (1-p)^{min(n,\nu)}$ On constate que $p = 1 - (1 - P_{\max})^{1/\nu}$; ainsi $P_{\max}$ permet de définir de façon univoque p à partir de ν et vice-versa.

6.3.4 Résultats des simulations

Pour chaque valeur de $P_{\max}$, nous avons tracé la valeur moyenne de l'évolution de ρ sur 50 simulations figure 6.9. Tous les modèles convergent vers un même état final $\rho(\infty) \leq P_{\max}$.

Nous avons simulé les différentes règles de transmission sur nos réseaux réels pour différentes valeurs de $P_{\max}$. Pour évaluer l'influence de ce paramètre nous choisissons les valeurs suivantes : $P_{\max} \in \{0.04, 0.4, 0.7, 0.99\}$. Pour construire la meilleure approximation possible des modèles TM et CM nous procédons à une simple homothétie des résultats empiriques originaux. Nous choisissons ensuite p et ν pour CM et TM en minimisant la distance quadratique entre le modèle empirique original (RM) et les distributions de la probabilité d'adoption totale associée à chaque modèle. P_{TM} et P_{CM} sont représentés figure 6.8 pour une valeur de $P_{\max} = 0.4$.

tel-00464316, version 1 - 16 Mar 2010

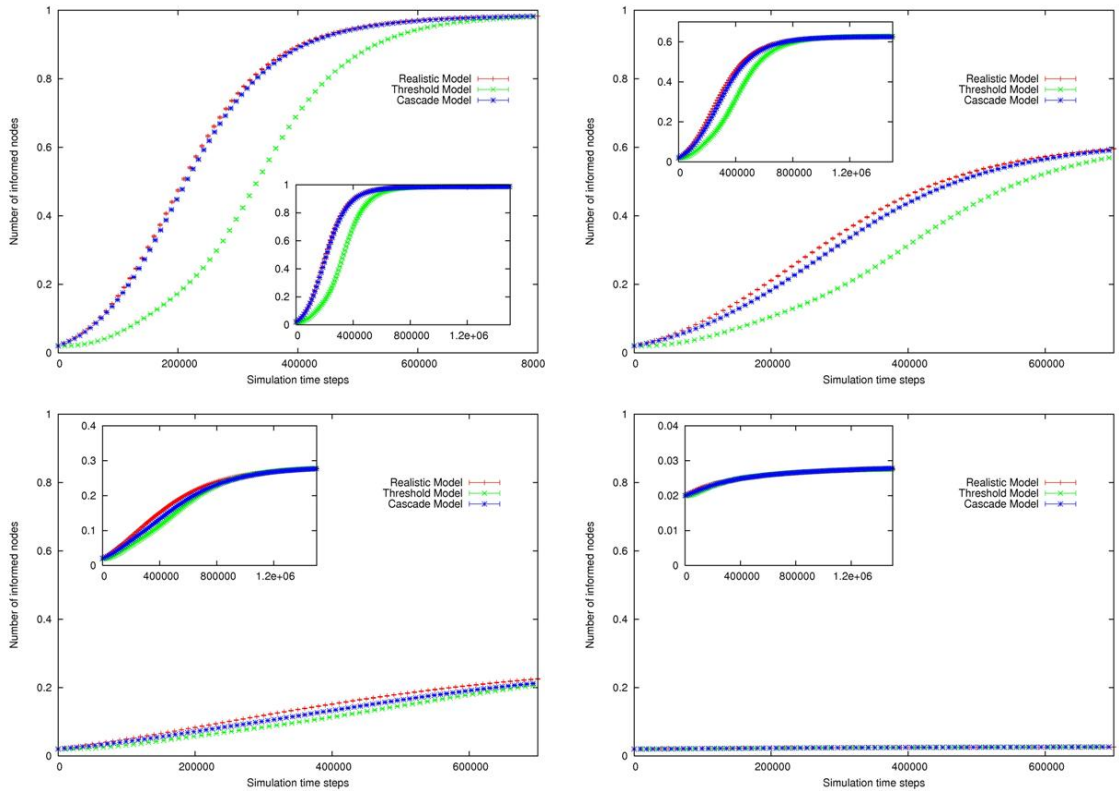


FIGURE 6.9: Evolution de ρ pour des hypothèses de transmission RM, TM et CM, pour 4 valeurs de $P_{\max}$, de gauche à droite et de haut en bas : $P_{\max} = 0.99$, $P_{\max} = 0.7$, $P_{\max} = 0.4$, $P_{\max} = 0.04$. Les encarts présentent les comportements asymptotiques.

$\rho(\infty)$ est naturellement décroissant avec $P_{\max}$. Les agents ayant une probabilité maximale d'être informés au cours de la simulation (certains agents peuvent ne jamais entrer en contact avec des voisins informés), on a également l'inégalité : $\rho(\infty) \leq (\lambda + (1-\lambda)P_{\max})$, qui devient une égalité pour $P_{\max} = 1$.

La vitesse de convergence est également d'autant plus rapide que $P_{\max}$ augmente, et ceci indépendamment du modèle de transmission choisi. Pour les $P_{\max}$ importants, le modèle TM échoue à reconstruire de façon satisfaisante la dynamique de diffusion réelle (RM), tandis que le modèle CM s'en rapproche plus fidèlement. On a représenté figure 6.10, l'erreur relative entre la dynamique observée avec l'hypothèse RM, TM ou CM (calculée selon la formule $\|\rho_{RM} - \rho_{TM}\|/\|\rho_{RM}\|$ et $\|\rho_{RM} - \rho_{CM}\|/\|\rho_{RM}\|$ respectivement). L'hypothèse à seuil TM a le taux d'erreur le plus important pour des valeurs de $P_{\max}$ importantes tandis que l'hypothèse cascade est de plus en plus proche de la réalité avec $P_{\max}$ croissant. TM semble néanmoins être la meilleure approximation possible pour $P_{\max} = 0.04$ (qui est proche de la valeur empirique observée dans les données de Leskovec et al. (2007b)).

D'autres améliorations des modèles à seuil et à cascade sont sans doute envi-

tel-00464316, version 1 - 16 Mar 2010

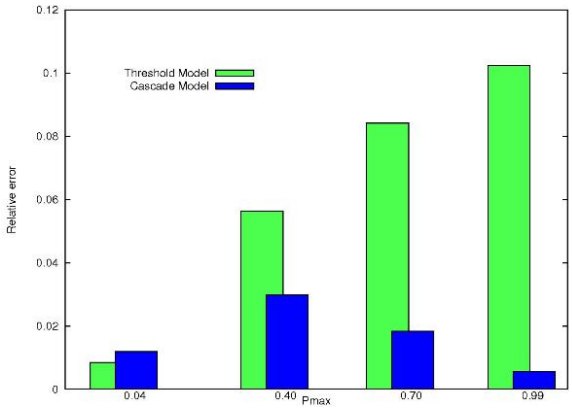


FIGURE 6.10: Erreurs relatives entre les modèles à seuil (TM) et à cascade (CM) comparés au modèle empirique.

sageables [12], mais il pourrait également s'avérer que certains processus de transmission ne puissent pas être modélisés ni par un modèle à seuil ni un modèle à cascade (par exemple, les courbes de probabilité d'achat de livres $P(n)$ en fonction d'un nombre de recommandations n calculées par Leskovec et al. (2007b) indiquent de façon surprenante une décroissance de la probabilité pour des valeurs importantes de n, suggérant que l'hypothèse d'une succession de probabilité $p(n)$ ne serait pas pertinente dans ce cadre [13].

12. Entre autres possibilités d'amélioration, on pourrait définir une probabilité d'adoption décroissante avec le nombre d'interactions (Kempe et al., 2003)). Les modèles TM et CM sont sans doute également perfectibles en attribuant respectivement à chaque agent ou à chaque lien entre agents, un seuil, ou une probabilité d'infection différente comme le suggère Valente (1996)

13. il faut néanmoins rester prudent quant à ce point, les courbes en question agrégeant un ensemble de produits et d'agents auxquels pourraient simplement correspondre des classes de comportements contrastés.

tel-00464316, version 1 - 16 Mar 2010

Résumé du chapitre:

Ce chapitre nous a permis à l'aide d'un protocole de simulation très simple et en comparant systématiquement nos résultats sur une série de réseaux, dont la topologie reproduit un certain nombre de caractéristiques structurelles de deux réseaux de terrain, de mettre en évidence des propriétés topologiques à même d'influer de manière significative sur la vitesse de diffusion. *A contrario* des observations classiques dans la littérature, la distribution de degré du réseau ne semble avoir qu'une importance très mineure vis-à-vis de notre protocole de diffusion. La cohésion locale, ou de façon plus large, la structure naturellement modulaire des réseaux sociaux pourrait jouer un rôle primordial.

Dans tous les cas de figure, quelle que soit la nature du réseau envisagé (social, infrastructure, biologique, etc.) et indépendamment de l'hypothèse de transmission retenue, on peut s'attendre à ce que le caractère modulaire des réseaux agisse comme un obstacle à la diffusion d'une entité. On pourrait même dire que les réseaux dotés d'une structure communautaire tendent à limiter la diffusion de l'information, en créant un nombre important de liens "redondants" qui compartimentent les informations. Cette redondance peut néanmoins s'avérer bénéfique lorsque l'adoption d'une information implique un processus d'évaluation qui peut nécessiter un recoupement ou une confiance dans les sources. De façon plus générale les résultats de ce chapitre engagent à une certaine prudence et humilité dans la modélisation de réseaux aléatoires. Alors que la grande majorité des études d'épidémiologie ou de diffusion sur les réseaux s'appuient sur des réseaux stylisés, on constate que les résultats dynamiques peuvent différer très sensiblement entre des réseaux aléatoires classiques et des réseaux réels.

.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

CHAPITRE 7

Cascades informationnelles

Sommaire

7.1 Cascades informationnelles et diffusion 208
7.1.1 Jeu de données empirique 208
7.1.2 Distance attentionnelle 209
7.1.3 Sous-graphes de diffusion 212
7.2 Relais d'information et attention 214
7.2.1 Premières transmissions 214
7.2.2 Petits-Fils 215
7.2.3 Secondes transmissions et attention 216
7.3 Courts-circuits informationnels 217
7.3.1 Secondes transmissions et edge range 217
7.3.2 Effets couplés 219
7.3.3 Conclusion 220

Nous avons jusqu'à maintenant abordé la question de la diffusion d'information dans un réseau social en postulant un processus de transmission inter-individuelle ou une topologie *a priori*. Cette approche nous a, notamment, permis de mettre en évidence un certain nombre de paramètres topologiques influençant ou non la vitesse de diffusion en fonction d'une hypothèse de transmission donnée.

Plus précisément, nous avons donc testé des modèles stylisés de réseau ou de processus de transmission inter-individuelle en les comparant à leur contrepartie "réelle". Néanmoins , nous n'avons pas pu pousser cette comparaison des modèles stylisés à un phénomène de diffusion réel dont (i) on puisse effectuer le recueil longitudinal de l'ensemble des événements de transmission interindividuelle, tout en (2) ayant connaissance de la topologie du réseau sous-jacent. À notre connaissance, à l'exception de (Adar et al., 2004a), il n'existe pas d'études portant sur un phénomène de diffusion réel avec un suivi exhaustif et longitudinal de la propagation d'une information incluant un relevé de l'ensemble des événements de transmission doublé d'une connaissance explicite du réseau social sur lequel se déploie la diffusion.

Or la compréhension de tels processus est capitale à plusieurs titres. D'une part il s'agit de mieux saisir la façon dont évolue la distribution des connaissances au

sein de ces systèmes, la reprise d'une information ou *l'imitation* d'un agent voisin étant un moteur essentiel dans la dynamique des communautés de savoirs. La compréhension d'un processus de diffusion empirique peut également éclairer sous un autre jour les structures mésoscopiques et positions des agents au sein du réseau social. Comment l'environnement relationnel d'un agent le rendra plus ou moins exposé à une information se propageant, et comment la position qu'il occupe dans le réseau rendra cet agent plus ou moins influent vis-à-vis de ses voisins ? Plus généralement, c'est la question de la co-évolution des contenus et de la topologie du réseau, entre dynamiques sociales et sémantiques, qui est investiguée. Nous avions une posture macroscopique dans le chapitre précédent, et nous étions concentrés sur une observable globale du réseau : le taux de diffusion, nous privilégierons, cette fois-ci, une approche plus locale fondée sur l'analyse des comportements individuels vis-à-vis de la diffusion.

Pratiquement, nous nous intéresserons à un processus de diffusion dans la blogosphère en tachant de suivre la transmission d'*URLs* au sein d'un réseau de blogs. Les billets produits par les blogueurs contiennent souvent un certain nombre de liens vers des ressources extérieures (une vidéo, une publication hébergée par un site institutionnel, etc.). Ces ressources s'accompagnent fréquemment d'une référence à un blog tiers ayant également pointé vers ces ressources. Nous ferons l'hypothèse que ces liens de citation permettent de reconstruire le chemin réellement emprunté par l'information. Cette hypothèse semble raisonnable dans le cas des blogs citoyens où la norme veut que l'on cite ses sources. Nous nous appuierons sur deux jeux de données : les échantillons de la blogosphère française américaine déjà présentés dans le chapitre 2. Ce travail sur les cascades informationnelles est en grande partie extrait d'une publication avec Camille Roth : (Cointet and Roth, 2009).

7.1 Cascades informationnelles et diffusion

Dans un premier temps nous introduisons un certain nombre d'éléments de formalisme et d'outils de mesure sur le réseau afin de caractériser à la fois le phénomène de propagation et le réseau social sous-jacent.

7.1.1 Jeu de données empirique

Les deux réseaux de blogs que nous étudierons sont constitués de l'ensemble de liens dynamiques de citation entre blogs C déjà introduits dans le chapitre 2. Pour rappel, on a extrait $11\ 552$ billets dans le cas français et $71\ 376$ dans le cas américain. Les blogs français ont été suivis sur une période de 6 mois entre le 1^{er} janvier 2007 et le 30 juin 2007, les blogs américains, pendant 4 mois du 1^{er} novembre 2007 au 29 février 2008. Le réseau $(\mathcal{B}, P)$ est constitué de l'ensemble des blogs $\mathcal{B}$ ($|\mathcal{B}| = 120$ dans le cas de la blogosphère française et $|\mathcal{B}| = 1\ 066$ dans

tel-00464316, version 1 - 16 Mar 2010

le cas de la blogosphère américaine), et de l'ensemble des liens de citation datés entre blogs formalisés grâce à la matrice d'adjacence dynamique $P, P_t(i,j) = 1$ si le blog i cite le blog j au temps t. On s'appuiera également sur le réseau dynamique agrégé lié à la matrice d'adjacence $\mathbf{P}_t = \sum_{t' \leq t} P_{t'}$ (les variables rendant compte de l'agrégation temporelle sont en gras) qui permet d'apprécier à un temps t donné le poids des relations qu'entretiennent les blogs les uns avec les autres. On compte $2\,229$ liens dans le cas français (dont 850 sont non répétés) et $229\,736$ dans le cas américain (pour à peine $15\,032$ liens uniques).

D'autre part, on définit, pour chaque jeu de données, l'ensemble des URLs mentionnées dans l'ensemble des billets extraits sur l'ensemble de la période d'observation dans les deux réseaux. Cet ensemble d'URLs est noté $\mathcal{U}$. On exclut naturellement de cet ensemble les liens de citation vers des blogs appartenant à nos échantillons de blogosphères, et les URLs comportant moins de 10 caractères. On dénombre respectivement $3\,140$ et $96\,637$ URLs dans nos deux systèmes. On définit également l'ensemble des matrices U_t dont les coordonnées $U_t(i,u)$ valent 1 si le blog i mentionne la ressource $u \in \mathcal{U}$ au temps t [1] .

Nous définissons maintenant un certain nombre de mesures sur les nœuds du réseau social à même de nous éclairer sur la façon dont certaines caractéristiques structurelles des agents peuvent être liées à la dynamique des épisodes de diffusion empiriques que nous souhaitons observer.

7.1.2 Distance attentionnelle

Nous souhaitons décrire les phénomènes d'influence au niveau individuel et notamment la façon dont certains paramètres structurels sont pertinents pour prédire la longévité d'un épisode de diffusion transitant à travers un nœud donné. À ce titre, l'attention qu'exerce un blog sur son environnement nous paraît cruciale pour comprendre les dynamiques d'influence.

Etant donné le réseau agrégé, $\mathbf{P}_t$, on définit l'*attention* $\mathbf{a}$ en normalisant chaque colonne de $\mathbf{P}_t$ selon la formule suivante :

$$\mathbf{a}_t(i,j) = \frac{\mathbf{P}_t(i,j)}{\sum_{j=1}^{|\mathcal{B}|} \mathbf{P}_t(i,j)}$$

L'attention $\mathbf{a}_t(i,j)$ d'un blog i vis-à-vis d'un blog j est donc simplement définie au temps t comme la proportion de liens de i vers j parmi tous les liens sortants de i observés jusqu'à t. L'attention permet de quantifier "l'importance" de j pour i indépendamment du profil d'activité de i (la distribution des degrés sortants étant fortement hétérogène, une mesure brute du nombre de liens sortants aurait tendance à rendre incommensurable les attentions de deux blogs d'activité très

1. Ce sont ces données : relationnelles et d'usage : U_t et $\mathbf{P}_t$ qui ont servi dans le chapitre 6, paragraphe 6.3.2, à calculer les probabilité d'adoption d'une URL en fonction de sa présence dans le voisinage d'un blog

tel-00464316, version 1 - 16 Mar 2010

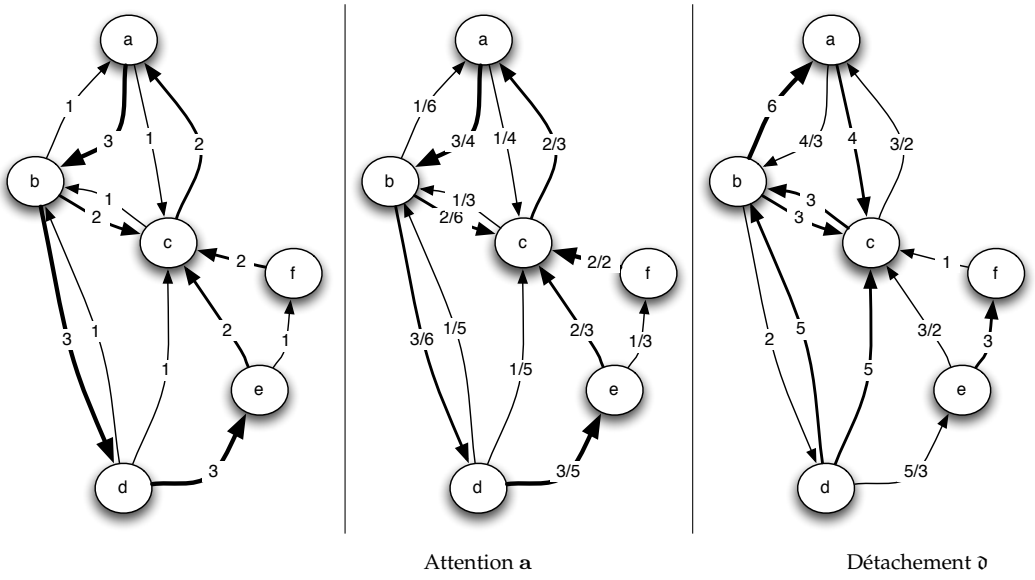


FIGURE 7.1: *À gauche :* Un exemple de réseau de citation pondéré $\mathbf{P}$ (on a omis la dépendance temporelle par souci de clarté) : les poids des liens correspondent simplement au nombre de liens produits entre deux blogs au temps t. *Au centre* et *à gauche :* les liens sont respectivement étiquetés par les valeurs d'attention et de détachement correspondantes. Par exemple, le blog b a cité c deux fois sur un total de $1 + 2 + 3 = 6$ liens de citation, son attention vis-à-vis de c est donc $\mathbf{a}(b, c) = \frac{2}{6}$. *le détachement* $\mathfrak{d}(b, c)$ est égal à l'inverse de l'attention entre b et c soit 3. L'*attention totale* $\alpha(c)$ exercée par la blog c agrège l'ensemble des attentions des blogs b, a, d, e et f vis-à-vis de c, et vaut 2.45. La distance attentionnelle entre les blogs b et c correspond au chemin de distance minimale dans le réseau de détachement, ainsi $\partial(b, e) = \mathfrak{d}(b, d) + \mathfrak{d}(d, e) = \frac{11}{3}$

différente). Une valeur importante de l'attention de i vis-à-vis de j indique que j focalise une grande part de l'attention de i. Une notion similaire, la "matrice d'influence" a récemment été introduite par Java et al. (2006).

