Nouvelles perspectives en sciences sociales

Reconstruction en sciences sociales : le cas des réseaux de savoirs

Camille Roth

Volume 2, numéro 2, mars 2007

URI : https://id.erudit.org/iderudit/602460ar
DOI : https://doi.org/10.7202/602460ar

Aller au sommaire du numéro

Éditeur(s)

Prise de parole

ISSN

1712-8307 (imprimé)
1918-7475 (numérique)

Découvrir la revue

Citer cet article

Roth, C. (2007). Reconstruction en sciences sociales : le cas des réseaux de savoirs. *Nouvelles perspectives en sciences sociales*, *2*(2), 59–101. https://doi.org/10.7202/602460ar

Résumé de l'article

Des agents produisant et échangeant des connaissances constituent un système complexe socio-sémantique, dont l'étude représente un défi à la fois théorique, dans la perspective de résoudre un problème de reconstruction en sciences sociales, et pratique, avec des applications permettant aux agents de connaître la dynamique du système dans lequel ils évoluent. Nous montrons que plusieurs aspects significatifs de la structure d'une communauté de savoirs sont principalement produits par la dynamique d'un réseau épistémique où co-évoluent agents et concepts. La structure est principalement décrite par la taxonomie de communautés de savoirs à partir de simples relations entre agents et concepts et de treillis de Galois; nous obtenons une description historique se rapportant à la progression des champs, leur déclin, leur spécialisation ou leurs interactions. Nous micro-fondons ensuite ces phénomènes en exhibant et en estimant empiriquement des processus d'interaction au niveau des agents, en co-évolution avec les concepts au sein du réseau épistémique, qui rendent compte de la morphogenèse et de l'émergence de plusieurs faits stylisés structurels de haut-niveau.

Reconstruction en sciences sociales : le cas des réseaux de savoirs

Camille Roth
University of Surrey, Guilford (Royaume-Uni)
CREA, CHRS / École polytechnique, Paris, France

Des agents produisant, manipulant et échangeant des connaissances constituent un système complexe socio-sémantique : ils sont totalement immergés dans un flot d'informations sur lequel ils peuvent aussi rétroagir. Cet objet d'étude n'est pas nouveau : la psychologie sociale et l'épistémologie, par exemple, s'intéressent depuis longtemps aux propriétés de ce type de communautés de savoirs[1]. Toutefois, la disponibilité massive des contenus informationnels, puis l'augmentation conséquente du potentiel d'interaction ont transformé ce qui était vu comme de simples « groupes de savoirs » en « société de savoirs » globalisée et généralisée. Simultanément, le savoir étant distribué et appréhendé de manière plus horizontale, en réseau ; ce changement d'échelle a requis l'utilisation de nouvelles méthodes afin de caractériser des phénomènes différents.

L'étude de ces communautés a ainsi connu un engouement et une attention sans précédent, dans une perspective à la fois théorique et

[1] Frederik Schmitt (dir.), *Socializing epistemology: The social dimensions of knowledge,* Lanham (MD), Rowman & Littlefield, 1995.

pratique. Sur le plan théorique, il est devenu possible de développer davantage le projet de naturalisation des sciences sociales. Sur le plan pratique, diverses applications sont envisageables – notamment pour la politique de la recherche, parce que les scientifiques eux-mêmes forment une communauté de savoirs, mais aussi, par exemple, en tant qu'outil de prospective ou d'amélioration de la diffusion des innovations.

Reconstruire. Nous nous situons ici dans le cadre de ce programme de recherche. Plus précisément, nous voulons connaître et modéliser le comportement et la dynamique de ces communautés. Nous abordons ainsi plus largement la question de la reconstruction en sciences sociales, un problème inverse consistant fondamentalement à reproduire divers faits stylisés observés empiriquement dans un système donné. L'on distingue traditionnellement le « bas-niveau » des objets microscopiques (agents, interactions, etc.) et le « haut-niveau » des descriptions macroscopiques (communautés, structures globales). Il s'agit ainsi :

(i) de déduire une certaine observation de haut-niveau pour ce système à partir de phénomènes essentiellement de bas-niveau ; et

(ii) de reconstruire l'évolution des observations de haut-niveau en inférant « la dynamique » des objets de bas-niveau.

Par exemple, les sociologues utilisent de plus en plus souvent l'analyse de réseaux sociaux pour appréhender des phénomènes de haut-niveau traditionnellement décrits de façon agrégée : qualifier la cohésion d'une communauté, trouver les causes d'une crise, etc. En procédant ainsi, ils abordent clairement le premier problème, « (i) » : ils exhibent une relation formelle entre des objets de haut et bas-niveau – ils reconstruisent la « structure sociale » qu'ils valident par le biais de descriptions de haut-niveau classiquement admises[2]. En ce sens, ils font l'hypothèse que le bas-niveau choisi (tel un réseau social) contient suffisamment d'information sur le phénomène en question, l'avantage étant souvent que les données de bas-niveau sont plus faciles à recueillir et plus robustes. Formellement, le premier problème est équivalent à la question suivante : soit un phénomène de haut-niveau H, et des objets

[2] Linton C. Freeman, « Social networks and the structure experiment », dans Linton C. Freeman, Douglas R. White et A. Kimball Romney (dir.), *Research methods in social network analysis*, Fairfax (VA), George Mason University Press, 1989, p. 11-40.

de bas-niveau L, existe-t-il une projection P telle que $P(L) = H$, pour toute paire empiriquement valide L et H ? Si oui, comment la trouver ? Le haut niveau H peut par exemple renvoyer à une certaine configuration des communautés présentes, à certaines régularités et motifs sociaux, ou à toute autre description sociologique qui puisse aussi s'exprimer en termes formels. Le bas-niveau L peut, quant à lui, se rapporter à l'historique des interactions entre agents (un réseau social) ou à diverses observations individuelles décrivant le statut des agents. De manière générale, il s'agira de descriptions au niveau des agents, formalisables, à nouveau. La projection P peut alors être vue dans certains cas comme une agrégation du microscopique vers le macroscopique. Nous illustrerons ces notions dans des cas concrets ci-après.

En tout état de cause, cette approche doit aussi être correcte dans un cadre évolutionnaire : soit des dynamiques empiriques λe et ηe sur le bas-niveau L et le haut-niveau H respectivement, il faut s'assurer que la projection P soit telle que $P(Lt + \Delta t) = Ht + \Delta t$, c'est-à-dire qu'il doit être possible de décrire l'observation finale de H à partir de l'évolution empirique de L. En d'autres termes, il faut que le résultat projeté d'une évolution de bas-niveau (des agents) corresponde à l'évolution de haut-niveau (du système). Ceci revient plus formellement à exiger que $Po\lambda e = \eta eoP$: le schéma de reconstruction détaillé à la figure 1 forme un diagramme commutatif, qui est fréquemment employé dans le contexte des systèmes dynamiques[3]. Ensuite, une fois que P est définie et valide, le second problème « (ii) » revient à montrer qu'une dynamique de bas-niveau permet la reconstruction de la dynamique de haut-niveau. Cette approche est généralement le cœur de tout travail de modélisation, cependant nous insistons ici sur le fait que les objets de bas-niveau doivent jouer un rôle central[4]. Ainsi, le second problème revient à trouver une dynamique λ qui reproduise la dynamique empirique de haut-niveau ηe, via P. Cette dynamique est stylisée : en tant que tels, les

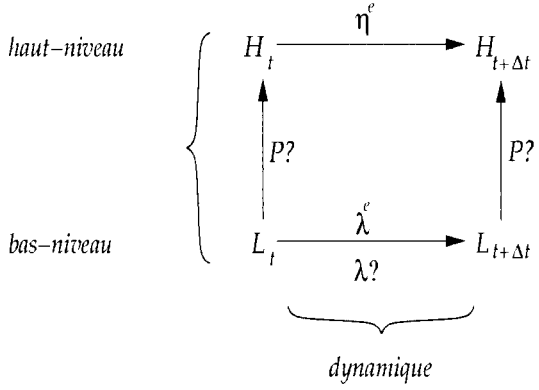
[3] Voir Alexander Rueger, « Robust supervenience and emergence », *Philosophy of science*, vol. 67, n° 3, 2000, p. 466-489 ; Martin Nilsson-Jacobi, « Hierarchical organization in smooth dynamical systems », *Artificial life*, vol. 11, n° 4, 2005, p. 493-512 ; Heather Turner *et al.*, « Rule migration: Exploring a design framework for emergence », *International journal of unconventional computing*, vol. 3, n° 1, 2007, à paraître.

[4] Eric Bonabeau, « Agent-based modeling: Methods and techniques for simulating human systems », *PNAS*, vol. 99, n° 3, 2002, p. 7280-7287.

objectifs du modèle sont restreints à la reconstruction du haut-niveau, c'est-à-dire que la dynamique du modèle λ peut ne pas être intégralement fidèle à la dynamique empirique ηe. Il est envisageable que l'état final empirique $L^{e}_{t+\Delta t}$ ne soit pas égal à l'état final calculé avec la dynamique du modèle $\lambda(Lt)$. Il suffit en effet que les faits stylisés de haut niveau soient correctement reconstruits : $Po\lambda = Po\lambda_{e}$, même si $\lambda \neq \lambda_{e}$. De fait, il n'est pas nécessaire que λ soit un modèle parfait de λ_{e}, il suffit qu'elle soit valide via P.

Figure 1

Le *problème de la reconstruction* revient à trouver 1) une projection P valide et 2) une dynamique artificielle λ satisfaisante, connaissant les dynamiques empiriques η^{e} et λ^{e}.



Ceci permet une reconstruction réussie même lorsqu'il est impossible de décrire totalement la dynamique empirique λ_{e} ou quand l'état de bas-niveau L est imparfaitement connu (il se peut que les données au sujet de certains agents soient incorrectes) – seules les descriptions de haut-niveau à reconstruire doivent être correctes. Par exemple, l'impossibilité de prédire le nombre réel d'amis d'un agent donné (un fait spécifique sur L) n'interdit pas de reconstruire le fait que la distribution des amis suit une loi puissance (un fait spécifique sur H). Si la dynamique est stylisée, elle n'est, par contre, pas artificielle et la construction de λ doit malgré tout être basée sur des observations empiriques.

Reconstruire une communauté de savoirs. Nous pouvons à présent nous focaliser sur le système complexe socio-sémantique mentionné plus

haut, une communauté de savoirs, pour lequel nous allons précisément résoudre un problème de reconstruction. Quelles interprétations sociologiques ce type de formalisme permet-il d'envisager ? Nous reconstruirons en effet plusieurs aspects de la structure d'une telle communauté, qui constitueront les phénomènes de haut-niveau, au premier rang desquels figure la description de la communauté en sous-communautés plus petites et plus précises. Ici, le concept de « communauté épistémique » est purement descriptif : il ne s'agit pas d'une coalition d'individus qui ont intérêt à rester alliés dans cette coalition, mais simplement d'un ensemble d'agents qui partagent les mêmes problématiques.

Nous voulons vérifier l'hypothèse suivant laquelle la structure d'une communauté de savoirs est principalement produite par la co-évolution des agents et des concepts. Dans un premier temps, nous exhibons une projection (P) qui fournit la structure de n'importe quelle communauté de savoirs (H) à partir de descriptions au niveau des agents et des concepts (L) : ceci correspond au premier problème. Une méthode adéquate et efficace pour achever cette tâche consiste à utiliser les treillis de Galois. Les historiens des sciences décrivent et classent traditionnellement les diverses communautés d'un champ de savoirs au sein de taxonomies auxquelles devront justement correspondre les taxonomies produites par notre méthode. Mieux, pour tout temps t, notre projection P permettra d'obtenir H_t à partir de L_t et ainsi, étant donné une dynamique empirique de bas-niveau λ_e, nous reproduirons la dynamique empirique de haut-niveau η_e. En conséquence, il s'agit aussi d'une description formelle d'une partie de la scientométrie et de l'épistémologie appliquée : décrire les champs scientifiques et l'évolution paradigmatique à partir de données quantitatives de bas-niveau.

Dans un second temps, nous procéderons à la micro-fondation des phénomènes de haut-niveau dans la dynamique du bas-niveau des agents et concepts ; ceci résoudra le second problème de la reconstruction. Plus précisément, nous introduirons un cadre co-évolutionnaire basé sur un réseau social, un réseau sémantique et un réseau socio-sémantique, c'est-à-dire un réseau épistémique constitué d'agents, de concepts et des liens entre eux. Nous montrerons ainsi que la dynamique de ce réseau épistémique suffit à reproduire divers faits stylisés pertinents. Étant donné H et une dynamique empirique η_e sur H, nous proposerons ainsi des méthodes pour concevoir un λ à partir

de données empiriques concernant L (au niveau du réseau épistémique) de sorte que $Po\lambda\ (L) = \eta_e oP\ (L)$. Nous soutiendrons ainsi notre affirmation selon laquelle les communautés épistémiques sont produites par la co-évolution entre agents et concepts.

1. Structure des communautés de savoirs

Les scientifiques, les journalistes, les communautés socioculturelles représentent diverses instances d'une société de savoirs, en étant des « sous-sociétés » plus petites, emboîtées, avec des sujets plus précis ; elles sont partiellement indépendantes, partiellement imbriquées. Toute communauté de savoirs semble structurée en diverses sous-communautés implicites, pendant que l'expertise est distribuée de manière hétérogène sur tous les agents : des frontières apparaissent entre les sous-groupes, à la fois horizontalement, avec différents domaines de compétence, et verticalement, avec différents niveaux de spécificité. Un tel « système complexe épistémique » réalise un travail de cognition sociale à grande échelle, les concepts étant introduits et manipulés par les agents de façon décentralisée, collective, interactive et en réseau.

