Analyse de Réseaux Sociaux : Détection et Qualification de communautés

Cécile Bothorel

IMT ATLANTIQUE

HABILITATION À DIRIGER DES RECHERCHES

Spécialité

Informatique

Analyse de Réseaux Sociaux : Détection et Qualification de communautés

---

Présentée par

Cécile BOTHOREL

Préparée à

Département logique des usages, des sciences sociales et de l'information (LUSSI)
Equipe DECIDE - Lab-STICC - CNRS, UMR 6285
IMT Atlantique

5 Juillet 2023

Université de Bretagne Occidentale

HABILITATION A DIRIGER DES RECHERCHES

Avis de soutenance

Madame BOTHOREL CECILE

présentera ses travaux en vue de l'habilitation à diriger des recherches, sur le sujet suivant :

"Analyse de Réseaux Sociaux : Détection et Qualification de communautés."

Le mercredi 5 juillet 2023 à 14h

à l'UBO, salle de conférences, Pôle numérique.

Le jury sera ainsi composé :

- M. AMBLARD FREDERIC, Professeur des universités
Université Toulouse 1 Capitole 2 - TOULOUSE
- MME BALAGUE CHRISTINE, Professeure des universités
Institut Mines-Télécom Business School - EVRY-COURCOURONNES
- M. BILLOT ROMAIN, Professeur
IMT Atlantique - PLOUZANE
- M. CHERIFI HOCINE, Professeur des universités
Université de Bourgogne - DIJON
- MME LARGERON CHRISTINE, Professeure des universités
Université Jean Monnet - SAINT-ETIENNE
- M. SINGHOFF FRANK, Professeur des universités
Univ. de Bretagne Occidentale - BREST

A BREST, le 22 juin 2023
Le Président de l'Université de Bretagne Occidentale,

P. OLIVARD

Résumé

Analyse de Réseaux Sociaux : Détection et Qualification de communautés

by Cécile BOTHOREL

Le travail présenté en vue de mon HDR offre un panorama des activités de recherche menées au cours des douze dernières années. Mon parcours scientifique, depuis mon doctorat, en passant par la R&D chez Orange Labs, a été varié et a abordé des problématiques différentes. Il apparaît rétrospectivement que toutes mes recherches s'articulent en réalité autour des trois concepts clés que sont la détection de communautés dans les réseaux sociaux, l'utilisabilité d'algorithmes par un décideur et l'interdisciplinarité.

Les médias sociaux offrent un terrain de recherche riche. De part l'accessibilité des données et leur nature intrinsèquement complexe, de nombreux défis théoriques et algorithmiques restent à relever. Ces dernières années, je me suis intéressée à la structuration en communautés des grands graphes réels (Complex Networks). La plupart des méthodes de clustering de graphes disponibles alors, ne prenaient que très peu en compte les métadonnées caractérisant les individus. J'ai proposé des algorithmes pour graphes avec attributs, tenant compte à la fois de la topologie des relations mais également des profils des nœuds, pour découvrir des communautés à la fois denses et homogènes.

Le deuxième verrou que j'ai adressé est celui du manque d'utilisabilité des méthodes de détection de communautés. Alors que le concept même de communauté ne fait pas consensus, et que les algorithmes développés pour les expliciter sont basés chacun sur leurs propres hypothèses, j'ai cherché à caractériser les partitions obtenues. J'ai proposé des méthodes, des métriques et une étude exhaustive d'une douzaine d'algorithmes sur plus d'une centaine de jeux de données réels. J'ai ainsi pu dégager des familles d'algorithmes, mais également apporter un éclairage sur les différences et points communs des différents types de réseaux sociaux étudiés, selon la nature de leur organisation communautaire.

Enfin, au-delà de la comparaison des algorithmes eux-mêmes, se pose la question du choix de l'algorithme à utiliser. J'ai travaillé avec des chercheurs en Sciences Humaines Sociales et proposé une méthodologie de sélection d'algorithme. Grâce à ce travail, je peux affirmer que l'interdisciplinarité n'est pas une fin en soi. Il ne s'agit pas de fournir des modèles et outils d'une disciplinaire à l'autre, mais les échanges entre disciplines se nourrissent mutuellement et deviennent un moyen de créer de la synergie et d'adresser de nouveaux challenges scientifiques et techniques.

Remerciements

Je tiens à remercier Christine Largeron, Hocine Cherifi et Frédéric Amblard d'avoir accepté la tâche fastidieuse de rapporter ce mémoire, et de m'avoir fait part de leurs conseils. Je remercie également Christine Balagué, Frank Singhoff et Romain Billot d'avoir participé à mon jury et rendu ce moment convivial et enrichissant.

Depuis mes débuts en thèse, j'ai eu l'occasion de travailler avec de nombreux co-auteurs, sans qui mes travaux n'auraient pas vu le jour et que je tiens à saluer. En particulier, je remercie l'ensemble des doctorants, post-doctorants et stagiaires que j'ai pu encadrer pour leur collaboration active et qui reconnaîtront une grande partie des résultats que je présente ici.

Merci également à toutes celles et ceux dont les remarques ou suggestions m'ont ouvert de nouvelles perspectives. Je réserve une pensée spéciale pour mes collègues passés et présents qui évoluent dans d'autres disciplines (sociologie et économie) pour m'avoir offert des terrains d'application aussi riches que variés et donné un fil conducteur à mes travaux.

Je voudrais ensuite remercier tous les membres du Département LUSSI d'IMT Atlantique pour l'ambiance de travail particulièrement agréable dans laquelle j'ai la chance d'évoluer. Mes remerciements vont également à l'ensemble des membres de l'équipe DECIDE du Laboratoire CNRS LAB-STICC, en espérant avoir l'occasion de partager encore de nombreux séminaires et les moments décontractés qu'ils offrent. Et bien sûr, je n'oublie pas les collègues d'Orange Labs puis d'IMT Atlantique : je saisis l'occasion pour leur exprimer mon plaisir de travailler parmi eux.

La mémoire me fait défaut pour citer nommément toutes les personnes qui ont jalonné mon parcours, tant personnel que professionnel, mais rien n'aurait été possible sans eux.

Je voudrais remercier enfin mes collègues Laurent, Inna, Nicolas et Nicolas avec qui je me projette dans la suite des travaux décrits dans ce manuscrit.

Un remerciement tout particulier à mes coachs Nicolas et Sophie qui ont su chacun à leur manière me motiver et m'aider à prendre (presque toujours) plaisir à écrire.

Et bien sûr, merci Hugo, Barbara, et Sophie pour tout le reste, sans vous rien n'aurait été possible.

Table des matières

Introduction

L'analyse de réseaux sociaux est historiquement une approche en Sciences Sociales où les acteurs et leurs relations sociales sont modélisés par des réseaux. Moreno a probablement été le premier en 1934 à utiliser un graphe, le *sociogramme*, pour étudier les affinités de 506 jeunes filles d'un même pensionnat, et ainsi analysé le fonctionnement de groupes sociaux, non pas en tant que somme d'individus, mais comme une organisation régie par des relations inter-personnelles (Moreno, 1934). Le simple fait de modéliser les relations par un graphe permet d'en analyser la structure. La forme du réseau renseigne sur son efficience d'un point de vue global, mais aussi, selon la position des acteurs qui la composent, détermine leurs opportunités ou leurs contraintes. Par exemple, selon la théorie bien connue de la *force des liens faibles* (Granovetter, 1973), un réseau offrant des zones de densité faible entre des zones plus denses (les *cercles* de Mark Granovetter) permet une certaine ouverture et par exemple une richesse d'information pour les acteurs situés aux frontières de ces zones, qui en plus de leur relations « fortes », ont aussi de simples « connaissances » qui apportent de la diversité. L'analyse structurale accorde beaucoup d'importance à la qualification de positions-clé et associe la notion de pouvoir par exemple à des acteurs centraux dans un réseau. Parmi les autres travaux emblématiques, on peut citer l'expérience de Stanley Milgram à l'origine du concept de réseaux *petit monde* (Milgram, 1967) ou encore les notions de *trous structuraux* et de *capital social* introduites par Ronald Burt (Burt, 1992).L'objectif de l'analyse structurale est double : il s'agit, pour Michel Forsé de « comprendre en quel sens une structure sociale contraint formellement des comportements, tout en résultant des interactions entre les éléments qui la constituent » (Forsé, 2008).

Avec l'accès à des grands jeux de données, l'analyse des grands graphes terrains ou réseaux complexes (en anglais, *Complex Networks*) est devenue une discipline à part entière depuis le tout début des années 2000, à la croisée de l'Informatique et de la Physique. Tandis que [1] les physiciens recherchent les lois régissant ces réseaux complexes et proposent des modèles de génération de graphes ou encore des benchmarks basés sur ces modèles génératifs ; citons par exemple les travaux d'Albert et Barabási définissant statistiquement les réseaux aléatoires, petit monde ou encore scale-free (Albert, Jeong et Barabási, 1999 ; Barabási et Albert, 1999). Les informaticiens proposent quant à eux des algorithmes pour calculer sur ces grands jeux de données des positions ou des zones remarquables, respectivement des *centralités* et des *communautés*, ou encore ils cherchent à prédire l'apparition de liens dans un réseau ou inférer des nœuds manquants dans des collectes partielles de données ; des ingénieurs, travaillant pour un moteur de recherche, inventent la mesure de centralité du *PageRank* (Page et al., 1999) qui a fait la fortune de Google ; Girvan et Newman mettent en évidence la structuration en *communautés*, communes aux grands graphes

1. Dichotomie un peu abusive, j'en conviens. Bien sûr les frontières sont poreuses et deviennent probablement moins nettes au fil du temps.

réels, et proposent alors un algorithme (Edge Betweenness) pour les détecter (Girvan et Newman, 2002).

De leur côté, les chercheurs en marketing utilisent ces outils pour détecter des acteurs influents, conçoivent des modèles de diffusion et procèdent à des simulations sur des réseaux synthétiques générés selon les modèles de graphes complexes ci-dessus. Des sociologues (Traud, Mucha et Porter, 2012) ou des biologistes (Enright, Van Dongen et Ouzounis, 2002) utilisent la détection de communautés pour mieux comprendre la structure des objets qu'ils étudient. En management, c'est l'aspect organisationnel des équipes qui est étudié, et par exemple leur efficacité ou encore la qualité de leur production. Gerald C. Kane et Sam Ransbotham ont montré que certaines organisations mises en œuvre par les contributeurs de Wikipédia conduisent à des articles de qualité. Ces mêmes auteurs ont également analysé la centralité des articles modélisés par un réseau et montré que la position d'un article est corrélée avec sa qualité (Kane et Ransbotham, 2012 ; Kane et Ransbotham, 2016).

Sans aller jusqu'à faire une exhaustive étude historique (voire sociologique, anthropologique, philosophique, etc.) de l'analyse de réseaux au sein de différentes disciplines, il me semble que l'on passe d'une relation « outil-domaine d'application » ou « question de recherche-outils » à une relation plus symétrique, moins utilitaire.

La fertilisation croisée entre disciplines, en particulier l'intelligence artificielle et les humanités est bien une tendance actuelle. Plus qu'une rupture scientifique, le succès actuel de l'intelligence artificielle est lié à un développement spectaculaire des masses de données et des moyens de calcul, et plus généralement, comme le souligne Cédric Villani, à une « mise en données du monde (datafication) ». L'interdisplinarité est un des axes majeurs pour une « recherche agile et diffusante », attractive en cerveaux et fédératrice de travaux réunissant acteurs académiques et industriels. [2].

Ainsi de nombreux projets interdisciplinaires sont encouragés par les financeurs, notamment français, mais également à l'échelle internationale. Des numéros spéciaux de revues dédiées à l'intelligence artificielle ou en informatique, mais aussi en Sciences Humaines et Sociales mettent en avant ces collaborations. Il en est de même au niveau des conférences, workshops, écoles d'été où la diversité des disciplines est encouragée. L'objectif reste bien sûr d'apporter des terrains applicatifs aux algorithmes mis en place, mais l'enjeu, me semble-t'il, est également d'identifier de nouvelles problématiques, mettre en place de nouvelles méthodologies et d'enrichir mutuellement les travaux traditionnellement mono-disciplinaires.

Comme l'expliquent Loet Leydesdorff et Inga Ivanova, l'interdisciplinarité n'est pas un objectif en soi, mais un moyen de créer de la *synergie*. Synergie signifie que le tout offre plus de possibilités que la somme de ses parties (Leydesdorff et Ivanova, 2020). Le travail décrit dans ce mémoire s'inscrit indubitablement et résolument dans cette tendance.

Nous avons constaté que les techniques de détection de communautés, pourtant un sujet mature dans le domaine des réseaux complexes, ne sont à l'heure actuelle pas encore si démocratisées en Sciences Humaines et Sociales. Les humanités se sont emparées des métriques locales et manipulent les différents type de centralités couramment pour détecter différents rôles ; elles utilisent également les métriques globales pour quantifier le diamètre, la densité ou encore le coefficient de clustering et en déduire des caractéristiques globales de cohésion par exemple. Mais la détection

2. Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A. C., Levin, F., & Rondepierre, B. (2018). Donner un sens à l'intelligence artificielle

de structures à un niveau intermédiaire, dit meso, est sous-exploitée. L'objectif de la *détection de communautés* est la détection de ces structures. Il n'y a pas de consensus sur la définition de ce qu'est une communauté. Il est cependant communément admis qu'il s'agit d'une partie dense d'un réseau (Fortunato, 2010). Il existe bien sûr des applications réelles, telles que la recherche d'organisations criminelles (Yang et al., 2012), la détection de spam (Cao et al., 2012), ou encore pour étudier comment l'homophilie structure un réseau tel que Facebook en sociologie (Traud, Mucha et Porter, 2012). Mais en général, une seule méthode est utilisée sans vraiment s'interroger sur la philosophie régissant son principe de résolution, et sans se demander si elle est la plus adaptée au besoin et à la nature des regroupements d'acteurs visés.

Au cours de ces dernières années, j'ai travaillé principalement sur ces deux axes et contribué à résoudre deux verrous scientifiques qui peuvent se résumer ainsi :

— Le manque d'expressivité des graphes exploités par les méthodes de détection de communautés
— Le manque d'utilisabilité des méthodes de détection de communautés

Détecter des zones denses revient, dans la plupart des méthodes, à ne s'intéresser qu'à la structure, à la topologie des relations. Or de nombreuses questions cherchent, comme c'est le cas des applications d'apprentissage non supervisé, à trouver des acteurs proches mais *similaires*, que l'on peut décrire et des groupes homogènes que l'on peut interpréter. Dans un premier temps, je focaliserai mon propos sur la détection de communautés *socio-sémantiques*, ou encore la détection de communautés dans des graphes avec attributs (*attributed networks* en anglais). J'illustrerai avec une étude marketing de la blogosphère culinaire et montrerai la pertinence de l'utilisation conjointe de relations sociales et d'attributs pour cibler des influenceurs sensibles à une thématique donnée. J'ai pu appliquer également cette association de données complémentaires dans le cadre de la recommendation de contenus.

Du fait de la variété des contextes d'analyse, les définitions des communautés sont multiples. Creusefond, 2017 donne l'exemple d'une entreprise cherchant à diffuser un message publicitaire dans certaines communautés, et pointe le fait qu'alors, les communautés devront être définies par les facteurs facilitant la transmission. Si l'on reprend la problématique de la détection de spam, il s'agit plutôt de se baser sur des critères de cohésion interne, où les membres d'une communauté doivent avoir des caractéristiques proches (Largillier, Peyronnet et Peyronnet, 2010).

Mais alors, quel algorithme appliquer pour quel type de communautés recherchées ? Comment définir ce qu'est une communauté ? Comment utiliser les algorithmes et comparer leurs résultats ? C'est ce que j'aborderai dans un second temps. Je balaierai les différentes mesures de qualité des communautés, et au-delà de leur propre interprétation, je proposerai à travers un exemple appliqué à une plate-forme de financement participatif, comment choisir une méthode et utiliser le résultat d'une analyse bivariée pour faire ce choix éclairé de méthodes.