On peut également définir l'attention totale exercée par un blog j comme la somme des attentions qu'il exerce sur l'ensemble des blogs :

$$\alpha_t(j) = \sum_i \mathbf{a}_t(i, j)$$

Nous définissons maintenant la notion inverse de l'attention que nous appellerons *détachement* , en considérant simplement les valeurs inverses de $\mathbf{a}$: On peut interprèter le détachement $\mathfrak{d}_t(i, j)$ de i vis-à-vis de j comme le coût relatif à la circulation d'une information de i vers j. Par convention, le détachement entre deux blogs non connectés vaut l'infini. Ces mesures de détachement nous permettent de construire le graphe de détachement, dont les blogs constituent les nœuds et dont les liens (i, j) sont pondérés par les valeurs de détachement $\mathfrak{d}_t(i, j)$.

Enfin, nous définissons la notion de distance de détachement comme la distance minimale (que l'on calcule avec l'algorithme de dijkstra (Dijkstra, 1959))

tel-00464316, version 1 - 16 Mar 2010

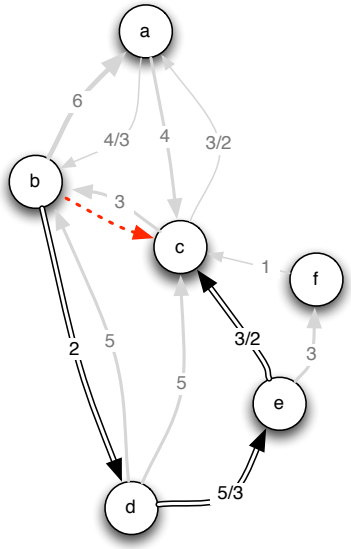


FIGURE 7.2: *Calcul du "edge-range" :* pour calculer l'edge-range $\mathbf{r}(b,c)$, on supprime d'abord le lien de b vers c avant de calculer le chemin de coût minimum entre b et c, en utilisant les valeurs de détachement $\mathfrak{d}$ comme poids sur les arêtes du graphe $\mathbf{P}$. Sur cet exemple, le chemin $(b-d-e-c)$ a une distance $\mathbf{r}(b,c) = \frac{31}{6}$. Des chemins avec moins d'intermédiaires comme $(b-a-c)$ ont en fait un coût attentionnel plus grand (distance 10 dans ce cas).

dans le graphe de détachement. On note $\partial_t(i,j)$ cette distance, on peut l'interpréter comme une mesure d'éloignement, au sens attentionnel entre deux blogs i et j. Une distance attentionnelle importante entre deux blogs signifie que le coût nécessaire pour faire transiter une information d'un blog à un autre est important. Elle peut-être considérée comme une forme de quantification d'un "temps" caractéristique nécessaire au transit d'une information d'un bog à un autre. La figure 7.1 illustre, par un exemple, l'ensemble des mesures introduites : attention (a), attention total (α), détachement ($\mathfrak{d}$) et enfin distance attentionnelle (∂) .

"Edge-range" distance La notion d'*edge range* a été introduite par Watts (1999) pour un lien (i,j) comme la distance entre deux nœuds i et j après suppression du lien entre i et j. Elle a notamment été utilisée récemment dans des études de diffusion par Kossinets et al. (2008).

Nous étendons cette notion au cas d'un graphe pondéré, dont les poids sont constitués par les valeurs de détachement. Formellement, nous définissons l'edge-range pondéré $\mathbf{r}(i,j)$ du lien (i,j) comme la distance pondérée minimale dans le graphe de détachement, dont le lien (i,j) a été supprimé (ou de façon équivalente en remplaçant la valeur $\mathfrak{d}_t(i,j)$ par ∞). En d'autres termes, c'est la somme minimale des valeurs de détachement le long du plus court chemin indirect de i vers j, c'est à dire l'attention totale requise pour faire transiter une information de j vers i si le lien de i vers j était omis (la circulation de l'information "remonte" les liens

tel-00464316, version 1 - 16 Mar 2010

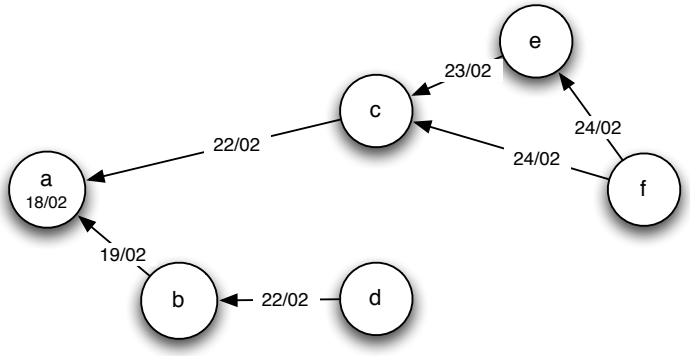


FIGURE 7.3: Illustration d'un sous-graphe de diffusion σ_{u_0}. Les étiquettes temporelles indiquent la date à laquelle les blogs ont simultanément mentionné u_0 et cité un blog de $\mathcal{B}$ ayant déjà mentionné u_0. NB : Les sous-graphes de diffusion ne forment pas nécessairement un arbre, comme en atteste le blog f sur cet exemple qui cite simultanément c et e comme sources.

de citation). Naturellement, l'edge-range pondéré $\mathbf{r}(i,j)$ n'est défini que sur l'ensemble des paires de nœuds (i,j) déjà connectées dans $\mathbf{P}$, *i.e.* telles que $\mathbf{P}(i,j) > 0$. Un exemple de calcul de $\mathbf{r}(i,j)$ est présenté figure 7.2.

L'ensemble de ces mesures sur nos réseaux de blogs est par la suite considéré en tenant compte de la totalité des citations créées sur l'intégralité de la période d'observation *i.e.* à $t_f = 181$ pour la blogosphère politique française, et à $t_f = 121$ dans le cas de la blogosphère américaine. Par la suite, on notera donc nos mesure sans mentionner le paramètre temporel. Ainsi, $\mathbf{r}(i,j) = \mathbf{r}_{t_f}(i,j)$ et $\alpha(i) = \alpha_{t_f}(i)$.

7.1.3 Sous-graphes de diffusion

Nous nous concentrons sur les épisodes de diffusion liés à une ressource mise en circulation au sein d'un réseau social. Nous introduisons la notion de *sous-graphe de diffusion* pour définir les motifs créés par la circulation de ressources. Un sous-graphe de diffusion regroupe l'ensemble des agents ayant mentionné une ressource donnée - une URL dans un billet - et étant reliés les uns aux autres au travers de l'ensemble des liens dirigés (i,j) tels que i a mentionné dans un même billet la ressource en question et cité un agent j ayant mentionné la ressource antérieurement.

La figure 7.3 propose un exemple de sous-graphe de diffusion. Dans cet exemple, une URL u_0 est en premier lieu mentionnée par le blog a le 18 février. Elle est ensuite mentionnée par b le 19 février, qui cite a concomitamment. u_0 diffuse ensuite vers d depuis b et vers c depuis a le 22 février, etc.

Plus formellement, étant donnée une ressource $u \in \mathcal{U}$, on définit le sous-graphe de diffusion de u : $\sigma_u \in \mathcal{P}(\mathcal{B}) \times \mathcal{P}(\mathcal{B} \times \mathcal{B})$ comme la combinaison :

tel-00464316, version 1 - 16 Mar 2010

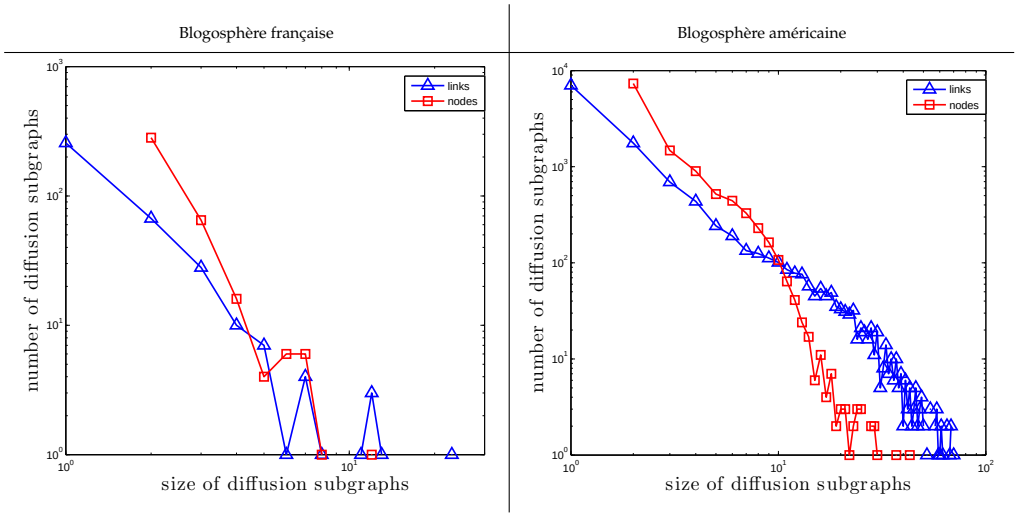


FIGURE 7.4: Distributions des tailles des sous-graphes de diffusion en nombre de nœuds (*carrés rouges*) et en nombre de liens (*triangles bleus*), dans nos deux cas d'étude.

- des agents mentionnant u
- des liens dirigés (i, j) dans $\mathbf{P}$ tels que i cite simultanément j et u dans un même billet, *après* que j a cité u.

On peut également écrire que les sous-graphes de diffusion σ_u sont constitués de l'ensemble des liens dirigés (i, j) et des nœuds associés tels que :

- $\mathbf{P}_t(i, j) > \mathbf{P}_{t-1}(i, j)$ (*i.e.* il existe un nouveau lien dans $\mathbf{P}_t$ de i vers j au temps t),
- $U_t(i, u)$=1 (i mentionne la ressource u au temps t),
- $\exists t' \leq t, U_{t'}(j, u) = 1$ (j a mentionné u antérieurement).

Un sous-graphe de diffusion est considéré comme *trivial* s'il ne contient aucun lien, *i.e.* si la ressource correspondante n'est impliquée dans aucun événement de transmission explicite entre deux agents. Sur les $3\ 140$ URLs de la blogosphère française, seules 381 correspondent à des sous-graphes de diffusion non-triviaux. Dans le cas de la blogosphère américaine, ces chiffres sont naturellement plus élevés : on dénombre $11\ 709$ sous-graphes de diffusion non triviaux différents soit également un peu plus de 10% des ressources dénombrées. Dans la suite nous nous concentrerons uniquement sur l'ensemble des sous-graphes non triviaux.

Nous avons représenté la distribution des tailles de l'ensemble des sous-graphes de diffusion figure 7.4. Les distributions de taille sont sensiblement hétérogènes aussi bien en terme de nombre de liens que de nombre de nœuds. Cette observation est cohérente avec l'observation faite sur la distribution des tailles de cascades par Leskovec et al. (2007a), même si la définition des cascades adoptée par ces auteurs est très différente de la nôtre[2]. La majorité des sous-graphes ne

2. Dans ces travaux les cascades sont des sous-graphes du réseau de billets indépendamment de

tel-00464316, version 1 - 16 Mar 2010

contient donc qu'un unique événement de transmission sur un total de 631 (respectivement $39\ 540$) événements de transmissions. Ces sous-graphes les plus simples représentent environ $1/4$ (resp. $1/6$) du nombre total de liens de transmission observés dans nos deux systèmes (blogosphère française et américaine).

7.2 Relais d'information et attention

La capacité d'un blog à être une source de diffusion privilégiée, est souvent considérée comme directement liée à la quantité de ses liens entrants (Gill, 2004; Java et al., 2006), les blogueurs ayant la plus large audience étant considérés comme les plus influents. Cette hypothèse peut être simplement vérifiée en examinant la corrélation existante entre la tendance d'un blog à accaparer l'attention de blogs tiers et sa *capacité de diffusion* réelle.

7.2.1 Premières transmissions

Dans une première approche nous évaluons la corrélation entre *l'attention totale* d'un blogueur et les événements de transmission qu'il génère. La figure 7.5 (courbes rouges) représente la distribution du nombre moyen de transmissions provenant d'un blog ayant une attention totale α donnée dans le réseau de citation [3].

On observe que les valeurs associées à de forts α sont effectivement corrélées avec un plus grand nombre d'événements de transmission.

Cette observation peut paraître tautologique — les liens de transmission constituant par nature un sous-ensemble du réseau de citation. Nénamoins cette observation reste stable si l'on calcule les attentions sur le réseau des commentaires ou de blogroll de la blogosphère française ; ce qui indiquerait que ce sont bien des effets d'audience qui induisent cette corrélation : les blogs plus "influents" sont basiquement ceux ayant un lectorat actif plus étendu.

La notion de lectorat *actif* qui est liée à la régularité des relations entre blogs pourrait bien être décisive ici, l'audience brute d'un site (en terme de lecteurs blogueurs et non blogueurs ou plus simplement en nombre de "hits", soit le nombre de visiteurs d'un site) n'étant pas nécessairement une mesure fiable de son influence. Ainsi Iribarren and Moro (2007) suggèrent d'après des données de diffusion liées à la recommandation d'un produit, que le nombre de recommandations

toute transmission de ressource.

3. Bien que non présentées ici, nous avons observé des corrélation similaires entre le nombre de liens de transmissions induits et une mesure d'audience, au sens large, mesurée directement par le nombre de liens entrants dans l'ensemble des réseaux. Nous considérons cependant que l' *attention totale* mesure plus précisément les effets d'audience car elle est basée sur les profils attentionels des individus qui permettent de remplacer le nombre de liens reçus par un blog par l' *importance relative* qu'il représente pour les blogueurs s'y référant.

tel-00464316, version 1 - 16 Mar 2010

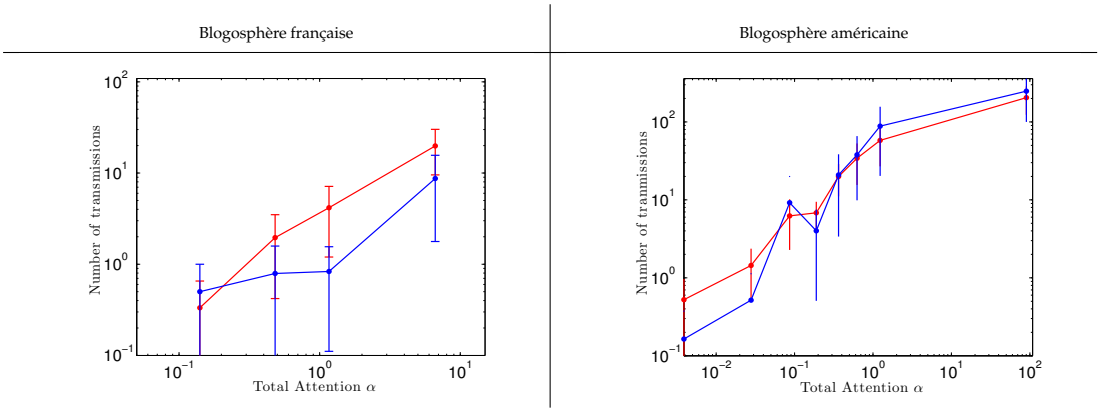


FIGURE 7.5: Nombres moyens de premières (points rouge) et secondes (points bleus) transmissions initiées par un blog d'attention totale α, accompagnés par leurs intervalles de confiance respectifs. (NB : les distributions ont été tracées avec une échelle logarithmique et en utilisant respectivement 4 et 8 quantiles des valeurs de α compte tenu de l'hétérogénéité de leur distribution).

envoyées par un individu à son entourage n'est pas nécessairement liée au nombre de ses contacts. Friggeri et al. (2009) présentent des données de diffusion d'une applet sur le web qui montrent l'absence de corrélation entre la capacité de transmission d'un site (nombre de transmissions effectives de l'applet depuis ce site) et son audience mesurée comme un nombre de hits sur sa page. De façon connexe, Lento et al. (2006) insistent, dans une étude sur des communautés de blogueurs hébergés sur une plateforme d'édition de blogs (Wallop), sur la nature des liens qui lient les nouveaux arrivants aux anciens utilisateurs ; ainsi, le nombre de liens semble moins déterminant pour prédire l'activité futur d'un utilisateur que son degré d'engagement dans la communauté mesuré à travers la *force* des liens (au sens où ils sont répétés et symétriques) qu'il entretient avec d'autres utilisateurs actifs.

7.2.2 Petits-Fils

Si les blogs les plus lus induisent un plus grand nombre d'événements de transmission, nous restons toujours ignorants quant à la structure des sous-graphes de diffusion au delà d'une perspective purement locale. Nous abordons maintenant une autre caractérisation de plus longue portée de ces sous-graphes, à savoir le nombre de *petits-fils*. Plus précisément nous aimerions savoir si une information transmise à partir d'un blog et transitant par un relais aura plus ou moins tendance à être à nouveau "diffusée" en fonction de paramètres structurels liés au blog à l'origine du premier lien de transmission. Autrement dit, étant donné un blog i, nous cherchons à déterminer si certaines caractéristiques topologiques portant sur i sont à même d'induire un nombre plus ou moins élevé d'événements dits de "secondes transmissions", soit de motifs du type $i \leftarrow j \leftarrow k$ dans les sous-

tel-00464316, version 1 - 16 Mar 2010

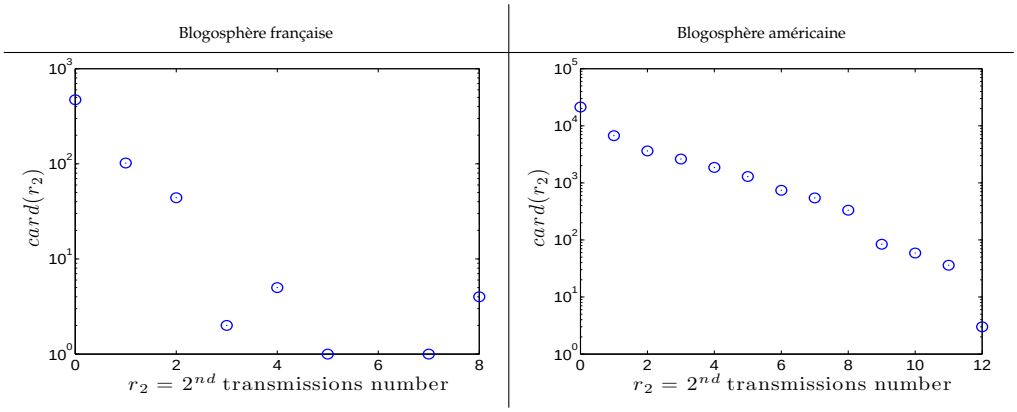


FIGURE 7.6: Distribution du nombre $_2$ de petits-fils associés à chaque transmission.

graphes de diffusion (comme le triplet $a \leftarrow b \leftarrow d$, $a \leftarrow c \leftarrow f$ ou $a \leftarrow c \leftarrow e$ sur notre exemple figure 7.3).