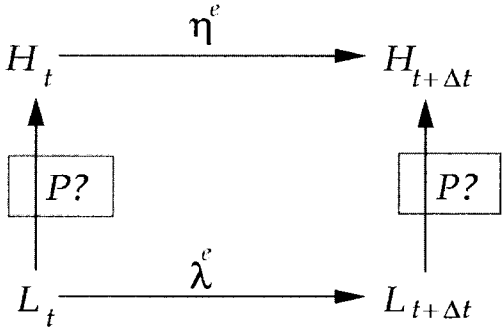
Toutefois, alors que les agents peuvent potentiellement accéder à une large portion des savoirs produits par la communauté épistémique tout entière, ils n'en connaissent en fait qu'une petite partie, principalement à cause de limitations cognitives et physiques. Plus précisément, les agents ont une représentation implicite de la structure de la communauté globale à laquelle ils participent : les embryologistes connaissent les fondements de la biologie moléculaire, de la biologie et de la science en général. Cette connaissance est toutefois limitée et subjective, similaire à une taxonomie populaire (*folk taxonomy*), au sens anthropologique[5], c'est-à-dire une taxonomie basée sur l'expérience individuelle à l'inverse des taxonomies scientifiques, considérées objectives et systématiques[6]. Ainsi, les historiens des sciences ont-ils souvent le dernier mot dans l'élaboration et dans la validation de ce type de métaconnaissances : les taxonomies d'experts sont prodigieusement plus

[5] Alejandro Lopez *et al.*, « The tree of life: Universal and cultural features of folkbiological taxonomies and inductions », *Cognitive psychology*, vol. 32, n° 3, 1997, p. 251-295.

[6] Brent Berlin, *Ethnobiological classification – Principles of categorization of plants and animals in traditional societies,* Princeton, Princeton University Press, 1992.

sûres que les taxonomies populaires ; elles sont toutefois très coûteuses à réaliser et rarement exhaustives.

Figure 2
Le premier problème de la reconstruction :
connaissant empiriquement les états L_t et la dynamique λ^e de bas-niveau,
quelle projection P produirait des observations de haut-niveau réalistes ?
Ici, le bas-niveau décrit l'usage, par les agents, de certains concepts ;
λ^e est ainsi simplement l'évolution empirique de cet usage.



Il est donc intéressant de permettre aux agents de comprendre la structure et l'activité de leur communauté de savoirs, à tout niveau de spécificité ou de généralité. Nous proposons une méthode pour créer automatiquement une taxonomie des champs de connaissances en construisant, en ordonnant et en manipulant l'hypergraphe épistémique d'une communauté donnée. L'hypergraphe épistémique est un graphe de communauté de savoirs, où chaque nœud est une communauté rassemblant à la fois des agents et des concepts. Nous appelons communauté de savoirs, ou « communauté épistémique », tout type de groupe d'agents intéressés par des problématiques communes[7] : un groupe de recherche sur un sujet précis, un champ entier, un paradigme, une discipline scientifique. En outre, cette notion n'est pas

[7] Peter Haas, « Introduction: Epistemic communities and international policy coordination », *International organization*, vol. 46, n° 1, hiver 1992, p. 1-35 ; Robin Cowan *et al.*, « The explicit economics of knowledge codification and tacitness », *Industrial & corporate change*, vol. 9, n° 2, 2000, p. 212-253.

nécessairement limitée à des groupes académiques ni à des « communautés de pratique »[8].

L'hypergraphe épistémique devrait rendre compte 1) des champs et des tendances qui sont présentes et 2) du type de relations qui les lient. En retour, cette taxonomie doit corroborer la perception intersubjective des agents : ainsi le *H* empirique donné par les experts sera comparé au *P(L)* produit par la méthode. Enfin, connaître la taxonomie à n'importe quelle période nous permet de décrire l'évolution du système et de reconstruire l'histoire de la communauté sur des bases objectives.

Nous présentons d'abord le cadre formel requis pour catégoriser hiérarchiquement les communautés épistémiques à travers une représentation de la communauté tout entière sous forme de treillis. Ensuite, nous montrons comment construire des taxonomies réduites (hypergraphes épistémiques) et suivre leur évolution, dans une approche tenant à la fois de l'épistémologie appliquée et de la scientométrie.

1.1. Cadre formel

Contexte. Divers cadres formels et procédés automatiques ont été proposés pour analyser les communautés de savoirs et trouver des groupes d'agents ou de documents liés par des notions communes. La plupart des travaux proviennent des disciplines liées à la découverte de connaissances dans les bases de données[9], parallèlement au développement important du contenu informationnel disponible sous forme électronique (en particulier les données scientifiques) et de la scientométrie qui a développé un riche ensemble de méthodes visant à caractériser spécifiquement de telles communautés[10]. En étudiant à la fois les articles, leurs auteurs et les concepts qu'ils utilisent, le but est de rendre compte de l'évolution des paradigmes[11] en utilisant, entre autres,

---

[8] Jean Lave et Etienne Wenger, *Situated learning: Legitimate peripheral participation*, Cambridge, Cambridge University Press, 1991.

[9] Anil K. Jain *et al.*, « Data clustering: A review », *ACM Computing Surveys*, vol. 31, n° 3, 1999, p. 264-323.

[10] Loet Leydesdorff, « In search of epistemic networks », *Social studies of science*, vol. 21, 1991, p. 75-110.

[11] Michel Callon *et al.*, *Mapping the dynamics of science and technology*, London, MacMillan Press, 1986 ; Katherine W. McCain, « Cocited author mapping as a valid representation of intellectual structure », *Journal of the American society for information*

les données de co-citation[12] ou de co-occurrence[13], afin de produire des cartes bidimensionnelles. Néanmoins, la plupart des approches sont fondées soit sur les relations sociales, avec des méthodes d'extraction de communautés issues de la théorie des graphes appliquée aux réseaux sociaux[14], soit sur la similarité sémantique, notamment des méthodes de *clustering* appliquées aux bases de données de documents où chacun d'entre eux est un vecteur dans un espace sémantique[15]. Il y a eu peu de tentatives de relier les aspects sociaux et sémantiques, alors que la notion de communauté épistémique est précisément duale ; d'un côté, un groupe d'agents qui, de l'autre côté, partagent des concepts communs.

Avec cette profusion de méthodes de découvertes de communautés, souvent issues de l'intelligence artificielle, il devient intéressant de savoir comment représenter les communautés de manière ordonnée. Globalement, diverses techniques permettent de produire et de représenter des structures catégorielles dont, notamment, le *clustering* hiérarchique[16], l'analyse de concepts formels (*formal concept analysis*)[17], les applications de

science, vol. 37, n° 3, 1986, p. 111-122.

12 Henry Kreuzman, « A co-citation analysis of representative authors in philosophy: Examining the relationship between epistemologists and philosophers of science », *Scientometrics*, vol. 51, n° 3, 2001, p. 525-539.

13 Ed C. M. Noyons et Anthony F. J. van Raan, « Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research », *Journal of the American society for information science*, vol. 49, n° 1, 1998, p. 68-81.

14 Stanley Wasserman et Katherine Faust, *Social network analysis: Methods and applications*, Cambridge, Cambridge University Press, 1994.

15 Gerard Salton *et al.*, « Vector space model for automatic indexing », *Communications of the ACM*, vol. 18, n° 11, 1975, p. 613-620.

16 Stephen C. Johnson, « Hierarchical clustering schemes », *Psychometrika*, vol. 2, 1967, p. 241-254.

17 Rudolf Wille, « Restructuring lattice theory: An approach based on hierarchies of concepts », dans Ivan Rival (dir.), *Ordered sets*, Dordrecht-Boston, Reidel, 1982, p. 445-470.

la théorie des graphes[18], les réseaux neuronaux[19]. Ici, la notion de taxonomie – un ensemble ordonné de catégories, ou taxons – est particulièrement pertinente pour les communautés de savoirs ; elle est par ailleurs fréquemment utilisée en biologie[20], en psychologie cognitive[21] et en ethnographie et anthropologie[22]. Parallèlement, alors que les taxonomies ont initialement été fabriquées au travers d'approches subjectives, les méthodes sont à présent devenues davantage formelles et statistiques[23]. Cependant, la construction de taxonomies elle-même est en général rarement étudiée et souvent limitée à la création de dendrogrammes ou d'arbres ; en outre, la dynamique des taxonomies a été relativement négligée. Notre intention est ainsi de nous attaquer à ces deux sujets : construire d'abord une taxonomie pertinente pour décrire les communautés épistémiques, puis suivre leur évolution (figure 2).

Communautés épistémiques : définitions. Nous cherchons à savoir 1) quels agents partagent les mêmes concepts, et 2) quels sont ces concepts. Ainsi, notre définition d'une communauté épistémique est simplement caractérisée par des problématiques communes et ne doit pas nécessairement être une communauté sociale.

Définition 1 (communauté épistémique). Soit un ensemble d'agents S et les concepts qu'ils ont en commun ; nous appelons « communauté épistémique de S » le plus grand ensemble d'agents qui utilisent ces concepts.

Considérer la communauté épistémique (CE) d'un ensemble d'agents étend cet ensemble au plus grand groupe d'agents qui partagent ses

18 Harrison C. White *et al.*, « Social-structure from multiple networks. I: Blockmodels of roles and positions », *American journal of sociology*, vol. 81, 1976, p. 730-780 ; Mark E. J. Newman, « Detecting community structure in networks », *European physical journal B*, vol. 38, 2004, p. 321-330.

19 Teuvo Kohonen, *Self-organizing maps*, 3e édition, Berlin, Springer, 2000.

20 Robert H. Whittaker, « New concepts of kingdoms of organisms », *Science*, vol. 163, 1969, p. 150-160.

21 Eleanor Rosch et Barbara L. Lloyd, « Cognition and categorization », *American psychologist*, vol. 44, no 12, 1978, p. 1468-1481.

22 Brent Berlin, *op. cit.*

23 Robert R. Sokal et Peter H. A. Sneath, *Principles of Numerical Taxonomy*, San Francisco (CA), W. H. Freeman, 1963.

concepts. Cette notion est proche de l'« équivalence structurelle[24] » : les CE sont des groupes d'agents liés de manière équivalente à certains concepts. Formellement, nous lions les agents aux concepts grâce à une relation binaire R entre l'ensemble de tous les agents S et l'ensemble de tous les concepts C. Ici, $R \subseteq \mathrm{S} \times \mathrm{C}$ exprime n'importe quel type de lien entre un agent s et un concept c : dans notre cas, le lien représente le fait que s a utilisé c. Ensuite, nous introduisons l'opération « ^ » telle que pour tout sous-ensemble $S \subseteq \mathrm{S}$, on dénote par $S^{\wedge}$ l'ensemble des éléments de **C** qui sont R-liés *à tout* élément de S, c'est-à-dire : $s^{\wedge} = \{c \in \mathbf{C} \mid sRc\}$ et $S^{\wedge} = \{c \in \mathbf{C} \mid \forall s \in S, sRc\}$. De façon similaire, « * » est l'opération duale telle que $\forall c \in \mathbf{C}$, $\forall C \subseteq \mathbf{C}$, $c^{*} = \{s \in \mathbf{S} \mid sRc\}$ et $C^{*} = \{s \in \mathbf{S} \mid \forall c \in C, sRc\}$. Par définition $(\varnothing)^{\wedge} = \mathbf{C}$ et $(\varnothing)^{*} = \mathbf{S}$.

Autrement dit, $S^{\wedge}$ dénote l'*intension* d'un ensemble d'agents S, soit l'ensemble des concepts utilisés par chaque agent de S ($\forall s \in S$) ; et C^{*} l'*extension* d'un ensemble de concepts C, c'est-à-dire les agents utilisant chaque concept de C. Ce formalisme est une manière robuste de rendre compte de notions abstraites, au sens philosophique[25], caractérisées par leur extension (implémentation physique) et leur intension (contenu interne) : les concepts sont des propriétés des auteurs qui les utilisent et les auteurs sont les *loci* des concepts.

Opération de clôture, hypergraphe épistémique. L'opération jointe « $^{\wedge *}$ » est une opération de clôture[26], à savoir qu'elle est 1) extensive ($S \subseteq S^{\wedge *}$), 2) idempotente ($(S^{\wedge *})^{\wedge *} = S^{\wedge *}$) et 3) croissante ($S \subseteq S' \Rightarrow S^{\wedge *} \subseteq S'^{\wedge *}$). $S^{\wedge *}$ est appelé la clôture de S. L'extensivité signifie que la clôture n'est jamais moins grande que l'ensemble initial et l'idempotence implique qu'appliquer « $^{\wedge *}$ » plus d'une fois ne change plus la clôture. Finalement, la croissance de « $^{\wedge *}$ » indique que la clôture d'un ensemble plus grand est, elle aussi, plus grande.

Plus simplement, appliquer ainsi « $^{\wedge *}$ » à S retourne tous les agents utilisant le même ensemble de concepts que les agents de S ont en commun : « $^{\wedge *}$ » fournit la CE de n'importe quel ensemble d'agents, une

[24] François Lorrain et Harrison C. White, « Structural equivalence of individuals in social networks » *Journal of mathematical Sociology*, vol. 1, 1971, p. 49-80.

[25] Rudolf Wille, « Concept lattices and conceptual knowledge systems », *Computers mathematics and applications*, vol. 23, 1992, p. 493-515.