Le décideur est bien au cœur de mes préoccuppations. A l'ère de la démocratisation de l'intelligence artificielle, il reste encore beaucoup de chemin à parcourir pour rendre nos travaux accessibles. Que ce soit nos méthodes, algorithmes ou résultats, tant dans notre propre domaine scientifique que pour répondre à des questions diverses et variées. Evidemment, je ne cherche pas à produire des méthodes universelles ; en jetant mon dévolu sur les terrains en Sciences Sociales, j'ai conscience de l'ampleur de la tâche, mais j'affiche néanmoins l'ambition, en collaboration avec mes collèges, de paver la discipline des *Sciences Sociales Computationnelles*.

Chapitre 1

Réseaux sociaux avec attributs

Un réseau social est l'ensemble des relations sociales entre les membres d'un groupe. Il est communément défini par des connexions entre acteurs ; les acteurs peuvent être des personnes ou des organisations, les liens peuvent être de plusieurs types : amitié, affiliation, dépendance ou encore communication. L'analyse de réseaux sociaux considère ces relations comme des graphes, où les acteurs sont représentés par des nœuds et les relations par des arêtes. Le médecin psychiatre Jacob L. Moreno a introduit le sociogramme en 1933. Représenter un réseau social par un graphe objective les relations interpersonnelles d'un groupement social, ce qui permet d'en étudier la structure des relations : les liens d'influence, les individus en position de pouvoir, la dynamique du groupe, etc.

Toutefois, à l'heure actuelle, les jeux de données décrivant des réseaux sociaux contiennent également des données décrivant chaque acteur, par exemple ses préférences de lecture, son âge, son emplacement géographique, ou par exemple d'autres informations relatives au contexte du réseau (en ligne ou hors ligne). Ces informations additionnelles, non structurelles, sont souvent sous-exploitées. Les réseaux sociaux enrichis de la sorte seront appelés dans la suite de ce document *réseaux sociaux augmentés*, ou *réseaux sociaux avec attributs* (*attributed graphs* en anglais), pour les différentier des réseaux sociaux purement structurels modélisés par des sociogrammes.

Ce travail présente un cadre général pour faciliter l'intégration des différents types d'information présents dans un réseau social. Notre proposition est de combiner la composante structurelle avec des données décrivant les acteurs, puis exécuter différentes analyses sur ces réseaux sociaux augmentés. Au-delà de la formalisation d'un modèle de réseau social avec attributs, et de la méthode d'intégration des différents types de données, cette proposition inclut également un modèle visuel centré sur l'interaction entre communautés, permettant d'analyser finement les rôles des individus clés y participant.

1.1 Les composantes d'un réseau social augmenté

La typologie des réseaux sociaux proposée par Wasserman structure notre réflexion concernant l'analyse de réseaux sociaux augmentés (Wasserman et Faust, 1994). Ceux-ci peuvent être décrits à travers trois *variables* : (1) la variable structurelle qui contient l'information des liens entre les acteurs, (2) la variable de composition, qui décrit chaque acteur, par exemple avec un profil, et (3) la variable d'affiliation, utilisée pour décrire les nœuds selon leur affectation à différents groupes, événements ou division d'une entreprise.

Dans la Figure 1.1 est présenté un exemple de réseau social d'entreprise. Cet exemple corporatif est simple, et surtout de petite taille, mais il aide à présenter les variables introduites : la variable structurelle décrit les collaborations entre les employés, elle est représentée par les arêtes ; la variable de composition contient le nom, ainsi que la localisation du poste occuppé, voire éventuellement des informations sur l'expertise, l'âge, l'ancienneté, etc. ; la variable d'affiliation est ici dérivée de l'affectation de chaque acteur à l'une des deux agences de l'entreprise et est représentée par la forme des nœuds : les nœuds carrés représentent des personnes de New York tandis que les nœuds ronds représentent des personnes du bureau de Boston

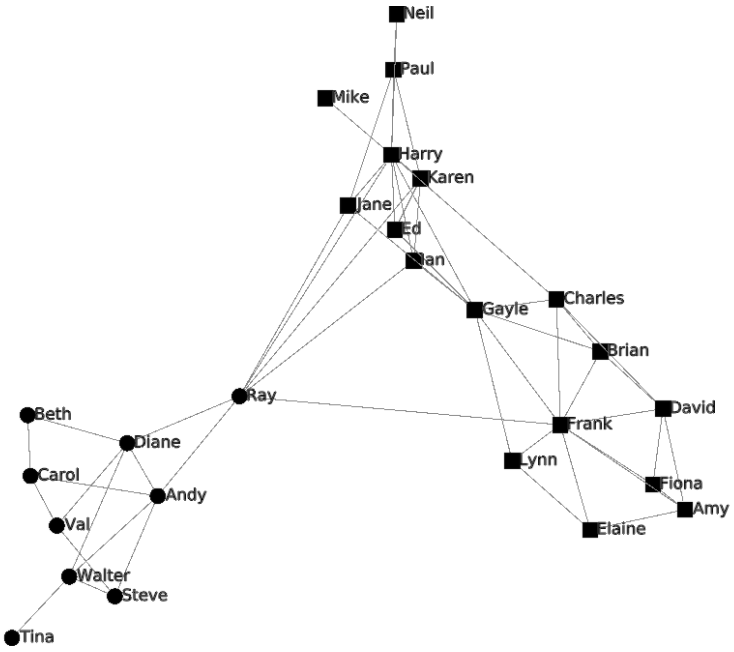


FIGURE 1.1 – Exemple d'un réseau social corporatif basique

A noter que la terminologie de *variable* introduite par Wasserman en 1994 introduit du flou dans un contexte de data science appliquée aux graphes (notre domaine), où une variable a une granularité plus faible. Nous préférerons ici la notion de *composante* à variable.

L'analyse des réseaux sociaux cherche à extraire de nouvelles connaissances de chacune des trois composantes présentées ci-dessus. Toutefois ce traitement n'utilise qu'un seul type d'information à la fois en général. Par exemple, en n'utilisant que la composante structurelle, il est possible d'identifier des nœuds centraux ou effectuer d'autres mesures sur la topologie du graphe. En revanche, avec la composante de composition, les chercheurs peuvent appliquer des techniques de fouille sur les données décrivant les individus, et analyser une partition, des motifs et des règles d'association par exemple.

La composante d'affiliation peut nourrir l'étude de l'appartenance (ou la co-appartenance) des acteurs à certains groupes ou associations. Cette composante a cette particularité qu'elle peut être aussi générée à partir de l'une ou l'autre des deux autres composantes. L'affectation à un groupement social peut résulter d'un processus de détection de communautés qui produit une partition du graphe et classe les nœuds dans des groupes découverts par un algorithme. La variable d'affiliation n'est autre qu'une représentation d'un graphe biparti (Wasserman et Faust, 1994).

1.1.1 Approche centrée décideur

Dans ces quelques exemples d'analyse cités ci-dessus, chaque composante est traitée séparément, rejetant une portion de l'information au détriment d'analyses plus complètes, tenant compte de l'ensemble de la richesse des données disponibles.

Une des questions à prendre en compte avant d'analyser les données de composition, est la diversité des valeurs qu'elle peut contenir. Par exemple cette composante contient l'information du genre, de l'âge, de la taille et d'autres caractéristiques de l'acteur, comme les préférences personnelles en termes de sports, livres, art, tout ceci pouvant être décrits par des mots clés relatifs à des publications. Malgré le fait que cette variété permet de bien décrire chaque acteur, elle engendre des difficultés pour analyser les acteurs comme un ensemble. Ce phénomène est bien connu sous le nom de *malédiction de la dimension* (Bellman, 1961). A mesure que le nombre de dimensions augmente, les données deviennent très vite très éparses dans l'espace, ce qui met les modèles d'apprentissage en difficulté ; de plus, utiliser toutes les caractéristiques qui décrivent un acteur peut introduire du bruit pour certaines applications, et tout simplement ne pas être pertinent pour l'analyste qui cherche, au final, à étudier l'impact de certaines caractéristiques bien choisies sur la structuration du réseau social.

Une approche pour éviter ce phénomène est de réduire la dimension. Plutôt que de la réduire de manière automatique (on sort du cadre de ce travail précis, et on y reviendra dans les perspectives, avec la notion de plongement), nous proposons de confier ici cette tâche à l'analyste du réseau social. En fonction du but recherché, il sera amené à diviser le profil des acteurs en des sous-ensembles qui ont un sens pour lui : l'idée est d'imaginer ici une *méthode générique d'analyse de réseau exploitant la sélection, opérée par un décideur, d'un sous-profil pertinent* dans le contexte d'une analyse. On pourra ainsi par exemple analyser un réseau social en ne considérant que les loisirs ou bien en ne considérant que les compétences académiques des acteurs.

L'analyse de réseau social augmenté proposée ici porte principalement sur la détection de communautés pour laquelle on souhaite *classer, au sein d'un même groupe, les acteurs à la fois proches dans la structure du graphe et proches du point de vue de leur profil*. Cette analyse automatique s'accompagne d'une analyse visuelle avec la proposition d'une visualisation adaptée.

L'idée est à terme de pouvoir approfondir l'étude d'un réseau social en montrant l'impact de données décrivant les acteurs sur la structuration du réseau. L'un des premiers cas d'usage proposés est de comparer les différents regroupements obtenus par la sélection de différents sous-ensemble de données, appelés ici sous-profils. Nous parlerons ici de "point de vue" pour désigner la sélection d'un sous-profil (Cruz Gomez, Bothorel et Poulet, 2011c). Ce sous-profil conditionnera l'ensemble du processus d'analyse, et peut être vu alors comme un paramètre de l'analyse. A noter que le "point de vue" prend tout son sens dans un outil interactif d'analyse visuelle où le décideur est amené à dérouler plusieurs scénarios. Ce "point de vue"

peut également être "vide" si l'on ne le considère pas du tout, et dans ce cas on se ramène à l'analyse de la composante structurelle seule.

Nous définissons ci-après la notion de *point de vue* et la manière dont nous modélisons les réseaux sociaux, en augmentant le graphe social par des variables de composition (variables prises ici dans le sens de la data science).

1.1.2 Définitions

L'analyse d'un réseau social intégrant l'ensemble des données qui le composent n'est pas un processus simple. Cette intégration exige d'établir des relations entre elles, mais aussi de créer un mécanisme exploitant leur unification afin d'extraire de nouvelles connaissances, en complément des mécanismes d'analyse basés sur chacune d'elles prise séparément.

Le problème peut être divisé en deux sous problèmes (Figure 1.2). D'abord, comment représenter les composantes de telle façon qu'elles puissent être utilisées et analysées de façon holistique, et par quelle méthode ? Ensuite, comment utiliser un modèle de visualisation pour représenter l'intégration des trois composantes concernées ?

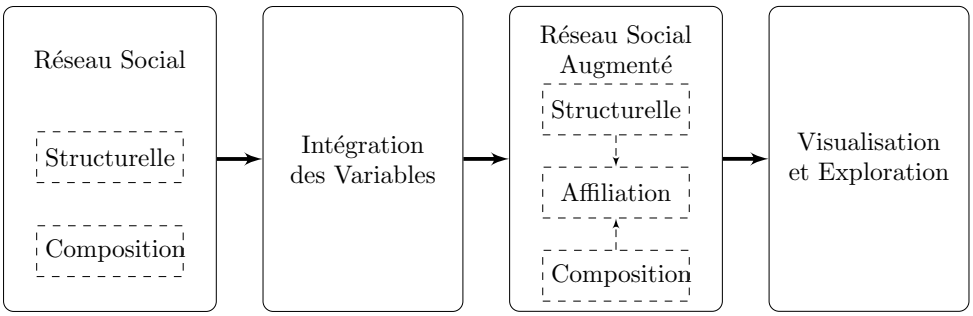


FIGURE 1.2 – Exemple d'un réseau social corporatif basique

Pour faire cela nous introduisons ici les définitions des différentes *composantes* et de *graphe augmenté*, ainsi que la notion de *point de vue*. Soit $G(V, E)$ un graphe non orienté représentant la structure du réseau social, où V est l'ensemble des nœuds et E est l'ensemble des arêtes ; ce graphe contient uniquement l'information concernant les connections des nœuds, c'est-à-dire la composante structurelle. La composante de composition de ce réseau peut être définie comme un vecteur F_i^* modélisant les variables de chaque acteur i : ce vecteur peut être défini dans un espace $d-$dimensionnel où d peut varier selon la nature du profil décrivant les acteurs.

La composante structurelle représente l'ensemble des relations entre les acteurs, soit le graphe social G. La composante compositionnelle représente les variables (*features* en anglais) F^* qui décrivent chaque acteur de façon individuelle selon d dimensions, par exemple les profils des acteurs. Enfin, la composante d'affiliation $\mathcal{A}$ représente l'appartenance des acteurs à des sous-ensembles, par exemple des clubs ou des communautés. La composante d'affiliation peut être dérivée soit à partir de la composante structurelle, soit à partir de la composante compositionnelle, soit d'une combinaison des deux telle que proposée dans ce travail.

Étant donné F un sous-ensemble de f caractéristiques compositionnelles pris dans F^*, un point de vue $PoV_F = \{f_1, ..., f_n\}^T \in \mathcal{R}^{nxf}$, avec $f \leq d$, décrit les n nœuds du graphe G.

Ainsi, nous pouvons définir un réseau social augmenté $\mathcal{S}^+$. Etant donné un réseau social $\mathcal{S}(G, PoV_F)$ où G est un graphe représentant la composante structurelle

et PoV_F représentant la composante de composition sous la forme d'un point de vue, le réseau social augmenté est défini par $\mathcal{S}^+(G, PoV_F, \mathcal{A})$, où $\mathcal{A}$ est la composante d'affiliation dérivée des autres deux composantes.

Grâce au point de vue, le réseau social augmenté peut être décrit par des perspectives différentes qui permettent de réaliser une analyse pour chaque perspective. Par exemple, les nœuds peuvent être groupés de plusieurs manières : une par point de vue, en rendant possible les contrastes selon les caractéristiques sélectionnées dans F (Cruz Gomez, Bothorel et Poulet, 2010 ; Cruz Gomez, Bothorel et Poulet, 2011c).

1.2 Détection de communautés dans les graphes avec attributs

L'une des hypothèses de ce travail est de considérer que la détection de communautés constitue un type d'analyse permettant de mettre en œuvre les trois composantes décrites précédemment. Nous ambitionnons de classer, au sein d'un même groupe, les acteurs à la fois proches dans la structure du graphe et proches du point de vue de leur profil.

1.2.1 Détection de communautés

L'analyse des réseaux sociaux est en général dédiée à la composante structurelle. On trouve par exemple l'identification des nœuds importants ou des relations non triviales. Les différentes métriques d'analyse sont en général calculées sur le graphe pris comme un tout, avec la notion structurante de chemin. La détection de communauté est l'une de ces analyses désormais clé lorsque l'on traite de grands (voire très grands) graphes réels. Elle permet de trouver de manière automatique des zones denses, topologiquement parlant, dans le graphe social.

Il s'agit d'un processus de clustering basé sur une distance entre les nœuds définie par les chemins du graphe. Il tient compte du nombre important d'arêtes tissant le réseau d'un sous-ensemble de nœuds en comparaison avec le réseau global. L'idée de base est de trouver des partitions dont le nombre d'arêtes à l'intérieur des groupes est supérieur au nombre d'arêtes en dehors des groupes. Par conséquent, chaque groupe ainsi identifié possède une forte connectivité à l'intérieur et une faible connectivité avec les autres groupes.