Les destinataires de la ressource à distance 2 seront appelés les *petits-fils* du blog source i (toujours dans notre exemple, d, e et f sont les petit-fils de a). La distribution du nombre moyen de secondes transmissions provoquées à la suite de chaque événement de transmission noté r_2 est représentée figure 7.6 (dans notre exemple, l'événement de transmission $a \leftarrow b$ induit un petit-fils d soit un r_2 de 1, la transmission $a \leftarrow c$ induit deux petits-fils, $r_2 = 2$). Cette distribution est fortement décroissante (apparemment exponentiellement dans le cas de la blogosphère américaine, les données sont trop bruitées dans le cas de la blogosphère française pour pouvoir conclure), on constate que plus de la moitié des événements de transmission sont sans suite (pourtant les sous-graphes de diffusion avec un seul lien même s'ils sont majoritaires dans la population des sous-graphes représentent à peine $1/5$ du total des liens de transmissions) ce qui semble indiquer que les sous-graphes de diffusion sont de faible longueur [4]. Contrairement à Leskovec et al. (2007b,a), nous ne nous intéressons pas ici à la topologie précise de ces cascades de diffusion, mais bien à la mesure et la compréhension de l'influence locale exercée par les blogs sur leur environnement.

7.2.3 Secondes transmissions et attention

Pour appréhender cette influence à plus longue distance, nous avons calculé le nombre total de petits-fils "contaminés" par une ressource provenant d'un blog doté d'une attention totale α donnée. La distribution de cette observable en fonction de l'attention totale des blogs est tracée figure 7.5 (courbes bleues). Les résultats sont extrêmement bruités pour la blogosphère française. Néanmoins, nous ne notons pas de variation significative entre la distribution du nombre de "fils" (pre-

4. Les sous-graphes de diffusion n'étant pas des arbres mais des graphes orientés acycliques, il serait plus correct de parler d'un faible diamètre.

tel-00464316, version 1 - 16 Mar 2010

mières transmissions, en rouge), et la distribution de "petits-fils". Dans les deux cas, nous observons une augmentation mécanique du nombre de descendants en fonction de l'attention totale du blog sans que le passage d'une génération à l'autre ne montre que l'attention totale α du blog source joue un rôle particulier sur les secondes transmissions.

7.3 Courts-circuits informationnels

Des mesures purement locales semblent peu propices pour nous éclairer sur la dynamique globale des cascades informationnelles. Quels paramètres topologiques sont à même d'expliquer qu'une information se propagera à plus ou moins longue distance dans un environnement donné ? Existe-t-il des déterminants structurels, propres au réseau, qui augmentent localement la probabilité d'une information à être reprise par la suite ? Pour aller au-delà de la caractérisation d'effets simplement liés à l'audience ou au lectorat d'un blog, nous nous sommes penchés sur la façon dont certaines propriétés dyadiques telles que le *edge range* pouvaient être corrélées à la probabilité de relayer une information. Nous nous concentrerons donc sur les "secondes transmissions" dans les sous-graphes de diffusion. Pour rappel, les *secondes transmissions* correspondent aux transmissions depuis un blog qui est déjà *un relais* pour une ressource. Autrement dit, les secondes transmissions sont liées à la *longévité* d'un épisode de diffusion. La question que nous posons est la suivante : une fois une ressource transmise d'un blog source à un blog relais, comment se poursuit la diffusion de la ressource dans la blogosphère depuis ce blog relais ?

Nous utilisons notre mesure d'*edge-range* pour caractériser la nature du premier lien de transmission entre le blog source, et le blog relais. Une valeur d'edge-range importante met en contact deux blogs qui seraient très distants en terme de temps de circulation de l'information si ce lien était absent. Un edge-range important signale donc un *lien faible* (Granovetter, 1973) au sens où il joue le rôle d'un raccourci informationnel entre deux zones du réseau relativement déconnectées.

7.3.1 Secondes transmissions et edge range

Nous souhaitons vérifier l'hypothèse selon laquelle une information transitant à travers un lien faible pourrait se propager à plus de blogs par la suite, le lien faible faisant figure de raccourci informationnel. Pour tester cette hypothèse nous mesurons le nombre de liens de transmission dans chaque sous-graphe de diffusion en fonction du edge-range du lien par lequel la ressource a initialement transité. Autrement dit, étant donné un blog i qui participe à un sous-graphe σ_u (on appellera i l'initiateur), un second blog $j \in \sigma_u$ qui cite l'initiateur i tout en mentionnant u (j joue le rôle du blog relais), on examine alors le nombre de blogs

tel-00464316, version 1 - 16 Mar 2010

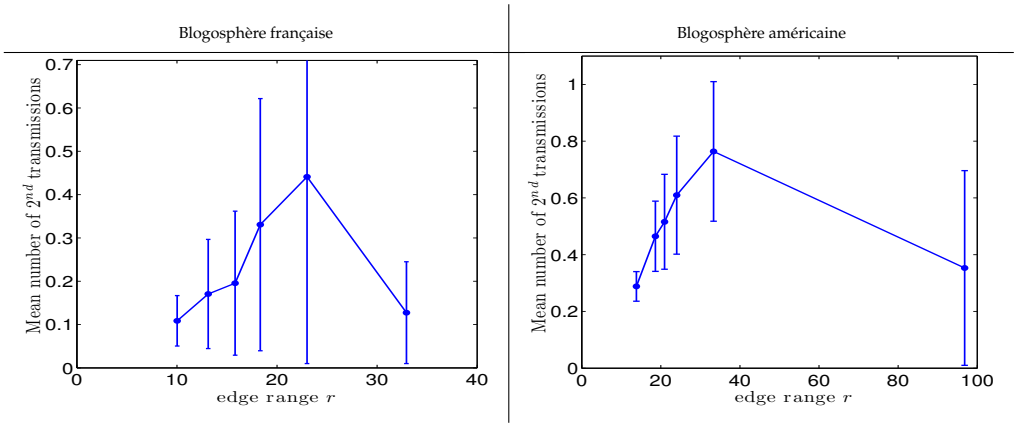


FIGURE 7.7: Nombre moyen de secondes transmissions (k, j) en fonction de la valeur d'edge range pondéré $\mathbf{r}(j, i)$ de la première transmission. Pour pallier le manque de données, nous avons divisé les différentes valeurs de $\mathbf{r}$ en 6 quantiles pour les deux réseaux.

k dans σ_u tel que $(k, j) \in \sigma_u$, en fonction de la *faiblesse* du lien entre le blog relais et l'initiateur : $\mathbf{r}(j, i)$. Ainsi, et toujours en suivant notre exemple figure 7.3, on cherche à corréler le nombre de petits fils induits par a à travers les blogs relais b et c (respectivement $r_2 = 1$ et $r_2 = 2$) aux valeurs d'edge-range respectives $\mathbf{r}(b, a)$ et $\mathbf{r}(c, a)$.

On a représenté, figure 7.7, le nombre moyen de secondes transmissions induites par une information ayant transité par un lien d'un edge-range donné (les valeurs de $\mathbf{r}$ ont été regroupées en 6 quantiles). Les intervalles de confiance obtenus dans le cas de la blogosphère française rendent délicate toute interprétation définitive, mais les résultats sont plus concluants concernant notre second jeu de données. Les ressources ayant transité à travers des liens dont les $\mathbf{r}$ sont plus importants tendent à diffuser vers un nombre plus grand de petits-fils. Les liens faibles agiraient ainsi comme des catalyseurs pour la diffusion. Néanmoins cette observation doit être nuancée pour les très fortes valeurs de edge-range, pour lesquelles on constate un décrochage du nombre de secondes transmissions [5].

Ce résultat est également cohérent avec le fait qu'une information qui transite par un lien *fort* entre un blog i et j ($\mathbf{r}(j, i)$ faible) a également de fortes chances d'être directement et rapidement accessible à des blogs tiers dans le voisinage immédiat de j (qui ont une forte probabilité d'appartenir également au voisinage de i) directement à la source : i. Le "coût de circulation de l'information" pour des voisins de j est clairement supérieur dans le cas où celle-ci a emprunté un lien faible, car les blogs à même de propager l'information dans le voisinage de j sont alors plus distants de i. Le blog j a ainsi produit un court-circuit informationnel entre ses voisins et i. La notion de court-circuit informationnel peut également être

5. Les ressources ayant été relayées à travers un lien particulièrement faible pourraient manquer de pertinence dans l'environnement relationnel **et** thématique du blog relais.

tel-00464316, version 1 - 16 Mar 2010

rapprochée du concept de trous structuraux (Burt, 1992). Dans cette perspective, le *capital social* d'un agent n'est pas mesuré à l'aune du nombre de ses relations mais en fonction de sa capacité à créer des ponts de part et d'autre de *trous structuraux*. Cette théorie recoupe également le concept d'intermédiarité développé par Freeman (1979) visant à mesurer la capacité d'un agent dans une position clé à constituer un point de passage obligé pour l'ensemble des chemins d'un réseau. De nombreuses études empiriques ont montré l'importance de cette notion : rôle des liens faibles pour trouver un travail (Granovetter, 1973), analyse de l'accroissement de l'influence et de la centralité de la famille Médicis à travers une "stratégie" systématique d'occupation des trous structuraux grâce aux mariages (Padgett and Ansell, 1993), créativité accrue de managers en situation d'intermédiaires (brokerage) entre plusieurs groupes dans une compagnie d'électronique (Burt, 2004) etc.

Ici nous pouvons donner une mesure quantitative de l'avantage pour un blog d'être connecté à un blog par un lien faible (ou suffisamment faible, *i.e.* un *lien moyen*). Les ressources transitant par ce type de lien ont tendance à être plus souvent reprises par des blogs tiers, et augmentent, ce faisant, l'influence du blog relais.

7.3.2 Effets couplés

Nous avons mis en évidence le rôle de l'attention totale dans la production de liens de transmissions, et avons constaté qu'une information ayant préalablement transité via un "lien moyen" avait plus de chance de se propager (d'un facteur supérieur à 2 lorsqu'on compare, , dans le cas de la blogosphère américaine, les deux configurations "extrêmes" suivantes : $\mathbf{r} < 20$ ou $\mathbf{r} \approx 40$).

Nous pouvons illustrer les effets combinés de α et $\mathbf{r}$ en représentant figure 7.8 le nombre moyen de secondes transmissions observées en fonction de la valeur d'edge-range $\mathbf{r}$ de la première transmission et de l'attention α du blog relais. Ces courbes montrent à nouveau l'importance de l'attention totale d'un blog vis-à-vis de son influence. Nous observons également que le edge range peut très sensiblement catalyser ou contrecarrer l'effet de l'attention du blog relais en augmentant sensiblement le nombre de petits-fils après qu'une information a transité à travers un "lien moyen".

Si l'influence d'un agent peut réellement être mesurée à travers le nombre de transmissions dont il est la source, alors on peut conclure, au regard de cette dernière figure (et principalement à l'aide du profil de la blogosophère américaine), que l'influence d'un blog est à la fois fonction de l'attention qu'il suscite dans son environnement - simple capital social agrégé - et de sa capacité à servir de relais d'information entre des zones relativement éloignées du réseau. Cette dernière caractérisation de l'influence nous rapproche de la vison du capital social de Burt (2000, 2004). Nous avions également déjà signalé, dans le chapitre précédent, la pertinence de la notion connexe de liens faibles dans la compréhension de la dy-

tel-00464316, version 1 - 16 Mar 2010

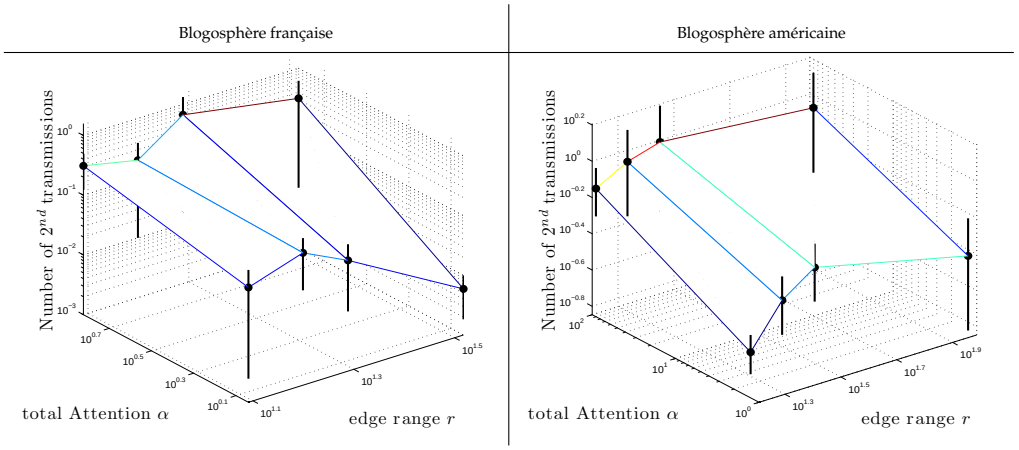


FIGURE 7.8: Nombre moyen de secondes transmissions $j \leftarrow k$ en fonction de la valeur d'edge-range $\mathbf{r}(j,i)$ de la première transmission et de l'attention du blog relais $\alpha(j)$. Les valeurs de α et de $\mathbf{r}$ correspondent à nouveau aux quantiles de leur distribution respective.

namique d'une diffusion sur un réseau social (section 6.2.1).

7.3.3 Conclusion

Nous avons montré comment la circulation d'information pouvait être affectée par des paramètres locaux (microscopiques) du réseau social supportant la diffusion : au niveau égo-centré, l'attention totale qu'exerce un blog sur son environnement augmente son influence. Identifier les blogs les plus influents est notamment une question stratégique pour le marketing viral dont l'approche classique (largement hérité du "2-step flow model" de Katz and Lazarsfeld (1955) consiste à cibler des "leaders d'opinion" censés relayer le message initial au plus grand nombre.) Cette vision a été fortement critiquée notamment par Watts and Dodds (2007) car elle met uniquement en exergue le début du phénomène complexe que constitue le processus diffusion. Ainsi d'autres propriétés, mesurées à un niveau moins égocentré et tenant compte de la structure environnante du réseau et des transmissions ultérieures, telles que la corrélation apparente entre edge-range et secondes transmissions, offrent une illustration quantitative de la capacité des "liens moyens" à significativement augmenter les capacités de diffusion d'une ressource dans ces communautés en ligne. L'influence d'un blogueur ne peut donc pas être réduite à son audience active, la structure du réseau l'environnant, et même l'origine des informations qu'il véhicule, semblent avoir un effet tout aussi capital sur sa capacité d'influence.

Au delà de la caractérisation plus ou moins complexe de l'influence des blogs en fonction de propriétés structurelles, nous avons laissé de côté une question primordiale dans la compréhension de ces dynamiques d'influence : la coévolution entre la structure du réseau social et les motifs d'influence entre blogs. Nous

tel-00464316, version 1 - 16 Mar 2010

avons, en première approximation, dans cette étude mesuré l'ensemble des propriétés structurelles des agents en fonction d'un réseau agrégé sur l'ensemble de la période d'observation. Pourtant, ce réseau est dynamique ; les épisodes de diffusion font par nature évoluer le réseau social des blogs. Diffusion et dynamique du réseau sont en quelque sorte co-constitutifs. il faudrait dès lors étudier non seulement la façon dont l'influence d'un acteur sur son environnement évolue dans le temps en fonction de la dynamique du réseau, mais aussi la façon dont cette influence est susceptible de déformer à plus ou moins longue distance le réseau autour de cet acteur.

Nous pouvons également nous interroger sur la stabilité de ces résultats en fonction du type de réseau sur lequel la diffusion se déroule, ainsi qu'en fonction du type de ressource qui y transite. Nous nous attendons à observer une certaines variabilité dans les comportements individuels selon que l'on observe la diffusion d'une recommandation, d'une innovation, ou même d'une nouvelle. Certains propriétés structurelles des réseaux sociaux pourraient donc s'avérer plus ou moins "propices" en fonction du type de diffusion et un agent influent vis-à-vis d'une catégorie pourrait s'avérer peu influent vis-à-vis d'une autre.

On peut également s'interroger sur le rôle des processus de diffusion dans la circulation d'idées et de concepts en Science. Malheureusement, comme on l'a déjà signalé, les réseaux sociaux que l'on peut reconstruire à partir de l'activité scientifique sont relativement pauvres : on ne peut, pour l'instant, que reconstruire des réseaux de collaboration ou des réseaux de citation qui ne traduisent pas l'ensemble des influences intellectuelles auxquelles les scientifiques sont soumis (qui peuvent transiter *via* des lectures, des conférences auxquelles les chercheurs participent, etc.). Et même si nous considérons ce type de réseau, sans doute lacunaire, deux obstacles importants se dressent pour définir des protocoles expérimentaux susceptibles de retracer des épisodes de diffusion empiriques tels que nous les avons observés dans la blogosphère.

D'une part, il faut identifier des d'entités suffisamment stables et atomiques (le pendant de nos URLs) afin de pouvoir établir sans ambiguïté le moment où un individu adopte. D'autre part, il s'agit de repérer les événements de transmission entre individus afin de pouvoir retracer la circulation de cette entité à travers le réseau social. Même si la citation peut parfois jouer ce rôle de marqueur d'une transmission d'information, il sera sans doute nécessaire d'employer des méthodes d'inférence pour reconstruire l'intégralité du parcours d'une information. Concernant la nécessité de définir des entités atomiques traçables, le langage naturel, aussi technique soit-il et hormis quelques exceptions assez limitées, semble, pour l'instant, relativement peu propice à cet exercice. Il est sans doute nécessaire dans ce cas d'en appeler à une modélisation de la diffusion différente, faisant appel, non pas à la circulation de concepts, mais à des observables d'un autre ordre. En déplaçant quelque peu notre ambition, on pourrait ainsi examiner dans ce type de communauté de savoirs, la façon dont la participation à une conférence

tel-00464316, version 1 - 16 Mar 2010

se propage au sein d'un réseau de collaboration, ou encore, la façon dont les individus adoptent ou non les marqueurs textuels propres à un champ épistémique (tel qu'on les a décrit dans le chapitre 4) en fonction de leur environnement social.