[26] Garrett Birkhoff, *Lattice theory*, Providence (RI), American mathematical society, 1948.

fois pour toutes. Soit deux sous-ensembles $S \subseteq \mathbf{S}$ et $C \subseteq \mathbf{C}$, un couple (S, C) est dit clos si et seulement si $C = S^{\wedge}$ et $S = C^{*}$. Comme $(S^{\wedge *}, S^{\wedge})$ est la communauté épistémique basée sur S, tout couple clos est une communauté épistémique. Nous pouvons maintenant introduire la notion d'« hypergraphe épistémique ».

Définition 2 (graphe, hypergraphe). Un graphe **G** est un couple (V, E) où V est un ensemble de sommets et $E \subseteq V \times V$ un ensemble d'arêtes liant des paires de sommets. Un hypergraphe **hG** est un couple (V, hE) où V est un ensemble de sommets et hE un ensemble d'hyper-arêtes connectant un ensemble de sommets. hE est donc fondamentalement un sous-ensemble de $\wp(V)$, l'ensemble des parties de V.

Définition 3 (hypergraphe épistémique). L'hypergraphe épistémique de S est l'hypergraphe de CE $(\mathbf{S}, \{S^{\wedge *} \mid S \subseteq \mathbf{S}\})$ avec des hyper-arêtes reliant des groupes d'agents appartenant à une même CE.

Chaque hyper-arête peut être étiquetée avec l'ensemble de concepts correspondant à l'ensemble d'agents qu'elle relie, $S^{\wedge}$. Notons que toutes ces propriétés sont similaires et, en fait, duales en considérant les CE basées sur des concepts, obtenues grâce à « $*^{\wedge}$ ». Un hypergraphe épistémique pourrait, de façon équivalente, être basé sur des concepts : $(\mathbf{C}, \{C^{*\wedge} \mid C \subseteq \mathbf{C}\})$, avec des hyper-arêtes liant les concepts d'une même CE.

Plus généralement, en s'éloignant du strict symbolisme formel, l'hypergraphe épistémique est une structure qui permet d'encoder les diverses CE propres à un groupe socio-sémantique donné. Il reste à montrer en quoi cette structure peut nous aider à représenter la structure taxonomique d'une communauté de savoirs.

1.2. Treillis de Galois : des relations aux taxonomies dynamiques

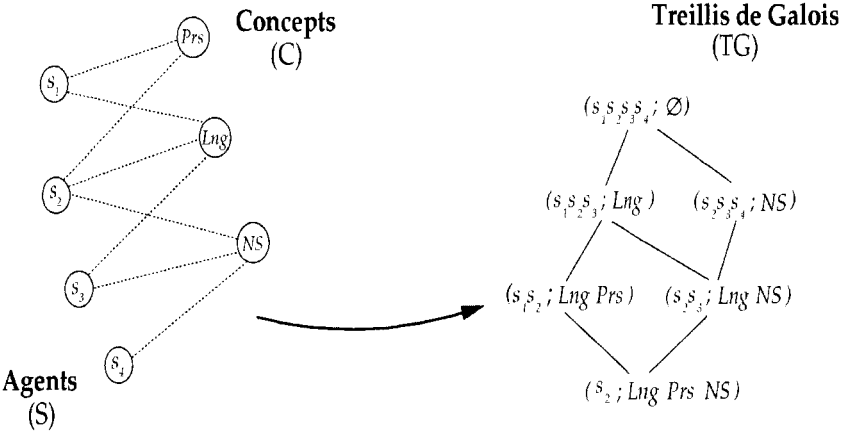
Taxonomies et treillis. Une relation entre agents et concepts est en effet suffisante pour capturer les communautés sous-jacentes d'un champ scientifique donné. Il faut néanmoins encore hiérarchiser l'ensemble brut des CE pour construire une taxonomie. L'approche canonique aristotélicienne pour ranger les catégories consiste à utiliser des arbres : les catégories sont des nœuds, et les sous-catégories sont les fils de leur

unique catégorie-parent. Dans ce cas, il est difficile, voire impossible, de gérer les objets appartenant à des catégories multiples ou paradigmatiques. Le treillis est une amélioration immédiate de la structure d'arbre en permettant notamment de représenter le recouvrement de catégories (les taxons peuvent avoir plus d'un ascendant). Afin de représenter les CE hiérarchiquement dans une taxonomie à base de treillis, nous introduisons d'abord un *ordre partiel* « $\angle$ » entre CE en permettant de classer certaines CE les unes par rapport aux autres.

En particulier, une CE $(S, S^{\wedge})$ est un sous-champ d'un champ $(S', S'^{\wedge})$ si son intension est plus précise que celle du champ : $(S, S^{\wedge}) \angle (S', S'^{\wedge}) \Leftrightarrow S \subseteq S'$. Nous pouvons ainsi rendre compte à la fois de la généralisation et de la spécialisation d'un couple clos car $(S, S^{\wedge})$ peut être vu comme une spécification de $(S, S^{\wedge})$ (plus de concepts, moins d'agents) et inversement $(S, S^{\wedge})$ est un « super-champ » ou une généralisation de $(S, S^{\wedge})$. Le treillis de Galois (TG) est alors une structure naturelle pour représenter les CE (figure 3).

Figure 3

Création d'un treillis de Galois
de six CE (à droite) à partir d'une communauté (à gauche)
contenant les agents s_1, s_2, s_3, s_4 et les concepts « linguistique » (*Lng*),
« neuroscience » (*NS*), « prosodie » (*Prs*).
Les CE sont des couples (extension, intension)=(S, C) avec $S^{\wedge}=C$ et $C^{*}=S$.
Une CE plus près du sommet est plus générale : la hiérarchie reflète la relation de généralisation/spécialisation induite par « $\angle$ ».



Définition 4 (treillis de Galois). Le treillis de Galois G**S**,**C**,R est l'ensemble de tous les couples clos (S,C) $\subseteq$ **S** x **C** par la relation R : G**S**,**C**,R={(S^, S^*) | S $\subseteq$ **S**}, partiellement ordonné par « $\angle$ ».

Élaguer le treillis. La taille des TG est néanmoins un désavantage majeur, étant potentiellement exponentielle et pouvant déjà atteindre plusieurs centaines de milliers de CE avec peu d'agents et de concepts. Un TG contient toutes les CE parmi lesquelles plusieurs ne correspondent pas à un champ de savoirs véritable ou pertinent : comment produire, alors, une représentation utile et utilisable ? En d'autres termes, nous voulons sélectionner et extraire les CE majeures à partir d'un TG potentiellement énorme, tout en excluant les CE non significatives afin d'être plus près des taxonomies d'expert. Formellement, l'hypergraphe épistémique partiel des CE extraites est un ensemble partiellement ordonné qui se superpose à la structure du treillis et qui bénéficie encore des propriétés taxonomiques qui nous intéressent.

Nous devons formuler des critères permettant de distinguer les CE utiles en vue de décrire concisément la taxonomie. L'importance de ce processus de sélection a été jusqu'ici relativement sous-estimée dans l'étude des TG, une grande part des travaux se focalisant sur le calcul et la représentation[27], tandis que peu d'auteurs insistent sur le besoin d'interprétation et d'approximation pour gérer la complexité combinatoire des TG[28]. *A priori*, nous voudrions certainement garder les CE les plus peuplées : si un ensemble de concepts correspond à un champ, son extension devrait être assez grande. Toutefois, certaines CE sont trop spécifiques alors que de petites CE près du sommet sont probablement intéressantes en tant que champs minoritaires. Nous proposons

[27] Robert Godin *et al.*, « Design of class hierarchies based on concept (Galois) lattices », *Theory and practice of object systems (TAPOS)*, vol. 4, n° 2, 1998, p. 117-134 ; Sergei O. Kuznetsov et Sergei A. Obiedkov, « Comparing performance of algorithms for generating concept lattices », *Journal of experimental and theoretical artificial intelligence*, vol. 14, n^{os} 2-3, 2002, p. 189-216.

[28] Gerd Stumme *et al.*, « Computing iceberg concept lattices with TITANIC », *Data and Knowledge Engineering*, vol. 42, 2002, p. 189-222 ; Frederik J. Van Der Merwe et Derrick G. Kourie, « Compressed pseudo-lattices », *Journal of experimental and theoretical artificial intelligence*, vol. 14, n^{os} 2-3, 2002, p. 229-254 ; Vincent Duquenne *et al.*, « Structuration of phenotypes and genotypes through Galois lattices and implicationsé », *Applied artificial intelligence*, vol. 17, n° 3, 2003, p. 243-256.

plusieurs critères de sélection : 1) la taille de l'extension, 2) le niveau (plus petite distance au sommet), 3) la spécificité (taille de l'intension), 4) les sous-communautés (nombre de descendants). Ensuite, nous fabriquons plusieurs heuristiques de sélection attribuant un score à chaque CE en combinant ces critères, de sorte que nous ne gardions que les meilleures CE : par exemple, en favorisant 1) les grandes CE, 2) près du sommet, 3) anormalement spécifiques, 4) ayant peu de descendants. Suivant l'objectif de la reconstruction, le réglage de ces heuristiques s'appuie fortement sur le contexte empirique.

Évolution des taxonomies. Nous voulons aussi pouvoir retrouver l'histoire d'un champ en étudiant longitudinalement les taxonomies. Un premier ensemble de motifs dynamiques permet de décrire l'évolution épistémique : 1) le progrès ou le déclin d'un champ, 2) l'enrichissement ou l'appauvrissement d'un champ (réduction ou extension de l'ensemble des concepts qui y sont liés), et 3) la fusion ou la scission de disciplines (l'émergence ou la disparition de CE constituées de plusieurs champs). En termes de changements entre deux périodes successives, ces motifs se traduisent simplement par une variation dans la population d'une CE donnée, variation dont l'interprétation dépend finalement de la position de la CE dans l'hypergraphe épistémique partiel : suivant 1) qu'il y ait simplement une variation de population, 2) qu'elle ait lieu pour un sous-champ ou 3) que ce sous-champ soit en fait un sous-champ commun à plusieurs CE (ayant, par exemple, deux CE pour parents).

1.3. Étude de cas

Protocole empirique. Nous avons étudié la communauté des embryologistes travaillant sur l'animal-modèle *zebrafish* pendant la période de 1990 à 2003, considérée par les experts du domaine comme le début de la croissance majeure de la communauté. Nous utilisons des données indiquant quand un agent s se sert d'un concept c, et nous adoptons des hypothèses linguistiques simplistes en supposant qu'un terme lemmatisé correspond à une notion (un terme lemmatisé est un terme dont on ne conserve que la racine : le lemme de « général » est « gener », tout comme le lemme de « générales » – ceci permet de ne pas distinguer les différentes variations d'un terme). À l'aide de notre expert, nous restreignons le dictionnaire aux 70 mots les plus utilisés et signifiants de

la communauté afin d'éviter les termes neutres et rhétoriques (tels que *the, however* et *study*) et paradigmatiques (tels que *pattern* et *development* qui ne sont pas porteurs de sens au sein de cette communauté).

Nous attribuons ainsi une notion à un agent dès qu'un mot lemmatisé apparaît dans le titre ou dans le résumé de l'un de ses articles. Notre principale source de données est *MedLine*, un recueil de références bibliographiques produit par la Librairie nationale de médecine des États-Unis. Nous divisons la base de données en plusieurs périodes et construisons une série de matrices de relation (R) agrégeant tous les événements de chaque période correspondante. Avant de procéder, nous devons aussi spécifier la façon dont nous choisissons :

i. la largeur de la période, c'est-à-dire la taille d'une période, qui doit être assez grande pour récolter suffisamment d'information et assez petite pour rendre compte précisément des évolutions ;
ii. l'incrément de temps entre deux périodes, qui définit le rythme d'observation et doit être plus court que la largeur de la période.

Nous examinons finalement trois périodes : 1990-1995, 1994-1999 et 1998-2003 (période de six ans et incrément de quatre ans). Pour limiter le temps de calcul, nous avons aussi considéré un échantillon aléatoire de 255 auteurs pour chaque période. Avec cette taille fixe d'échantillon, nous pouvons comparer l'importance relative de chaque champ par rapport aux autres au sein de la taxonomie en évolution.

Reconstruire l'histoire. En observant les résultats de la figure, nous pouvons suggérer l'histoire suivante : 1) l'étude du cerveau et de la moelle épinière a décru, avec moins de liens vers les aspects ventraux/dorsaux, 2) la communauté a commencé à s'intéresser aux relations entre le signal, les chemins (*pathway*) et les récepteurs (tous liés en réalité aux messages biochimiques) et, finalement, 3) il y a eu un intérêt massif à l'égard des sujets liés à l'homme et des nouveaux liens avec l'étude des gènes homologues et des vertébrés, ce qui souligne la croissance des études comparatives et interdisciplinaires. Le point 2) implique davantage que la simple émergence de nombreuses sous-communautés communes : toutes les paires de concepts dans l'ensemble {*growth, pathway, receptor, signal*} forment une clique de communautés jointes, un motif qui peut être interprété comme l'émergence d'un sous-paradigme (figure 4). Nous avons comparé avec succès ces résultats à des taxonomies empiriques provenant à la fois d'experts, membres de

la communauté en question, de la littérature[29] et, dans une moindre mesure, d'actes de la conférence de référence de cette communauté (Cold Spring Harbor Laboratory).

1.4. Une histoire des sciences formelle

Jusqu'ici la détection de catégories et, par ricochet, de communautés, a été principalement étudiée en informatique théorique[30] (avec des liens peu évidents avec ce que les sciences sociales appellent des communautés) et en sociologie, laquelle introduit des hypothèses et des outils propres aux réseaux sociaux[31] et fournit des méthodes de catégorisation mieux adaptées à la détection de groupes sociaux. Toutefois, la plupart de ces méthodes ne permettent pas aux agents d'appartenir à plusieurs communautés se chevauchant sans être encastrées – les agents doivent au mieux faire partie d'une lignée croissante de communautés. Ce problème disparaît en utilisant des treillis. En outre, nos communautés épistémiques auraient été difficiles à découvrir en utilisant des méthodes fondées sur un unique réseau, par exemple seulement un réseau social : les agents d'une même CE ne sont pas nécessairement liés socialement.