Des revues de littérature inventorient et proposent des typologies des (très) nombreuses méthodes de détection de communautés existantes (Fortunato, 2010 ; Fortunato et Hric, 2016), y compris pour les méthodes produisant des communautés avec recouvrement où un nœud peut appartenir à plusieurs groupes (Chen et Wei, 2019). Nous avons nous-mêmes étudié de manière approfondie une douzaine d'entre elles parmi les méthodes bien connues (Dao, Bothorel et Lenca, 2020), le Chapitre 3.5 détaillera ces travaux.

Mais les méthodes pointées par ces états de l'art n'utilisent que la composante structurelle du réseau social. La communauté scientifique n'a abordé ce type d'analyse sur les graphes avec attributs que dans les débuts des années 2010. Nous avons assez tôt proposé une revue de littérature sur ce type de réseaux (Bothorel et al., 2015). Depuis lors, nous pouvons citer (Chunaev, 2020), dans laquelle trois types d'approches sont proposées par la communauté scientifique :

1. le *pré-traitement préalable des attributs*, qui modifie le réseau pour encoder les différentes composantes dans les arêtes, permet ensuite d'appliquer des algorithmes de détection de communautés classiques ;

2. une intégration en *post-traitement*, qui à partir de deux clusterings de nœuds, l'un structurel, l'autre compositionnel, fusionne les résultats pour produire une seule partition hybride ;
3. sans pré- ni post-traitement, la *fusion des composantes* se fait de manière simultanée lors de la découverte des communautés dans le réseau.

Nous avons contribué à ces trois approches. La section 1.2.2 détaille notre méthode avec pré-traitement, qui enchaîne un clustering d'attributs (grâce ici à la méthode d'apprentissage non supervisé des cartes de Kohonen) puis la détection de communautés par l'algorithme de Louvain dans un graphe dont les arêtes encodent la simlarité de profil des noeuds (Cruz Gomez, Bothorel et Poulet, 2011c ; Cruz Gomez, Bothorel et Poulet, 2013a ; Cruz Gomez, Bothorel et Poulet, 2014).

Concernant la deuxième approche (section 1.2.3), le clustering des deux composantes structurelle et compositionnelle sont traitées en parallèle, puis l'intégration des composantes se fait en post-traitement (Cruz Gomez, Bothorel et Poulet, 2013b ; Cruz Gomez et Bothorel, 2013).

La section 1.2.4 expose quant à elle l'intégration simultanée des composantes dans une variante de l'algorithme de Louvain optimisant, à chaque itération, non seulement la modularité, mais également l'entropie relative aux attributs des nœuds (Cruz Gomez, Bothorel et Poulet, 2011a). Cette méthode, très simple à comprendre, implémenter, et rapide à exécuter, est régulièrement utilisée comme baseline pour de nombreux travaux.

1.2.2 Intégration des composantes dans des réseaux sociaux augmentés : une méthode avec pré-traitement

L'intégration des composantes structurelles et de composition permet de guider le processus d'identification des communautés en ajoutant la similarité (compositionnelle) de chaque nœud aux critères de la structure utilisés pendant l'identification des communautés. Pour ce faire, selon l'approche avec pré-traitement, nous divisons le processus de détection de communautés en deux phases : d'abord, un clustering de nœuds selon leur similarité de profil, puis, après avoir modifié les poids des arêtes du graphe pour refléter une distance composite des noœuds, un algorithme de clustering structurel influencé de telle manière à ce que les groupes contiennent des nœuds similaires et connectés (Cruz Gomez, Bothorel et Poulet, 2011c ; Cruz Gomez, Bothorel et Poulet, 2013a ; Cruz Gomez, Bothorel et Poulet, 2014).

Première phase : clustering de composition

Étant donné un point de vue dérivé d'un ensemble PoV_F, chaque nœud peut être caractérisé par son vecteur d'attributs ou une instance u du point de vue. Il est possible d'utiliser ces vecteurs en entrée d'un algorithme de classification non supervisée comme les cartes auto-organisatrices ((Kohonen, 1997)). Cela permet de créer des groupes de nœuds suivant la similarité de leurs attributs, c'est-à-dire les instances de u sont les données en entrée de l'algorithme. L'avantage de cet algorithme est que, à la différence des approches comme les $k-$means, l'utilisateur n'a pas besoin de fixer a priori le nombre final de groupes.

L'algorithme de cartes de Kohonen utilisé a un réseau $\mathcal{N}$ basé sur une grille rectangulaire de taille de $f \times f$ neurones, avec $f = |PoV_F|$, le nombre d'attributs utilisés dans le point de vue. Les valeurs initiales des poids sont tirées aléatoirement. Les poids des neurones sont ajustés selon leur proximité au neurone gagnant. Un taux

d'apprentissage η est utilisé pour éviter les maxima locaux et des convergences prématurées. Après chaque itération, le taux d'apprentissage est réduit par un facteur $\varepsilon, 0 < \varepsilon < 1$. Le voisinage est calculé avec une taille t et le neurone gagnant est de centre c.

La sortie est alors une partition $\mathbf{C}_{SOM}$ formée par des groupes de nœuds similaires en termes du point de vue choisi. Pour mesurer la qualité de la partition nous utilisons la distance moyenne entre les points de chaque groupe, laquelle a été mise à l'échelle pour avoir des valeurs entre 0 et 1.

Deuxième phase : influence de la composition sur le clustering structurel

Une fois que la partition compositionnelle $\mathbf{C}_{SOM}$ a été calculée, on peut alors entrer dans la seconde phase de la méthode. Dans cette étape, on utilise un algorithme classique de détection de communautés, l'algorithme Fast Unfolding, connu aussi sous le nom d'algorithme de Louvain, proposé par (Blondel et al., 2008). Cet algorithme utilise un processus de Monte-Carlo pour optimiser la modularité Q, présentée par (Newman et Girvan, 2004).

Avant l'exécution de la méthode de Louvain, on inclut les informations obtenues lors de la première phase. Cela est effectué par la modification des poids des arêtes en fonction de la partition obtenue $\mathbf{C}_{SOM}$. Pour chaque paire de sommets $v_i, v_j \in V$, $\forall v_i \neq v_j$, le poids de l'arête $e\left(v_i, v_j\right)$ est modifié par la distance euclidienne des instances du point de vue correspondant à chaque nœud :

$$w_{ij} = 1 + \alpha(1 - d(\mathcal{N}_{ij}))\delta_{ij} \tag{1.1}$$

avec $\alpha \geq 1$ une constante, $d\left(\mathcal{N}_{ij}\right)$ la distance entre les neurones i et j, et $\delta_{ij} = 1$ si v_i et v_j appartiennent au même cluster dans $\mathbf{C}_{SOM}$ et 0 sinon.

Une fois que les poids sont modifiés selon l'équation 1.1, une partition, $\mathbf{C}_{SOM-FU}$ est calculée en utilisant le Fast Unfolding (FU).En modifiant les poids du graphe avec l'équation 1, le graphe devient pondéré et les arêtes avec un poids plus grand ont une probabilité plus élevée d'être affectées à la même communauté.

La composante d'affiliation $\mathcal{A}_{\mathcal{S}^+}$ résultante reflète l'ensemble des communautés avec des groupes de nœuds similaires et connectés, qui intègrent l'information de la structure et des attributs.

1.2.3 Intégration des composantes dans des réseaux sociaux augmentés : une méthode avec post-traitement

Dans ce travail, nous présentons une méthode qui intègre les composantes en post-traitement. L'idée est de combiner, après coup, des partitions structurelles et compositionnelles existantes, de bonne qualité car obtenues sans compromis de manière indépendante.

La motivation est double. Premièrement, il s'agit d'utiliser des techniques spécialisées pour chaque type de composante, ce qui produit des clusters de bonne qualité, des communautés topologiques bien nettes d'une part, et des groupes de nœuds homogènes du point de vue de leurs attributs d'autre part. Deuxièmement, une méthode offrant la réutilisation des partitions calculées au préalable, peut s'avérer précieuse en termes de complexité de calcul, lorsque l'on manipule des grands réseaux.

Nous nous appuyons sur une matrice de contingence avec les groupes structurels en lignes et les groupes compositionnels en colonnes (Cruz Gomez, Bothorel et

Poulet, 2013b ; Cruz Gomez et Bothorel, 2013). Le problème revient alors à manipuler les lignes et les colonnes de la matrice pour obtenir une nouvelle partition qui maintienne un bon compromis entre les deux dimensions.

La matrice de contingence telle que présentée dans le tableau 1.1, est une matrice où chaque entrée n_{ij} représente le nombre de nœuds communs entre les clusters $u_i \in \mathbf{C}_G$ et $v_j \in \mathbf{C}_{F^*}$.

		Partition $\mathbf{C}_{F^*}$				
	Class	v_1	v_2	...	v_r	Somme
	u_1	n_{11}	n_{12}	...	n_{1r}	$n_{1\cdot}$
	u_2	n_{21}	n_{22}	...	n_{2r}	$n_{2\cdot}$
Partition $\mathbf{C}_G$	⋮	⋮	⋮	⋱	⋮	⋮
	u_m	n_{m1}	n_{m2}	...	n_{mr}	$n_{m\cdot}$
	Somme	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot r}$	n

TABLE 1.1 – Matrice de contingence de deux partitions : $\mathbf{C}_G$, obtenue à partir de la composante structurelle et $\mathbf{C}_{F^*}$, obtenue à partir des attributs

Le principe de notre algorithme est de manipuler la configuration des lignes afin de décomposer la partition structurelle $\mathbf{C}_G$ en fonction de la partition compositionnelle $\mathbf{C}_{F^*}$ en colonne.

La division de chaque ligne peut être effectuée selon différents critères dépendant uniquement de la configuration de la matrice de contingence, en particulier des colonnes de la matrice. Nous proposons deux stratégies pour contrôler la combinaison :

— Naïve : pour chaque ligne i, si $n_{ij} > 0$ alors les nœuds appartenant au groupe structurel u_i et au groupe de composition v_j formeront une communauté, c'est-à-dire que pour chaque entrée supérieure à 0 dans la matrice de contingence il y aura une communauté dans la partition finale combinée.

— Basée sur la variance : pour chaque ligne i, si $\frac{(n_{ij}-\mu_i)}{\sigma_i} \geq 1$, cet ensemble n_{ij} constituera une nouvelle communauté. Ici, μ_i et σ_i sont respectivement la moyenne et l'écart-type de la ligne i. Ainsi, les communautés structurelles sont réparties en fonction de la représentativité des catégories de composition, c'est-à-dire celles qui présentent une variance positive plus importante.

Ces deux stratégies nous permettent d'évaluer les groupes structurels à la lumière des groupes compositionnels et de les décomposer si la condition est remplie.

Pour illustrer notre proposition, nous utilisons un petit réseau social composé de 24 nœuds et de 63 arêtes. Ce réseau a été divisé en quatre groupes structurels et trois groupes de composition, qui pourraient refléter 3 classes d'âge ou toute autre clusters pré-calculés à partir de profils plus élaborés. La première étape consiste donc à construire la matrice de contingence, présentée dans le tableau 1.2.

L'étape suivante est le traitement de chaque ligne de la matrice de contingence $\mathcal{C}$. Le critère utilisé dans cet exemple est le critère naïf. La nouvelle partition $\mathbf{C}^*_{\text{Naïve}}$ est composée de 9 groupes comme présenté dans le tableau 1.3. Enfin, il s'agit de reconstruire la matrice d'affiliation.

Le tableau 1.4 présente un résumé des résultats pour cet exemple de base. Nous utilisons trois mesures pour comparer les résultats finaux. Premièrement, nous utilisons l'ARI pour mesurer la similarité entre les partitions, deuxièmement, la densité, qui mesure la partition du point de vue de la partition structurelle et enfin, l'entropie, qui mesure l'ordre, c'est-à-dire l'homogénéité des groupes du point de vue des attributs dans la partition finale.

	$\mathbf{C}_{F^*}$		
$\mathbf{C}_G$	3	3	0
	2	3	1
	3	2	1
	0	0	6

TABLE 1.2 – Matrice de contingence $\mathcal{C}$ pour notre exemple jouet de réseau social

L'ARI entre la nouvelle partition et la partition de composition originale montre que la distance entre elles a été réduite, ce qui signifie que les nouveaux groupes sont davantage alignés avec la variable de composition que les groupes découverts sur la seule base de la structure des relations. Ceci est également observé sur la valeur de l'entropie, qui tombe à 0 indiquant que chaque nouveau groupe est composé de nœuds ayant des attributs similaires ; cependant, le coût de la réduction de l'entropie est la perte de la densité, ce qui implique une réduction de la qualité de la partition en termes structurels.

		$\mathbf{C}_{F^*}$		
$\mathbf{C}_G$	u_0	3	0	0
		0	3	0
	u_1	2	0	0
		0	3	0
		0	0	1
	u_2	3	0	0
		0	2	0
		0	0	1
	u_3	0	0	6

TABLE 1.3 – Matrice de contingence résultant du processus de séparation des communautés structurelles. Les 4 groupes structurels initiaux u_i ont été subdivisés en 9 nouveaux groupes.

Partition	**Groupes**	**ARI** $(_{\text{w.r.t}\,\mathbf{C}_{F^*}})$	**Densité**	**Entropie**
$\mathbf{C}_G$	4	0.1998	0.9365	4.9467
$\mathbf{C}^*_{\text{Naïve}}$	9	0.4232	0.4444	0
$\mathbf{C}^*_{\text{Variance}}$	6	0.3229	0.6508	2.6593

TABLE 1.4 – Summary of results of the algorithm for the example social network

Lorsque le critère de variance est utilisé, la densité de la partition est supérieure à celle du cas naïf, ce qui fait que les nœuds forment des communautés topologiquement mieux formées. Les groupes sont également plus similaires que dans la partition structurelle pure (comme attendu), ce qui permet d'obtenir l'effet souhaité.

Les partitions unifiées présentent ainsi des propriétés intéressantes, telles que des groupes d'acteurs cohérents et homogènes. La méthode offre un contrôle précis

du processus de combinaison, donnant de nouvelles possibilités d'exploration aux analystes sans avoir à recalculer les partitions orignales.

1.2.4 Intégration des variables dans des réseaux sociaux augmentés : fusion simultanée des variables

La troisième approche pour intégrer structure et attributs est de fusionner ces deux composantes dans un processus unifié. Contrairement aux deux catégories précédentes qui utilisent des algotithmes classiques de détection de communautés en fin ou au début du processus, la fusion simultanée, au contraire, nécessite des mécanismes bien souvent différents.

Dans sa revue de littérature, Petr Chunaev recense une bonne trentaine de méthodes dans cette catégorie (Chunaev, 2020). Nous trouvons des méthodes basées sur :

— la factorisation matricielle non négative et un processus d'optimisation où l'objectif est de trouver une matrice d'appartenance à des communautés en minimisant une fonction qui contraint les nœuds similaires et fortement connectés à converger vers les mêmes communautés.
— les modèles probabilistes génératifs, par exemple des modèles génératifs bayésiens ou modèles de bloc stochastique : il s'agit de déduire statistiquement un modèle du réseau attribué en supposant que la topologie et la sémantique sont générées en fonction de certaines distributions paramétriques.
— la modification des fonctions objectif d'algorithmes connus de détection de communautés (par exemple Louvain, Normalised Cut) ou de clustering tels k-means, k-medoids ou kNN.
— des méta-heuristiques et des algorithmes génétiques cherchant l'optimisation multi-objectif (e.g. modularité et similartité d'attributs).

Pour notre part, nous avons adapté la méthode de Louvain (Blondel et al., 2008) pour optimiser, à chaque itération de l'algorithme, non seulement la modularité locale, mais également l'entropie, garantissant ainsi un compromis entre la qualité structurelle de la partition et l'homogénéité des attributs au sein des communautés agrégées au terme de chaque étape (Cruz Gomez, Bothorel et Poulet, 2011a).