L'avantage de telles approches est qu'elle nous permettent également de quantifier le "coût" des déplacements opérés. Alors qu'il est, *a prioiri*, peu évident, en l'état, de quantifier la pertinence d'une information (ou d'une URL) dans une sphère informationnelle donnée (en somme d'évaluer sa *fitness* pour reprendre le vocabulaire de la mimétique), la métrique que nous avons définie lors de notre travail de cartographie des sciences pourrait être d'une grande aide pour quantifier les déplacements individuels (section 4.6.2). Dès lors, on pourrait mesurer l'influence qu'un individu exerce sur un autre non pas comme une mesure brute du nombre de ressources qui transitent de l'un à l'autre, mais de façon plus géométrique, en quantifiant directement les déplacements opérés par chacun au sein d'un paysage conceptuel.

tel-00464316, version 1 - 16 Mar 2010

Résumé du chapitre:

Dans ce chapitre, nous nous sommes attachés à définir un protocole expérimental d'observation de cascades informationnelles. En suivant la reprise d'URLs au sein d'une communauté de blogueurs, nous avons pu mettre en évidence un certain nombre de corrélations entre la structure locale du réseau qui supporte le processus de diffusion, et la longévité d'un épisode de diffusion.

Globalement notre objectif a été de trouver les déterminants structurels de l'influence qu'exerce un blog sur son environnement. Nous avons observé dans un premier temps que les blogs les plus populaires, au sens où ils attirent naturellement une plus grande attention, sont également les plus à même d'induire un grand nombre de transmissions.

Au delà de ces effets de lectorat actif, nous avons également mesuré quantitativement le rôle des liens faibles dans la propagation d'une ressource. Une ressource ayant transité *via* un lien faible (ou plus précisément un *lien moyen*) a une probabilité plus importante d'être à nouveau propagée que si elle avait transité *via* un lien fort. L'influence d'un blog dépend donc, à la fois, de son audience active, et de sa capacité à produire des "courts-circuits informationnels" entre deux zones du réseau "distantes" l'une de l'autre.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

Conclusion

NOUS avons, au cours de cette thèse, exploré les dynamiques sociales, socio-sémantiques et sémantiques des communautés de savoirs. Celles-ci ont été définies comme des espaces hybrides. Elles mettent en jeu, d'une part, des individus qui interagissent au sein d'un *réseau social* et qui produisent des contenus liés à un domaine donné, et d'autre part, des entités sémantiques liées au sein d'un *réseau sémantique* doté d'une structure autonome. Les dynamiques des dimensions sociale et sémantique sont soumises à des couplages à différents niveaux, et en premier lieu au niveau micro *via* le *réseau socio-sémantique* qui décrit l'usage des concepts par les individus. Les trois réseaux, que nous venons de mentionner, sont regroupés au sein d'un *réseau épistémique* qui nous a permis de modéliser les communautés de savoirs en intégrant à notre analyse l'ensemble des interactions existant entre les différentes entités. À ce titre, le réseau épistémique permet d'appréhender les dynamiques sociales et culturelles d'une façon aussi symétrique que possible, tout en offrant les conditions nécessaires à leur couplage.

Deux terrains d'étude nous ont permis d'interroger ces dynamiques entremêlées : des communautés scientifiques dont les membres produisent individuellement et collectivement des énoncés scientifiques au sein de publications, ainsi que des communautés de blogueurs formant une arène de discussion virtuelle, la blogosphère politique, qui se définit à la fois comme un territoire d'expression dans l'espace public et comme un lieu d'échanges et de mises en relation entre blogueurs. Ces deux communautés peuvent être définies comme des systèmes socio-cognitifs distribués, *i.e.* les dynamiques qui les animent — qu'elles mobilisent des entités sociales, sémantiques, ou les deux — sont toujours d'ordre local. Pour autant, ces systèmes ne sont pas dénués de propriétés structurelles émergentes.

Notre exploration de ces communautés de savoirs s'est appuyée sur la collecte des traces textuelles résultant de l'activité des individus qui animent ces espaces. À partir de ces traces empiriques, notre objectif a été de reconstruire la phénoménologie de ces communautés de savoirs en nous efforçant d'analyser leurs dynamiques à différents niveaux. Nous avons ainsi distingué, au *niveau micro*, les dynamiques individuelles qui modifient les arrangements locaux du réseau épistémique, et au *niveau macro*, les dynamiques des motifs émergeant des interactions locales et qui structurent le réseau épistémique. En plus de ces interrogations portant sur la morphogenèse du réseau épistémique, nous avons également appréhendé les propriétés dynamiques émergentes de ces communautés en tentant de mieux saisir les processus de diffusion d'information qu'elles supportent.

Morphogenèse des réseaux. Les questions de morphogenèse (partie II) se sont articulées en deux approches selon que nous prenions comme point de départ les dynamiques micros ou macros de nos communautés de savoirs. À partir de l'observation des blogosphères politiques américaine et française (chapitre 3) nous avons caractérisé la morphogenèse des communautés de savoirs en nous focalisant, au *niveau micro*, sur les régularités auxquelles sont soumises les dynamiques locales du réseau épistémique. Nous avons pu identifier certains comportements individuels réguliers pouvant participer à la stabilisation de certains motifs émergents dont nous avons également évalué la stabilité. Ainsi, nous avons montré que le capital social, mais également le capital sémantique, des agents augmentent leur attractivité vis-à-vis de nouveaux liens de citation, cette propriété étant susceptible d'expliquer la distribution hétérogène du capital social dans ces espaces ainsi que sa stabilité. On a également observé une tendance à l'*homophilie sémantique* ainsi qu'une propension à interagir avec des agents proches dans le réseau social, qui peuvent tendre à renforcer l'homogénéité des agrégats sociaux (*transitivité*), et socio-sémantiques existants. Au delà de ces effets de "sélection", nous avons également mis en évidence de façon quantitative les conséquences de processus locaux d'*influence sociale* sur les profils sémantiques des blogueurs. L'ensemble de ces observations a donc permis d'illustrer la richesse des corrélations existantes entre dimensions sociale et sémantique, accréditant notre thèse d'une co-évolution entre tissu relationnel inter-individuel et production de contenus.

L'originalité de ces résultats tient à la façon dont ils montrent combien les dynamiques individuelles sont intimement liées à des paramètres ne rentrant pas classiquement dans le cadre de l'analyse des réseaux sociaux. Tout en restant fidèles à l'impératif anticatégorique qui caractérise cette approche, nous avons montré que l'introduction d'une dimension sémantique, plutôt que de complexifier la description du système, l'a simplifiée. Comparativement à un examen inhérent à la seule sphère sociale, le cadre des réseaux épistémiques, en nous permettant d'identifier ces régularités d'ordre sociosémantique dans le comportement des acteurs, a permis de rendre les dynamiques micros plus *prédictibles*.

Phylogenèse des réseaux. Au *niveau macro* (chapitre 4) nous nous sommes efforcés de caractériser les motifs structurels que constituent les agrégats d'entités densément interconnectées. La structure et l'articulation de ces ensembles mésoscopiques, qui émergent des dynamiques individuelles (dans les réseaux social, socio-sémantique, et sémantique), révèlent l'organisation de haut-niveau de nos communautés de savoirs. Nous nous sommes penchés tout particulièrement sur la reconstruction des dynamiques multi-échelles des communautés scientifiques à partir de statistiques d'occurrences et de cooccurrences de termes extraites de grands corpus de publications. Nous avons proposé un ensemble de méthodologies, qui nous a permis de reconstruire, de façon statique d'abord, à différentes échelles, des cartes des sciences structurées prenant la forme de réseaux de champs

tel-00464316, version 1 - 16 Mar 2010

épistémiques, puis de façon dynamique, le réseau phylogénétique de ces champs épistémiques révélant les dynamiques mésoscopiques de transformations conceptuelles et de fertilisation croisée entre champs qui font évoluer ces paysages sémantiques. La structure de ces réseaux phylogénétiques semble également riche d'information et fait apparaître un certain nombre de motifs dynamiques robustes remarquables. Enfin, nous avons proposé une méthode de reprojection des individus immergés dans ces espaces. Nous avons ainsi pu montrer comment la topologie formée par un paysage sémantique pouvait rétroagir du niveau macro au niveau micro en "orientant" le déplacement des individus plongés dans ce paysage.

Ces méthodes de reconstruction ouvrent la voie vers une épistémologie quantitative qui permettra, à terme, d'observer et de questionner les dynamiques scientifiques avec des outils d'une précision inédite. La détection automatique de l'organisation interne de ces espaces permet d'en simplifier grandement la représentation. Nous pouvons maintenant imaginer de nouveaux modes de navigation nous permettant d'évoluer à travers la structure multi-échelle et mouvante de très grands corpus de textes. Ces méthodes pourront naturellement équiper des théoriciens des sciences, des gestionnaires, ou même des chercheurs désireux de mieux comprendre les creux et les reliefs des paysages sémantiques dans lesquels ils sont immergés. Dotés de ce type d'outils, les agents d'un système seront à même d'observer les conséquences parfois insoupçonnées de leurs propres actions. Il semble alors pertinent d'interroger les effets de cette opération réflexive sur les dynamiques futures des agents et du système dans son entier.

A plus long terme, ces reconstructions pourraient aussi servir de laboratoire d'expérimentation pour éclairer certaines questions que posent le domaine de l'évolution culturelle. Ainsi, la trajectoire d'un champ épistémique peut très bien se lire comme la manifestation d'un phénomène culturel transitoire, la reconstruction de ses dynamiques, mais également, la connaissance de ses " traits d'histoire de vie", offrent de nombreuses opportunités pour en interroger les déterminants. Quels sont les critères de viabilité d'un champ épistémique ? Peut-on rendre compte de la dynamique d'un paysage conceptuel et des individus qui s'y déplacent avec des modèles évolutionnistes ou mémétiques ? Au-delà de l'interrogation des dynamiques scientifiques, ces questions peuvent et doivent être étendues à d'autres espaces de production de connaissance. Observerons-nous les mêmes lois d'évolution, la même richesse dynamique si nous appliquons, par exemple, ces méthodes à des corpus de textes légaux, ou à des corpus textuels collectés dans les media sociaux ?

Dynamiques d'influence et diffusion des contenus. Enfin, nous nous sommes focalisés dans la partie III sur les processus de diffusion qui animent nos communautés de savoirs et qui forment une des modalités de couplage possible entre dimensions sociale et sémantique. Nous avons adopté deux approches pour com-

tel-00464316, version 1 - 16 Mar 2010

prendre les phénomènes de diffusion dans les communautés de savoirs.

Dans un premier temps, chapitre 5, nous nous sommes attachés à la question de l'influence entre différents groupes de blogs marqués politiquement et exposés au flux d'actualités provenant de la presse. Nous avons cherché à détecter des motifs inter-temporels systématiques de reprise de tel ou tel concept à partir de l'analyse longitudinale des contenus produits par ces sources. Nous avons ainsi pu résumer, au sein d'un *diagramme d'influence* synthétique, l'ensemble des corrélations liant dynamiquement les activités de production d'un groupe de sources à l'activité de production future d'un autre groupe de sources. L'originalité de cette analyse réside dans le fait qu'elle a permis de reconstruire une grande variété de modalités de mise sous influence entre groupes de sources, ne se limitant pas à des corrélations classiques deux à deux.

Dans un second temps, chapitres 6 et 7, nous avons envisagé la question de la diffusion sur un réseau social en nous interrogeant sur l'influence des structures (locales et globales) du réseau social vis-à-vis des processus de diffusion. Dans une perspective macroscopique, nous avons proposé un protocole simulatoire (chapitre 6) afin de comprendre en quoi la topologie d'un réseau affecte la dynamique globale de cette diffusion. En comparant systématiquement les dynamiques observées sur une série de réseaux, dont la topologie reproduit un certain nombre de propriétés structurelles de deux réseaux de terrain réels, nous avons identifié les propriétés topologiques à même d'influer de manière significative sur la vitesse de diffusion. Nous avons ainsi montré, que pour une hypothèse de transmission inter-individuelles donnée, la vitesse de diffusion est indépendante de la forme de la distribution de degré du réseau support du processus de diffusion. *A contrario*, la cohésion locale, ou de façon plus large, la structure naturellement modulaire des réseaux sociaux joue un rôle primordial vis-à-vis de la diffusion. Les zones densément connectés du réseau, isolées du reste du réseau, tendront, soit à emprisonner l'information qu'elle détiennent, soit à ne participer que plus tardivement à la diffusion. De façon générale, nous avons observé que les réseaux réels étaient, vis-à-vis du mécanisme de transmission simulé beaucoup plus lents que d'autres réseaux aléatoires possédant la même densité de liens.

Enfin, nous avons analysé, dans le chapitre 7, un processus de diffusion réel à un niveau plus micro, grâce au suivi *in-vivo* d'épisodes de diffusion d'URLs observés au sein des blogosphères politiques française et américaine. Notre protocole expérimental nous a permis de construire l'un des premiers jeux de données permettant d'observer simultanément un réseau social et le déroulement détaillé des cascades de diffusion qui s'y déploient. Cette analyse nous a permis de mettre en évidence un certain nombre de corrélations entre la structure locale du réseau, et la longévité d'un épisode de diffusion. Nous avons ainsi montré que l'influence d'un agent dépend, à la fois, de son audience active, et de sa capacité à produire des "courts-circuits informationnels" entre des zones distantes du réseau.

Les observations que nous avons tirées de cette partie consacrée aux proces-

tel-00464316, version 1 - 16 Mar 2010

sus de diffusion dans les réseaux sociaux nous ramènent aux questions de morphogenèse traitées dans la partie précédente. La "résistance" qu'opposent les réseaux réels aux processus de diffusion étudiés est-elle la conséquence d'un principe d'optimisation de la structure du réseau vis-à-vis de ces types de processus ? Si la structure globale des réseau sociaux est effectivement optimale vis-à-vis de certaines propriétés, quels mécanismes locaux de morphogenèse peuvent alors en expliquer l'émergence ? Cette question est peut-être liée aux "stratégies" locales (conscientes ou inconscientes) que peuvent déployer les acteurs pour se positionner au sein d'un espace social donné afin de maximiser leur influence vis-à-vis de leur environnement. À ce titre, il serait intéressant d'étendre notre étude sur les cascades informationnelles à un cadre co-évolutionnaire plus large, permettant de coupler les dynamiques de diffusion aux dynamiques du réseau social, afin de chercher à déterminer si ces dernières tendent vers une optimisation locale de l'influence des acteurs, ou vers une optimisation globale de la forme des cascades de diffusion.

Une sociologie quantitative des traces. Plus largement, cette thèse a permis d'entrouvrir les potentialités offertes par un travail systématique de reconstruction des dynamiques soicosémantiques à partir des traces de la vie sociale. Les perspectives ouvertes par l'observation des systèmes sociaux *in-vivo*, sur Internet en particulier, nous rapprochent d'une *sociologie quantitative des traces*, dont Tarde était l'annonciateur. Néanmoins, la disponibilité de ces données massives ne garantit pas, en soi, un chemin direct vers la compréhension des faits sociaux. Il s'agit également d'apprendre à décrire, détecter, modéliser les structures remarquables que revêtent ces traces à travers un travail de reconstruction phénoménologique ardu qui appelle au concours de nombreuses compétences (modélisation, mathématiques, informatique mais également économie, histoire, anthropologie, sociologie).

L'analyse des systèmes sociaux *in-vivo* offre également l'opportunité d'interroger les opérations de cognition sociale au sens large. La *cognition sociale* est un processus de traitement de l'information qui, par comparaison à la cognition individuelle, s'applique de manière distribuée sur l'ensemble des membres d'une communauté en interaction au sein d'un réseau social. Ce traitement distribué, même s'il peut agir comme une contrainte, permet de réaliser des tâches inaccessibles à une traitement cognitif purement individuel. Le traitement des informations présentes dans l'environnement par le réseau social est susceptible de produire de nouvelles configurations sémantiques, mais aussi, de catalyser la création ou la disparition de certaines interactions inter-individuelles. C'est donc l'opération de cognition sociale qui transforme simultanément les réseaux social et sémantique. Notre travail durant cette thèse peut donc être lu comme une mise en évidence des lois et motifs remarquables que laisse apparaître cette opération de cognition sociale sur les deux types de systèmes envisagés.

tel-00464316, version 1 - 16 Mar 2010

Interroger plus globalement les processus de cognition sociale qui s'appliquent à d'autres types de systèmes appelle au développement de protocoles d'observation originaux. Les différentes reconstructions possibles de ces dynamiques pourraient maintenant être confrontées à l'aide de protocoles s'appuyant sur des données d'observation communes et s'accordant sur des critères de validité partagés. Au-delà de l'enjeu purement épistémique, il s'agit également de concevoir des infrastructures de communication et de gestion des connaissances susceptibles, par exemple, de rendre plus robustes ces processus de cognition sociale. Quel que soit la nature de ces systèmes, les media digitaux constituent certainement, vis-à-vis de la cognition sociale, une opportunité pour collecter des données massives à la fois sur les processus et les produits qu'elle génère. La connaissance et l'analyse conjointe du réseau social, support de toute opération de cognition sociale, et des contenus et décisions élaborées localement et collectivement peut, à ce titre, catalyser un *tournant cognitif* qui engage l'ensemble des sciences sociales.