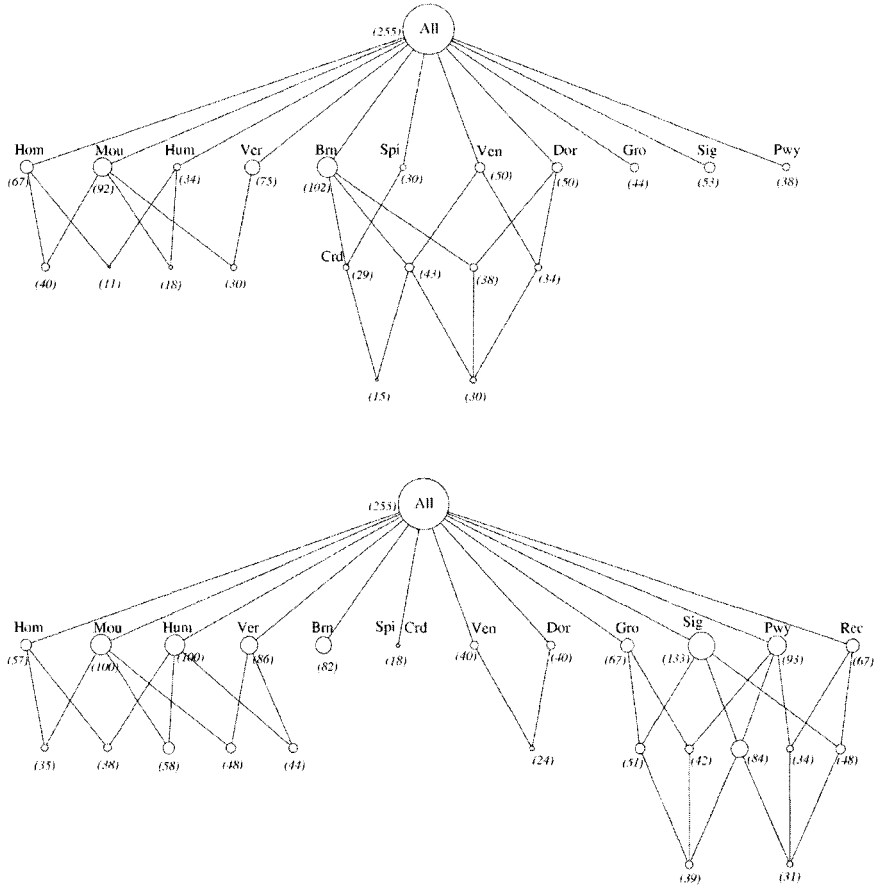
Nous avons proposé une méthode pour créer une taxonomie pertinente d'une communauté de savoirs. Nous avons montré que les TG permettent d'organiser automatiquement et hiérarchiquement une communauté en champs et en sous-champs, en rendant compte des recouvrements entre communautés épistémiques, communément appelés champs interdisciplinaires. Cependant, les TG ne réduisent pas beaucoup les données et nous avons ainsi introduit des critères pour discriminer les CE intéressantes, produisant de la sorte un hypergraphe

[29] Kimberly Dooley et Leonard I. Zon, « Zebrafish: A model system for the study of human disease », *Current opinion in genetics & development*, vol. 10, n° 3, 2000, p. 252-256 ; David J. Grunwald et Judith S. Eisen, « Headwaters of the zebrafish – Emergence of a new model vertebrate », *Nature reviews. Genetics*, vol. 3, n° 9, 2002, p. 717-724 ; Jane Bradbury, « Small fish, big science », *PLoS Biology*, vol. 2, n° 5, 2004, p. 568-572.

[30] John A. Hartigan, *Clustering algorithms*, New York (NY), Wiley, 1975 ; Mark E. J. Newman, *op. cit.*, 2004.

[31] Linton C. Freeman, « A set of measures of centrality based on betweenness », *Sociometry*, vol. 40, 1977, p. 35-41 ; Ronald S. Burt, « Cohesion versus structural equivalence as a basis for network subgroups », *Sociological methods and research*, vol. 7, 1978, p. 189-212 ; Stanley Wasserman et Katherine Faust, *op. cit.*

Figure 4
Deux hypergraphes épistémiques partiels
représentant la communauté à la fin de 1995 (haut) et à la fin de 2003 (bas).
Le nombre d'agents de chaque CE est donné entre parenthèses.



Légende : All : la communauté entière, Hom : *homologue / homologous*, Mou : *mouse,* Hum : *human,* Ver : *vertebrate,* Brn : *brain/ neural/ nervous/ neuron,* Spi : *spinal,* Crd : *cord,* Ven : *ventral,* Dor : *dorsal,* Gro : *growth,* Sig : *signal,* Pwy : *pathway,* Rec : receptor.

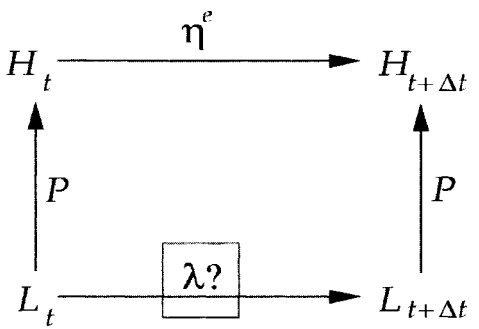
épistémique partiel qui est ainsi une représentation utilisable et informative de la structure de la communauté. L'étude longitudinale rend possible la description historique en capturant des faits stylisés liés à l'évolution épistémique, tels que le progrès, le déclin et l'interaction d'un champ. Nous avons finalement appliqué notre méthode à la sous-communauté des embryologistes travaillant sur l'animal-modèle *zebrafish* et validé nos résultats avec des taxonomies d'expert. En d'autres termes, nous avons conçu une fonction de projection P valide du bas-niveau des relations entre agents et concepts vers le haut-niveau des descriptions épistémologiques. En particulier, les deux hypergraphes épistémiques partiels peuvent être vus comme $P(L_{1995})$ et $P(L_{2003})$, corroborant les H_{1995} et H_{2003} fournis par les experts. La transition de H_{1995} à H_{2003} (η_e) est aussi reproduite : la dynamique épistémique reconstruite (η) est valide.

Plus généralement, on peut remplacer les auteurs par des objets, et les concepts par des propriétés : les TG constituent une méthode générique pour produire et analyser les taxonomies de nombreux autres domaines. Au tout premier plan, ils peuvent aider les historiens des sciences, notamment lorsqu'il y a beaucoup de données. Cette étude peut être considérée comme la première étude non subjective de la communauté « *zebrafish* ». En plus de l'épistémologie, de la scientométrie, de la sociologie et de l'histoire en général, d'autres domaines d'application et de validation sont possibles : l'économie (entreprises et technologies), la linguistique (mots et contextes). Des résultats significatifs dans divers domaines distincts renforceraient la pertinence de l'utilisation des TG pour ce type de tâches.

2. Micro-fondations des réseaux épistémiques

Dans la section précédente, nous avons considéré la structure des CE comme un fait stylisé de haut-niveau d'un système complexe socio-sémantique. Ici, nous « micro-fondons » cette caractéristique, c'est-à-dire que nous la reconstruisons à partir du bas-niveau des interactions entre agents au sein d'un réseau épistémique. Plus généralement, ceci revient à trouver la solution d'un problème inverse : quelles sont les dynamiques (éventuellement minimales) qui permettent de reproduire la structure d'un tel système évolutionnaire ? En d'autres termes, nous cherchons un modèle de morphogenèse du réseau épistémique qui

Figure 5
Le second problème de la reconstruction :
quelle dynamique de bas-niveau λ produirait, via *P*,
des observations de haut-niveau empiriquement valides ?



corrobore les observations empiriques. Dans le cadre plus général du problème de reconstruction, ceci revient à trouver une dynamique λ décrivant les actions et l'évolution des agents telle que pour η_e et P données, on ait $P \circ \lambda = \eta_e \circ P$: la dynamique λ suffit à reconstruire la dynamique de haut-niveau empirique η_e.

Nous faisons ainsi l'hypothèse que modéliser le niveau des interactions entre agents co-évoluant avec les concepts qu'ils manipulent suffit à réaliser la reconstruction micro-fondée de ce système complexe social. Cette question est liée plus largement à un problème récent en sociologie structurale – la modélisation des réseaux sociaux – qui met en jeu diverses disciplines, de la théorie des graphes (utilisée à la fois en informatique et en physique statistique) à la sociologie mathématique en passant par l'économie[32]. Ce récent engouement provient essentiellement de l'observation que la structure des réseaux sociaux réels diffère fortement de celle des graphes aléatoires uniformes à la Erdós-Rényi (ER)[33] et donc que les agents interagissent de manière non aléatoire, en

[32] Brian Skyrms et Robin Pemantle, « A dynamic model of social network formation », *PNAS*, vol. 97, n° 16, 2000, p. 9340-9346 ; Réka Albert et Albert-Laszlo Barabási, « Statistical mechanics of complex networks », *Reviews of modern physics*, vol. 74, 2002, p. 47-97 ; Patrick Cohendet et al., « Émergence, formation et dynamique des réseaux – modèles de la morphogenèse », *Revue d'économie industrielle*, vol. 103, n^{os} 2-3, 2003, p. 15-42.

[33] Paul Erdós et Alfred Rényi, « On random graphs », *Publicationes mathematicæ*, vol. 6, 1959, p. 290-297.

fonction de préférences hétérogènes pour les autres nœuds du réseau. Alors que ce comportement était déjà abondamment documenté en sciences sociales[34], les modèles de réseaux ont pourtant été longtemps limités à des graphes aléatoires, notamment « ER »[35]. De fait, de nombreux travaux récents mettent en œuvre des modèles fondés sur des mécanismes d'interaction non uniforme et de croissance du réseau, afin de reconstruire des structures dont le réalisme est évalué à travers le prisme d'un large ensemble de paramètres statistiques[36].

L'objectif de cette section est double : d'abord, concevoir des outils pour mesurer empiriquement les phénomènes de bas-niveau à l'origine de l'évolution de ces réseaux, afin de définir des comportements d'interaction non arbitraires. Ensuite, utiliser ces mesures pour introduire un modèle qui reproduise des faits stylisés pertinents observés dans le réseau épistémique réel des scientifiques travaillant sur le *zebrafish*.

2.1. Réseaux

De la mesure au modèle. Les réseaux (ou graphes) sont omniprésents dans le monde réel : du plus bas niveau des interactions physiques à des niveaux de description tels que la biologie, la sociologie, l'économie et la linguistique. Pendant longtemps, cependant, l'approche des réseaux a été restreinte à des travaux essentiellement abstraits en théorie des graphes et à des études empiriques de petite taille, tandis que les modèles de réseau étaient limités au travail séminal d'Erdós et Rényi[37], dont on considérait assez souvent qu'il était réaliste pour la plupart des applications. Récemment, l'augmentation des performances computa-

[34] John C. Touhey, « Situated identities, attitude similarity, and interpersonal attraction », *Sociometry*, vol. 37, 1974, p. 363-374 ; Miller McPherson *et al.*, « Birds of a feather: Homophily in social networks ». *Annual review of sociology*, vol. 27, 2001, p. 415-444.

[35] Robert K. May, « Will a large complex system be stable ? », *Nature*, vol. 238, 1972, p. 413-414 ; Andrew Barbour et Denis Mollison, « Epidemics and random graphs », dans J.-P. Gabriel, C. Lefevre et P. Picard (dir.), *Stochastic processes in epidemic theory*, printemps 1990, Lecture notes in biomaths, 86, p. 86-89 ; Stanley Wasserman et Katherine Faust, *op. cit.*.

[36] Tom A. Snijders, « The statistical evaluation of social networks dynamics », *Sociological methodology*, vol. 31, 2001, p. 361-395 ; Sergey N. Dorogovtsev et José F. F. Mendes, *Evolution of networks – From biological nets to the Internet and WWW*, Oxford, Oxford University Press, 2003.

[37] Paul Erdós et Alfred Rényi, *op. cit.*

tionnelles a rendu possible l'usage de méthodes quantitatives sur de grands réseaux, amenant de nouveaux résultats qui révèlent les défauts des anciens modèles. Trois paramètres statistiques notamment ont permis d'avoir un point de vue absolument nouveau sur la topologie des réseaux : 1) le coefficient d'agrégation (ou *clustering coefficient* – la proportion de voisins d'un nœud qui sont aussi connectés entre eux, soit une mesure de la transitivité), 2) la distance moyenne (la longueur du plus court chemin entre deux nœuds) et 3) la distribution des degrés (ou connectivité des nœuds). D'autres paramètres statistiques pertinents ont été proposés et mesurés empiriquement, menant à des modèles de morphogenèse corroborant les données empiriques, allant ainsi à l'encontre du modèle ER afin de le remplacer[38].