Algorithm 1 Augmented Graph Clustering Algorithm

Require: ε, i_{max}, PoV_F
1: $i \leftarrow 0$
2: $Q_{actual} \leftarrow$modularityCalc()
3: $Q_i \leftarrow$modularityOptimizationStep()
4: $\mathcal{C}_{\mathcal{H}} \leftarrow$entropyOptimization ($\mathcal{C}_i$)
5: $\mathcal{C}_0 \leftarrow \mathcal{C}_{\mathcal{H}}$
6: **while** $Q_{actual} - Q_i > \varepsilon$ **and** $i < imax$ **do**
7: communityAggregation($\mathcal{C}_i$)
8: $Q_{actual} \leftarrow Q_i$
9: $i \leftarrow i + 1$
10: $\mathcal{C}_i \leftarrow$modularityOptimizationStep()
11: $Q_i \leftarrow$modularityCalc()
12: **end while**
13: **return** $\mathcal{C}$

Le processus s'arrête, comme la méthode originale, lorsque la modularité ne peut être augmentée. Nous avons introduit un nombre d'itérations maximum i_{max} dans

les cas où la modularité ne convergerait pas. En effet, lors d'une itération (lignes 6-12 dans l'algorithme 1), l'optimisation de l'entropie se fait après l'optimisation de modularité, dégradant ainsi plus ou moins la qualité de la partition courante. Si les données ne répondent pas au principe « *Qui se ressemblent s'assemblent* », un compromis en similarité d'attributs et modularité peut ne pas être trouvé. Mais cela signifie que ni notre méthode, ni aucune autre technique, ne pourra trouver de communautés à la fois homogènes et denses !

1.3 Visualisation de graphes de communautés

Un graphe de communautés est un type de graphe hiérarchique dont la distance entre deux nœuds de l'arbre d'inclusion est au maximum égale à un (Eades et Feng, 1997).

La Figure 1.3a présente un exemple de graphe de communautés et la Figure 1.3b présente l'arbre d'inclusion de la partition. Cet arbre est utilisé pour visualiser la structure hiérarchique des différents niveaux qui représentent les groupes.

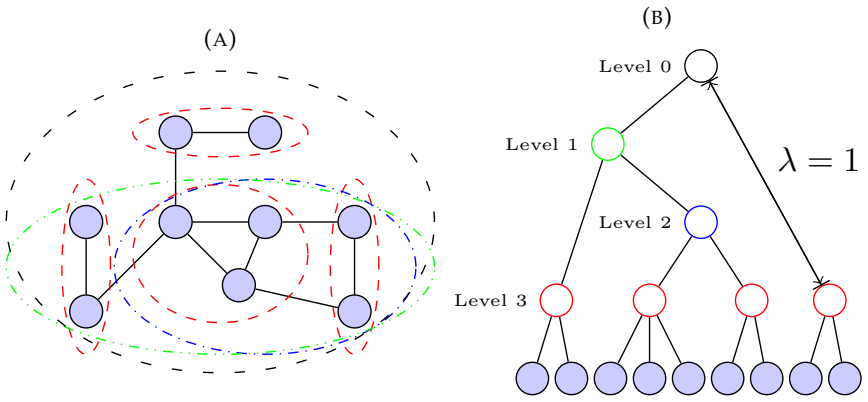


FIGURE 1.3 – Exemple d'un graphe de communautés et son arbre d'inclusion associé

En général les méthodes de visualisation de graphes de communautés ont été conçues pour mettre en évidence les différences entre chaque groupe, ce qui implique de présenter chaque groupe éloigné des autres (par exemple (Tamassia, 1987 ; Bourqui, Auber et Mary, 2007 ; Giacomo et al., 2007)). A contrario (Santamaría et Therón, 2008) proposent une méthode pour dessiner des communautés non disjointes. Mais ces algorithmes n'utilisent que la structure du graphe pour placer les nœuds et il n'est pas facile de leur ajouter d'autres critères, comme la similarité entre nœuds, pour trouver la position de chacun des nœuds.

Nous avons proposé à travers une dizaine de publications (dont les principales sont (Cruz Gomez, Bothorel et Poulet, 2011b ; Cruz Gomez, Bothorel et Poulet, 2012 ; Cruz Gomez, Bothorel et Poulet, 2013a ; Cruz Gomez, Bothorel et Poulet, 2014 ; Cruz Gomez, Bothorel et Poulet, 2013c)) :

— Un modèle pour visualiser les communautés d'une partition décrite par la composante d'affiliation $\mathcal{A}$ dérivée, soit de la composante structurelle, soit de la composante de composition, soit de notre processus d'intégration de ces deux composantes ;

— Ce modèle de visualisation, en plus d'exhiber classiquement les groupes, préserve la similarité des nœuds, quelle soit purement structurelle ou issue de l'intégration des composantes ;

— Ce modèle de visualisation propose de manière originale de focaliser l'analyse sur les interactions entre communautés ;
— De part un calcul de rôles d'interactions, le modèle facilite l'analyse des interactions entre communautés.

1.3.1 Modèle de visualisation de communautés

Le modèle présenté est conçu pour mettre en évidence les interactions entre les communautés et pour révéler des rôles importants de nœuds impliqués dans ces interactions. Ces rôles sont définis selon la partition et la structure locale de chaque nœud et le modèle visuel est utilisé pour aider à leur identification.

La première étape consiste à trouver les nœuds en charge de la connexion des groupes, c'est-à-dire, les nœuds qui sont connectés avec des nœuds d'autres groupes ; ces nœuds peuvent être vus comme des ponts entre les communautés. Pour ce faire l'ensemble des nœuds est divisé en deux catégories, comme le montre la Figure 1.4 : les nœuds dont les liens commencent et finissent dans leur propre communauté, dits *nœuds intérieurs*, et les nœuds avec au moins un lien qui commence ou finit dans une autre communauté, appelés *nœuds frontières*.

Les nœuds seront placés selon une mesure de similarité avec laquelle deux nœuds avec des voisinages similaires seront proches l'un de l'autre. Cette mesure de similarité utilise la proportion de voisins commun entre les deux nœuds, i.e. ces nœuds sont similaires si ses voisins sont similaires.

En utilisant une partition de k communautés, la division des nœuds produit $k+1$ sous-ensembles : un avec les nœuds frontières provenant de toutes les communautés et k avec les nœuds intérieurs provenant de chaque communauté. L'algorithme calcule une matrice de similarité pour chacun des $k+1$ sous-ensembles et à partir de chacune de ces matrices la position de chaque nœud est établie. Les Figure 1.4a et 1.4b présentent des exemples de chaque type de nœud dans un réseau avec deux groupes.

(A) Exemple de nœuds frontières (B) Exemple de nœuds intérieurs

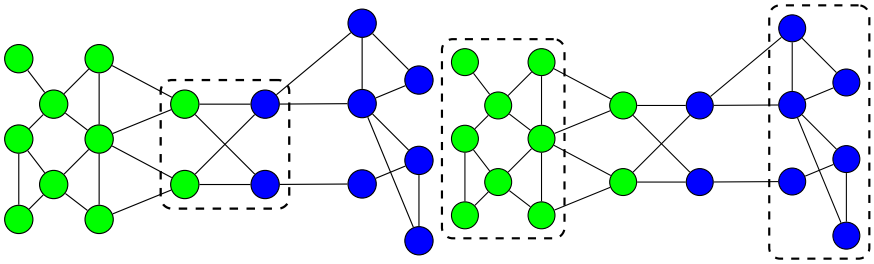


FIGURE 1.4 – Exemple de la localisation de chacun des types des nœuds dans un réseau social. Chacune des couleurs représente un groupe de la partition

1.3.2 Algorithme de placement des nœuds

L'objectif de l'algorithme est de placer les nœuds de façon telle que leur proximité indique la similarité existant entre eux. L'algorithme de tracé (*layout* en anglais) proposé est divisé en deux étapes : d'abord, placer les nœuds frontières, puis, dans un deuxième temps, placer les nœuds intérieurs (Cruz Gomez, Bothorel et Poulet,

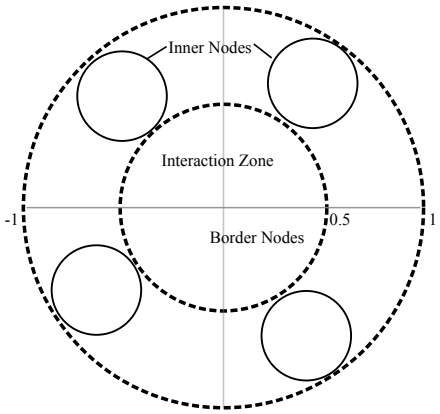
2013c ; Cruz Gomez, Bothorel et Poulet, 2013a). Ainsi l'analyste pourra axer son exploration du graphe sur les nœuds intervenant dans les relations inter-communautés d'une part, et en se focalisant sur une zone, frontière ou intérieure, aura la garantie que les nœuds proches le sont du fait d'une similarité combinant attributs et relations sociales.

Pour calculer le placement des nœuds frontière dans une zone d'interaction, nous utilisons la technique du *multi-dimensional scaling* (MDS) dédiée à la représentation visuelle des objets en fonction de leur similarité ou de leur dissemblance, et son implémentation, l'algorithme SMACOF (Scaling by MAjorizing a COmplicated Function) (Ingram, Munzner et Olano, 2009).

Les nœuds frontières sont placés dans un cercle. Comme la dissemblance entre nœuds est calculée selon leur voisinage, les positions proches permettent de voir leur proximité structurelle dans le graphe G. Ainsi des nœuds proches dans cette zone auront des rôles similaires de médiation par exemple.

Les nœuds intérieurs quant à eux sont regroupés par communauté. Chacun de ces sous-ensembles doit être placé près de leurs nœuds frontières déjà placés, qui appartiennent à la même communauté. Pour ce faire, nous définissons le centre de chaque communauté, et de la même façon que précédemment, utilisons SMACOF pour que les nœuds intérieurs de chaque communauté sont placés selon leur ressemblance de voisinage et face aux nœuds frontières de leur communauté.

(A) Localisation des éléments du modèle de dessin



(B) Localisation des nœuds intérieurs selon la position des nœuds frontières

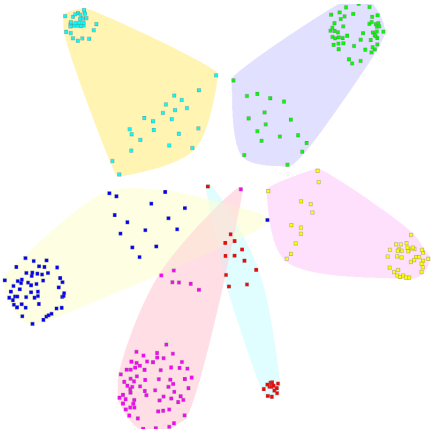


FIGURE 1.5 – Localisation des éléments et exemple du modèle de visualisation

La Figure 1.5a présente la disposition visuelle des différents éléments du modèle et la Figure 1.5b montre un exemple de la localisation finale des nœuds avec l'algorithme décrit.

La complexité de l'algorithme complet est en $O\left(2 \cdot n^2\right)$, où n est le nombre d'éléments de l'ensemble. Nous avons réalisé quelques expérience de passage à l'échelle mais également, nous avons voulu tester (visuellement) sa capacité à séparer des nœuds impliqués dans les interactions de ceux qui n'y participent pas.

Le Tableau 1.5 présente les graphes utilisés pour l'évaluation de l'algorithme de layout. La dernière colonne du tableau est le temps d'exécution de l'algorithme pour placer les nœuds de chaque réseau.

La Figure 1.6a présente le résultat du layout du réseau d'interactions en utilisant l'algorithme Fruchterman & Reingold. Dans ce cas les nœuds sont placées avec

Graphe	Nœuds	Arêtes	Nb. Comm	Q	Temps(s)
Réseau d'interaction de protéines	19928	82406	56	0,6493	1021
Réseau DBLP	10000	65734	843	0,8324	346
Réseau Twitter	5389	46440	12	0,5728	89
Réseau Facebook	334	5394	6	0,7728	36

TABLE 1.5 – Description des jeux de donnés utilisés pour tester l'algorithme de layout

(A) Layout du réseau d'interactions avec Fruchterman & Reingold (B) Layout du réseau d'interactions avec notre algorithme

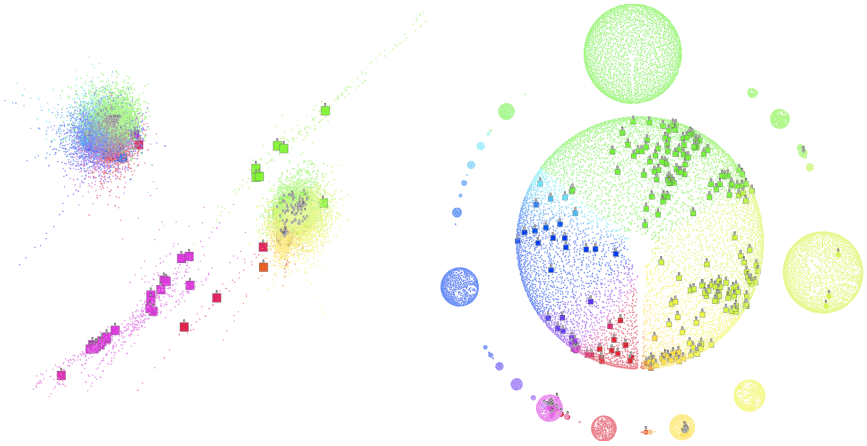


FIGURE 1.6 – Comparaison des résultats du layout pour le réseau d'interaction de protéines

une approche basée sur les forces. L'information communautaire n'est pas prise en compte, du moins pas de manière explicite. L'idée est en effet de minimiser les croisements d'arêtes tout en occuppant au maximum l'espace disponible. Cela pour conséquence de répartir au mieux les zones denses dans l'espace 2D. Lorsque le graphe est par nature constitué de zones denses bien nettes, cela a pour effet indirect de bien mettre en évidence les communautés. Mais dans la plupart des graphes réels, les nœuds frontières sont nombreux et ne permettent pas à l'algorithme basé sur les forces de rendre les groupes bien séparés. Cela fait que les communautés sont « mélangées », et que le graphe entier a tendance à être dessiné dans une zone d'interaction géante qui occuppe une grande partie de l'espace disponible. Il est difficile d'identifier tous les nœuds avec un rôle spécifique (et de manière générale la conformation en communautés). La Figure 1.6b présente le layout du même réseau avec notre algorithme, qui lui, a pour objectif de dessiner de manière nette les communautés. Nous pouvons y voir facilement les différences de taille entre groupes. Les nœuds frontières impliqués dans les interactions sont également clairement identifiables, et via la zone d'interaction, nous pouvons aussi avoir une idée de la part de chaque communautés dans les échanges avec l'extérieur. La taille des noeuds (taille du carré) reflète leur degré (ce n'est pas la même échelle d'une représentation à l'autre) et la couleur la communauté d'appartenance.

La Figure 1.7a présente le résultat du layout pour le réseau Twitter avec l'algorithme de Fruchterman & Reingold. Ce dessin montre les nœuds distribués partout

(B) Layout du réseau Twitter avec notre algorithme

(A) Layout du réseau Twitter avec Fruchterman & Reingold

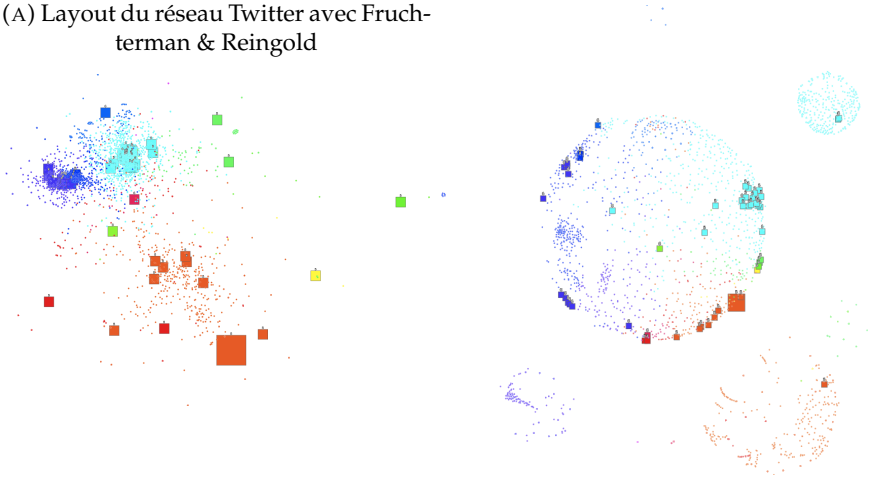


FIGURE 1.7 – Comparaison des résultats du layout pour le réseau Twitter

dans l'espace ; le nœud le plus gros (rouge) a un nombre élevé de connexions, toutefois la localisation ne donne pas beaucoup d'information en ce qui concerne ses liens : sont-ils externes ou internes à sa communauté d'appartenance ? Ce même nœud dans la Figure 1.7b, est placé dans la zone d'interaction, ce qui nous renseigne sur le fait qu'il s'agit d'un nœud frontière. Cette visualisation nous permet d'aller plus loin dans l'analyse : ce nœud est placé à côté d'autres nœuds rouges à fort degré, ce qui rend la communauté rouge moins vulnérable s'il venait à disparaître : ses voisins pourraient assurer la médiation entre la communauté rouge et les autres ommunautés.