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

ANNEXE A

Liste des termes associés à la blogosphère politique française

A.1 Liste des 190 syntagmes utilisés définir le bagage sémantique des blogs politiques français :

35 heures ; action publique ; aide au développement ; Allemagne ; altermondialisme ; antisémitisme ; baisse des prélèvements ; baisses d'impôts ; banlieue ; Banque européenne ; Bayrou ; blogosphère ; blogueurs ; bouclier fiscal ; Bové ; bravitude ; budget de la recherche ; budget de l'Etat ; capitalisme financier ; carte scolaire ; chiffres du chômage ; Chirac ; chômage ; classes moyennes ; CO2 ; collectivités locales ; écolo ; écologie ; communautarisme ; comptes publics ; Conseil d'analyse économique ; contrat de travail ; contribuables ; criminalité ; croissance ; débat public ; dette publique ; déficit budgétaire ; DIABOLISER ; dialogue social ; discrimination positive ; démocratie participative ; démocratie sociale ; Don Quichotte ; dépense publique ; drapeau français ; droit au logement ; éducation ; développement durable ; effet de serre ; emploi ; encadrement militaire ; endettement ; enseignement supérieur ; entreprises ; EPR ; Eric Besson ; Etats membres ; Europe ; finances publiques ; financier ; fiscal ; fiscalité ; FN ; fonction publique ; fonctionnaires ; François Bayrou ; François Hollande ; gauche antilibérale ; gaz à effet de serre ; hausse des prix ; hausse des salaires ; hausse du smic ; heures supplémentaires ; Hollande ; identité française ; identité nationale ; impôt sur les successions ; insécurité ; internautes ; Internet ; intérêt général ; islam ; islamisme ; Jacques Chirac ; Jean-Marie Le Pen ; Jospin ; jurys citoyens ; justice sociale ; Kärcher ; Kyoto ; législatives ; Lionel Jospin ; logement ; logement opposable ; logements sociaux ; LOLF ; maîtrise des dépenses ; MoDem ; monde agricole ; mondialisation ; Nicolas Sarkozy ; Olivier Besancenot ; pacte écologique ; pacte présidentiel ; pauvreté ; petites retraites ; peuple ; PIB ; plein emploi ; pouvoir d'achat ; prélèvements obligatoires ; productivité ; protection sociale ; prévention de la délinquance ; PS ; ps udf ; réchauffement climatique ; recettes fiscales ; recherche ; réforme des retraites ; régimes de retraite ; régimes spéciaux ; réforme ; référendum sur la Constitution ; régularisation ; Royal ; rural ; ruralité ; salaire minimum ; salariés ; sans-abri ; Sarkozy ; Sécurité sociale ; ser-

vice minimum ; service public ; Ségolène Royal ; socialiste ; solidarité ; sondages ; taux de chômage ; temps de travail ; territoire ; terrorisme ; Tony Blair ; traité constitutionnel ; travail ; TVA sociale ; UDF ; UMP ; Union européenne ; valeur travail ; Valérie Pécresse ; Verts ; vieillissement ; ville ; violences urbaines ; vote utile ; Xavier Bertrand ; zones rurales ; porte-avions ; homosexuel ; égalité des chances ; CSG ; droit de grève ; CNRS ; Gollnisch ; George Bush ; dialogue social ; contrat unique ; assurance maladie ; johnny ; technologies ; anti-Sarkozy ; troisième homme ; délocalisations ; précarité ; Frédéric Nihous ; Marseillaise ; Darfour ; chiffrage ; débats participatifs ; Villepinte ; Clearstream ; mai-68 ; outre-mer ; éléphants ; gare du Nord ; Iran ; Irak

A.2 Liste des termes associés à la blogosphère politique américaine

Liste des 79 syntagmes utilisés définir le bagage sémantique des blogs politiques américains :

iran ; global warming ; climate change ; primaries ; china ; dollar ; michigan ; john edwards ; gaza ; gay ; financial ; crime ; terror ; democracy ; british ; muslim ; faith ; palestin ; blog ; internet ; terrorist ; israel ; threat ; soldier ; iraq ; violen ; california ; afghan ; energy ; islam ; justice ; al qaeda ; clintons ; huckabee ; peace ; pakistan ; christian ; weapon ; immigrati ; woman ; mitt romney ; giuliani ; nuclear ; economy ; barack obama ; john mccain ; hillary clinton ; black ; security ; military ; democrats ; republicans ; money ; foreign policy ; jewish ; bush ; september ; abortion ; technology ; super tuesday ; national security ; middle east ; europe ; supreme court ; mexico ; human right ; kerry ; environment ; veterans ; george w bush ; war ; gun ; recession ; musharraf ; french ; tax cuts ; wall street ; vietnam ; africa

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

ANNEXE B

Corpus de termes des bases des domaines scientifiques explorés

B.1 Systèmes complexes

abduction ; accidents ; acid soils ; acoustics ; action ; Action observation ; adaptive control ; admissible controls ; admissible trajectory ; agent ; aggregate ; agriculture ; algorithms ; altruism ; analysis ; animal culture ; anticipations ; antisurvenance ; applied epistemology ; approximation algorithms ; architecture ; argumentation theory ; Artificial cell ; artificial intelligence ; assembly ; atmospheric physics ; automatic control ; automatic imitation ; autonomous agents ; avatar ; bayesian ; bayesian computing ; bayesian statistics ; behavioral economics ; behavioural theory ; belief ; best practice ; bifurcations ; biochemistry ; bioinformatics ; biosystems ; biotechnology ; border ; bounded rationality ; brain ; carbon dioxide ; cartography ; categorization ; category ; causal ; cell ; cellular networks ; chaos ; chemical ; chemical equilibria ; climatology ; coevolution ; cognitive economics ; cognitive ethology ; cognitive hierarchy ; cognitive psychology ; cognitive therapy ; collective ; collective action ; Collective Behaviours ; collective discovery ; collective intelligence ; collective memory ; collective rationality ; combinatorial optimisation ; communication ; community ; complex system ; complex systems ; complexity ; computational ; computational chemistry ; computer ; computer software ; computerized simulation ; concurrent engineering ; conformism ; consciousness ; constraints ; consumer behavior ; consumer sovereignty ; consumers ; contact ; contact geometry ; context ; continuity ; control ; critical phenomena ; cultural anthropology ; cultural evolution ; cultural learning ; cultural trait ; cultural traits ; cultural transmission ; cybernetics ; data acquisition ; data bases ; data Management ; data mining ; data modelling ; data security ; democracy ; density ; deontic logic ; dependability ; design ; development ; disasters ; distributed systems ; drosophila ; dynamic ; dynamic modelling ; dynamic reasoning ; dynamics ; ecological rationality ; ecology ; economic models ; econophysics ; electroencephalography ; embedded systems ; embodied agent ; embodied cognition ; emergence of cooperation ; emergent semantics ; empathy ; emulation ; enaction ; endogenous ; endogenous dynamics ; endogenous network ; endogenous networks ; endogenous preference ; energy ; energy consumption ; energy management system ; energy policy ; energy saving ; energy sources ; environment ; environment protection ; environmental technology ; epidemiology ; epilepsy ; epistemic authority ; ethnography ; evaluation ; evolution ; evolution of culture ; evolu-

tionary epistemology ; evolutionary games ; experimental development ; experimental economics ; finance ; finite element ; first person ; fitness ; forecasting techniques ; formal methods ; fossil fuels ; fractal ; framing effect ; freeware ; functional architecture ; functional programming ; fuzzy logic ; Galois lattices ; game theory ; gene expression ; general equilibrium ; genetic ; genetic algorithms ; geometry ; global network ; graph theory ; health ; herding behaviour ; heterogeneous agents ; heterogeneous networks ; heterogenous agents ; heuristic ; hiv ; human factors ; human modeling ; hybrid algorithms ; imitation ; immunogenetics ; inconsistency in reasoning ; individual ; inductive logic ; inference ; infinite dimensional ; informatics ; information system ; innovation ; integration ; integrative ; intelligence ; intelligent robot ; intentional action ; intentionality ; interactivity ; interface ; intersubjective relation ; intersubjectivity ; interview method ; intuitive experience ; invariant ; knowledge discovery ; knowledge managment ; learning ; life sciences ; logistics ; machine ; machine learning ; management ; manifold ; mapping ; mathematical models ; mathematical sociology ; maximizator ; measuring instruments ; meme ; memes ; memetics ; mereotopology ; metacognition ; mimetism ; mining technology ; minority game ; mirror neuron ; modeling ; modelling ; molecular ; monogenic ; monte carlo ; morphodynamic ; morphodynamics ; morphogenesis ; N400 ; nanotechnology ; nash ; natural language processing ; network formation ; network management ; networks ; neural network ; neural networks ; neurobiology ; neurolinguistics ; neuron ; neuroscience ; neurosciences ; noise ; non deterministic ; nonlinear control ; nuclear ; oil ; online ; open systems ; operational closure ; operations research ; optimisation ; optoelectronics ; ordinary differential equations ; organizational theory ; ozone ; P600 ; paleoclimatology ; parallel computing ; parallel processing ; parallel systems ; parallelism ; parameter convergence ; pareto ; participatory ; pattern formation ; percolation ; performance evaluation ; petroleum technology ; phase transition ; phenomenology ; phylogeny ; planning ; policy ; pollen ; population ; power generation ; power law ; prediction ; predictions ; preference ; pretend play ; priming ; prisoner's dilemma ; probability theory ; procedural invariance ; procedural rationality ; process management ; process quality ; processing ; product quality ; programming ; project management ; proof theory ; propagation ; prosociality ; prototyping ; Public Goods ; public services ; qualitative reasoning ; quantum physics ; radioactive waste ; rational imitation ; reasoning ; reconstruction ; reflective ; reflectivity ; reflexive ; regulons ; reliability ; religion ; renewable ; renormalization ; repair ; resource allocation ; resources substitution ; reverse engineering ; risk management ; risky choice ; ritual ; robotics ; robustness ; sacrifice ; scaling ; scientific discovery ; scientometrics ; second person ; security ; selfish ; sensors ; side effects ; sign ; signal processing ; signification ; simulation ; singularity ; situated agent ; small world ; social ; social aspects ; social belief ; social cognition ; social correlations ; social differenciation ; social dilemma ; social learning ; social networks ; software ; software engineering ; solar energy ; space ; spatial games ; stability ; state constraints ; statistical data ; statistical physics ; statistical testing ; stigmergic ; stochastic game ; stochastic game theory ; stochastic processes ; stochastic stability ; structuralism ; structure ; structuring principle ; subjective expe-

tel-00464316, version 1 - 16 Mar 2010

rience ; substantial rationality ; supervenience ; survenance ; sustainability ; sustainable development ; swarm intelligence ; symbiosis ; symbolic computation ; synthesis ; taxonomies ; theory of mind ; theory of uncertainty ; theory theory ; therapy ; thermodynamics ; topology ; traceability ; traffic simulation ; transfer functions ; transgenic ; trust ; tychastic ; Ultimatum Game ; uncertainty modelling ; verification ; viability ; victimary mechanism ; violence ; virtual ; vision ; vision system ; visual sensor ; visualisation ; waste management ; wavelets ; wiki ; wind energy ; workflow ; world wide web ; zebrafish

B.2 La métaphore réseau en biologie

ability ; absorption ; accumulation ; acid residues ; actin ; action potentials ; activate ; activated protein ; active site ; activity ; adaptation ; addition ; adhesion ; affinity ; aggregation ; algorithm ; allele ; alpha ; amino acid ; amplification ; anaphase ; annotation ; antibodies ; antibody ; antigen ; apoptosis ; approximation ; arabidopsis ; architecture ; arp2 ; array ; artificial neural network ; association ; atom ; attractor ; autonomous ; autoregulation ; auxin ; average ; axon ; bacillus ; background ; bacterium ; bank ; base ; bayesian ; bayesian network ; behavior ; bifurcation ; bind ; binding protein ; binding site ; biochemical pathways ; biodiversity ; bioinformatics ; biological data ; biological functions ; biological networks ; biological process ; biological systems ; biology ; biomass ; biomolecular ; biosynthesis ; blood flow ; body ; bond ; bond network ; brain ; branching ; budding ; budding yeast ; building ; bulk ; burst ; calcium ; cancer ; candidate ; capabilities ; capacity ; carbon ; carbon dioxide ; cardiac ; cascade ; catalysis ; cause ; cdc14 ; cdna ; cell ; cell cycle ; cell death ; cell differentiation ; cell division ; cell fate ; cell growth ; cell migration ; cell motility ; cell surface ; cell wall ; cellular ; cellular functions ; cellular networks ; cellular process ; center ; central nervous ; cerevisiae ; chain ; channel ; checkpoint ; chemistry ; chip ; chromatin ; chromosome ; circadian ; circadian clock ; circadian rhythms ; circuit ; circulation ; class ; classification ; clathrin ; climate ; clock ; clone ; cluster ; clustering ; code ; codon ; coexpression ; cofactor ; cognitive ; coherent ; coli ; collagen ; combine ; community ; comparing ; competition ; complete ; complex ; complex network ; complex systems ; complexes ; complexity ; component ; composition ; computation ; computational method ; computational model ; computer ; computing ; concentration ; conformation ; conformational change ; connecting ; connectivity ; conservation ; constraint ; construct ; control ; convergence ; cooling ; cooperative ; cooperativity ; coordination ; core ; correlation ; cortex ; coupling ; cross validation ; crystal structure ; cycle ; cyclin ; cycling ; cytochrome ; cytokine ; cytokinesis ; cytoplasm ; cytoskeleton ; data bank ; data mining ; data set ; database ; decay ; defective ; deformation ; degradation ; degrees c ; dehydrogenase ; density ; dependence ; detect ; determinants ; determine ; developing ; development ; dictyostelium ; differential equation ; differentiation ; diffraction ; diffusion ; dimer ; dioxide ; discrete ; discrimination ; disease ; disruption ; distribution ; diverse ; diversity ; divi-

tel-00464316, version 1 - 16 Mar 2010

sion ; dna binding ; dna damage ; dna interactions ; dna microarray ; dna repair ; dna replication ; dna sequence ; dopamine ; downstream ; drosophila ; drosophila melanogaster ; dynamical systems ; dynamics simulations ; ecology ; ecosystem ; edge ; effective ; effector ; efficient ; electron microscopy ; elegans ; elegans ; embryo ; emerge ; emergence ; encode ; endocytosis ; endogenous ; endoplasmic reticulum ; energy ; engineering ; enhancement ; entropy ; environment ; environmental conditions ; enzyme ; epidermal growth ; epistasis ; epithelial cells ; equation ; escherichia ; escherichia coli ; eukaryotes ; eukaryotic cells ; event ; evolution ; evolvability ; evolve ; exchange ; excitation ; exon ; experiment ; experimental evidence ; expressed genes ; expression ; expression data ; expression level ; expression pattern ; expression profiling ; extinction ; extracellular ; families ; feedback ; feedback control ; feedback inhibition ; feedback loop ; feedback mechanism ; feedback regulation ; feedforward ; fibroblasts ; filament ; firing ; fission ; fission yeast ; flexibility ; flexible ; flow ; fluorescent ; flux ; fold ; food web ; form ; formalism ; framework ; frequencies ; function ; functional annotation ; functional genomics ; functional modules ; fusion ; fuzzy ; gain ; galaxy ; gaussian ; genbank ; gene ; gene duplication ; gene expression ; gene expression data ; gene expression levels ; gene expression patterns ; gene expression profiles ; gene function ; gene interactions ; gene network ; gene ontology ; gene products ; gene regulation ; gene regulatory network ; gene transcription ; genes encoding ; genetic algorithm ; genetic interactions ; genetic network ; genetic regulatory network ; genetic variation ; genome ; genome scale ; genome sequence ; genome wide ; genomic data ; genotype ; germ ; global network ; globular ; glutamate ; golgi ; golgi network ; gradient ; graph ; growth ; growth rate ; gtpase ; helix ; heterogeneity ; heterogeneous ; hidden markov ; hierarchical clustering ; hierarchy ; high resolution ; high throughput ; high throughput data ; higher order ; hippocampus ; histone ; hiv ; homeostasis ; homogeneous ; homologous ; homology ; hormone ; host ; hub ; hubble ; hubble space ; hubble space telescope ; human ; human brain ; human genome ; human protein ; hybrid ; hybridization ; hydrogen bond network ; hydrogen bonding ; image ; imaging ; immune ; immune response ; immunity ; implementation ; in vitro ; inactivation ; independent ; indirect ; infection ; influences ; information ; information processing ; inhibitor ; input ; instability ; insulin ; integrate ; integrative ; interact ; interacting ; interaction ; interaction data ; interaction map ; interaction network ; interactive ; intercellular ; interface ; interference ; intermolecular ; interneurons ; interplay ; intracellular ; intracellular signaling ; involve ; kappa ; kappa b ; kinase ; knockout ; large scale ; laser ; lattice ; layer ; learning ; level ; ligand ; ligand binding ; limb ; lineage ; linear ; link ; linkage ; lipid ; liver ; local structure ; localization ; locomotor ; locus ; long range ; loop ; lymphocytes ; machine learning ; machinery ; macromolecular ; magnetic resonance ; mammalian ; management ; mapping ; marker ; markov ; mass spectrometry ; mathematical model ; mating ; matrix ; maturation ; mechanism ; melanogaster ; membrane ; messenger ; messenger rna ; metabolic control ; metabolic network ; metabolic pathway ; metabolite ; methane ; mice ; microarray ; microarray data ; microarray datasets ; microarray experiments ; microarray gene ; microscopy ; microtubule ; migration ; mitochondria ; mitosis ; model ; mo-

tel-00464316, version 1 - 16 Mar 2010

delling ; modification ; modular ; modulate ; modulation ; module ; molecular ; molecular basis ; molecular biology ; molecular dynamics ; molecular interaction ; molecular level ; molecular networks ; molecule ; monitoring ; monte carlo ; morphogenesis ; morphology ; motif ; motility ; motor ; mouse ; mrna ; multicellular ; multicellular organisms ; mutagenesis ; mutant ; mutation ; myosin ; natural selection ; negative feedback ; nematoda ; nervous ; network ; network architecture ; network model ; network motifs ; network structure ; network topology ; networking ; neural network ; neuron ; neurotransmitter ; neutron ; nf kappa ; nf kappa b ; niche ; nitrogen ; node ; noise ; non coding ; nonlinear ; notch ; nuclear ; nucleation ; nucleotide ; nucleotide sequence ; nucleus ; nutrient ; occurrence ; oligonucleotide ; ontology ; open ; operate ; operon ; optimization ; organism ; organization ; organizing ; oscillation ; output ; overexpression ; overlapping ; oxygen ; pairwise ; paradigm ; parameter ; participate ; pathogen ; pathway ; pattern ; pattern formation ; pattern recognition ; patterning ; peptide ; perform ; performance ; persistence ; perturbation ; petri ; phage ; phase ; phenotype ; phosphatase ; phosphate ; phosphorylation ; physics ; physiology ; plant ; plasma membrane ; plasmodium ; plasticity ; polar ; polarity ; polymer ; polymerase ; polymerization ; polymorphisms ; pool ; population ; position ; positive ; positive feedback ; power law ; predicting ; prediction ; prediction accuracy ; prediction methods ; presence ; probability ; process ; produce ; profiling ; program ; progression ; project ; prokaryotes ; proliferation ; promote ; promoter sequence ; propagation ; property ; protein ; protein complex ; protein complexes ; protein data ; protein data bank ; protein dna ; protein domain ; protein folding ; protein function ; protein interaction ; protein interaction data ; protein interaction network ; protein kinase ; protein network ; protein protein interaction ; protein sequence ; protein structure ; proteome ; pulse ; putative ; qualitative ; quantify ; quantitative ; quantitative trait ; quantum ; random ; reaction ; reaction networks ; rearrangement ; receptor ; recognition ; recombinant ; recombination ; reconstruction ; recruitment ; recurrent ; reduction ; redundant ; region ; regression ; regulate ; regulated genes ; regulation ; regulatory elements ; regulatory genes ; regulatory interactions ; regulatory mechanisms ; regulatory network ; regulatory pathways ; regulatory proteins ; regulatory relationships ; regulon ; relationship ; reliability ; repair ; replication ; representation ; repression ; resistance ; resource ; response ; retina ; reverse engineering ; reversible ; rhythm ; rhythmicity ; ribosome ; rna binding ; rna polymerase ; robust ; robustness ; root ; rrna ; saccharomyces ; saccharomyces cerevisiae ; salt ; sampling ; scaffold ; scale ; scale free ; scoring ; screening ; secondary structure ; segmentation ; segregation ; selection ; selective ; selectivity ; self organizing ; sense ; sensitive ; sensitivity ; sensor ; sequence ; sequence alignment ; sequence data ; sequence database ; sequencing ; serine ; signal transduction ; signaling ; signaling molecules ; signaling network ; signaling pathway ; signalling ; significance ; silico ; simulating ; simulation ; simultaneous ; single cell ; site ; small world ; software ; species ; specification ; specificity ; spectrometry ; spectroscopy ; spectrum ; spike ; spindle ; splice ; spontaneous ; stability ; stabilization ; state ; statistics ; steady state ; stem ; stem cell ; stimulation ; stimulus ; storage ; strain ; strand ; strategy ; stream ; strength ; stress ; structu-

tel-00464316, version 1 - 16 Mar 2010

ral basis ; structural features ; structural genomics ; structural information ; structure ; structure prediction ; subcellular ; subcellular localization ; subgraphs ; subnetworks ; subset ; substrate ; subunit ; support vector machine ; suppressor ; surface ; susceptibility ; switch ; synapse ; synaptic plasticity ; synchronization ; system ; systems biology ; t cell ; target ; target gene ; targeting ; telomere ; temperature ; thaliana ; theories ; thermostability ; threshold ; time ; time scales ; time series ; time series data ; timing ; tissue ; tolerance ; topology ; trafficking ; trait ; trans ; trans golgi ; trans golgi network ; transcription ; transcription networks ; transcriptional regulation ; transcriptome ; transduction ; transduction pathways ; transfer ; transformation ; transient ; transition ; transition state ; translocation ; transmembrane ; transmission ; transport ; tuberculosis ; tumor ; turnover ; tyrosine ; uncertainty ; unique ; universe ; upstream ; variability ; variance ; variation ; variety ; vascular ; vegetation ; velocity ; vertebrate ; vesicle ; vessels ; virulence ; virus ; visual cortex ; wide range ; wild ; x ray ; yeast ; yeast cell ; yeast saccharomyces ; yeast saccharomyces cerevisiae ; zebrafish