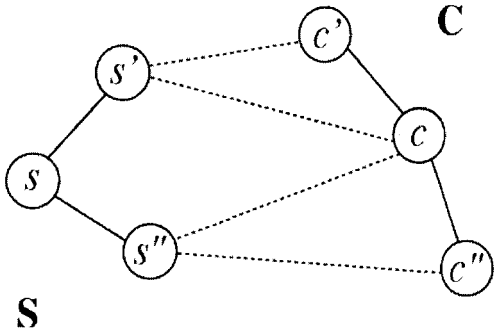
Plus spécifiquement, Barabási et Albert (BA)[39] ont souligné que la croissance du réseau et l'attachement préférentiel pouvaient être des processus clés : ils ont reconstruit une distribution de degrés en « loi-puissance », donc empiriquement réaliste, à l'aide d'un modèle où les nouveaux nœuds apparaissent à vitesse constante et s'attachent aux anciens nœuds proportionnellement à leur degré. Par la suite, de nombreux auteurs ont introduit des modèles de morphogenèse de réseaux fondés sur divers modes de création préférentielle de liens dépendant de diverses propriétés et de mécanismes, tandis que les processus de croissance consistent principalement en l'addition régulière de nœuds. L'idée générale est d'exhiber des paramètres statistiques de haut-niveau et de suggérer des processus au niveau du réseau qui permettent de déduire les premiers des seconds. Après avoir sélectionné un ensemble de faits stylisés pertinents à expliquer, la conception du modèle dépend en toute logique de deux sous-tâches : définir la façon dont les agents sont censés interagir ainsi que spécifier comment le réseau croît. Néanmoins, même dans les travaux récents, les hypothèses sur ce type de mécanismes sont rarement vérifiées empiriquement. Cette attitude, qui reste appropriée dans le cas de modèles normatifs, peut

[38] Duncan J. Watts et Steven H. Strogatz, « Collective dynamics of "small-world" networks », *Nature*, vol. 393, 1998, p. 440-442 ; Michalis Faloutsos *et al.*, « On power-law relationships of the Internet topology », *Computer communication review*, vol. 29, n° 4, 1999, p. 251-262 ; Albert-Laszlo Barabási et Réka Albert, « Emergence of scaling in random networks », *Science*, vol. 286, 1999, p. 509-512.

[39] Albert-Laszlo Barabási et Réka Albert, « Emergence of scaling in random networks », *op. cit.*

sembler téméraire dans le cas de modèles descriptifs. Nous nous attacherons ainsi 1) à exhiber des faits stylisés de haut-niveau propres aux réseaux épistémiques (notamment la structure de CE), 2) à suggérer des phénomènes de bas-niveau pertinents pour rendre compte de ces faits de haut-niveau, 3) à concevoir des outils de mesure pour appréhender ces phénomènes de bas-niveau et 4) à proposer un modèle de reconstruction basé sur la dynamique de bas-niveau ainsi observée qui reconstruise correctement les faits de haut-niveau voulus. Cet enchaînement de tâches décrit précisément une reconstruction dynamique réussie.

Figure 6
Exemple de réseau épistémique avec $S=\{s, s', s''\}$, $C=\{c, c', c''\}$, et les relations R_S, R_C (lignes pleines) et R (pointillés).



Réseaux épistémiques. Nous enrichissons d'abord le réseau socio-sémantique de la première partie avec un « réseau social » (liant les agents) et un « réseau sémantique » (liant les concepts). Un « réseau épistémique » est ainsi donné par ces trois réseaux, cruciaux pour expliquer l'influence réciproque et la co-évolution des auteurs et concepts (figure 6). Ainsi, un réseau épistémique est constitué d'un réseau social, d'un réseau sémantique et d'un réseau socio-sémantique. Les nœuds dans le réseau social S sont des agents, et les liens représentent la co-occurrence de deux agents au sein d'un événement. Plus formellement, $S = (S,E_S)$, où S est l'ensemble des agents et E_S l'ensemble des liens non dirigés. Le réseau sémantique $C=(C,E_C)$ est le

réseau des co-occurrences des concepts au sein des événements : C dénote l'ensemble des concepts, E_C dénote les liens entre concepts. Le réseau socio-sémantique G_{SC} est composé d'agents de S, de concepts de C, et de liens entre eux, $E_{SC} = R$, dénotant l'usage des concepts par les agents.

2.2. Caractéristiques de haut-niveau

Suivant le contexte de la section 1, les événements sont des articles, les agents sont leurs auteurs et les concepts sont choisis parmi une sélection de mots des résumés. Comme précédemment, les données empiriques proviennent de données bibliographiques concernant des embryologistes travaillant sur le *zebrafish*, ici durant la période 1997-2004. L'échantillon contient environ 10 000 auteurs, 6 000 articles et 70 concepts. Les 70 concepts sont identiques à ceux qui ont été choisis précédemment, et ils sont donnés *a priori* : dans le réseau sémantique, seuls de nouveaux liens peuvent apparaître, pas de nouveaux concepts (par contre, les concepts peuvent être de plus en plus utilisés).

Nous ajoutons à présent à la description de haut-niveau de la section 1 les paramètres statistiques suivants spécifiques à ce réseau – de fait, il s'agit principalement de paramètres bipartis (c'est-à-dire propres à un réseau socio-sémantique), car de nombreuses caractéristiques traditionnelles des réseaux monopartis (réseau social seul, ou réseau sémantique simple) sont déjà abondamment documentées par ailleurs.

Distributions de degrés. Dans un réseau épistémique, des liens apparaissent dans les réseaux social, sémantique et socio-sémantique ; il faut ainsi s'intéresser à quatre distributions de degrés :

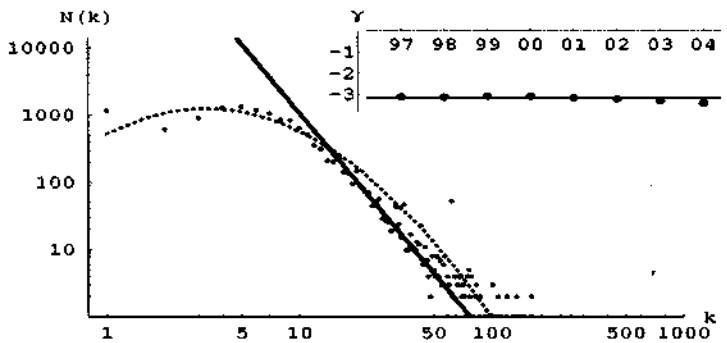
i. les degrés k du réseau social : de nombreux agents sont liés à peu d'agents, tandis que quelques-uns seulement le sont à beaucoup d'autres. On considère traditionnellement que cette distribution suit une « loi-puissance », même s'il semblerait davantage qu'il s'agisse d'une loi « log-normale » (figure 7). En anglais, ce type de réseau est fréquemment appelé ***scale-free*** (sans échelle).
ii. les degrés k_c du réseau sémantique : l'on observe que les concepts sont progressivement tous connectés entre eux.

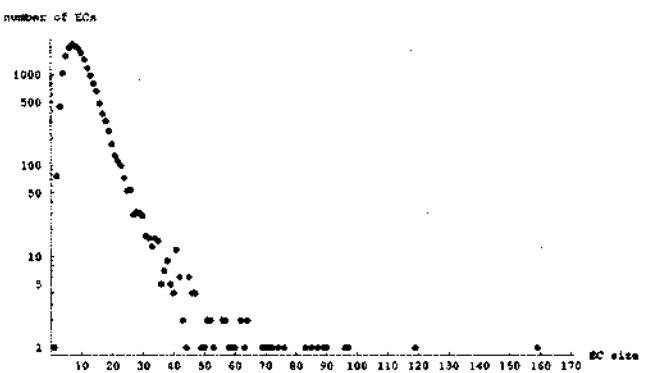
iii. les degrés des agents vers les concepts ($k_{a \to c}$) : cette distribution suit aussi une loi-puissance. Peu d'agents utilisent de nombreux concepts, beaucoup d'agents utilisent quelques concepts.

iv. les degrés des concepts vers les agents ($k_{c \to a}$) : ici, peu de concepts sont utilisés par beaucoup d'agents - parce qu'ils sont soit pertinents, soit populaires - et la plupart des concepts sont utilisés par un nombre moyen d'agents. Il n'y a pas de concept peu usité dans notre sélection.

Figure 7

Au-dessus, distribution de degrés pour le réseau social $N(k)$ (points), et approximation en loi « puissance » (courbe pleine, $\propto k^{\gamma}$, $\gamma = -3{,}39$) et « log-normale » (courbe pointillée, $P(k) \propto k^{\beta 2 \log k + \beta 1}$).
En dessous, distribution des tailles des CE pour un TG calculé sur un échantillon aléatoire de 250 agents et 70 concepts.





Coefficients d'agrégation. Ce qui est traditionnellement appelé « coefficient d'agrégation » (*clustering coefficient*)[40] est concrètement une mesure de la transitivité, décrivant de quelle manière les voisins d'un nœud donné sont connectés entre eux (« les amis de mes amis sont aussi mes amis »). Ce coefficient est généralement anormalement élevé dans les réseaux sociaux. Toutefois, il tend aussi à être nécessairement élevé dans les réseaux construits à partir d'événements réunissant plusieurs agents, comme la rédaction d'articles en commun, où l'on considère que tous les agents participant à l'événement sont connectés entre eux. Le réseau se construit donc à partir d'additions de cliques[41], qui ajoutent naturellement de nombreux « triangles », augmentant artificiellement la transitivité. Le coefficient d'agrégation constitue ainsi un paramètre pauvrement informatif. À cet égard, dans le cas d'un réseau socio-sémantique, il semble davantage pertinent de s'intéresser au coefficient d'agrégation biparti qui dénombre la proportion de « losanges »[42]. En d'autres termes, ce coefficient constitue une mesure intéressante de la façon dont deux agents connectés à un même concept sont susceptibles d'être aussi connectés à d'autres concepts en commun ; de même, il mesure la façon dont deux concepts liés à un même agent peuvent aussi avoir d'autres agents en commun (il s'agirait ici d'une variété très locale d'équivalence structurelle). Ce coefficient est d'un ordre de grandeur plus élevé dans notre réseau réel que dans des réseaux aléatoires conservant la même « loi-puissance » : qualitativement, on peut en déduire que les couples d'agents liés à un même concept partagent aussi d'autres concepts anormalement souvent.

Structure épistémique. Nous l'avons souligné dans la section 1, les communautés épistémiques ont une structure spécifique. Nous observons, en particulier, qu'il y a de nombreuses communautés peu et très peu peuplées, et beaucoup moins de communautés très peuplées, comme l'illustre la distribution des populations des CE à la figure 7. Il

---

[40] Duncan J. Watts et Steven H. Strogatz, *op. cit.*

[41] Jean-Loup Guillaume et Matthieu Latapy, « Bipartite structure of all complex networks », *Information processing letters*, vol. 90, n° 5, 2004, p. 215-221.

[42] Gary Robins et Malcolm Alexander, « Small worlds among interlocking directors: Network structure and distance in bipartite graphs », *Computational and mathematical organization theory*, vol. 10, 2004, p. 69-94 ; Pedro G. Lind *et al.*, « Cycles and clustering in bipartite networks », *Physical review E*, vol. 72, 056127, 2005.

s'agit là d'un fait stylisé-clé qu'un modèle de réseau épistémique adéquat doit reconstruire.

En outre, tout comme nous avons observé le *clustering* biparti entre agents et concepts, nous aimerions aussi voir dans quelle mesure les agents sont proches sémantiquement les uns des autres et, plus spécifiquement, de quelle manière ils sont semblables à leur voisinage social. À cet effet, nous introduisons une « distance sémantique » entre agents, c'est-à-dire une fonction d'une paire d'agents qui a pour propriété de décroître (ou de croître) avec le nombre de concepts que les deux agents partagent (ou pas), qui vaut 0 lorsque les agents sont liés au mêmes concepts et 1 lorsqu'ils n'ont aucun concept en commun. Techniquement, nous utilisons une distance basée sur le coefficient de Jaccard[43] qui respecte cette propriété. Nous mesurons ensuite la distribution des distances sémantiques dans le réseau : alors que les agents aux profils similaires sont rares, les résultats sont radicalement différents en considérant le voisinage social seulement, car les voisins sont à une distance sémantique fortement plus faible (figure 8).

2.3. Dynamique de bas-niveau

Concevoir un modèle de morphogenèse crédible des faits de haut-niveau présentés ci-dessus requiert de comprendre les mécanismes à la fois d'interaction et de croissance. Nous montrons comment il est possible de fabriquer une telle dynamique de bas-niveau λ à partir de données empiriques.

2.3.1. Mesure du comportement d'interaction

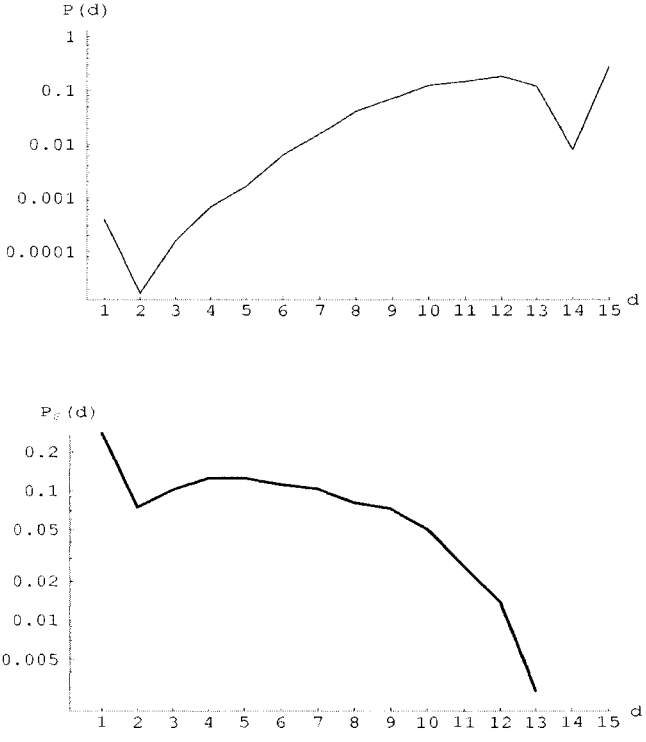
Formellement, l'attachement préférentiel (AP) est la propension pour un agent de participer à une interaction avec un autre agent en fonction de certaines des propriétés de cet agent. Il s'agit donc d'homophilie au sens large, c'est-à-dire que l'AP peut aussi devenir une mesure de l'hétérophilie. De manière générale, les estimations quantitatives de cet attachement réduisent souvent l'AP à une quantité scalaire, en calculant directement les moyennes empiriques ou en adoptant des approches

[43] Vladimir Batagelj et Matevz Bren, , Comparing resemblance measures », *Journal of classification*, vol. 12, n° 1, 1995, p. 73-90.

d'estimation économétriques ou à base de modèles de Markov[44] ; ou bien elles sont liées à l'AP classique relatif au degré des agents[45]. Dans ce dernier cas, la mesure d'homophilie est fréquemment très détaillée, mais focalisée uniquement sur la connectivité des agents.