1.3.3 Rôles dans les réseaux de communautés

Un aspect important de l'analyse des réseaux sociaux est la mesure de l'importance de certains acteurs du réseau. Cette importance peut être la capacité de certains nœuds pour déconnecter deux ou plusieurs groupes dans le réseau, ou pour concentrer le flux de messages entre différentes équipes de travail dans une entreprise. Cet aspect a été étudié dans des espaces académiques autant qu'industriels comme présenté par (Aldrich et Herker, 1977), par (Guimera et Amaral, 2005) et par (Cross et Parker, 2004). Le Tableau 1.6 présente un résumé comparatif des rôles trouvés dans les réseaux de communautés.

La dernière colonne spécifie le type de nœud qui peut être trouvé dans le rôle (I pour des nœuds intérieurs, F pour des nœuds frontières). Les rôles variés de Guimerà et Amaral offrent un large potentiel pour différencier les nœuds (Cruz Gomez, Bothorel et Poulet, 2013c ; Cruz Gomez, Bothorel et Poulet, 2013a), mais parfois, la typologie proposée par Cross et Parker est suffisante et plus facile à intuiter.

Avant de détailler un cas d'application explicitant l'utilisation de nos méthodes dans la section suivante, nous avons mené des expériences qualitatives pour voir l'influence des points de vue sur les partitions et les rôles détectés.

Aldrich and Herker	Cross and Parker	Guimerà and Amaral	Type
N/D	Personnes périphériques	Nœuds ultra périphériques	I
		Nœuds périphériques	F
	N/D	Nœuds connecteurs non centraux	F
	N/D	Nœuds sans proches non centraux	F
Nœuds d'ouverture types I et II	Nœuds centraux	Nœuds centraux provinciaux	I, F
	Nœuds d'ouverture	Nœuds connecteurs centraux	F
	Nœuds d'intermédiation	Nœuds sans proches centraux	F

TABLE 1.6 – Résumé comparatif des définitions des rôles dans les réseaux de communautés et leur utilisation potentielle pour nos deux types de nœuds (I pour les nœuds intérieurs, F pour les nœuds frontières)

1.3.4 Illustration de la méthode de visualisation pour comparer deux points de vue

Nous utilisons ici le réseau Facebook déjà utilisé précédemment (Table 1.5) et les rôles de Cross et Parker. Pour cette expérimentation (Cruz Gomez, Bothorel et Poulet, 2012), le jeu de donnés représente un réseau social personnel récolté avec NameGenWeb (Hogan, 2011) et inclut des personnes proches de son propriétaire : famille, amis, collègues et anciens collègues. Au total le réseau a 334 nœuds et 5 394 arêtes. Avec ce graphe nous avons défini deux points de vue. Un sans attributs, PoV_{NULL}, qui reflète une analyse classique de réseau social, et un autre point de vue, PoV_{COMP}, qui décrit les compétences de chaque acteur. Les compétences sont au nombre de 7 : 1. Math & Science, 2. Business Administration, 3. Law, 4. Social Sciences, 5. Software engineering, 6. Other fields, 7. Arts.

PoV	Groupes	Modularité	Distance (±écart type)
NULL	6	0,7728	Par rapport à PoV_{COMP} : 0,4689 (±0,0201)
COMP	10	0,8138	0,2820 (±0,0431)

TABLE 1.7 – Partitions issues des 2 points de vue NULL et COMP. La modularité mesure la qualité de la partition en termes de connectivité. La distance, calculée sur la composante compositionnelle, est la distance moyenne, normalisée, entre les points de chaque groupe dans la carte de Kohonen, cf. Section 1.2.2.

Concernant le point de vue PoV_{COMP}, la valeur de modularité est supérieure à celle du PoV_{NULL}. La qualité de la partition au sens strictement structurel s'est détériorée lorsqu'on a modifié les poids des arêtes pour intégrer la similarité de compétences entre nœuds. Inversement, la distance entre les profils de compétences intra communautaires a baissé, proposant des groupes un peu moins « nets » en termes de modularité, mais plus homogènes concernant les attributs, ce qui est confirmé par la distribution des compétences par communauté (cf Figure 1.8).

Dans (Cruz Gomez, Bothorel et Poulet, 2012), nous explorons les différents types de nœuds classés selon les rôles de Cross et Parker. Si l'on fait simplement ici un focus sur les nœuds dits d'intermédiation, nous pouvons voir sur la Figure 1.9 que le

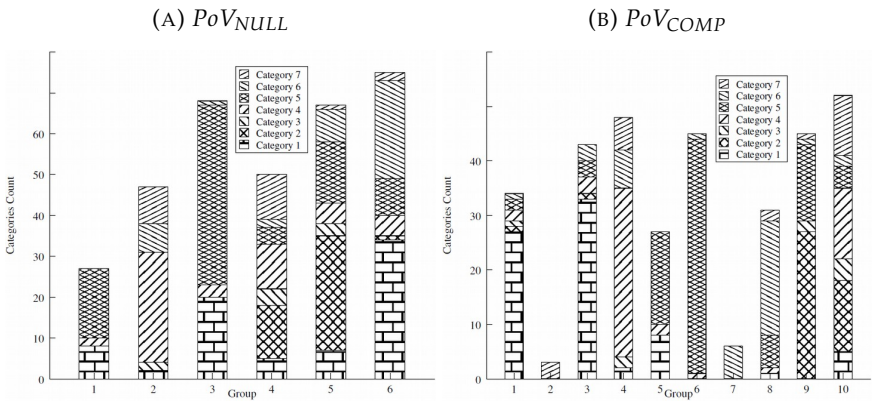


FIGURE 1.8 – Effet des points de vue sur la distribution des catégories (ici compétences) dans les communautés calculées sur le réseau Facebook

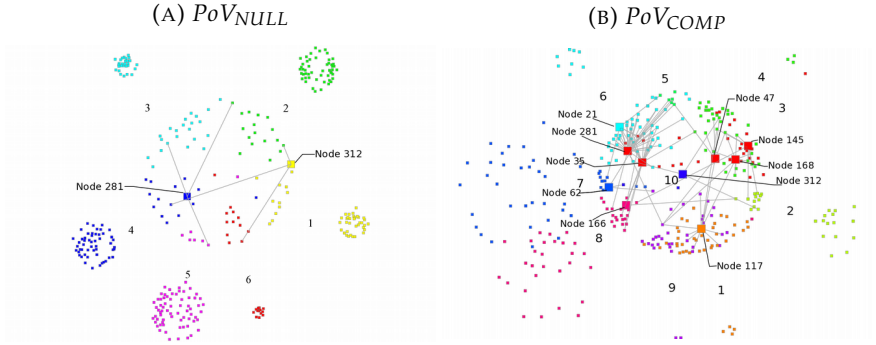


FIGURE 1.9 – Comparaison des points de vue du layout de notre algorithme pour le réseau Facebook

point de vue modifie la classification des nœuds clés. Les nœuds possédant ce rôle connectent différentes communautés dans la zone d'interaction, c'est-à-dire qu'ils établissent un pont entre deux communautés ou plus. La partition NULL contient uniquement deux nœuds d'intermédiation, qui connectent les différents groupes du graphe, tandis que dans la partition COMP, il existe 10 nœuds de ce type dont deux sont les mêmes que dans la partition NULL et 8 sont nouveaux. Croiser relations sociales et compétences permet ainsi de détecter des intermédiaires inter-communautés de manière plus fine, et contextuelle au partage d'expertise.

1.4 Application à l'étude de la blogosphère de la cuisine

Pour ce cas d'application, nous utilisons à nouveau la méthode de détection de communautés avec pré-traitement (Section 1.2.2). L'objectif est de cartographier la blogosphère Française de l'univers de la cuisine. Nous analysons d'abord les blogs, (i) en considérant leur caractéristiques en nombre de publications, les partages vers les plateformes de réseaux sociaux et leurs thématiques, et (ii) la structure du réseau lui-même, les nœuds étant ici les blogs, et les arêtes, les liens inter-blogs. Ensuite, nous recherchons une partition composite pour regrouper les blogs en communautés densément connectées et homogènes du point de vue de leurs caractéristiques. Enfin, nous visualisons la partition pour mettre en évidence les interactions entre les

communautés, et nous identifions les blogs ayant des rôles et des positions clés dans l'écosystème de la cuisine virtuelle.

Ce travail, réalisé en 2014 dans le cadre de la Chaire Réseaux Sociaux de l'IMT [1], a un objectif marketing. L'idée est de proposer une nouvelle segmentation des communautés virtuelles de l'univers culinaire ouvrant ainsi des perspectives en termes d'outils de gestion de la relation client (ciblage, ou gestion de communautés par exemple) et plus généralement de contribuer à de nouvelles méthodes d'études marketing à partir des données de réseaux sociaux.

1.4.1 Jeu de données

Nous avons travaillé sur le dataset Cuisineblog [2] qui rassemble un ensemble de blogs relatifs à des recettes de cuisine, mais aussi d'autres sites tels que des pages Facebook ou des sites dédiés à l'achat d'ingrédients. Le graphe, notre composante structurelle, est constitué des liens hypertextes entre sites, et le poids des arêtes représente le nombre d'interactions : les différents posts en effet se mentionnent les uns les autres très régulièrement. Les caractéristiques des sites, notre composante de composition, concernent 5 variables : le nombre de likes Facebooks et le nombre d'événements mentionnant le blog sur Twitter, reflétant l'activité sociale autour des blogs ; le type du site (nos Blogs, mais aussi des sites de ressources, ou encore des sites « hors base » que la collecte nous a amenés car pointés par les blogs) ; le nombre de posts reflétant l'activité éditoriale ; et enfin la thématique du blog. La thématique est un vecteur de dimension 7, il est défini par le nombre de posts classifiés dans chaque thème parmi :

1. Corps gras alimentaire
2. Céréale et produit céréalier
3. Sucre et produit sucré
4. Produit de pêche
5. Viande
6. Produit laitier
7. Légume et fruit

	Catégories						
Site web	**1**	**2**	**3**	**4**	**5**	**6**	**7**
http ://academiedesvinsanciens.org/	31	1	42	100	103	61	124
http ://cookiesdelice.canalblog.com/	95	48	70	63	60	60	125

TABLE 1.8 – Exemple de distribution des catégories pour deux sites web

La Table 1.8 présente deux exemples de sites web et leurs catégories. A noter qu'un message peut contribuer à différentes catégories. Le premier site est principalement consacré à l'appréciation et à l'évaluation des vins. Cela se reflète dans le fait que les catégories Fruits et légumes, Poisson et Viande (7, 4 et 5) ont des valeurs les plus élevées. Le deuxième site web est principalement consacré aux desserts, la pâtisserie et la boulangerie avec des préoccupations liées aux Corps gras alimentaires

1. https://chairereseaux.wp.imt.fr/
2. Dataset issu du projet Open Food System, financé par l'état français dans le cadre du programme des investissements du futur, www.openfoodsystem.fr

et aux Fruits et légumes (1 et 7). Cependant, d'autres types de recettes peuvent être trouvées, c'est pourquoi toutes les catégories sont représentées.

Si l'on considère tous les types de sites, nous obtenons un graphe de 11390 nœuds et 45594 arêtes. Après une analyse préliminaire des attributs des nœuds, nous avons constaté qu'environ 93% des sites web sont extérieurs à la blogosphère culinaire (type HORS BASE) et sont rarement liés à l'alimentation ou à la cuisine. Cela rend de ce fait les catégories caduques et nous a conduit à ne garder que les sites de types BLOG que nous avions initialement ciblés. Nous obtenons un graphe composé uniquement de 661 nœuds et 6844 arêtes.

1.4.2 Détection de communautés dans le graphe avec attributs

Nous avons mis en œuvre la méthode avec pré-traitement décrite dans la Section 1.2.2 dans le cadre de cette étude applicative. En raison de la petite taille de notre dataset et la faible dimensionalité des attributs, nous avons utlisé l'algorithme k-means pour produire les clusters compositionnels de nos blogs. L'analyse de la silhouette nous a conduit à rechercher 9 clusters. Parmi ceux-ci, 2 groupes concentrent 53% et 23% des blogs : le premier (G5) est dédié aux produits céréaliers et sucrés et recueillent beaucoup de likes sur Facebook (présence très faible sur Twitter) ; le deuxième groupe (G7) concerne les mêmes thématiques, avec un focus non négligeable sur les matières grasses également, mais, à l'inverse, ne suscite que très peu d'activités sur les réseaux sociaux. Les 7 autres groupes sont de taille réduite et ne contiennent qu'entre 0.15 et 12% des blogs.

Après modification des poids des arêtes, l'algorithme de Louvain détecte 5 communautés. S'il on regarde la distribution des groupes sur les communautés, Table 1.9, nous remarquons qu'ils sont pour la plupart répartis sur plusieurs communautés, sauf les plus petits groupes (G4, G6 et G9). Et inversement, les communautés hébergent une grande variété de blogs. Seule la communauté 3 se distingue car elle n'a regroupé que des blogs liés au produits céréaliers, au sucre et aux matières grasses, mais à l'activité sociale variable. On obtient globalement des communautés pour lesquelles la motivation des membres n'est pas la spécialisation thématique. Les blogs font preuve d'ouverture thématique dans leurs référencements mutuels.

Si l'on regarde la modularité de la partition (Table 1.10), nous constatons que celle-ci est vraiment très faible, indiquant un découpage structurel peu efficace. La densité confirme cela : 2/3 des arêtes sont internes aux communautés, ce qui signifie qu'1/3 d'entre elles relient les communautés. La Figure 1.10a nous confirme visuellement que la zone d'interaction est très conséquente. Les communautés ne sont donc pas nettes d'un point de vue topologique non plus. D'ailleurs, pour mieux comprendre à quel point notre combinaison de variables dégrade la modularité, nous avons réalisé la détection de communautés sur le graphe d'origine. La partition obtenue avant modification des arêtes, sur le graphe d'origine, présente elle-même une modularité extrêmement faible. Nous avons en parallèle observé que distribution des degrés des nœuds n'est pas en loi de puissance : notre réseau n'est pas de type *scale-free*, ce qui peut expliquer la nature non communautaire de notre graphe.

1.4.3 Détection de rôles

Le sentiment d'appartenance communautaire est l'un des moteurs qui incite les internautes à partager sur les réseaux sociaux (Beaudouin, Legon et Pasquier, 2016 ; Helme-Guizon, Magnoni et al., 2013 ; Richardson, 2015). Le graphe des blogs étudié ici, représente en lui-même une communauté à part entière de blogueurs culinaires.