B.3 Développement durable - CAB

Electricity ; rural areas ; risk assessment ; river water ; intensification ; cost analysis ; reports ; topography ; classification ; nitrates ; damage ; methodology ; technical progress ; zoning ; grasslands ; population density ; production ; soil types ; residential areas ; sugarcane ; drought ; international comparisons ; rice ; trade relations ; economic theory ; carbon ; angling ; human activity ; grazing ; botanical composition ; flooding ; ownership ; species diversity ; local population ; water allocation ; biomass production ; groundwater ; sociology ; nitrate ; satellite imagery ; development aid ; rotations ; social impact ; protection of forests ; accounting ; models ; irrigation ; manures ; fuel crops ; soil fertility ; cattle farming ; monitoring ; small farms ; seasonal variation ; natural resources ; coastal areas ; institution building ; rural tourism ; economic development ; human ecology ; government policy ; irrigation water ; bioremediation ; international trade ; technology ; bioenergy ; world ; tropical forests ; nutrients ; industry ; water availability ; plant pests ; renewable resources ; temporal variation ; polluted water ; economic growth ; pollutants ; water distribution ; greenhouse gases ; air pollution ; forestry development ; hunting ; mathematical models ; investment ; emission ; forest products ; information ; mapping ; ethics ; refuse ; remote sensing ; conservation areas ; weeds ; carbon dioxide ; research ; innovation adoption ; directives ; landsat ; salinity ; access ; spatial distribution ; ecological disturbance ; migration ; drinking water ; water quality ; carrying capacity ; groundwater pollution ; composting ; international organizations ; population dynamics ; optimization ; cropping systems ; chemical composition ; plant water relations ; harvesting ; projects ; yields ; multipurpose trees ; gender relations ; vegetation types ; wildlife ; animal husbandry ; shifting cultivation ; parity ; mountain forests ; econometric models ; range management ; land ; pastoral society ; pig farming ; resources ; biomass ; biodiversity ; rehabilitation ; land

tel-00464316, version 1 - 16 Mar 2010

ownership ; political power ; woodlands ; economic impact ; animal manures ; traditional society ; agriculture ; development programmes ; mangrove forests ; water resources ; environmental factors ; outdoor recreation ; atmosphere ; public health ; european union ; geographical distribution ; taxes ; hydrology ; use efficiency ; streams ; pastoralism ; medicinal plants ; forest recreation ; diversification ; animal production ; supply balance ; forecasts ; waste water ; government ; degraded land ; biology ; data analysis ; seeds ; roads ; valuation ; social participation ; environmental protection ; forests ; fuelwood ; tourism development ; wetlands ; tourist industry ; support measures ; production structure ; settlement ; soil water ; poverty ; transition economies ; environmental management ; right of access ; flood control ; private ownership ; forest fires ; law ; drainage systems ; heritage areas ; coffee ; feeds ; land diversion ; land resources ; farming ; economic policy ; economic viability ; biogas ; projections ; techniques ; cooperation ; local government ; climate ; agricultural trade ; property ; ethnic groups ; rain ; resettlement ; watershed management ; farming systems ; aquatic environment ; cadastres ; participation ; water harvesting ; marketing ; legislation ; conservation ; water systems ; indigenous knowledge ; project implementation ; dynamic models ; forest fragmentation ; non-governmental organizations ; change ; knowledge ; welfare economics ; spatial variation ; organizations ; cap ; subsidies ; water conservation ; community involvement ; nitrogen ; transgenic plants ; alternative farming ; environmental policy ; climatic change ; households ; common lands ; human diseases ; visitor behaviour ; insect pests ; world trade organization ; equipment ; farm size ; agricultural research ; farm management ; grassroots organizations ; genetic engineering ; learning ; stand structure ; economic analysis ; soil ; geological sedimentation ; livestock ; carbon sequestration ; agricultural economics ; farmers' associations ; water flow ; semiarid zones ; food supply ; privatization ; wheat ; eu regulations ; agricultural situation ; wild animals ; production costs ; community action ; arable land ; losses ; capital ; water use ; rangelands ; solid wastes ; agricultural policy ; history ; marine parks ; mangroves ; soil management ; partnerships ; afforestation ; estuaries ; herbicides ; social change ; pollution control ; nature tourism ; ecotourism ; silvopastoral systems ; nitrogen fertilizers ; comparisons ; trickle irrigation ; energy sources ; trees ; culture ; horticulture ; production possibilities ; dams ; maintenance ; polluted soils ; less favoured areas ; conflict ; trends ; dairy farms ; public domain ; lakes ; farm surveys ; simulation ; landscape conservation ; cotton ; transport ; grassland management ; visits ; regulations ; farms ; acreage ; economic situation ; non-market benefits ; canals ; eutrophication ; control ; stream flow ; crop yield ; quality ; decision making ; management ; energy conservation ; tourists ; introduced species ; property rights ; fuels ; common property resources ; databases ; land classification ; resource conservation ; leases ; efficiency ; estimation ; expenditure ; statistical analysis ; habitats ; development projects ; savannas ; energy policy ; macroeconomics ; networking ; regression analysis ; cost benefit analysis ; waste water treatment ; waste management ; phosphorus ; demand ; crops ; land markets ; groundwater recharge ; incentives ; economics ; forest products industries ; willingness to pay ; landscape ; training ; community forestry ; globaliza-

tel-00464316, version 1 - 16 Mar 2010

tion ; tourist attractions ; plant diseases ; forest ecology ; industrial wastes ; low input agriculture ; catchment hydrology ; logging ; natural resource economics ; plant genetic resources ; villages ; resource management ; information systems ; data collection ; mining ; agricultural households ; forest policy ; environmental degradation ; uncertainty ; woman's status ; application rates ; milk production ; ecosystems ; crop production ; saline water ; empowerment ; land management ; islands ; open spaces ; trade policy ; profits ; environmental assessment ; forest plantations ; water costs ; agroforestry ; population pressure ; water pollution ; irrigation equipment ; arid zones ; reviews ; cultural values ; upland areas ; deforestation ; membership ; fish farming ; tillage ; policy ; water use efficiency ; finance ; production functions ; rural development ; land evaluation ; fishery resources ; economic evaluation ; tourism impact ; trade liberalization ; certification ; contracts ; environmental education ; wildlife management ; water policy ; infrastructure ; intellectual property rights ; suburban areas ; forecasting ; non-wood forest products ; innovations ; thematic mapper ; floodplains ; air pollutants ; production economics ; resource utilization ; decentralization ; demography ; drainage ; research projects ; aquaculture ; land reform ; pastures ; regional development ; social forestry ; modernization ; consumer preferences ; industrialization ; energy resources ; agricultural land ; land use planning ; silviculture ; maps ; energy ; funding ; evaluation ; structural change ; agricultural structure ; air quality ; costs ; services ; marine environment ; public services ; pest control ; land capability ; contingent valuation ; land productivity ; supply ; recycling ; environmental legislation ; development planning ; land transfers ; prices ; development ; global warming ; technology transfer ; education ; erosion control ; precipitation ; arable farming ; land consolidation ; pollution ; amenity and recreation areas ; pesticides ; nature conservation ; case studies ; forest economics ; recreation ; rivers ; fishery policy ; organic farming ; operating costs ; forest trees ; regions ; simulation models ; price policy ; agricultural entomology ; farm structure ; constraints ; barley ; compensation ; wildlife conservation ; parks ; leaching ; terms of trade ; watersheds ; forest management ; contamination ; indicators ; sanitation ; cereals ; productivity ; administration ; socioeconomics ; water recreation ; rural environment ; tropical rain forests ; vegetables ; community development ; regional planning ; forestry ; biotechnology ; land development ; risk ; rural communities ; methane ; guidelines ; profitability ; ethanol ; soil conservation ; assessment ; tenure systems ; algorithms ; farm income ; tourism ; potatoes ; water balance ; development policy ; soil pollution ; agricultural development ; roles ; urban development ; private sector ; cultural heritage ; resource allocation ; water management ; fishery management ; urban areas ; commercialization ; international cooperation ; fallow ; natural disasters ; surface water ; farm inputs ; forest resources ; urbanization ; household surveys ; landowners ; animal welfare ; waste disposal ; cultivation ; waste utilization ; farmers' attitudes ; erosion ; effluents ; agricultural production ; environment ; standards ; water ; food safety ; national parks ; linear programming ; theory ; surveys ; returns ; irrigated farming ; landscape ecology ; waste treatment ; rent ; computer software ; invasions ; visitors ; energy consumption ; trade ; floods ; farmers ; tourism policy ; dairy

tel-00464316, version 1 - 16 Mar 2010

farming ; social development ; soyabeans ; consumer attitudes ; plant communities ; soil degradation ; irrigation systems ; salinization ; tariffs ; mountain areas ; communication ; food production ; politics ; land use ; forest ownership ; trade agreements ; design ; attitudes ; plant breeding ; behaviour ; runoff ; fresh water ; markets ; information services ; opportunity costs ; geographical information systems ; cooperatives ; communities ; agricultural sector ; nature reserves ; health ; wild birds ; public opinion ; fees ; temperature ; utilization ; endangered species ; abandoned land ; motivation ; imports ; institutions ; externalities ; soil organic matter ; labour ; cultivars ; environmental impact ; ecology ; livestock farming ; diffusion of information ; land prices ; ranching ; multiple use ; riparian vegetation ; fisheries ; international agreements ; weed control ; reservoirs ; income ; population growth ; water supply ; rural economy ; agroforestry systems ; conservation tillage ; consumption ; women ; highlands ; tropics ; rain forests ; maize ; timbers ; destinations ; food security ; agrarian reform ; vegetation ; human population ; habitat destruction ; planning ; sediment ; genetic resources ; private forestry ; marine areas ; reserved areas ; land policy ; water table ; exports ; aquifers ; water reuse ; fishing ; usage ; feasibility studies ; reclamation ; growth ; employment ; land degradation ; species richness ; extension ; regional policy ; wilderness ; weather ; desertification ; heavy metals ; biosafety ; fertilizers

tel-00464316, version 1 - 16 Mar 2010

tel-00464316, version 1 - 16 Mar 2010

ANNEXE C

Requête développement durable

La requête précise construite par Marc Barbier et Andreï Mogoutov est la suivante :

("farming systems" OR "farming system") OR TS=("cropping systems" OR "cropping system") OR TS=("agricultural systems" OR "agricultural system") OR TS=("agricultural knowledge") OR TS=("farmers participation") OR TS=("natural resource management") OR TS=("nature conservation") OR TS=(small scale farm* OR smallholder farm* OR family farm*) OR TS=("livestock systems" OR "livestock system") OR TS=("organic agriculture") OR TS=("livestock farming system" OR "livestock farming systems") OR TS=("rural system" OR "rural systems") OR TS=("agrarian system" OR "agrarian systems") OR TS=("local food" OR "local foods") OR TS=("pluriactivity" OR "pluriactivities") OR TS=("social learning" OR "social learnings") OR TS=("Farm management")) OR TS=(("livelihood" or "livelihoods" or "system approach" or "systems approach"or "household" or "households" or "R&D" or "research developement" or "extensions systems" or "extension system") AND TS=(agricult* farm or farming or rural)

tel-00464316, version 1 - 16 Mar 2010

Bibliographie

Abrahamson, E., Rosenkopf, L., 1997. Social network effects on the extent of innovation diffusion : A computer simulation. Organization Science 8 (3), 289–309.

Adamic, L. A., Adar, E., 2005. How to search a social network. Social Networks 27 (3), 187–203.

Adamic, L. A., Glance, N., 2005a. The political blogosphere and the 2004 us election : divided they blog. Proceedings of the 3rd international workshop on Link discovery, 36–43.

Adamic, L. A., Glance, N., 2005b. The political blogosphere and the 2004 u.s. election : divided they blog. Proceedings of the 3rd international workshop on Link discovery, 36–43.

Adar, E., Zhang, L., Adamic, L. A., Lukose, R., 2004a. Implicit structure and the dynamics of blogspace. Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference.

Adar, E., Zhang, L., Adamic, L. A., Lukose, R. M., 2004b. Implicit structure and the dynamics of blogspace. Proceedings.

Albert, R., Jeong, H., Barabási, A.-L., 2000. Error and attack tolerance of complex networks. Nature.

Ali-Hasan, N. F. A.-H. N. F., Adamic, L. A., 2007. Expressing social relationships on the blog through links and comments. .

Almaas, E., Barabási, A., 2005. Power laws in biological networks. Power laws, scalefree networks and genome biology. Landes Bioscience.

Amblard, F., Deffuant, G., 2004. The role of network topology on extremism propagation with the relative agreement opinion dynamics. Physica A 343, 725–738.

Amin, A., Roberts, J., 2006. Communities of practice : Varieties of situated learning. Dynamics of Institutions and Markets in Europe research paper.

Amin, A., Roberts, J., 2008. Knowing in action : Beyond communities of practice. Research Policy 37 (2), 353–369.

Archer, M. S., 1996. Culture and agency : The place of culture in social theory. Cambridge Univ Pr.

Axelrod, R., 1997a. The Complexity of Cooperation : Agent-Based Models of Competition and Collaboration. Princeton University Press, bouquin d'axelrod, disseminating culture.

Axelrod, R., 1997b. The dissemination of culture : A model with local convergence and global polarization. Journal of conflict resolution, 203–226.

Backstrom, L., Huttenlocher, D., Kleinberg, J. M., Lan, X., 2006a. Group formation in large social networks : membership, growth, and evolution. Proceedings, 44–54.

Backstrom, L., Huttenlocher, D., Kleinberg, J. M., Lan, X., 2006b. Group formation in large social networks : membership, growth, and evolution. Proceedings of the 12th ACM SIGKDD international conference ....

Bala, V., Goyal, S., 1998. Learning from neighbours. Review of Economic Studies 65 (3), 595–621.

Balog, K., Mishne, G., de Rijke, M., 2006. Why are they excited ? identifying and explaining spikes in blog mood levels. Proceedings 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006), April.

Bar-Yossef, Z., Rajagopalan, S., 2002. Template detection via data mining and its applications. Proc. 11th Intl. World-Wide Web Conference,, 580–591.

Barabási, A.-L., 2005. The origin of bursts and heavy tails in human dynamics. Nature 435 (7039), 207–11.

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509.

Barabási, A.-L., Albert, R., Jeong, H., 1999. Mean-field theory for scale-free random networks. Physica A 272, 173–187.

Barbier, M., Garandel-Batifol, V., Bompart, M., 2008. A textual analysis and scientometric mapping of the dynamic of knowledge in and around the ifsa community. The 8th IFSA European Symposium, 6-10 July 2008, Clermont Ferrand –France, 1–20.

Barbour, A., Mollison, D., 1990. Epidemics and random graphs. Stockholm University, 86–89.

Barrat, A., Barthélémy, M., Pastor-Satorras, R., Vespignani, A., 2004. The architecture of complex weighted networks. PNAS 101 (11), 3747–3752.

Barthélémy, M., Barrat, A., Pastor-Satorras, R., et al., 2005. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. Journal of Theoretical Biology.

tel-00464316, version 1 - 16 Mar 2010

Batagelj, V., Mrvar, A., 1998. Pajek-program for large network analysis. Connections 21 (2), 47–57.

Bearman, P. S., Moody, J., Stovel, K., 2004. Chains of affection : The structure of adolescent romantic and sexual networks 1. American Journal of Sociology 110 (1), 44–91.

Beck, U., 1992. Risk society : towards a new modernity. Sage Publications Ltd.

Bentley, R. A., Ormerod, P., 2009. Were people imitating others or exercising rational choice in on-line searches for'swine flu' ? arxiv.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech 10008.

Boguna, M., Pastor-Satorras, R., 2002. Epidemic spreading in correlated complex networks. Physical Review E 66, 047104.

Bollobás, B., 1985. Random Graphs. Cambridge University Press.

Boltanski, L., Thévenot, L., 1991. De la justification : les économies de la grandeur. Gallimard Paris.

Bonaccorsi, A., 2008. Search regimes and the industrial dynamics of science. Minerva 46 (3), 285–315.

Bonneuil, C., Joly, P.-B., Marris, C., 2008. Disentrenching experiment : The construction of gm–crop field trials as a social problem. Science, Technology & Human Values 33 (2), 201.

Borgatti, S. P., Mehra, A., Brass, D. J., Labianca, G., 2009. Network analysis in the social sciences. Science 323 (5916), 892.

Börner, K., Scharnhost, A., 2009. Visual conceptualizations and models of science. Journal of Informetrics.

Bourigault, D., Fabre, C., Frérot, C., Jacques, M.-P., Ozdowska, S., Recourcé, G., 2005. Syntex, analyseur syntaxique de corpus. Actes des 12èmes journées sur le Traitement ....

Boyack, K. W., Klavans, R., Börner, K., 2005. Mapping the backbone of science. Scientometrics 64 (3), 351–374.

Braam, R. R., Moed, H. F., van Raan, A. F. J., 1991. Mapping of science by combined cocitation and word analysis. ii. dynamical aspects. Journal American Society Information Science 42 (4), 252–266.

tel-00464316, version 1 - 16 Mar 2010

Breiger, R. L., 1974. The duality of persons and groups. Social Forces 53 (2), 181–190.

Bryant, S., Forte, A., Bruckman, A., 2005. Becoming wikipedian : transformation of participation in a collaborative online encyclopedia. Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work.

Burk, W. J., Kerr, M., Stattin, H., 2008. The co-evolution of early adolescent friendship networks, school involvement, and delinquent behaviors. Revue française de sociologie 3 (49), 499 à 522.

Burt, R. S., 1978. Cohesion versus structural equivalence as a basis for network subgroups. Sociological Methods and Research 7, 189–212.

Burt, R. S., 1987. Social contagion and innovation : Cohesion versus structural equivalence. American Journal of Sociology 92 (6), 1287–1335.

Burt, R. S., 1992. Structural holes : The social structure of competition. Harvard Univ Pr.

Burt, R. S., 1997. A note on social capital and network content. Social Networks 19 (4), 355–373.

Burt, R. S., 2000. The network structure of social capital. Research in organizational behavior 22 (2), 345–423.

Burt, R. S., 2004. Structural holes and good ideas. American Journal of Sociology 110 (2), 349–399.

Buter, R., Noyons, E. C. M., 2002. Using bibliometric maps to visualise term distribution in scientific papers. Sixth International Conference on Information Visualisation (IV'02), 697–702.

Callaway, D., Newman, M. E. J., Strogatz, S. H., Watts, D. J., 2000. Network robustness and fragility : Percolation on random graphs. Physical Review Letters E 85, 5468–5471.

Callon, M., 1994. Is science a public good ? fifth mullins lecture, virginia polytechnic institute, 23 march 1993. Science, Technology & Human Values 19 (4), 395.

Callon, M., 2006. Can methods for analysing large numbers organize a productive dialogue with the actors they study ? Eur Manage Rev 3 (1), 7–16.

Callon, M., Courtial, J.-P., Laville, F., 1991. Co-word analysis as a tool for describing the network of interaction between basic and technological research : The case of polymer chemistry. Scientometrics 22 (1), 155–205.

tel-00464316, version 1 - 16 Mar 2010

Callon, M., Courtial, J.-P., Turner, W. A., Bauin, S., 1983. From translations to problematic networks : An introduction to co-word analysis. Social Science Information 22 (2), 191–235.