Figure 8

Au-dessus : distribution des distances sémantiques sur le réseau entier.
En dessous : distribution des distances sémantiques en ne considérant que le voisinage social des agents.



[44] Tom A. Snijders, *op. cit.* ; Roger Guimera *et al.*, « Team assembly mechanisms determine collaboration network structure and team performance », *Science*, vol. 308, 2005, p. 697-702 ; Walter W. Powell *et al.*, « Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences », *American journal of sociology*, vol. 110, n° 4, 2005, p. 1132-1205.

[45] Albert-Laszlo Barabási *et al.*, « Evolution of the social network of scientific collaborations », *Physica A*, vol. 311, 2002, p. 590-614 ; Sidney Redner, « Citation statistics from 110 years of physical review », *Physics today*, vol. 58, 2005, p. 49-54.

Dans un modèle de reconstruction, nous pouvons vouloir à la fois connaître l'AP par rapport à diverses propriétés dont nous pensons qu'elles sont responsables de la structure observée et qu'il faut reconstruire, mais aussi connaître le vaste spectre des comportements d'attachement préférentiel possibles, c'est-à-dire ne pas uniquement savoir que la distance sémantique joue un rôle important, voire plus important que le degré. Nous considérons à cet effet que les points suivants sont cruciaux : 1) les propriétés basées strictement sur la topologie du réseau peuvent ne pas rendre compte de phénomènes sociaux complexes : par exemple l'AP lié à l'homophilie[46] demande de qualifier les agents en utilisant des données non structurelles ; 2) les quantités scalaires simples ne peuvent pas exprimer la riche hétérogénéité du comportement d'interaction ; et 3) les modèles supposent souvent que les propriétés ne sont pas corrélées, ce qui peut parfois revenir à compter deux fois le même effet. À notre connaissance, s'il existe bien de nombreuses études vérifiant plusieurs de ces points, il semble ne pas exister de travaux résolvant simultanément ces trois points.

Nous concevons ici l'AP comme un mécanisme flexible et général basé sur des propriétés à la fois topologiques et non topologiques, décrivant exhaustivement l'étendue des interactions possibles (sous forme d'un histogramme plutôt que d'une seule quantité scalaire), prenant en compte les influences croisées des différentes propriétés.

AP monadique et dyadique. Nous distinguons, en outre, les propriétés d'un seul agent, ou propriétés « monadiques » (telles que le degré, l'âge, etc.) des propriétés définies sur une paire d'agents, ou dyadiques (distance sociale, dissimilarité, etc.). En effet, en travaillant avec des propriétés monadiques, il s'agit de connaître la propension qu'ont certains types d'individus à participer à une interaction davantage que d'autres types d'agents. À l'inverse, avec les dyades, il s'agit de savoir si une interaction aura plus facilement lieu et sera plus fréquente en fonction du type des couples d'agents.

Formellement, nous supposons que l'AP des agents par rapport à une propriété monadique donnée m peut être décrit par une fonction f de m, la propension d'interaction, indépendante de la distribution des agents de type m au sein de la population : $f(m)$ est simplement la

[46] Miller McPherson *et al.*, *op. cit.*

probabilité conditionnelle $P(L \mid m)$ qu'un agent de type m reçoive un lien L. Ainsi, il est $f(m)$ fois plus probable qu'un agent de type m participe à une interaction. Par exemple, l'hypothèse sur l'AP relatif au degré traditionnellement utilisée dans les modèles de type Barabási-Albert revient, en fait, à une hypothèse sur f équivalente à $f(k) \propto k$: la propension d'attachement préférentiel est proportionnelle au degré des agents.

En pratique, on montre que l'on peut estimer f grâce à $\hat{f}(m) = v(m)/P(m)$ si $P(m) > 0$, 0 sinon, où $v(m)$ est le nombre de nouvelles extrémités de liens qui s'attachent à des nœuds de type m au long d'une période, et $P(m)$ dénote typiquement la distribution des nœuds de type m. Lorsqu'une propriété n'a pas de sens pour un nœud unique, comme la proximité, la similarité ou les distances en général, on adopte un point de vue dyadique en supposant de manière similaire qu'il existe un comportement essentiel d'interaction dyadique décrit par une fonction $g(d)$ pour une propriété dyadique donnée d, définie sur des paires d'agents. $g(d)$ correspond à la probabilité conditionnelle $P(L \mid d)$ et, à nouveau, g est estimée par $\hat{g}(d) = v(d)/P(d)$. Le comportement d'attachement préférentiel mesuré et décrit par $\hat{f}$ (ou $\hat{g}$) peut ensuite être utilisé pour fabriquer les hypothèses de modélisation, soit en prenant directement la fonction estimée empiriquement, soit en stylisant l'allure de $\hat{f}$ (ou $\hat{g}$) en vue de trouver potentiellement des solutions analytiques. Remarquons toutefois que $\hat{f}$ peut toujours varier en fonction de propriétés globales du réseau (taille, diamètre, etc.). Montrer que $\hat{f}$ est indépendant de ce genre de propriétés impose de comparer les différentes allures de $\hat{f}$ pour diverses périodes et configurations.

2.3.2. AP empirique

À l'aide de ces outils, nous mesurons l'AP relatif 1) à une propriété monadique, le degré, et 2) à une propriété dyadique, la distance sémantique utilisée précédemment, afin de rendre compte de l'homophilie. Ceci permet d'estimer si le comportement des agents révèle des régularités quant au choix des collaborateurs. En particulier, nous faisons l'hypothèse que, d'une part, le comportement est influencé de façon significative par ces propriétés et que, d'autre part, un modèle fondé sur des hypothèses forgées autour de ces propriétés suffit à reconstruire les faits stylisés de haut-niveau présentés plus haut.

AP lié au degré. Nous calculons $\mathcal{f}(k)$ pour l'AP lié à une propriété monadique standard, le degré k, et empiriquement nous vérifions essentiellement l'hypothèse classique $f(k) \propto k$: statistiquement, les agents ont une propension k fois plus grande à collaborer avec des agents de degré k[47]. Ce résultat précis n'est pas nouveau et est en accord avec les travaux précédents sur ce type d'AP[48]. Toutefois, en nous intéressant à l'activité des agents, nous remettons en question la métaphore classique du *rich-get-richer* (les riches s'enrichissent) suivant laquelle les agents les plus connectés sont les plus attractifs, reçoivent davantage de connexions et sont donc davantage connectés.

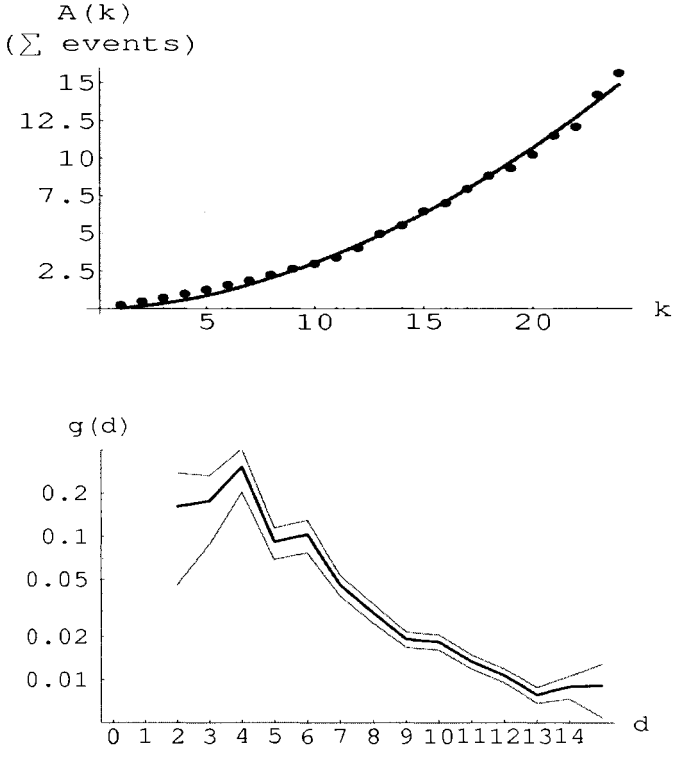
En fonction de k, les agents « riches » sont, en fait, proportionnellement plus actifs que les agents « pauvres » (ils participent à davantage d'événements, figure 9), et de fait participent à davantage d'interactions. Dans ce cas, le comportement sous-jacent est simplement une activité et non une attractivité linéaire : *rich-work-harder* (« les riches travaillent plus », mais ne sont pas plus attractifs) – les agents n'ont pas réellement de préférence ni de volonté d'interagir avec les agents les plus connectés. Quoique les deux interprétations soient équivalentes en ce qui concerne la mesure de l'AP, elles ne sont pas comportementalement équivalentes et ont des répercussions différentes sur la modélisation, notamment dans les modèles à base d'événements, tels que celui que nous allons présenter. Dans ce cas, en effet, le modélisateur est amené à affiner le comportement d'interaction en incluant à la fois la participation dans des événements et le nombre d'interactions par événements, plutôt que simplement des interactions préférentielles : un groupe, plutôt qu'une paire d'agents seulement, choisit d'interagir

[47] Plus précisément, les données sont approchées de façon très satisfaisante par une simple courbe linéaire – la meilleure approximation non linéaire dévie néanmoins légèrement : $f(k) \; ak^{0,97}$. Cependant, l'intervalle de confiance de l'exposant est [0,6] ; 1,34], soit un intervalle beaucoup trop étendu pour prétendre à une précision sufisante. Lorsque les données sont bruitées, comme c'est le cas ici, il est utile de calculer la propension cumulée pour laquelle la meilleure approximation non linéaire est $\mathcal{f}(k)a \; k^{1,83}$ ($\pm$ 0,05), ce qui confirme une légère déviation de l'hypothèse d'une préférence strictement linéaire, qui devrait donner $\mathcal{f}(k)a \; k^2$.

[48] Mark E. J. Newman, « Clustering and preferential attachment in growing networks », *Physical review letters E*, vol. 64, 025102, 2001 ; Hawoong Jeong *et al.*, « Measuring preferential attachment for evolving networks », *Europhysics letters*, vol. 61, n° 4, 2003, p. 567-572.

Figure 9

Au-dessus, activité cumulée $A(k) = \Sigma_{\text{de k'=1 à k}} a$ (k) en termes d'articles par période (événements par période) par rapport au degré de l'agent ; et meilleure approximation non linéaire ($k^{1,88} \pm 09$, ligne pleine).
En dessous, propension d'interaction homophilique $\hat{g}$ par rapport à la distance sémantique $d \in \{0,...,15\}$ (trait épais)
et intervalle de confiance pour $p < 0,05$ (trait fin).



conjointement[49]. Plus généralement, une telle caractéristique soutient l'idée selon laquelle les événements, et non pas les liens, sont le niveau correct de modélisation des réseaux sociaux : un événement réunit un nombre arbitraire d'agents interagissant ensemble et conjointement, ce qui se résume dans certains cas particuliers à une interaction dyadique.

[49] José J. Ramasco *et al.*, « Self-organization of collaboration networks », *Physical review E*, vol. 70, 036106, 2004 ; Roger Guimera *et al.*, *op. cit.*

AP lié à l'homophilie. Nous évaluons aussi l'homophilie des agents (leur propension à interagir avec des agents semblables) en utilisant la distance sémantique présentée dans la section 2.2. Les résultats empiriques présentés à la figure 9 montrent que, alors que les agents favorisent les interactions avec des agents très légèrement différents, ils préfèrent tout de même fortement les agents semblables. En outre, l'allure exponentielle de $\hat{g}$, suggère que l'homophilie a une influence plus sensible que la connectivité sur le comportement d'interaction des agents. Ceci souligne l'importance de caractéristiques sémantiques et, plus généralement, non structurales, pour modéliser de tels réseaux.

Nous vérifions l'indépendance de ces deux propriétés – degré et distance sémantique –, c'est-à-dire qu'un agent de faible degré n'est pas plus souvent à une plus grande distance sémantique des autres agents que les agents de plus haut degré. Bien que nous ayons examiné un exemple restreint sur ces deux propriétés, il pourrait être aussi très pertinent de mesurer l'AP par rapport à d'autres paramètres, tels que la distance sociale, le nombre d'amis communs, etc. Cependant, l'objectif est autant d'exhiber des propriétés crédibles comportementalement que non corrélées les unes aux autres, si possible. En ce sens, ni le nombre d'amis communs ni la distance sociale[50] ne semblent être de bons candidats.