TABLE 1.9 – Distribution des groupes sur les communautés

	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**
Comm. 1	1.6	1.2	12.0	0.2	54.2	-	25.7	4.9	0.2
Comm. 2	2.0	1.0	12.1	-	67.7	-	16.2	1.0	-
Comm. 3	-	-	50.0	-	-	-	50.0	-	-
Comm. 4	1.1	1.1	7.5	-	63.4	1.1	25.8	-	-
Comm. 5	-	4.0	20.0	-	56.0	-	12.0	8.0	-

(A) Communautés

(B) Provincial et kinless hubs

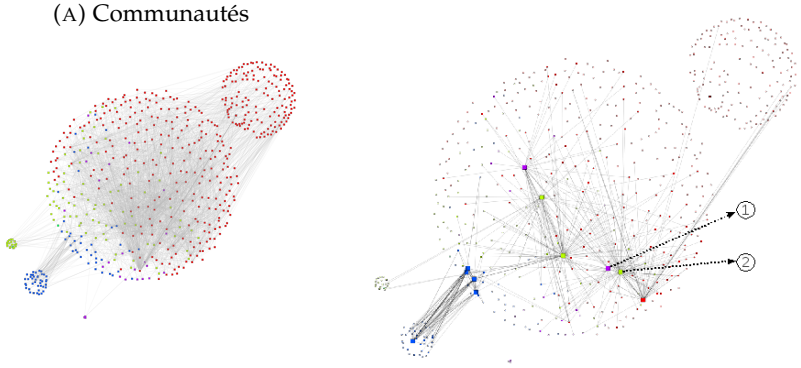


FIGURE 1.10 – Visualisation du réseau de blogs culinaires

Nous venons de voir que les blogueurs n'avaient que peu tendance à s'enfermer dans une thématique, et au contraire, à se montrer ouverts à l'ensemble des domaines liés à la cuisine. En se référençant mutuellement, ils démontrent à la fois de cette ouverture thématique mais également d'un sentiment d'appartenir à la même communauté.

D'un point vue marketing, pour trouver les influenceurs culinaires, nous pourrions en tant qu'analyste de réseaux, trouver les nœuds présentant les centralités les plus fortes. Pour trouver les experts sur une thématique, nous pourrions caractériser ces nœuds par des mots-clés, i.e. nos catégories. Comment faire la sélection de ces nœuds ? Devons-nous faire un classement des plus centraux puis filtrer par thème ? Quitte à prendre de nœuds peu centraux au final mais « experts » ? Devons-nous plutôt relâcher la contrainte de l'expertise et privilégier les plus centraux ? Si l'on ajoute une autre information, qui est la popularité du nœud, i.e. l'activité sociale que ses posts engendrent sur Facebook et Twitter, cela complexifie d'avantage le choix de la stratégie.

Avec nos travaux, l'intégration des composantes structurelles et topologiques nous permet d'obtenir un graphe tel que les individus centraux le seront dans un modèle où les distances reflètent, via un compromis, l'ensemble des dimensions considérées, i.e. la topologie, les thèmes et la popularité. Il est alors possible de rechercher « facilement » les nœuds qui influencent les plus grandes communautés sensibles à telles ou telles thématiques et dont les membres suscitent le plus d'activité sur les réseaux sociaux visés par la campagne marketing que nous cherchons à mettre en place.

Au-delà des mesures de centralités classiques (intermédiarité, degré, PageRank,

TABLE 1.10 – Results of the community detection algorithm for the blogs social network

	Communautés	Q	**Densité**	**Entropie**
Partition structurelle	8	0.3718	0.5662	9.4883×10^{-3}
Partition combinée	5	0.1021	0.6619	5.8121×10^{-3}

proximité, valeurs propres), nous avons voulu tester la détection de rôles telle que proposée par Guimera et Amaral (Guimera et Amaral, 2005) sur notre graphe de communautés socio-thématico-populaires.

Nous n'allons pas lister ici toutes les caractéristiques de toutes les communautés ni tous les rôles détectés. Parmi les rôles importants pour le marketing, nous pouvons citer les *provincial hubs* qui sont représentatifs de leur communauté avec un degré interne important, mais ils ont aussi un index de participation superieur à 0, signifiant qu'ils touchent potentiellement d'autres communautés. Ils appartiennent en majorité à la communauté 4 qui contient 14% des nœuds. Ils ne sont pas très spécialisés ni très actifs sur les réseaux sociaux, mais sont très ancrés dans leur communauté. Ils agissent en tant que hubs et peuvent ainsi potentiellement toucher beaucoup d'autre blogs sur d'autres thématiques. Les *kinless hubs* sont, quant à eux, moins implantés dans leur propre communauté (degré interne faible) mais participent d'avantage aux relations inter-communutés : plus de la moitié de leurs relations sont externes. Ils sont encore moins spécialisés, mais propablement plus populaires au sein de la sphère culinaire car encore plus fortement connectés à des communautés de blogs très variés en terme de thématiques. Certains d'entre eux sont également très relayés sur les réseaux sociaux, tel que le nœud 1 (laviede- choupette.calnalblog.com) représenté sur la Figure 1.10b. Le nœud 2 (christelle56.overblog.com) est aussi un kinless hub. Il placé près du nœud 1 dans notre visualisation, cela signifie qu'ils partagent des voisins communs, bien qu'ils appartiennent à des communautés différentes et aient un profil d'activité sociale et thématique également différent. Solliciter le nœud 1 peut permettre de toucher à la fois les acteurs de référence, aux expertises variées, de la blogosphère de cuisine, via d'autres hubs, mais cela permet également de générer des partages sur Facebook et Twitter, touchant ainsi les amateurs et le grand public.

1.5 Conclusion

Les travaux présentés dans ce chapitre proposent un cadre et des méthodes d'analyse de communautés dans des réseaux sociaux augmentés (avec attributs sur les nœuds), exploitant de manière combinée l'information provenant à la fois des composantes structurelle et de composition présentes dans un réseau social.

Les contributions réalisées dans ce travail interviennent selon 5 axes principaux :

— Une méthode centrée décideur qui définit, via la notion de *point de vue*, le sous-ensemble des variables décrivant les acteurs, en lien avec son étude ou analyse ;

— Trois variantes d'algorithmes de détection de communautés qui permettent d'intégrer les composantes structurelle et de composition pour trouver des communautés de nœuds à la fois fortement connectés et similaires du point de vue de leurs attributs ;

— Un algorithme générique de visualisation de communautés qui exploite la variable d'affiliation : celle-ci peut être fournie a priori dans les données, ou issue de mécanismes de détection de communautés ;
— La visualisation proposée est centrée sur l'interaction entre les communautés, produisant ce que nous avons appelé *la zone d'interaction* d'une part, et les regroupements de nœuds intérieurs en périphérie du modèle de dessin ;
— Le plongement des nœuds dans l'espace 2D respectant la similarité de leur voisinage, permet, notamment, via la détection de rôles, d'identifier la localisation des individus clés et de mieux comprendre lesquels participent aux interactions entre groupes, en lien avec leur profil si la variable d'affiliation provient de nos approches intégrant les composantes structurelle et compositionnelle des réseaux.

Dans ce travail, le point de vue est défini a priori, par exemple par un décideur qui souhaiterait explorer l'impact de tels ou tels attributs sur la configuration des communautés sociales en fonction de la question métier qu'il cherche à résoudre. Mais le point de vue pourrait également être généré de manière automatique, en ne gardant que les attributs maximisant des critères de qualité qu'il faudrait alors définir. Une fouille de données en amont, via une ACP par exemple, permettrait de chercher les attributs les plus discriminants par exemple. Concernant les techniques de détection de communautés, la tendance actuelle est de se tourner vers les techniques classiques de Machine Learning qui permettent un passage à l'échelle d'une part, et une généralisation des modèles d'autre part, permettant ainsi de prendre en compte la dynamique des réseaux et en particulier la classification de nouveaux nœuds, voire de nouveaux graphes à partir de modèles pré-entraînés (Transfer Learning) (Chunaev, 2020).

Ainsi l'approche maintenant populaire est de construire une représentation latente du graphe sous forme de matrice de similarité entre nœuds puis d'y appliquer une méthode de classification telle que les *k*-means, les cartes auto-organisatrices ou encore le clustering spectral. En général, le clustering spectral appliqué sur cette matrice de similarité est la méthode la plus répandue et celle qui obtient les meilleures performances.

Tout le challenge réside dans la constitution de cette matrice de similarité. Certains travaux utilisent des métriques *ad hoc* pour calculer la distance entre chaque paire de nœuds. Cette distance est souvent une combinaison linéaire entre une distance structurelle (longueur du plus court chemin par exemple) et une distance entre les vecteurs d'attributs (distance de Jaccard ou euclidienne) (Combe et al., 2012 ; Olteanu, Villa-Vialaneix et Cierco-Ayrolles, 2013 ; Falih et al., 2018). Cependant ces techniques ne reflètent pas nécessairement fidèlement les relations complexes du graphe d'origine et conduisent à un clustering sous-optimal (Shi et al., 2019).

C'est pourquoi la tendance actuelle est d'apprendre la similarité entre nœuds à partir des données de telle sorte à refléter la structure complexe du graphe d'origine (revue de l'état de l'art sur les embeddings de nœuds (Cui et al., 2019)).

Nous indiquons ici les deux types d'approches majeures, les approches »profondes » et d'autres approches dites « superficielles » (shallow) en anglais.

Méthodes profondes (Deep learning) Le deep learning ou apprentissage profond est appliqué à de nombreux domaines scientifiques (pour ne pas dire tous ;-)). Les réseaux complexes ne font pas exception. La promesse est d'adresser des gros volumes de données, et de produire des modèles prenant en compte des examples de graphes de taille et de topologie variées, et dont la nature discrète et complexe engendre des complexités algorithmiques pour des méthodes exhaustives classiques.

Les réseaux de neurones récursifs ont été introduits dans les années 90 pour traiter des structures d'arbre et ont été généralisés par les Graph Neural Networks (GNN) récurrents et les Graph Convolutional Networks (GCN) à propagation avant (feed-forward) pour manipuler tout types de graphes, avec ou sans cycles, dirigés ou non (cf. revue de l'état de l'art sur les graph neural networks (Wu et al., 2020) ou le deep learning appliqué au problème de la détection de communautés (Su et al., 2022)).

En matière de détection de communautés, les approches les plus courantes sont celles basées sur les autoencoders (GAE). L'encoder produisant une représentation latente du graphe est un GNN, de 2 ou 3 couches en général. Le principe est de voir le graphe comme une grille et de « dézoomer » à chaque couche, élargissant au fil des convolutions, le voisinage des nœuds (parallèle avec les pixels dans les images). En dézoomant, on compresse l'information (relation de voisinage mais aussi attributs) jusqu'à obtenir une représentation compacte à faible dimension synthétisant le graphe. Le GNN est couplé avec un decoder, en général un produit scalaire, pour reconstruire la matrice d'adjacence du graphe et ainsi, de façon auto-supervisée, construire le modèle neuronal (Cao et al., 2018). Différentes variantes existent : « variational graph autoencoder » (VGAE) (Kipf et Welling, 2016), « marginalized graph networks » (MGAE) (Wang et al., 2017). En particulier, avec les approches « adversarially regularized graph autoencoder » (ARGE) et « variational graph autoencoder » (ARVGE) (Pan et al., 2018), des données bruitées sont introduites pour rendre le pouvoir prédictif plus robuste. Il s'agit de générer des probabiltés de connexion entre nœuds puis dans la partie discrimative, étant données ces probabilités, retrouver les structures du départ. Ces méthodes sont souvent utilisées pour générer des graphes synthétiques plutôt réalistes, mais elles ne peuvent cependant que capturer les voisins de chaque nœud à deux ou trois sauts de distance et ne parviennent ainsi pas à bien saisir la structure globale des communautés de grands graphes. On peut également citer les architectures « attention networks » qui ont été introduites récemment (largement utilisées en traduction automatique) pour caractériser l'importance des voisins (Veličković et al., 2017).

Les méthodes citées ci-dessous n'opèrent pas le clustering, puisque l'idée est de générer des représentations latentes agnostiques de la tâche de data mining envisagée. Le clustering est alors appliqué en une deuxième phase à partir de la réprésentation latente profonde (couche cachée issue de l'encoder). La méthode « deep attentional embedded graph clustering » (DAEGC) est quant à elle dédiée au clustering, et propose de réaliser le plongement du graphe et le clustering de manière unifiée (Wang et al., 2019). Les caractéristiques des nœuds et les informations sur la structure topologique d'ordre 2 sont encodées dans la représentation latente via un graph attentional encoder. Des « soft labels » issus de l'embedding du graphe lui-même sont générés (k-means) pour auto-superviser le processus de clustering. Embedding et clustering auto-supervisé sont ensuite optimisés de manière conjointe pour intégrer structure et attributs et ainsi tirer profit de toutes les composantes pour produire des communautés pendant la phase d'entraînement du réseau de neurones.

D'une manière générale, il est bien connu que les modèles profonds demande un temps d'entraînement long, qu'ils ont un nombre important d'hyper-paramètres à ajuster et qu'ils ont donc une tendance à sur-apprendre, prenant ainsi mal en compte la dynamique des réseaux et en particulier la classification de nouveaux nœuds.

Mais il y a aussi des problèmes spécifiques aux graphes, comme le « phénomène d'over-smoothing », phénomène qui tend à faire converger les attributs qualifiant les nœuds vers un même vecteur après de multiples convolutions (les attributs sont considérés comme des signaux sur graphe et les convolutions agissent comme des

filtres passe-bas lissant les particularités). Même s'il existe des astuces pour contrecarrer ce phénomène, à l'heure actuelle, la profondeur n'apparaît pas comme un atout et bien au contraire, l'ultra-profondeur (centaine de couches) dégrade parfois sévèrement les performances obtenues comme par exemple pour la tâche de classification de noeuds (Rusch, Bronstein et Mishra, 2023).

Il reste des challenges pour bien comprendre ces phénomènes : pourquoi cela est-il efficace sur une image qui est une grid, donc un graphe particulier ? Est-ce parce que les graphes complexes sont de forme petit monde (faible diamètre) ce qui fait qu'en peu de convolutions tout le graphe est couvert ? La notion de voisinage n'est pas la même non plus. Reconnaître un objet sur une image se fait en analysant l'objet et l'ensemble de son contexte. Mais pour les réseaux complexes, comme les réseaux sociaux, un voisinage à faible distance (2 voire 3) n'a-t'il pas bien souvent plus de « sens » que de considérer l'ensemble du graphe ? Aussi, les GNN peu profonds, non seulement sur-apprennent moins mais procurent des résultats très bons avec seulement 2 couches (Zhou et al., 2020).

Méthodes non profondes (Shallow methods) L'idée est de se tourner vers des techniques dites « shallow », non ou très peu profondes, pour trouver les plongements de nœuds. On peut citer la factorisation de matrice (Yang et al., 2018 ; Shen et al., 2018) ou des méthodes basées sur les marches aléatoires (Perozzi, Al-Rfou et Skiena, 2014 ; Grover et Leskovec, 2016), ou encore l'utilisation simple de convolutions sur le graphe (une seule couche d'un GNN) combinées à du clustering spectral (Kang et al., 2022).

Toutes ces nouvelles approches basées sur l'apprentissage de représentation permettent de dépasser les limitations bien connues des algorithmes traitant directement les graphes, à savoir leur complexité, leur faible parallélisation et elles permettent d'appliquer des méthodes d'apprentissage classiques. Notons cependant que les clusters obtenus en fin de processus peuvent regrouper des nœuds initialement déconnectés dans le graphe d'origine.

Le paysage de la détection de communauté sur les réseaux avec attributs évolue très vite, et j'invite le lecteur à consulter les récentes revues de littérature. En particulier une synthèse centrée sur l'apprentissage de représentation quelles soient profondes, ou basées sur des modèles probabilistes, plus anciens comme les stochastic block models (Jin et al., 2021). Chunaev propose quant à lui une synthèse plus large, qui décrit une très grande variété de méthodes dont la plupart de celles décrites dans ce chapitre (Chunaev, 2020).