Callon, M., Ferrary, M., 2006. Les réseaux sociaux à l'aune de la théorie de l'acteur-réseau. Sociologies Pratiques 13, 37–43.

Callon, M., Latour, B., 1981. Unscrewing the big leviathan : how actors macrostructure reality and how sociologists help them to do so. Advances in Social Theory and Methodology : Toward an Integration of Micro-and Macro-sociologies, Knorr-Cetina, K. and Cicourel, A.V., Routledge, 277–303.

Callon, M., Law, J., Rip, A., 1986. Mapping the dynamics of science and technology. London, Macmillan.

Cambrosio, A., Limoges, C., Courtial, J.-P., Laville, F., 1993. Historical scientometrics ? mapping over 70 years of biological safety research with coword analysis. Scientometrics.

Capocci, A., Servedio, V., Caldarelli, G., Colaiori, F., 2005. Detecting communities in large networks. Physica A : Statistical Mechanics and its Applications.

Cardon, D., Delaunay-Teterel, H., 2006. La production de soi comme technique relationnelle. Réseaux 138 (2006/4), 15–71.

Cattuto, C., 2006. Semiotic dynamics in online social communities. The European Physical Journal C-Particles and Fields.

Cha, M., Mislove, A., Adams, B., Gummadi, K. P., 2008. Characterizing social cascades in flickr. Proceedings of the first workshop on Online social networks, 13–18.

Chateauraynaud, F., Trabal, P., 2003. Internet à l'épreuve de la critique.

Chavalarias, D., Cointet, J.-P., 2008. Bottom-up scientific field detection for dynamical and hierarchical science mapping - methodology and case study. Scientometrics 75 (1).

Chavalarias, D., Cointet, J.-P., 2009. The reconstruction of science phylogeny. Arxiv preprint arXiv :0904.3154.

Chen, C., 2004. Searching for intellectual turning points : Progressive knowledge domain visualization. PNAS.

Chen, C., 2006. Citespace ii : Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science 57 (3), 359–377.

tel-00464316, version 1 - 16 Mar 2010

Chen, C., Cribbin, T., Macredie, R., Morar, S., 2002. Visualizing and tracking the growth of competing paradigms : Two case studies. Journal of the American Society for Information Science 53 (8), 678–689.

Chi, Y., Zhu, S., Song, X., Tatemura, J., Tseng, B. L., 2007. Structural and temporal analysis of the blogosphere through community factorization. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 163–172.

Christakis, N. A., Fowler, J. H., 2007. The spread of obesity in a large social network over 32 years. The New England Journal of Medicine, 10.

Christakis, N. A., Fowler, J. H., 2008. The collective dynamics of smoking in a large social network. New England Journal of Medicine 358 (21), 2249.

Chubin, D. E., 1976. The conceptualization of scientific specialties. The sociological quarterly 17 (4), 448–476.

Clauset, A., Newman, M. E. J., Moore, C., 2004. Finding community structure in very large networks. Physical Review E.

Cohendet, P., Créplet, F., Dupouet, O., 2003. Innovation organisationnelle, communautés de pratique et communautés épistémiques : le cas de linux. Revue française de gestion 146 (2003/5), 99–121.

Cointet, J.-P., 2008. Reconstruction multi-echelle des dynamiques scientifiques. Journées Jeunes Chercheurs du Département des Sciences Sociales de l'INRA, Montpelier, 1–17.

Cointet, J.-P., Chavalarias, D., 2008. Multi-level science mapping with asymmetric co-occurrence analysis : Methodology and case study,. NETWORKS AND HETEROGENEOUS MEDIA 3 (2), 267–276.

Cointet, J.-P., Faure, E., Roth, C., 2007. Intertemporal topic correlations in online media. Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007).

Cointet, J.-P., Roth, C., 2007. How realistic should knowledge diffusion models be ? Journal of Artificial Societies and Social Simulation.

Cointet, J.-P., Roth, C., 2009. Socio-semantic dynamics in a blog network. SocialCom09, 1–8.

Coleman, J. S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Coleman, J. S., Katz, E., Menzel, H., 1957a. The diffusion of an innovation among physicians. Sociometry 20 (4), 253–270.

tel-00464316, version 1 - 16 Mar 2010

Coleman, J. S., Katz, E., Menzel, H., 1957b. The diffusion of an innovation among physicians. Sociometry.

Conein, B., 2003. Communauté épistémique et réseaux cognitifs : coopération et cognition distribuée. Internet. Une utopie limitée. Nouvelles régulations, nouvelles solidarités, Conein, Bernard, Massit-Folléa, Françoise et Proulx, Serge (dir.).

Conein, B., 2004. Cognition distribuée, groupe social et technologie cognitive. Réseaux 124, 53–79.

Cowan, R., David, P., Foray, D., 2000. The explicit economics of knowledge codification and tacitness. Industrial and corporate change 9 (2), 211–253.

Cowan, R., Jonard, N., 2004a. Network structure and the diffusion of knowledge. Journal of Economic Dynamics and Control.

Cowan, R., Jonard, N., 2004b. Network structure and the diffusion of knowledge. Journal of Economic Dynamics and Control 28, 1557–1575.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. M., Suri, S., 2008a. Feedback effects between similarity and social influence in online communities. ACM New York, NY, USA.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. M., Suri, S., 2008b. Feedback effects between similarity and social influence in online communities. ACM New York, NY, USA.

Crane, D., 1972. Invisible Colleges : Diffusion of Knowledge in Scientific Communities. University of Chicago Press.

Crépey, P., Alvarez, F. P., Barthélémy, M., 2006. Epidemic variability in complex networks. Physical Review E 73 (4), 046131.

Crutchfield, J. P., Young, J., 1989. Inferring statistical complexity. Physical Review Letters 63 (2), 105–108.

Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A., 2005. Comparing community structure identification. Journal of Statistical Mechanics : Theory and Experiment.

Davis, J. A., 1963. Structural balance, mechanical solidarity, and interpersonal relations. American Journal of Sociology, 444–462.

Davis, J. A., Leinhardt, S., 1972. The structure of positive interpersonal relations in small groups. J. Berger (Ed.), Sociological Theories in Progress. Vol. 2. Boston-Houghton Mifflin.

tel-00464316, version 1 - 16 Mar 2010

de Solla Price, D., 1965. Networks of scientific papers. Science.

de Solla Price, D., 1976. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27 (4), 292–306.

Deffuant, G., 2006. Comparing extremism propagation patterns in continuous opinion models. Journal of Artificial Societies and Social Simulation 9 (3), 8.

Deffuant, G., Amblard, F., Weisbuch, G., Faure, T., 2002. How can extremism prevail ? a study based on the relative agreement interaction model. Journal of Artificial Societies and Social Simulation 5 (4), 1.

Deroian, F., 2002. Formation of social networks and diffusion of innovations. Research Policy 31, 835–846.

Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numerische Mathematik, 269–271.

Dillenbourg, P., Poirier, C., Carles, L., 2003. Communautés virtuelles d'apprentissage : e-jargon ou nouveau paradigme ? in A. Taurisson et A. Sentini Pédagogies.net, Montréal, Presses.

Dorat, R., Latapy, M., Conein, B., Auray, N., 2007. Multi-level analysis of an interaction network between individuals in a mailing-list. Annales des télécommunications.

Dorogovtsev, S. N., Goltsev, A. V., Mendes, J. F. F., 2007. Critical phenomena in complex networks. arxiv cond-mat.stat-mech.

Dorogovtsev, S. N., Mendes, J. F. F., 2003. Evolution of Networks : From Biological Nets to the Internet and WWW. Oxford University Press, USA.

Dupuy, J.-P., 2004. Vers l'unité des sciences sociales autour de l'individualisme méthodologique complexe. Revue du MAUSS 24 (2004/2), 310–328.

Eguiluz, V. M., Klemm, K., 2002. Epidemic threshold in structured scale-free networks. Physical Review Letters 89, 108701.

Eisenberg, E., Levanon, E. Y., 2003. Preferential attachment in the protein network evolution. Journal reference : Phys. Rev. Lett Phys Rev Lett 91, 138701.

Ellison, G., Fudenberg, D., 1995. Word-of-mouth communication and social learning. Quarterly Journal of Economics 110 (1), 93–125.

Emirbayer, M., Goodwin, J., 1994. Network analysis, culture, and the problem of agency. American Journal of Sociology 99 (6), 1411–1454.

tel-00464316, version 1 - 16 Mar 2010

Erdös, P., Rényi, A., 1959. On random graphs. Publicationes Mathematicae 6, 290–297.

Farkas, I. J., Ábel, D., Palla, G., Vicsek, T., 2007. Weighted network modules. New Journal of Physics.

Fischer, G., 2001. Communities of interest : Learning through the interaction of multiple knowledge systems. 24th Annual Information Systems Research Seminar In Scandinavia (IRIS'24), 1–14.

Flichy, P., 2000. Internet or the ideal scientific community. Réseaux 7 (2), 155–182.

Flichy, P., 2008. «internet, un outil de la démocratie ?». La vie des idées.

Freeman, L. C., 1979. Centrality in social networks : Conceptual clarification. Social Networks 1 (3), 215–239.

Freeman, L. C., 2004. The Development of Social Network Analysis : A Study in the Sociology of Science. Empirical Press.

Friedberg, E., 1997. Le pouvoir et la règle. Dynamiques de l'action organisée, Paris, Seuil.

Friggeri, A., Cointet, J.-P., Latapy, M., 2009. A real-world spreading experiment in the blogosphere. happyflu.com.

Funabashi, M., Chavalarias, D., Cointet, J.-P., 2009. Order-wise correlation dynamics in text data. Complex Networks : Results of the 1st International Workshop on Complex Networks (CompleNet 2009), 161.

Gallos, L. K., Song, C., Havlin, S., Makse, H. A., 2007. Scaling theory of transport in complex biological networks. PNAS 104 (19), 7746.

Ganesh, A., Massoulié, L., Towsley, D., 2005. The effect of network topology on the spread of epidemics. Proceedings 2, 1455–1466.

Garas, A., Argyrakis, P., 2008. A network approach for the scientific collaboration in the european framework programs. arxiv physics.soc-ph.

Garfield, E., 2004. Historiographic mapping of knowledge domains literature. Journal of Information Science 30 (2), 119–145.

Giddens, A., 1981. Agency, insitution, and time-space analysis. Advances in Social Theory and Methodology : Toward an Integration of Micro-and Macro-sociologies, Knorr-Cetina, K. and Cicourel, A.V., Routledge, 8.

Gilbert, M., 2003. Marcher ensemble. Essais sur le fondements des phénomènes collectifs, Paris : PUF.

tel-00464316, version 1 - 16 Mar 2010

Gilbert, N., Pyka, A., Ahrweiler, P., 2001. Innovation networks – a simulation approach. Journal of Artificial Societies and Social Simulation 4 (3), 8.

Gill, K. E., 2004. How can we measure the influence of the blogosphere. WWW 2004 Workshop on the Weblogging Ecosystem : Aggregation, Analysis, and Dynamics.

Girvan, M., Newman, M. E. J., 2002. Community structure in social and biological networks. PNAS 99, 7821–7826.

Glance, N., Hurst, M., Tomokiyo, T., 2004. Blogpulse : Automated trend discovery for weblogs. WWW 2004 Workshop on the Weblogging Ecosystem : Aggregation.

Goldenberg, J., Libai, B., Muller, E., 2001. Talk of the network : A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12 (3), 211–223.

Goldman, A. I., 2008. The social epistemology of blogging. Information Technology and Moral Philosophy.

Goodman, L. A., 1964. Mathematical methods for the study of systems of groups. American Journal of Sociology.

Gotz, M., Leskovec, J., McGlohon, M., Faloutsos, C., 2009. Modeling blog dynamics. Proceedings of the Third International ICWSM Conference (2009).

Granovetter, M. S., 1973. The strength of weak ties. The American Journal of Sociology 78 (6), 1360–1380.

Granovetter, M. S., 1978a. Threshold models of collective behavior. American Journal of Sociology 83 (6), 1420–1443.

Granovetter, M. S., 1978b. Threshold models of collective behavior. The American Journal of Sociology.

Granovetter, M. S., 1985. Economic action and social structure : The problem of embeddedness. The American Journal of Sociology 91 (3), 481–510.

Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A., 2005. The predictive power of online chatter. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 78–87.

Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A., 2004. Information diffusion through blogspace. Proceedings of the 13th international conference on World Wide Web, 491–501.

tel-00464316, version 1 - 16 Mar 2010

Guillaume, J.-L., Latapy, M., 2004. Bipartite structure of all complex networks. Information Processing Letters 90 (5), 215–221.

Gulati, R., 1995. Social structure and alliance formation patterns : A longitudinal analysis. Administrative Science Quarterly 40 (4).

Haas, P. M., 1992. Introduction : epistemic communities and international policy coordination. International Organization 46 (1), 1–35.

He, Q., 1999. Knowledge discovery through co-word analysis. Library Trends.

Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., Yu, N., 2005. Conversations in the blogosphere : An analysis "from the bottom up". Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05).

Hethcote, H. W., 2000. The mathematics of infectious diseases. SIAM Review 42 (4), 599–653.

Heylighen, F., Heath, M., Overwalle, F. V., 2004. The emergence of distributed cognition : a conceptual framework. Proceedings of Collective Intentionality IV, Siena (Italy).

Hindman, M., Tsioutsiouliklis, K., Johnson, J., 2003. Googlearchy : How a few heavily-linked sites dominate politics on the web. Annual Meeting of the Midwest Political Science Association 4.

Hine, C., 2005. Virtual methods : Issues in social research on the Internet. Berg Publishers.

Holland, P. W., Leinhardt, S., 1976. Local structure in social networks. Sociological Methodology, 1–45.

Holme, P., Edling, C., Liljeros, F., 2004. Structure and time evolution of an internet dating community. Social Networks 26 (2), 155–174.

Holme, P., Kim, B. J., 2002. Growing scale-free networks with tunable clustering. Physical Review E 65 (2), 1–4.

Hopcroft, J., Khan, O., Kulis, B., Selman, B., 2004a. Tracking evolving communities in large linked networks. PNAS 101 Suppl 1, 5249–53.

Hopcroft, J., Khan, O., Kulis, B., Selman, B., 2004b. Tracking evolving communities in large linked networks. PNAS 101 (1), 5249–5253.

Huberman, B. A., Romero, D. M., Wu, F., 2008. Social networks that matter : Twitter under the microscope. arxiv.

tel-00464316, version 1 - 16 Mar 2010

Huelsenbeck, J. P., Rannala, B., 1997. Phylogenetic methods come of age : Testing hypotheses in an evolutionary context. Science.

Hull, D. L., 1988. Science as a process : an evolutionary account of the social and conceptual development of science. University of Chicago Press.

Hull, D. L., 2001. Science and Selection : Essays on Biological Evolution and the Philosophy of Science. Cambridge University Press.

Hutchins, E., 1996. 2 learning to navigate. Understanding practice : Perspectives on activity and context, 35.

Iribarren, J. L., Moro, E., 2007. Information diffusion epidemics in social networks. arxiv.

Java, A., Kolari, P., Finin, T., Oates, T., 2006. Modeling the spread of influence on the blogosphere. Proceedings of the 15th International World Wide Web, 7.

Java, A., Song, X., Finin, T., Tseng, B., 2007. Why we twitter : understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 56–65.

Jeong, H., Neda, Z., Barabási, A.-L., 2003a. Measuring preferential attachment for evolving networks. Europhys. Lett.

Jeong, H., Néda, Z., Barabási, A.-L., 2003b. Measuring preferential attachment for evolving networks. Europhysics Letters 61 (4), 567–572.

Johnson, D. H., Sinanovic, S., 2001. Symmetrizing the kullback-leibler distance. IEEE Transactions on Information Theory.

Johnson, J., 2006. Hypernetworks for reconstructing the dynamics of multilevel systems. submitted to the European Conference on Complex Systems, Oxford.

Jones, B., Wuchty, S., Uzzi, B., 2008. Multi-university research teams : Shifting impact, geography, and stratification in science. Science.

Jones, J. H., Handcock, M. S., 2003. An assessment of preferential attachment as a mechanism for human sexual network formation. Proceedings of the Royal Society B : Biological Sciences 270 (1520), 1123–1128.

Kadushin, C., 1966. The friends and supporters of psychotherapy : on social circles in urban life. American Sociological Review, 786–802.

Katz, E., Lazarsfeld, P. F., 1955. Personal influence. Free Press.

Keller, E. F., 2005. Revisiting "scale-free" networks. Bioessays 27 (10), 1060–1068.

tel-00464316, version 1 - 16 Mar 2010

Kempe, D., Kleinberg, J. M., 2002. Protocols and impossibility results for gossip-based communication mechanisms. Proceedings, 471–480.

Kempe, D., Kleinberg, J. M., Tardos, E., 2003. Maximizing the spread of influence through a social network. Proceedings, 137–146.

Kempe, D., Kleinberg, J. M., Tardos, E., 2005. Influential nodes in a diffusion model for social networks. Proceedings 3580, 1127–1138.

Kleinberg, J. M., 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM.

Kleinberg, J. M., 2000. Navigation in a small world. Nature 406 (6798), 845–845.

Kleinberg, J. M., 2003. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery 7 (4), 373–397.

Kleinberg, J. M., 2005. Temporal dynamics of on-line information streams. Data Stream Management : Processing High-Speed Data Streams, Springer.

Klemm, K., Eguiluz, V. M., 2001. Highly clustered scale-free networks. the American Physical Society 65, 1231–1235.2002.

Klemm, K., Eguiluz, V. M., Toral, R., Miguel, M. S., 2005. Globalization, polarization and cultural drift. Journal of Economic Dynamics and Control 29, 321–334.

Knorr-Cetina, K. D., 1982. Scientific communities or transepistemic arenas of research ? a critique of quasi-economic models of science. Social Studies of Science 12 (1), 101–130.

Knorr-Cetina, K. D., 1995. Laboratory studies : The cultural approach to the study of science. Handbook of Science and Technology Studies, 140–166.

Kossinets, G., Kleinberg, J. M., Watts, D. J., 2008. The structure of information pathways in a social communication network. arxiv physics.soc-ph.

Kossinets, G., Watts, D. J., 2006. Empirical analysis of an evolving social network. Science.

Kuhn, T. S., 1970a. The Structure of Scientific Revolutions. University of Chicago Press.

Kuhn, T. S., 1970b. The Structure of Scientific Revolutions, Postscript. University of Chicago Press.

Kullback, S., Leibler, R., 1951. On information and sufficiency. The Annals of Mathematical Statistics, 79–86.

tel-00464316, version 1 - 16 Mar 2010

Kuperman, M., Abramson, G., 2001. Small world effect in an epidemiological model. Physical Review Letters 86 (13), 2909–2912.

Lahire, B., 1998. L'homme pluriel : les ressorts de l'action. Nathan.

Lancichinetti, A., Fortunato, S., Kertesz, J., 2009. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11 (033015), 033015.

Latapy, M., Magnien, C., Vecchio, N., 2008. Basic notions for the analysis of large two-mode networks. Social Networks 30 (1), 31–48.

Latour, B., 1987. Les "vues" de l'esprit. Réseaux.

Latour, B., 2001. Gabriel tarde and the end of the social. The social and its problems.