AP envers les concepts. Nous nous intéressons finalement à la façon dont les concepts sont choisis : les concepts les plus connectés sont-ils utilisés plus fréquemment, « interagissant » ainsi avec encore plus d'auteurs ? Empiriquement, on observe que les concepts sont choisis proportionnellement à leur degré socio-sémantique (soit le nombre d'agents qui les utilisent), qui reflète leur succès (figure 10) de manière plus ou moins semblable à l'attachement entre agents.

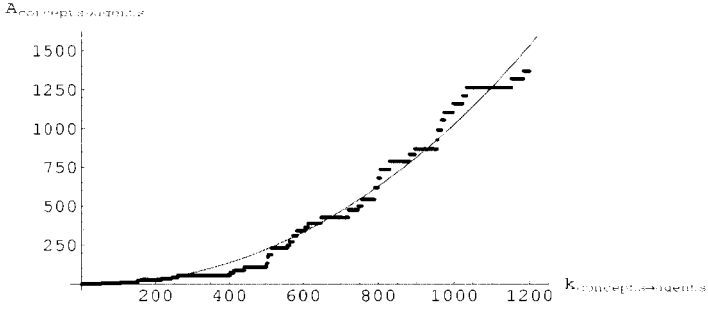
2.3.3. Paramètres liés à la croissance et aux événements

Pour compléter la description de la croissance du réseau, il est essentiel de connaître la structure des événements quant à leur composition en termes d'auteurs et de concepts. La dynamique est dirigée par l'apparition de nouveaux articles, qui sont régulièrement produits et mettent

[50] Mark E. J. Newman, « Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality », *Physical review E*, vol. 64, 016132, 2001.

Figure 10

Utilisation ou activité cumulée des concepts $A_{concepts\ ,agents}$, par rapport au degré socio-sémantique $k_{concepts\ ,agents}$. Une approximation non linéaire donne une activité cumulée $A_{concepts\ ,agents}(k_{c\ ,a}) \propto k_{c\ ,a}^{2,19}$, ce qui implique que l'activité elle-même soit légèrement supra-linéaire, c'est-à-dire $a_{concepts\ ,agents}(k_{c\ ,a}) \propto k_{c\ ,a}^{1,19}$, c'est-à-dire davantage que proportionnelle.



en jeu, d'une part, des auteurs déjà présents (anciens nœuds) et éventuellement une fraction de nouveaux auteurs et, d'autre part, des concepts apportés par les auteurs, ainsi que de nouveaux concepts.

Nous mesurons aussi les distributions d'agents et de concepts au sein des articles. La distribution du nombre d'agents par article suit approximativement une distribution géométrique, et la proportion de nouveaux auteurs montre que dans la plupart des cas les auteurs sont 1) soit tous nouveaux, 2) soit tous anciens, 3) soit pour moitié nouveaux et pour moitié anciens. La proportion de nouveaux auteurs dans tous les articles est stable quelle que soit la période. En outre, la distribution du nombre de concepts par article peut être approximée par une distribution géométrique. Ceci dit, bien que les anciens auteurs d'un article apportent une certaine partie de leurs concepts, certains concepts sont utilisés pour la première fois par chacun de ces auteurs. La distribution de la proportion de ces « nouveaux » concepts permet de distinguer les concepts pris parmi le bagage conceptuel des auteurs de ceux qui sont absolument nouveaux.

Ensuite, nous déterminons la croissance brute du réseau ; puisque l'ensemble des concepts est fixé *a priori*, seul le réseau social accueille de nouveaux nœuds. L'évolution ΔN_t de la taille du réseau social N_t est fortement liée au nombre d'articles n_t, puisque les nouveaux auteurs apparaissent au travers de nouveaux articles. À cet égard, la croissance du nombre d'articles est ici globalement arithmétique : $n_{t+1} = n_t + n_+$ (avec $n_+ \approx 96$). Le nombre moyen de nouveaux auteurs étant constant au cours du temps, ΔN et n ont un comportement similaire. De fait, N croît quadratiquement. Notons, toutefois, que ce phénomène de croissance du nombre d'articles par période est propre à l'évolution de cette situation empirique. L'évolution de n et N est une conséquence de cela. Ce n'est visiblement pas le cas dans n'importe quelle communauté, par exemple, si ce champ de recherches devait être abandonné.

2.4. Modèle de reconstruction

En nous appuyant sur des paramètres de bas-niveau empiriques (composition des articles, préférences d'interaction), nous pouvons concevoir un modèle visant à reconstruire une structure de haut-niveau compatible avec les faits stylisés observés (distributions de degrés et distances sémantiques, coefficients de *clustering*, structure des CE). Trois caractéristiques de modélisation sont implémentées : croissance du réseau à base d'événements, co-évolution entre agents et concepts, et descriptions de bas-niveau réalistes, en particulier en ce qui concerne les interactions. Les événements sont donc des articles mettant en jeu des agents (plus ou moins actifs suivant leur degré k, et se réunissant préférentiellement en fonction de leurs intérêts) et des concepts (plus ou moins populaires, suivant leur degré $k_{c,a}$). Notre modèle de réseau épistémique co-évoluant fonctionne ainsi (figure 11) :

i. Création et définition des événements. n_t articles sont créés à chaque période : $n_{t+1} = n_t + n_+$. La taille des ensembles d'auteurs et de concepts suit une loi géométrique ayant pour paramètre la moyenne observée empiriquement.
ii. Choix des auteurs. En stylisant les faits empiriques décrits plus haut, les articles mettent en jeu de manière équiprobable soit seulement des nouveaux auteurs, soit seulement des anciens, soit des anciens et des nouveaux en proportions égales. S'il existe au moins un ancien agent, un « initiateur » est choisi aléatoirement

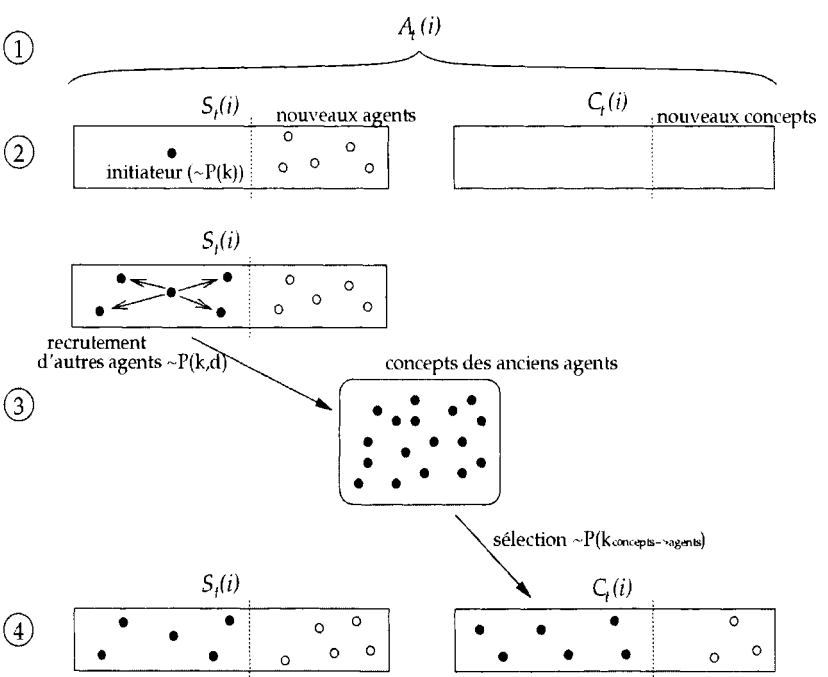
proportionnellement à son degré social k. Ensuite, d'autres anciens agents de degré k sont choisis suivant $P(L|k,d)=P(L|k)P(L|d)$, d étant la distance sémantique à l'initiateur. Finalement, de nouveaux nœuds sont ajoutés.

iii. Choix des concepts. Les nouveaux concepts (tels qu'aucun ancien agent ne les a utilisés) représentent une proportion fixe des concepts de l'article. Les autres concepts sont choisis parmi l'ensemble des concepts des auteurs, proportionnellement suivant leur degré $k_{c \to a}$.

iv. Mise à jour du réseau, lorsque les ensembles d'agents et de concepts sont définis.

Figure 11

Modéliser un événement en spécifiant le contenu de l'article i, $A_t(i) = (S_t(i),C_t(i))$, couple d'ensembles d'agents et de concepts.



Nous avons simulé le modèle de morphogenèse pour huit périodes, initialisé avec un réseau épistémique vide et un taux de croissance de 100 articles par période ($n_1=100$, $n_{+}=100$). Nous nous sommes intéressés aux réseaux finaux, dont la structure est en bonne adéquation avec le monde réel pour chaque fait stylisé : 1) la taille du réseau, 2) les distributions de degrés, 3) les coefficients d'agrégation et 4) la structure des CE. Ce modèle rend ainsi compte du phénomène de production des communautés épistémiques par la co-évolution des agents et des concepts. Non seulement la structure de haut-niveau est correctement recréée, mais les dynamiques de bas-niveau sont aussi cohérentes – ceci est crucial : il serait douteux de reconstruire des phénomènes de haut-niveau avec des dynamiques de bas-niveau incorrectes. La validité des descriptions doit concerner aussi bien le haut niveau que le bas niveau.

Enfin, il est intéressant de se demander quel rôle chaque hypothèse joue dans l'apparition de chacun des phénomènes de haut-niveau : notre modèle est-il un modèle minimal pour les faits stylisés sélectionnés ? Ces faits sont-ils toujours reproduits si on relâche certaines hypothèses ? Puisque de nombreuses combinaisons de modèles simplifiés sont envisageables, nous n'examinons que le relâchement d'une seule hypothèse à chaque fois. Dans ce cas, au moins un fait de haut-niveau n'est pas correctement reconstruit dès que l'on relâche n'importe quelle hypothèse du modèle (qu'il s'agisse de la modélisation à base d'événements, de l'AP lié au degré pour le choix des agents ou des concepts, ou de l'homophilie des agents).

2.5. Morphogenèse co-évolutive des réseaux de savoirs

Nous avons étudié la formation de la communauté scientifique *zebrafish* en supposant que nous pouvions microfonder l'évolution de la structure de ce système complexe social en modélisant des agents co-évoluant avec des concepts. Pour ce faire, nous avons dû concevoir une dynamique de bas-niveau λ du réseau épistémique qui soit en accord avec les données empiriques, et qui reconstruise correctement η_e via *P*. Ainsi, nous avons introduit des outils pour estimer les interactions de bas-niveau et les processus de croissance à partir des données passées : préliminaire indispensable pour réaliser un modèle réaliste et descriptif. *In fine*, nous avons défendu une approche empirique dans la conception des modèles : même si les faits stylisés désirés sont reproduits, il faut

que la dynamique de bas-niveau supposée soit fondée empiriquement. En évitant de recourir à un modèle uniquement normatif, cette méthodologie contribue à crédibiliser ce type d'approche prometteuse au sein des sciences sociales.

Le succès final de la reconstruction accrédite l'hypothèse initiale : la structure des communautés de savoirs est au moins produite par la co-évolution entre agents et concepts. Toutefois, nous soulignons qu'une telle co-évolution peut aussi dépendre de paramètres exogènes. Plusieurs phénomènes de bas-niveau pourraient être différents de façon significative dans d'autres groupes de recherches ou champs épistémiques. Prenons, par exemple, la croissance du domaine : pourquoi y aurait-il un intérêt croissant dans l'étude du *zebrafish* ? À l'avenir, l'étude de cet animal pourrait avoir des répercussions sur le traitement des cancers, attirant peut-être un grand nombre de chercheurs. Nous doutons fortement que ce type de paramètre puisse être endogénéisé dans un modèle. Plus généralement, l'incertitude sur la nouveauté et le nouveau savoir qui apparaît dans le système complexe social ne relève pas de l'incertitude sémantique : ce n'est pas un événement qui peut arriver ou pas, mais dont nous connaissons *a priori* la nature, donc probabilisable. Au contraire, il s'agit d'une incertitude radicalement différente, concernant l'ontologie même dont les agents vont disposer dans le futur[51], phénomène particulièrement sensible dans les systèmes sociaux. Nous devons ainsi faire face à une irréductibilité ontologique : un modèle ne peut pas exprimer et fournir quoi que ce soit de plus « neuf » que ce qui est déjà spécifié dans le langage et la grammaire du modèle, clos *a priori*. De fait, l'influence éventuelle de paramètres exogènes indécidables nous amène à modérer notre affirmation : alors que la reconstruction a visiblement été réussie, au sein d'une période donnée et de ses particularités, il est cependant probable que d'autres processus dans lequel le réseau épistémique est immergé pourraient aussi jouer un rôle significatif. En tant que tel, à condition que de tels paramètres soient stables pour l'échelle de temps considérée, nous avons explicitement montré qu'il était possible de reproduire une partie de la dynamique d'un système complexe social.

[51] David A. Lane et Robert R. Maxfield, « Ontological uncertainty and innovation », *Journal of evolutionary economics*, vol. 15, n° 1, 2005, p. 3-50.