Le benchmark Open Graph Benchmark (Hu et al., 2020) va fortement contribuer à l'enrichissement des connaissances en la matière, notamment en mettant à disposition de nombreux datasets complexes qui vont permettre de mettre en évidence le phénomène de sur-apprentissage et ainsi déclasser probablement des méthodes qui, en apparence, parce qu'évaluées sur des petits jeux de données simples, ont montré de bonnes performances au moment de leur publication.

Il est à espérer que pour la tâche de détection de communautés, qui maintenant a tendance à s'appeler « clustering » — vocabulaire du machine learning — dans ce nouveau paysage, de nouveaux jeux de données soient mis à disposition (ils sont encore trop peu à l'heure actuelle proposant une vérité terrain).

Chapitre 2

Qualité des communautés

La pertinence de méthodes de détection de communautés, comme en apprentissage non supervisé plus généralement, demande une vérité terrain pour confronter la partition détectée avec une classification connue a priori. Les données réelles sont souvent accompagnées de méta-données : par exemple, pour le réseau de produits achetés ensemble sur Amazon, les méta-données sont les catégories des produits ; pour le réseau de co-publication DBLP, les méta-données sont les conférences. Classiquement, les auteurs des méthodes calculent la précision, le rappel, le F1-score ou encore des mesures de validation telles que le Rand index ou l'information mutuelle, et évaluent alors la capacité de leur méthode à retrouver la répartition des nœuds en lien avec leurs méta-données. Cependant, en cohérence avec d'autres travaux (Hric, Darst et Fortunato, 2014 ; Peel, Larremore et Clauset, 2017), nous avons démontré qu'il n'est pas souhaitable d'utiliser les méta-données comme vérité terrain (*ground-truth* en anglais). En effet il y a une différence substantielle entre les communautés structurelles et les regroupements formés à partir des métadonnées. Nous trouvons de faibles scores de similarité entre ces deux types de regroupement, avec des taux de rappel et de précision très faibles. Les communautés détectées ont en effet un faible chevauchement avec les groupes de métadonnées, et vice versa. De plus, nous avons montré que les méthodes automatiques produisent de bien meilleures partitions si l'on considère des mesures telles que la modularité, la densité, la séparabilité, etc. (Dao, Bothorel et Lenca, 2017a)

En l'absence de vérité terrain, évaluer l'efficacité des méthodes de détection de communautés en terme de précision reste une question ouverte, comme c'est le cas dans la plupart des situations du monde réel (Nerurkar, Chandane et Bhirud, 2019).

Dans ce chapitre, nous abordons l'évaluation sous un autre angle. Notre objectif n'est pas de tester si tel ou tel algorithme découvre les "bonnes" communautés. Notre but est plutôt d'accompagner les décideurs ou les analystes de données dans leur choix d'une ou plusieurs méthodes parmi celles qui leur sont proposées dans les outils tels que Gephi ou autre librairie R ou Python. En effet, pour analyser un réseau

donné, l'analyste n'a pour l'instant pas d'autres possibilités que de se documenter sur le mécanisme intrinsèque implémenté par les méthodes mises à sa disposition, ou bien de les tester. Et bien souvent, en pratique, il réutilise celle(s) qu'il connaît déjà. Lorsqu'il en actionne plusieurs, il obtient plusieurs partitions, et se trouve alors démuni pour sélectionner celle ou celles qu'il va exploiter.

Notre parti pris est donc, non pas de comprendre la philosophie des techniques employées, i.e. quelle *fonction objectif* les différentes méthodes cherchent à optimiser, mais plutôt de comprendre la *nature des partitions produites*, d'un point de vue qualitatif et de fournir des outils descriptifs de ces partitions pour que l'analyste puisse opérer une sélection en connaissance de cause.

Nous étudions ici les résultats produits par 16 algorithmes populaires [1], obtenus sur plus d'une centaine de réseaux réels.

Nous procédons à une comparaison exhaustive de ces méthodes de façon à comprendre si elles produisent des résultats équivalents. Nous comparons bien sûr leur temps d'exécution, mais surtout, nous évaluons leur similarité en terme de mesures de validation telles que l'information mutuelle, leur co-performance en terme de modularité, ainsi que leur similarité en se basant sur la taille des clusters produits.

Sur la base de cette étude systématique, par une évaluation objective sur un nombre important de graphes réels et de nature variée, nous proposons une classification des méthodes elles-mêmes, débouchant sur un guide destiné à l'analyste lui permettant de faire un choix.

2.1 Méthodologie

2.1.1 Méthodes de partionnement

Nous présentons, dans cette section, quelques méthodes populaires de détection de communautés qui ont été largement utilisées et discutées dans la littérature. Notez que ces dernières années, un grand nombre de nouvelles méthodes ont été proposées, cependant, une analyse empirique et exhaustive de toutes les méthodes serait irréalisable. Nous avons sélectionné des méthodes importantes, parmi les plus représentatives et dont la version logicielle est disponible.

Il existe de nombreuses taxonomies possibles pour les méthodes de détection de communautés. Par exemple, on pourrait les classer en fonction des fonctions objectifs qu'elles optimisent, sur la base des hypothèses concernant la structure à trouver, les mesures de qualité attendues, le modèle théorique employé, etc. Il n'existe pas de consensus sur la façon dont les différentes méthodes sont similaires ni sur la façon dont elles peuvent être classées. Porter *et al.* utilise une classification basée sur des critères hétérogènes, comme le fait que les techniques soient locales, qu'elles soient basées sur la centralité, qu'elles optimisent la *modularité* [2] ou utilisent le clustering spectral (Porter, Onnela et Mucha, 2009). Dans (Fortunato, 2010 ; Fortunato et Hric, 2016), les auteurs regroupent les méthodes de détection de communautés en méthodes traditionnelles de clustering, méthodes basées sur la modularité, algorithmes spectraux, algorithmes dynamiques et méthodes basées sur l'inférence statistique. Coscia *et al.* classent la découverte de communautés en fonction de la distance entre nœuds, de la densité interne des clusters, de la détection des "ponts" (bridges), de

1. Nous nous plaçons dans le cadre de graphes sans attributs et restreignons l'étude aux méthodes bien connues, disponibles, qui produisent des partitions, i.e des clusters sans recouvrement.

2. Pour la première fois introduite par (Newman et Girvan, 2004) pour évaluer les niveaux de regroupement hiérarchique d'un algorithme de détection de communautés, la *modularité* est devenue la fonction objectif la plus populaire dans le contexte de la détection de communautés.

processus de diffusion, de modèle structurel, de regroupement des liens et de méta clustering (Coscia, Giannotti et Pedreschi, 2011). Dans le contexte des médias sociaux, Papadopulos *et al.* comparent des méthodes de détection de sous-structures, de regroupement de sommets, d'optimisation de la qualité des communautés, d'approches par division et d'approches basées sur des modèles (Papadopoulos et al., 2011). Bohlin *et al.* regroupent différentes approches en trois classes principales représentant différents modèles de réseau : modèles nuls, modèles de blocs et modèles de flux (Bohlin et al., 2014). Schaub *et al.* classifient les méthodes en quatre perspectives : basée sur la coupe, basée sur la densité interne de regroupement, basée sur l'équivalence stochastique et basée sur la dynamique montrant quatre facettes différentes de la structure de la communauté (Schaub et al., 2017). Enfin, Ghasemian *et al.* adoptent une classification expérimentale (Ghasemian, Hosseinmardi et Clauset, 2018) : les auteurs regroupent les méthodes de détection de communautés en familles distinctes, en fonction des résultats expérimentaux obtenus sur de nombreux réseaux du monde réel en utilisant une métrique de validation (un sujet que nous aborderons dans la section 2.4.2).

En ce qui nous concerne, nous avons choisi de classer les méthodes de détection des communautés en fonction de différentes approches théoriques, notamment la suppression d'arêtes clés, l'optimisation de la modularité, le processus dynamique et l'inférence statistique, à la manière de (Fortunato, 2010 ; Fortunato et Hric, 2016). Bien que toute taxonomie théorique puisse être discutable, cette catégorisation a pour but de soutenir nos analyses empiriques : nous cherchons à vérifier si la proximité théorique et conceptuelle peut engendrer, en pratique, une proximité des partitions produites. La Table 2.1 liste les méthodes que nous avons utilisées et indique l'implémentation utilisée.

— **Edge removal** : Dans cette approche, les arêtes inter-communautés d'un réseau sont progressivement supprimées afin de déconnecter les groupes densément connectés. Le problème de la détection des communautés se traduit par l'identification de candidats pour les arêtes inter-communautés sur la base de leurs positions topologiques. Parmi les techniques populaires, citons Edge Betweenness (GN dans le tableau 2.1) ou Edge Clustering Coefficient Coefficient, qui peut être basée sur des modèles triangulaires (RCCLP-3) ou quadrangulaires (RCCLP-4).

— **Modularity optimization** : Les méthodes de cette approche utilisent une fonction objectif commune appelée *modularité* (Newman et Girvan, 2004), mais ont des stratégies d'optimisation différentes. Appartiennent à cette catégorie les célèbres méthodes Louvain, Greedy optimzation (CNM) ou encore Spectral Method (SN).

— **Dynamic process** : Les méthodes de ce groupe n'utilisent pas directement les informations topologiques. Elles exploitent plutôt les informations stochastiques issues de divers modèles dynamiques régulés par la structure du réseau, afin de déduire la structure communautaire. On y trouve des méthodes basées sur les marches aléatoires (Walktrap) ou d'autres méthodes exploitant la théorie de l'information (Infomod et Infomap).

— **Statistical inference** : Cette approche prend en compte la signification statistique de la structure des communautés sur la base de différents modèles théoriques de réseau. Les méthodes optimisent généralement les fonctions de vraisemblance pour trouver la meilleure configuration correspondant aux hypothèses en utilisant différentes stratégies de recherche. On y trouve la célèbre méthode Stochastic Block Model (SBM), ou encore Oslom, qui mesure

Approche	Publication	Label	Complexité	Code
Edge removal	(Girvan et Newman, 2002)	GN	$\mathcal{O}(nm^2)$	igraph [a]
	(Radicchi et al., 2004)	RCCLP	$\mathcal{O}(m^4/n^2)$	Authors [b]
Modularity optimization	(Clauset, Newman et Moore, 2004)	CNM	$\mathcal{O}(m\log^2(n))$	igraph
	(Blondel et al., 2008)	Louvain	$\mathcal{O}(n\log(n))$	Authors [c]
	(Newman, 2006)	SN	$\mathcal{O}(nm\log(n))$	igraph
Dynamic process	(Pons et Latapy, 2005)	Walktrap	$\mathcal{O}(n)$	igraph
	(Rosvall et Bergstrom, 2007)	Infomod	NA	Authors [d]
	(Rosvall, Axelsson et Bergstrom, 2009)	Infomap	$\mathcal{O}(m)$	Authors [e]
Statistical inference	(Lancichinetti et al., 2011)	Oslom	$\mathcal{O}(n^2)$	Authors [f]
	(Riolo et al., 2017)	(DC)SBM	Parametric	Authors [g]
Other methods	(Reichardt et Bornholdt, 2006)	RB	$\mathcal{O}(n^2log(n))$	igraph
	(Raghavan, Albert et Kumara, 2007)	LPA	$\mathcal{O}(m)$	igraph
	(Xie et Szymanski, 2012)	SLPA	$\mathcal{O}(m)$	Authors [h]
	(Meo et al., 2014)	Conclude	$\mathcal{O}(n+m)$	Authors [i]

a. Published at http://igraph.org/
b. Published at http://homes.sice.indiana.edu/filiradi/resources.html
c. Published at https://sourceforge.net/projects/louvain/
d. Published at http://www.tp.umu.se/~rosvall/code.html
e. Published at http://www.mapequation.org/
f. Published at http://www.oslom.org/
g. Published at http://www-personal.umich.edu/~mejn/
h. Published at https://sites.google.com/site/communitydetectionslpa/
i. Published at http://www.emilio.ferrara.name/code/conclude/

TABLE 2.1 – Les méthodes de détection de communautés utilisées dans cette étude.

le niveau d'importance statistique d'une communauté en calculant la probabilité de trouver une communauté similaire dans un modèle nul.

— **Other methods** : Certaines approches définissent implicitement ou explicitement des exigences concernant la structure de la communauté ou mélangent différentes approches pour tirer parti des avantages de chacune d'entre elles. Pour simplifier la taxonomie théorique, nous les présentons dans un groupe à part. On y trouve LPA et SLPA, deux variantes basées sur la propagation de labels (mécanisme épidémiologique), Spin Glass model (RB) qui utilise un principe de physique théorique, les verres de spin, ou encore Conclude, qui combine marches aléatoires et optimisation de modularité.

Afin de maintenir la contrôlabilité de nos expériences et d'assurer la reproductibilité de l'analyse, toutes les méthodes présentées ci-dessus sont étudiées avec les paramètres par défaut déterminés par les auteurs.

2.1.2 Dataset experimental

Dans cette étude, nous considérons 108 réseaux différents, ce qui est relativement important. En référence, citons Orman *et al.* qui utilisent 6 réseaux pour évaluer la

Category	Size	Nodes	Edges	Notable networks
Biological	7	1860	10763	Yeast, brain, protein-protein interactions
Communication	9	39595	195032	Email, forums, message exchanges
Information	25	38358	159812	Amazon, DBLP, citation & education webs
Social	37	6888	49666	Facebook, Youtube, Google Plus networks
Technological	19	18431	48494	Internet, AS Caida, Gnutella P2P networks
Miscellaneous	11	4298	49033	Ecology, power-grid, synthetic networks
Total*	108			

TABLE 2.2 – Vue synthétique des réseaux utilisés dans notre analyse où "Size" est le nombre de réseaux utilisés dans catégorie. "Nodes" et "Edges" indiquent le nombre moyen de nœuds et d'arêtes présents dans les réseaux de chaque catégorie. *La dernière ligne montre le nombre total de graphes. Les données ont été collectées à partir de différentes sources, dont : http://networkrepository.com (Rossi et Ahmed, 2015), http://konect.uni-koblenz.de (Jerome, 2013), http://snap.stanford.edu (Leskovec et Krevl, 2014)

.

structure des communautés découvertes par plusieurs techniques de détection (Orman, Labatut et Cherifi, 2012), Lancichinetti *et al.* qui utilisent 15 réseaux pour caractériser les communautés structurelles (Lancichinetti et al., 2010) ; Hric *et al.* qui utilisent 16 réseaux pour révéler les différences entre les communautés structurelles et la vérité terrain (Hric, Darst et Fortunato, 2014) ; Leskovec *et al.* qui utilisent autour de 100 réseaux pour analyser le profil des communautés (Leskovec et al., 2008) et 230 réseaux pour évaluer la qualité des communautés de la vérité terrain dans les réseaux sociaux. Parmi les 230, 225 sont vraiment similaires, car issus de la plateforme de réseautage Ning [3] (Yang et Leskovec, 2015). Mentionnons enfin le travail connexe de Ghasemian *et al.* qui a introduit le grand corpus CommunityFitNet [4] contenant 572 réseaux du monde réel. La Table 2.2 résume la composition des réseaux que nous avons analysés, et la Figure 2.1 leurs propriétés structurelles.