Latour, B., 2005. Reassembling the Social : An Introduction to Actor-network-theory. Oxford University Press, USA.

Latour, B., 2006. Changer de société - refaire de la sociologie. Découverte.

Latour, B., 2007. Beware, your imagination leaves digital traces. Times Higher Literary Supplement, 6th April 2007.

Latour, B., Woolgar, S., 1986. Laboratory Life : the social construction of scientific facts. Vol. Princeton University Press. Princeton University Press.

Latour, B., Woolgar, S., 1988. La vie de laboratoire – La production des faits scientifiques. La Découverte.

Lave, J., Wenger, E. C., 1991. Situated learning : Legitimate peripheral participation. Cambridge University Press.

Lazarsfeld, P. F., Merton, R. K., 1954. Friendship as a social process : a substantive and methodological analysis. Freedom and control in modern society, 18–66.

Lazega, E., Jourda, M.-T., Mounier, L., Stofer, R., 2007. Des poissons et des mares : l'analyse de réseaux multi-niveaux. Revue française de sociologie.

Lento, T., Welser, H. T., Gu, L., Smith, M., 2006. The ties that blog : Examining the relationship between social ties and continued participation in the wallop weblogging system. 3rd Annual Workshop on the Weblogging Ecosystem.

Leskovec, J., Adamic, L. A., Huberman, B. A., 2007a. The dynamics of viral marketing. portal.acm.org.

Leskovec, J., Backstrom, L., Kleinberg, J. M., 2009. Meme-tracking and the dynamics of the news cycle. Proceedings of The Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09).

tel-00464316, version 1 - 16 Mar 2010

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., 2007b. Cascading behavior in large blog graphs. SIAM International Conference on Data Mining (SDM 2007).

Lew, B., 2000. The diffusion of tractors on the canadian prairies : The threshold model and the problem of uncertainty. Explorations in Economic History 37 (2), 189–216.

Leydesdorff, L., 1997. Why words and co-words cannot map the development of the sciences. Journal of the American Society for Information Science 48 (5), 418–427.

Leydesdorff, L., Rafols, I., 2009. A global map of science based on the isi subject categories. Journal of the American Society for Information Science 60 (2), 348–362.

Leydesdorff, L., Schank, T., 2008a. Dynamic animations of journal maps : Indicators of structural change and interdisciplinary developments. Journal of the American Society for Information Science 59 (11), 1810–1818.

Leydesdorff, L., Schank, T., 2008b. Dynamic animations of journal maps : Indicators of structural changes and interdisciplinary developments. Journal of the American Society for Information Science.

Licklider, J. C. R., Taylor, R., 1968. The computer as a communication device. Science and technology 76, 21–31.

Licoppe, C., Beaudoin, V., 2002. La construction électronique du social : les sites personnels. l'exemple de la musique. Réseaux 6 (116), 53–96.

Lieberman, E., Michel, J., Jackson, J., Tang, T., Nowak, M., 2007. Quantifying the evolutionary dynamics of language. Nature 449 (7163), 713–716.

Lih, A., 2003. Wikipedia as participatory journalism : Reliable sources ? metrics for evaluating collaborative media as a news resource. Nature 2004.

Lind, P. G., Gonzalez, M. C., Herrmann, H. J., 2005. Cycles and clustering in bipartite networks. Physical Review E 72, 056127.

Lloyd, A. L., May, R. M., 2001. How viruses spread among computers and people. Science 292 (5520), 1316–1317.

Lloyd, L., Kaulgud, P., Skiena, S., 2006. Newspapers vs. blogs : Who gets the scoop ? AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs.

Lorrain, F., White, H. C., 1971. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1 (49–80).

tel-00464316, version 1 - 16 Mar 2010

Marin, A., Wellman, B., 2010. Social network analysis. Forthcoming in Handbook of Social Network Analysis - Peter Carrington and John Scott, London, Sage, 1–23.

May, R. M., Lloyd, A. L., 2001. Infection dynamics on scale-free networks. Physical Review E 64 (6), 066112.

McPherson, M., Smith-Lovin, L., 2001. Birds of a feather : Homophily in social networks. Annual Review of Sociology 27, 415–440.

Menzel, H., Katz, E., 1955. Social relations and innovation in the medical profession : The epidemiology of a new drug. The Public Opinion Quarterly.

Michelet, B., 1988. L'analyse des associations. Unpublished Ph. D. Thesis, Université Paris VII, Paris.

Milgram, S., 1967. The small world problem. Psychology Today 2, 60–67.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U., 2004. Superfamilies of evolved and designed networks. Science 303 (5663), 1538–1542.

Mishne, G., de Rijke, M., 2006. Capturing global mood levels using blog posts. AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.

Molloy, M., Reed, B., 1995. A critical point for random graphs with a given degree sequence. Random Structures and Algorithms 161 (6), 161–179.

Moody, J., White, D. R., 2003. Structural cohesion and embeddedness : a hierarchical conception of social groups. American Sociological Review 68 (103–127).

Morris, S. A., 2000. Contagion. Review of Economic Studies 67 (1), 57–78.

Morris, S. A., der Veer Martens, B. V., 2008. Mapping research specialties. Annual Review of Information Science and Technology 42, 52.

Morris, S. A., Yen, G. G., 2004. Crossmaps : Visualization of overlapping relationships in collections of journal papers. PNAS 101 (Suppl 1), 5291–5296.

Motter, A. E., Zhou, C., Kurths, J., 2005. Network synchronization, diffusion, and the paradox of heterogeneity. Phys Rev E 71.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F., 2004. A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics 61 (1), 129–145.

tel-00464316, version 1 - 16 Mar 2010

Mulkay, M. J., 1976. The model of branching. Sociological Review 24, 125–133.

Mullins, N., 1972. The development of a scientific specialty : The phage group and the origins of molecular biology. Minerva 10 (1), 51–82.

Nei, M., 1996. Phylogenetic analysis in molecular evolutionary genetics. Annual Reviews in Genetics.

Newman, M., 2004a. Who is the best connected scientist ? a study of scientific coauthorship networks. Lecture notes in Physics 650, 337–370.

Newman, M. E. J., 2001. The structure of scientific collaboration networks. PNAS 98 (2), 404–409.

Newman, M. E. J., 2002. Spread of epidemic disease on networks. Physical Review E 66 (016128).

Newman, M. E. J., 2004b. Coauthorship networks and patterns of scientific collaboration. PNAS.

Newman, M. E. J., 2006. Finding community structure in networks using the eigenvectors of matrices. Physical Review E.

Newman, M. E. J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E.

Newman, M. E. J., Park, J., 2003a. Why social networks are different from other types of networks. Physical Review E 68 (3 Pt 2), 036122.

Newman, M. E. J., Park, J., 2003b. Why social networks are different from other types of networks. Physical Review E 68 (036122), arXiv :cond-mat/0305612.

Newman, M. E. J., Strogatz, S. H., Watts, D. J., 2001. Random graphs with arbitrary degree distributions and their applications. Physical Review E 64 (026118), arXiv :cond-mat/0007235.

Newman, M. E. J., Watts, D. J., Strogatz, S. H., 2002a. Random graph models of social networks. PNAS.

Newman, M. E. J., Watts, D. J., Strogatz, S. H., 2002b. Random graph models of social networks. PNAS.

Nowotny, H., Scott, P., Gibbons, M., 2001. Re-thinking science : Knowledge and the public in an age of uncertainty. Cambridge, Mass.

Nowotny, H., Scott, P., Gibbons, M., 2003. Introduction : 'mode 2' revisited : The new production of knowledge. Minerva 41 (3), 179–194.

tel-00464316, version 1 - 16 Mar 2010

Noyons, E. C. M., 2001. Bibliometric mapping of science in a policy context. Scientometrics 50 (1), 83–98.

O'Connor, R., 2009. Global : Facebook and twitter reshaping journalism as we know it. kauri.aut.ac.nz.

Onnela, J.-P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., Barabási, A.-L., 2007. Structure and tie strengths in mobile communication networks. PNAS 104 (18), 7332.

Origgi, G., 2006. Autorité épistémique et internet scientifique : la diffusion du savoir sur internet. halshs.archives-ouvertes.fr.

Padgett, J. F., Ansell, C. K., 1993. Robust action and the rise of the medici, 1400-1434. American Journal of Sociology 98 (6), 1259.

Palla, G., Barabási, A.-L., Vicsek, T., 2007a. Quantifying social group evolution. Nature 446 (7136), 664–667, supplementary materials.

Palla, G., Derenyi, I., Farkas, I. J., Vicsek, T., 2005a. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814.

Palla, G., Derenyi, I., Farkas, I. J., Vicsek, T., 2005b. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435 (7043), 814–8.

Palla, G., Farkas, I. J., Pollner, P., Derenyi, I., Vicsek, T., 2007b. Directed network modules. New Journal of Physics.

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Now Publishers.

Pastor-Satorras, R., Vespignani, A., 2001. Epidemic spreading in scale-free networks. Physical Review Letters 86 (14), 3200–3203.

Pattison, P., Wasserman, S., Robins, G., Kanfer, A. M., 2000. Statistical evaluation of algebraic constraints for social networks. Journal of Mathematical Psychology 44, 536–568.

Pestre, D., 2007. L'analyse de controverses dans l'étude des sciences depuis trente ans entre outil méthodologique, garantie de neutralité axiologique et politique. Mil neuf cent 1 (25), 29–43.

Pledel, I., 2008. Les nouvelles logiques d'expression : blogs et journalisme participatif, vers une e-démocratie ? recolecta.net.

Qazvinian, V., Rasoulian, A., Shafiei, M., Adibi, J., 2007. A large-scale study on persian weblogs. Proc. of Workshop on Text-Mining and Link-Analysis.

tel-00464316, version 1 - 16 Mar 2010

Robertson, T. S., 1967. The process of innovation and the diffusion of innovation. Journal of Marketing 31 (1), 14–19.

Robins, G., Alexander, M., 2004. Small worlds among interlocking directors : Network structure and distance in bipartite graphs. Computational & Mathematical Organization Theory.

Rogers, E. M., 1976. New product adoption and diffusion. The Journal of Consumer Research 2 (4), 290–301.

Rogers, E. M., 2003. Diffusion of Innovations. Free Press, 5th Edition.

Rosvall, M., Bergstrom, C. T., 2008a. Maps of random walks on complex networks reveal community structure. PNAS 105 (4), 1118–1123.

Rosvall, M., Bergstrom, C. T., 2008b. Maps of random walks on complex networks reveal community structure. PNAS.

Roth, C., 2005. Generalized preferential attachment : Towards realistic socio-semantic network models. ISWC 4th Intl Semantic Web Conference, Workshop on Semantic Network Analysis, Galway, Ireland 171, 29–42, also on arXiv :nlin.AO/0507021.

Roth, C., 2006. Co-evolution in epistemic networks – reconstructing social complex systems. Structure and Dynamics : eJournal of Anthropological and Related Sciences 1 (3), 3–163, Édition spÈciale consacrÈe à ma thËse.

Roth, C., 2008a. Co-évolution des auteurs et des concepts dans les réseaux épistémiques : le cas de la communauté zebrafish. Revue Francaise de Sociologie 48 (2), 333–367.

Roth, C., 2008b. Réseaux épistémiques : formaliser la cognition distribuée. Sociologie du Travail 50, 353–371.

Roth, C., Bourgine, P., 2006. Lattice-based dynamic and overlapping taxonomies : The case of epistemic communities. Scientometrics 69 (2), 429–447.

Roth, C., Cointet, J.-P., 2009. Social and semantic coevolution in knowledge networks. Social Networks.

Ruef, M., Aldrich, H. E., Carter, N. M., 2004. The structure of founding teams : Homophily, strong ties, and isolation among us entrepreneurs. American Sociological Review.

Ryan, B., Gross, N., 1943. The diffusion of hybrid seed corn in two iowa communities. Rural Sociology 8 (1), 15–24.

tel-00464316, version 1 - 16 Mar 2010

Salton, G., Wong, A., Yang, C. S., 1975. Vector space model for automatic indexing. Communications of the ACM 18 (11), 613–620.

Sewell, W. H. J., 1992. A theory of structure : Duality, agency, and transformation. American Journal of Sociology.

Shaikevich, I. M., 2005. Bibliometric maps of field of science. Infometrics 41 (6), 1534–1547.

Shalizi, C. R., 2001a. Causal architecture, complexity and self-organization in the time series and cellular automata. bactra.org.

Shalizi, C. R., 2001b. Causal architecture, complexity and self-organization in time series and cellular automata. Ph.D. thesis, chap. 11.

Shalizi, C. R., 2007. Social media as windows on the social life of the mind. arxiv.

Shalizi, C. R., Shalizi, K. L., 2004. Blind construction of optimal non-linear recursive predictors for discrete sequences. Proceedings, 504–511.

Shannon, C., Weaver, W., 2002. Mathematical Theory of Communication. University of Illinois Press.

Shi, X., Tseng, B., Adamic, L. A., 2007. Looking at the blogosphere topology through different lenses. Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007) 1001, 48109.

Simmel, G., 1898. The persistence of social groups. American Journal of Sociology 3 (5), 662.

Simmel, G., 1955. The web of group affiliations. Conflict and the web of group affiliations, 125–95.

Simmel, G., 1971. On individuality and social forms : Selected writings. University of Chicago Press.

Simondon, G., 1989. L'individuation psychique et collective : à la lumière des notions de forme, information, potentiel et métastabilité. Aubier.

Skupin, A., 2004. The world of geography : Visualizing a knowledge domain with cartographic means. PNAS.

Small, H. G., 1973. Co-citation in the scientific literature : A new measure of the relationship between two documents. Journal of the American Society for Information Science, 265–269.

Smith-Doerr, L., Powell, W. W., 2005. Networks and economic life. The handbook of economic sociology, 379–402.

tel-00464316, version 1 - 16 Mar 2010

Snijders, T. A. B., 2001. The statistical evaluation of social networks dynamics. Sociological Methodology 31, 361–395.

Snijders, T. A. B., Stokman, F. N., 1987. Extensions of triad counts to networks with different subsets of points and testing underlying random graph distributions. Social Networks 9, 249–275.

Soffer, S. N., Vázquez, A., 2005. Network clustering coefficient without degree-correlation biases. Phys. Rev. E 71 (5), 4.

Steglich, C., Snijders, T. A. B., Pearson, M., 2004. Dynamic networks and behavior : Separating selection from influence. Submitted for publication.

Steglich, C., Snijders, T. A. B., West, P., 2006. Applying siena : An illustrative analysis of the coevolution of adolescents' friendship networks, taste in music, and alcohol consumption. Methodology 2 (1), 48–56.

Steyvers, M., Tenenbaum, J. B., 2005. The large-scale structure of semantic networks : Statistical analyses and a model of semantic growth. Cognitive Science : A Multidisciplinary Journal.

Storer, N. W., 1966. The social system of science. New York : Holt, Rinehart and Winston.

Strogatz, S. H., 2001. Exploring complex networks. Nature.

Tarde, G., 1890. Les lois de l'imitation. Editions Kimé, 1993.

Tarde, G., 1898. Les lois sociales. classiques.uqac.ca.

Thelwall, M., 2006. Bloggers during the london attacks : Top information sources and topics. Proc. of the World Wide Web 2006 Workshop on the Weblogging, 8.

Thelwall, M., Price, L., 2006. Language evolution and the spread of ideas on the web : A procedure for identifying emergent hybrid word family members. Journal of the American Society for Information Science 57 (10), 1326–1337.

Thelwall, M., Vaughan, L., Bjorneborn, L., 2005. Webometrics. Annual Review of Information Science and Technology.

Turner, W. A., Chartron, G., Laville, F., Michelet, B., 1988. Packaging information for peer review : new co-word analysis techniques. Handbook of quantitative studies of science and technology (pp. 291-323). Netherlands : Elsevier Science Publishers.

Uchida, M., Shibata, N., Shirayama, S., 2007. Identification and visualization of emerging trends from blogosphere. Proceedings of International Conference on Weblogs and Social Media, 305–306.

tel-00464316, version 1 - 16 Mar 2010

Uzzi, B., Spiro, J., Murmann, R., Bothner, M., Zelek, M., et al., 2005. Collaboration and creativity : The small world problem. AJS.

Valente, T. W., 1995. Network Models of the Diffusion of Innovations. Hampton Press.

Valente, T. W., 1996. Social network thresholds in the diffusion of innovations. Social Networks 18, 69–89.

Viegas, F. B., Wattenberg, M., Kriss, J., van Ham, F., 2007a. Talk before you type : Coordination in wikipedia. Proceedings of the 40th Hawaii Intl Conf on System Sciences.

Viegas, F. B., Wattenberg, M., McKeon, M., 2007b. The hidden order of wikipedia. Lecture notes in Computer Science 4564, 445.

Wallsten, K., 2005. Political blogs and the bloggers who blog them : Is the political blogosphere and echo chamber. American Political Science Association's Annual Meeting.

Wang, Y., Chakrabarti, D., Wang, C., Faloutsos, C., 2003. Epidemic spreading in real networks : An eigenvalue viewpoint. Proceedings, 25–34.

Wasserman, S., Faust, K., 1994. Social Network Analysis : Methods and Applications. Cambridge University Press.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks : I. an introduction to markov graphs andp. Psychometrika 61 (3), 401–425.

Watts, D. J., 1999. Small World : The Dynamics of Networks between Order and Randomness. Princeton University Press.

Watts, D. J., 2002. A simple model of global cascades on random networks. PNASKljw dqsklmsdq.

Watts, D. J., Dodds, P. S., 2007. Influentials, networks, and public opinion formation. Journal of Consumer Research 34 (4), 441–458.

Watts, D. J., Dodds, P. S., Newman, M. E. J., 2002. Identity and search in social networks. Science 296, 1302–1305.

Watts, D. J., Strogatz, S. H., 1998. Collective dynamics of 'small-world' networks. Nature (393), 440–442.

Weiss, M., Moroiu, G., 2007. Ecology and dynamics of open source communities. scs.carleton.ca.

tel-00464316, version 1 - 16 Mar 2010

Wenger, E. C., Snyder, W. M., 2000. Communities of practice : The organizational frontier. Harvard business review 78 (1), 139–146.

White, H. C., Boorman, S. A., Breiger, R. L., 1976. Social structure from multiple networks. i : Blockmodels of roles and positions. The American Journal of Sociology 81 (4), 730–780.

Whittaker, J., 1989. Creativity and conformity in science : Titles, keywords and coword analysis. Social Studies of Science 19 (3), 473–496.

Wille, R., 1992. Concept lattices and conceptual knowledge systems. Computers Mathematics and Applications 23, 493.

Wu, F., Huberman, B. A., Adamic, L. A., Tyler, J. R., 2004. Information flow in social groups. Physica A 337, 327–335.

Zegura, E. W., Calvert, K. L., Bhattacharjee, S., 1996. How to model an internetwork. Proceedings 2, 594–602.

Zhang, P., Wang, J., Li, X., Li, M., Di, Z., Fan, Y., 2008. Clustering coefficient and community structure of bipartite networks. Physica A : Statistical Mechanics and its Applications 387 (27), 6869–6875.

Zhou, D., Song, Y., Zha, H., Zhang, Y., 2005. Towards discovering organizational structure from email corpus. Machine Learning and Applications.

Zitt, M., Bassecoulard, E., 2006. Delineating complex scientific fields by an hybrid lexical-citation method : An application to nanosciences. Information Processing & Management 42 (6), 1513–1531.

tel-00464316, version 1 - 16 Mar 2010