Conclusion

Nous avons couvert à la fois les enjeux théoriques liés à la reconstruction d'un système complexe social et l'étude pratique d'une communauté de savoirs réelle. Nous avons pu, en particulier, affirmer que les communautés épistémiques étaient principalement produites par la co-évolution entre agents et concepts. Plus précisément, dans la section 1, nous avons présenté une méthode permettant de décrire et de catégoriser les communautés de savoirs et de capturer des faits stylisés essentiels relatifs à leur structure. En particulier, nous avons reconstruit la taxonomie d'une communauté entière en utilisant des treillis de Galois. L'étude longitudinale de ces images statiques permet une description historique, en captant des faits stylisés tels que l'émergence d'un champ, son déclin, sa spécialisation et ses interactions avec d'autres champs (fusion, scission). La méthode est appliquée à des données empiriques et validée par les catégories et les histoires fournies par des experts du domaine. Nous concevons ainsi une fonction de projection P d'un bas-niveau défini par des liens entre agents et concepts vers le haut-niveau des descriptions épistémologiques. Dans la section 2, le principal objectif était de micro-fonder les propriétés de haut-niveau observées dans la section 1 : nous voulions connaître les processus responsables, au niveau des agents, de l'émergence de la structure des communautés épistémiques : ceci revient à proposer un modèle de morphogenèse. À cet effet, nous avons d'abord construit les outils permettant d'estimer les mécanismes d'interaction et de croissance à partir des données empiriques. Ensuite, en supposant qu'agents et concepts co-évoluent, nous avons reconstruit la structure d'une communauté scientifique réelle, pour une sélection pertinente de faits stylisés de haut-niveau.

Ce programme de recherche vise finalement à permettre aux agents de comprendre réflexivement la dynamique du système social global auquel ils participent. Plus largement, cela concourt à l'achèvement d'une société véritablement autonome, au sens de Castoriadis : une société qui, connaissant ses propres structures, dynamiques et représen-

tations, est capable de déterminer ses propres lois d'évolution, d'adapter son comportement précisément par rapport à sa propre dynamique[52].

Bibliographie

Albert, Réka et Albert-Laszlo Barabási, « Statistical mechanics of complex networks », *Reviews of modern physics*, vol. 74, 2002, p. 47-97.

Barabási, Albert-Laszlo, Hawoong Jeong, Erzsebet Ravasz, Zoltan Neda, Tamas Vicsek et Andras Schubert, « Evolution of the social network of scientific collaborations », *Physica A*, vol. 311, 2002, p. 590-614.

Barabási, Albert-Laszlo et Réka Albert, « Emergence of scaling in random networks », *Science*, vol. 286, 1999, p. 509-512.

Barbour, Andrew et Denis Mollison, « Epidemics and random graphs », dans J.-P. Gabriel, C. Lefevre et P. Picard (dir.), *Stochastic processes in epidemic theory*, printemps 1990, Lecture notes in biomaths, 86, p. 86-89.

Batagelj, Vladimir et Matevz Bren, « Comparing resemblance measures », *Journal of classification*, vol. 12, n° 1, 1995, p. 73-90.

Berlin, Brent, *Ethnobiological classification – Principles of categorization of plants and animals in traditional societies,* Princeton, Princeton University Press, 1992.

Birkhoff, Garrett, *Lattice theory*, Providence (RI), American mathematical society, 1948.

Bonabeau, Eric, « Agent-based modeling: Methods and techniques for simulating human systems », *PNAS*, vol. 99, n° 3, 2002, p. 7280-7287.

Bradbury, Jane, « Small fish, big science », *PLoS Biology*, vol. 2, n° 5, 2004, p. 568-572.

Burt, Ronald S., « Cohesion versus structural equivalence as a basis for network subgroups », *Sociological methods and research*, vol. 7, 1978, p. 189-212.

Callon, Michel, John Law et Arie Rip, *Mapping the dynamics of science and technology*, London, MacMillan Press, 1986.

Castoriadis, Cornélius, « La logique des magmas et la question de l'autonomie », dans Paul Dumouchel et Jean-Pierre Dupuy (dir.), *L'auto-organisation. De la physique au politique*, Paris, Seuil, 1983, p. 421-443.

Cohendet, Patrick, Alan Kirman et Jean-Benoît Zimmermann, « Émergence, formation et dynamique des réseaux – modèles de la morphogenèse », *Revue d'économie industrielle*, vol. 103, n°s 2-3, 2003, p. 15-42.

Cold Spring Harbor Laboratory, *Zebrafish development & genetics*, Cold Spring Harbor (NY), 1994, 1996, 1998, 2000, 2001, 2002, 2003.

Cowan, Robin, Paul A. David et Dominique Foray, « The explicit economics of knowledge codification and tacitness », *Industrial & corporate change*, vol. 9, n° 2, 2000, p. 212-253.

[52] Cornélius Castoriadis, « La logique des magmas et la question de l'autonomie », dans Paul Dumouchel et Jean-Pierre Dupuy (dir.), *L'auto-organisation. De la physique au politique*, Paris, Seuil, 1983, p. 421-443.

Dooley, Kimberly et Leonard I. Zon, « Zebrafish: A model system for the study of human disease », *Current Opinion in Genetics & Development*, vol. 10, n° 3, 2000, p. 252-256.

Dorogovtsev, Sergey N. et José F. F. Mendes, *Evolution of Networks – From biological nets to the Internet and WWW*, Oxford, Oxford University Press, 2003.

Duquenne, Vincent, Caroline Chabert, Ameziane Cherfouh, Anne-Lise Doyen, Jean-Maurice Delabar et Douglas Pickering, « Structuration of phenotypes and genotypes through Galois lattices and implications», *Applied artificial intelligence*, vol. 17, n° 3, 2003, p. 243-256.

Erdós, Paul et Alfred Rényi, « On random graphs », *Publicationes mathematicæ*, vol. 6, 1959, p. 290-297.

Faloutsos, Michalis, Petros Faloutsos et Christos Faloutsos, « On power-law relationships of the Internet topology », *Computer communication Review*, vol. 29, n° 4, 1999, p. 251-262.

Freeman, Linton C., « A set of measures of centrality based on betweenness », *Sociometry*, vol. 40, 1977, p. 35-41.

Freeman, Linton C., « Social networks and the structure experiment », dans Linton C. Freeman, Douglas R. White et A. Kimball Romney (dir.), *Research methods in social network analysis*, Fairfax (VA), George Mason university press, 1989, p. 11-40.

Godin, Robert, Hafedh Mili, Guy W. Mineau, Rokia Missaoui, Amina Arfi et Thuy-Tien Chau, « Design of class hierarchies based on concept (Galois) lattices », *Theory and practice of object systems (TAPOS)*, vol. 4, n° 2, 1998, p. 117-134.

Grunwald, David J. et Judith S. Eisen, « Headwaters of the zebrafish – Emergence of a new model vertebrate », *Nature reviews. Genetics*, vol. 3, n° 9, 2002, p. 717-724.

Guillaume, Jean-Loup et Matthieu Latapy, « Bipartite structure of all complex networks », *Information processing letters*, vol. 90, n° 5, 2004, p. 215-221.

Guimera, Roger, Brian Uzzi, Jarrett Spiro et Luis A. Nunes Amaral, « Team assembly mechanisms determine collaboration network structure and team performance », *Science*, vol. 308, 2005, p. 697-702.

Haas, Peter, « Introduction: Epistemic communities and international policy coordination », *International organization*, vol. 46, n° 1, hiver 1992, p. 1-35.

Hartigan, John A., *Clustering algorithms*, New York (NY), Wiley, 1975.

Jain, Anil K., M. Narasimha Murty et P. J. Flynn, « Data clustering: A review », *ACM computing surveys*, vol. 31, n° 3, 1999, p. 264-323.

Jeong, Hawoong, Zoltan Néda et Albert-Laszlo Barabási, « Measuring preferential attachment for evolving networks », *Europhysics letters*, vol. 61, n° 4, 2003, p. 567-572.

Johnson, Stephen C., « Hierarchical clustering schemes », *Psychometrika*, vol. 2, 1967, p. 241-254.

Kohonen, Teuvo, *Self-organizing maps*, 3e éd., Berlin, Springer, 2000.

Kreuzman, Henry, « A co-citation analysis of representative authors in philosophy: Examining the relationship between epistemologists and philosophers of science », *Scientometrics*, vol. 51, n° 3, 2001, p. 525-539.

Kuznetsov, Sergei O. et Sergei A. Obiedkov, « Comparing performance of algorithms for generating concept lattices », *Journal of experimental and theoretical artificial intelligence*, vol. 14, nos 2-3, 2002, p. 189-216.

Lane, David A. et Robert R. Maxfield, « Ontological uncertainty and innovation », *Journal of evolutionary economics*, vol. 15, no 1, 2005, p. 3-50.

Lave, Jean et Etienne Wenger, *Situated learning: Legitimate peripheral participation*, Cambridge, Cambridge University Press, 1991.

Leydesdorff, Loet, « In search of epistemic networks », *Social studies of science*, vol. 21, 1991, p. 75-110.

Lind, Pedro G., Marta C. Gonzalez et Hans J. Herrmann, « Cycles and clustering in bipartite networks », *Physical review E*, vol. 72, 056127, 2005.

Lopez, Alejandro, Scott Atran, John D. Coley, Douglas L. Medin et Edward E. Smith, « The tree of life : Universal and cultural features of folkbiological taxonomies and inductions », *Cognitive psychology*, vol. 32, no 3, 1997, p. 251-295.

Lorrain, François et Harrison C. White, « Structural equivalence of individuals in social networks », *Journal of mathematical sociology*, vol. 1, 1971, p. 49-80.

May, Robert K., « Will a large complex system be stable? », *Nature*, vol. 238, 1972, p. 413-414.

McCain, Katherine W., « Cocited author mapping as a valid representation of intellectual structure », *Journal of the American society for information science*, vol. 37, no 3, 1986, p. 111-122.

McPherson, Miller, Linn Smith-Lovin, James M. Cook, « Birds of a feather: Homophily in social networks », *Annual review of sociology*, vol. 27, 2001, p. 415-444.

Newman, Mark E. J., « Clustering and preferential attachment in growing networks », *Physical review letters E*, vol. 64, 025102, 2001.

Newman, Mark E. J., « Detecting community structure in networks », *European physical journal B*, vol. 38, 2004, p. 321-330.

Newman, Mark E. J., « Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality », *Physical review E*, vol. 64, 016132, 2001.

Nilsson-Jacobi, Martin, « Hierarchical organization in smooth dynamical systems », *Artificial life*, vol. 11, no 4, 2005, p. 493-512.

Noyons, Ed C. M. et Anthony F. J. van Raan, « Monitoring scientific developments from a dynamic perspective : self-organized structuring to map neural network research », *Journal of the American society for information science*, vol. 49, no 1, 1998, p. 68-81.

Powell, Walter W., Douglas R. White, Kenneth W. Koput et Jason Owen-Smith, « Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences », *American journal of sociology*, vol. 110, no 4, 2005, p. 1132-1205.

Ramasco, José J., Sergey N. Dorogovtsev et Romualdo Pastor-Satorras, « Self-organization of collaboration networks », *Physical review E*, vol. 70, 036106, 2004.

Redner, Sidney, « Citation statistics from 110 years of physical review », *Physics today*, vol. 58, 2005, p. 49-54.

Robins, Garry et Malcolm Alexander, « Small worlds among interlocking directors: Network structure and distance in bipartite graphs », *Computational and mathematical organization theory*, vol. 10, 2004, p. 69-94.

Rosch, Eleanor et Barbara L. Lloyd, « Cognition and categorization », *American psychologist*, vol. 44, n° 12, 1978, p. 1468-1481.

Rueger, Alexander, « Robust supervenience and emergence », *Philosophy of Science*, vol. 67, n° 3, 2000, p. 466-489.

Salton, Gerard, A. Wong et C. S. Yang, « Vector space model for automatic indexing », *Communications of the ACM*, vol. 18, n° 11, 1975, p. 613-620.

Schmitt, Frederik (dir.), *Socializing epistemology: The social dimensions of knowledge*, Lanham (MD), Rowman & Littlefield, 1995.

Skyrms, Brian et Robin Pemantle, « A dynamic model of social network formation », *PNAS*, vol. 97, n° 16, 2000, p. 9340-9346.

Snijders, Tom A., « The statistical evaluation of social networks dynamics », *Sociological methodology*, vol. 31, 2001, p. 361-395.

Sokal, Robert R. et Peter H. A. Sneath, *Principles of numerical taxonomy*, San Francisco (CA), W. H. Freeman, 1963.

Stumme, Gerd, Rafik Taouil, Yves Bastide, Nicolas Pasquier et Lotfi Lakhal, « Computing iceberg concept lattices with TITANIC », *Data and knowledge engineering*, vol. 42, 2002, p. 189-222.

Touhey, John C., « Situated identities, attitude similarity, and interpersonal attraction », *Sociometry*, vol. 37, 1974, p. 363-374.

Turner, Heather, Susan Stepney et Fiona Polack, « Rule migration: Exploring a design framework for emergence », *International journal of unconventional computing*, vol. 3, n° 1, 2007, à paraître. Presented at ECAL 2005.

Van Der Merwe, Frederik J. et Derrick G. Kourie, « Compressed pseudo-lattices », *Journal of experimental and theoretical artificial intelligence*, vol. 14, n°s 2-3, 2002, p. 229-254.

Wasserman, Stanley et Katherine Faust, *Social network analysis: Methods and applications*, Cambridge, Cambridge University Press, 1994.

Watts, Duncan J. et Steven H. Strogatz, « Collective dynamics of "small-world" networks », *Nature*, vol. 393, 1998, p. 440-442.

White, Harrison C., Scott A. Boorman et Ronald L. Breiger, « Social-structure from multiple networks. I: Blockmodels of roles and positions », *American journal of sociology*, vol. 81, 1976, p. 730-780.

Whittaker, Robert H., « New concepts of kingdoms of organisms », *Science*, vol. 163, 1969, p. 150-160.

Wille, Rudolf, « Concept lattices and conceptual knowledge systems », *Computers mathematics and applications*, vol. 23, 1992, p. 493-515.

Wille, Rudolf, « Restructuring lattice theory: An approach based on hierarchies of concepts », dans Ivan Rival (dir.), *Ordered sets*, Dordrecht-Boston, Reidel, 1982, p. 445-470.