2.2 Performances en temps de calcul

De manière assez classique, et parce que le temps de calcul est un critère de choix d'algorithme, nous avons comparé les méthodes selon leur performances computationnelles. Nous avons mesuré le temps pris pour calculer chaque partition sur chaque réseau (configuration par défaut des méthodes). Les calculs ont été effectués sur une machine équipée d'un processeur Intel Xeon CPU E5-2650 avec 32 cœurs 2.60 GHz et une capacité mémoire de 100 Go. En raison de la grande complexité de certaines méthodes, seuls les processus qui se terminent dans un délai raisonnable (moins de 4 heures) sont pris en compte. A titre de référence, nous avons laissé certains calculs plus longs se poursuivre. Par exemple, la méthode *Conclude* a mis environ 9 jours pour identifier les structures communautaires sur un réseau de 300 000 sommets et 1 million d'arêtes ; la méthode *GN* n'a pas terminé son calcul pour les réseaux de plus de 4 000 nœuds et 40 000 arêtes en 2 jours. Par conséquent, les expériences qui nécessitent théoriquement trop de temps sont négligées dans nos résultats. Il convient également de noter que les calculs des communautés sur les très

3. https ://www.ning.com/
4. https ://github.com/AGhasemian/CommunityFitNet

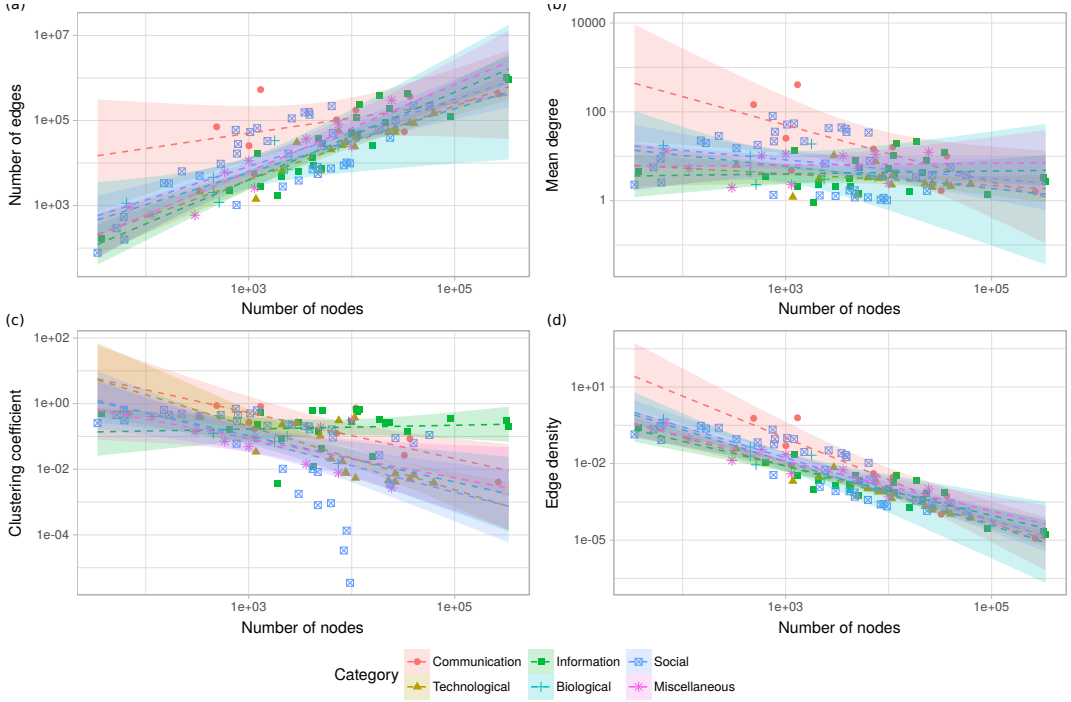


FIGURE 2.1 – De gauche à droite, de haut en bas, nous illustrons les caractéristiques structurelles des 108 réseaux : (a) Nombre d'arêtes en fonction du nombre de nœuds, (b) Degré moyen $\langle k \rangle$ en fonction du nombre de nœuds, (c) Coefficient de clustering en fonction du nombre de nœuds, (d) Densité en fonction du nombre de nœuds. La zone colorée correspond aux intervalles de confiance de 95% des relations entre variables estimées grâce à une régression linéaire pour chaque catégorie.

grands réseaux sont parfois limités par une mémoire restreinte. Ainsi, les calculs qui devraient être terminés en 4 heures mais qui nécessitent trop de mémoire ne peuvent pas non plus être présentés ici. Nous répétons les calculs 5 fois pour chaque paire graphe/méthode afin de réduire l'impact des fluctuations. En éliminant tous les cas qui ne satisfont pas à nos exigences, le taux de réussite final (nombre de partitions identifiées sur le nombre de tests possibles) s'établit à environ 44,72%, principalement en raison d'un dépassement de temps/mémoire.

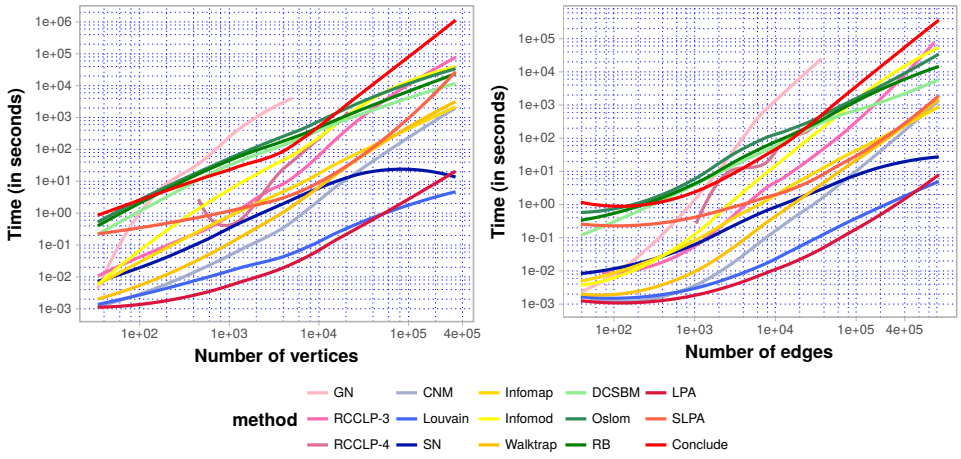


FIGURE 2.2 – Le temps d'exécution nécessaire à chaque méthode pour identifier les structures communautaires, estimé avec un modèle de régression local. Les méthodes de la même famille théorique sont représentées par des couleurs similaires.

Comme le montre la figure 2.2, à l'exception de *GN*, le temps d'exécution nécessaire pour toutes les autres méthodes est limité dans une plage qui augmente de façon polynomiale avec la taille du réseau, ce qui reflète les estimations théoriques (2.1). Cette plage est délimitée en haut par *Conclude*/*Oslom* et en bas par *LPA*, correspondant respectivement à la pire et à la meilleure méthode testée. Un autre fait important qui peut être déduit de cette figure est que, pour la plupart des réseaux du monde réel d'une taille inférieure à 1 million d'arêtes, choisir une méthode de détection rapide pourrait économiser 10^3 à 10^5 fois l'effort de calcul.

Nous montrons dans le tableau 2.3 le classement de ces méthodes selon nos tests. La première constatation est qu'au sein d'une même approche théorique, différents comportements cohabitent. Des méthodes passent à l'échelle, d'autres non, selon la fonction objectif, la mise en œuvre algorithmique ou encore l'implémentation dont elles font l'objet. Les approches par retrait d'arêtes (edge removal) que nous avons testées semblent toutefois peu efficaces : *GN* et *RCCLP-4* ne figurent même pas dans notre classement du fait de leur dépassement de temps de calcul pour les grands graphes.

Nous montrons à la fois les classements selon la moyenne et la médiane du temps. Le classement selon le temps moyen est fortement affecté par les mesures sur les grands graphes, il permet ainsi de départager les méthodes qui passent à l'échelle : *Louvain*, *LPA*, *SN* puis *Walktrap* sont à privilégier pour les grands graphes ; si le graphe dépasse les 10000 nœuds et 100000 arêtes, *Walktrap* est à éviter cependant. Le classement médian, quant à lui, reflète davantage la performance relative sur les petits et moyens graphes qui sont les plus nombreux.

TABLE 2.3 – Classement des méthodes en fonction du temps consommé. Les lignes sont regroupées selon notre taxonomie.

Méthode	Rang moyen	Rang médian	Passage à l'échelle
GN	-	-	-
RCCLP-3	9	8	Low
RCCLP-4	-	-	-
CNM	5	3	Medium
Louvain	1	2	High
SN	3	5	High
Walktrap	4	4	High
Infomod	12	9	Low
Infomap	6	7	Medium
Oslom	11	14	Low
DCSBM	8	12	Low
RB	10	13	Low
LPA	2	1	High
SLPA	7	6	Medium
Concude	13	11	Low

2.3 Taille des communautés

Après ces considérations de performance, nous nous concentrons sur la nature des résultats produits, c'est-à-dire les communautés elles-mêmes. Le nombre de communautés latentes qui devraient être induites à partir d'un réseau donné est l'une des questions majeures dans le contexte de la détection de communautés (Fortunato et Hric, 2016), (Riolo et al., 2017). Elle est équivalente au sujet du nombre attendu de clusters dans un problème classique de clustering. L'observation du nombre de communautés révèle des informations utiles sur la structure mésoscopique d'un réseau. La variation du nombre de communautés dans un réseau implique différents niveaux de résolution. Une façon analogue de décrire le concept de résolution est la distance d'observation d'un objet dans une scène : plus on se rapproche, plus on peut percevoir les détails de ses microstructures alors que, dans le même temps, les informations sur l'organisation globale tendent à être moins claires. Bien que plusieurs approches multi-résolution (Lambiotte, 2010 ; Pons et Latapy, 2011) incorporent des paramètres de résolution dans leurs solutions, fournissant des mécanismes plus flexibles et différentes échelles modulaires de réseaux, il n'est pas toujours évident de réguler ces paramètres de manière appropriée, ils sont à déterminer de façon *ad hoc* au contexte de l'étude. L'inclusion de paramètres multi-résolution élargit bien sûr la possibilité de comprendre les réseaux, mais au détriment de la commodité de l'automatisation, qui est parfois requise dans les problèmes de clustering.

2.3.1 Nombre des communautés produites

Nous avons calculé, pour toutes les méthodes, sur l'ensemble des communautés provenant des 108 partitions produites, le nombre moyen de communautés par partition. Si l'on utilise le nombre théorique k de communautés attendu dans le modèle *k-planted partition model* (Ames, 2013)), k tend vers $O(\sqrt{n})$, n étant le nombre de nœuds. On peut alors classifier les méthodes selon leur tendance à sur- ou sous-estimer ce nombre k. Le tableau 2.4 compile cette tendance. Il y a ici une certaine homogénéité parmi les approches théoriques : elles partagent des définitions proches de ce que doit refléter une communauté. Ce constat a également été fait par (Ghasemian, Hosseinmardi et Clauset, 2018).

TABLE 2.4 – Classement des méthodes selon leur nombre moyen de communautés détectées comparé avec le k-planted model. Une méthode sur-estime (over-fit) ce nombre si elle produit asymptotiquement plus que $\sqrt{n}$ clusters, et sous-estime (under-fit) dans le cas contraire.

Méthode	Nombre de communautés	Sur- ou sous-estimation
GN	Bigger	Over-fit
RCCLP-3	Bigger	Over-fit
RCCLP-4	Bigger	Over-fit
CNM	Close	Over-fit
Louvain	Close	Under-fit
SN	Smaller	Under-fit
Walktrap	Bigger	Over-fit
Infomod	Close	Under-fit
Infomap	Bigger	Over-fit
Oslom	Smaller	Under-fit
SBM	Smaller	Under-fit
DCSBM	Smaller	Under-fit
RB	Smaller	Under-fit
LPA	Bigger	Over-fit
SLPA	Bigger	Over-fit
Concude	Bigger	Over-fit

Cependant, bien qu'utile pour aider les praticiens à présumer du nombre attendu de clusters qu'une méthode détecterait par rapport à l'expérience théorique, il est toujours très difficile de décider quelle méthode utiliser, puisque la référence elle-même est basée sur les hypothèses du modèle sous-jacent, i.e k-planted partition model dans notre cas. Cela signifie également que si nous changeons la référence, le nombre attendu de communautés sera différent, et notre classification ne sera plus pertinente. C'est pourquoi, dans la section suivante, nous proposons une nouvelle technique permettant de comparer les méthodes, non pas sur le nombre de communautés produites, mais un critère corrélé, et tout aussi intuitif pour un décideur, la taille de celles-ci.

2.3.2 Comparaison des méthodes selon la taille des communautés produites

En tant qu'analyste, l'un des premiers critères intuitif et facile à appréhender est la taille des communautés produites. Selon l'étude, et la taille du réseau, il sera pertinent de minimiser le nombre de clusters, quitte à en augmenter la taille, ou au contraire, analyser finement des sous-groupes de taille raisonnable pour bien en comprendre la structure et focaliser sur des sous-ensembles de nœuds que l'on souhaite détailler.

Dans (Dao, Bothorel et Lenca, 2018), nous proposons une méthode de calcul de similarité entre méthodes selon la taille des communautés produites. Deux méthodes sont dites similaires si leurs distributions de densité exposent une grande zone d'intersection, comme le montre la Figure 2.3(a). En utilisant un estimateur de densité par noyau (lignes pleines de la Figure 2.3(b)), nous approximons la fraction commune des communautés de même taille par la zone de chevauchement de deux distributions continues correspondantes. Le principe de cette estimation est que deux méthodes similaires ne produisent pas toujours une grande partie des communautés de même taille exacte, mais plutôt une grande partie des communautés de taille comparable.

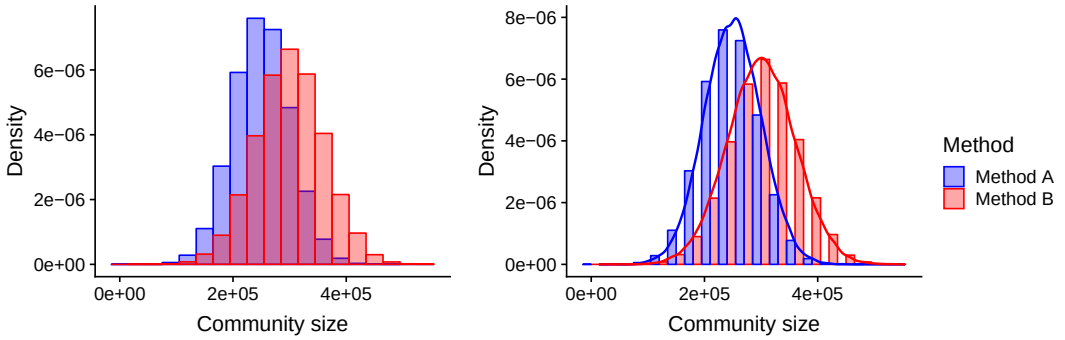


FIGURE 2.3 – La distribution des tailles des communautés détectées par deux méthodes différentes. À gauche (a) chevauchement à l'aide d'un histogramme, à droite (b) lorsque les tailles des communautés s'entrecroisent, la similarité est mieux estimée à l'aide d'un estimateur de densité par noyau.

La Figure 2.4 représente l'estimation des densités de taille de communautés, calculées en agrégeant l'ensemble des communautés obtenues sur les 108 partitions méthode par méthode. Comme nous pouvons le voir, il y a de nettes différences, ce qui démontre une diversité de stratégies de partitionnement. Toutefois, au sein d'une même famille théorique (comme indiqué dans le Tableau 2.1), nous pouvons constater une certaine homogénéité.

— Les méthodes de type *Edge removal* produisent plutôt de très petites communautés.
— Les méthodes optimisant la *modularité* génèrent quant à elles de très grandes communautés. Ce phénomène est bien connu, il s'agit d'un problème de limite de résolution de la modularité qui a tendance à agréger des petites communautés même bien dessinées en de plus grandes communautés (Fortunato et Barthelemy, 2006). Ces méthodes génèrent néanmoins des petites communautés d'une dizaine de personnes lorsque les réseaux sont de petite taille.
— Le groupe *Dynamic process* est le plus disparate, avec par exemple *Walktrap* et *Infomap* assez proche de la première famille et *Infomod* de la deuxième.

— Les méthodes du groupe *Statistical inference*, *SBM* et *DCSBM*, utilisent un processus d'échantillonnage de Monte Carlo, qui prend beaucoup de temps, afin de balayer l'espace des solutions. Ceci rend la méthode irréalisable, si le nombre maximum de clusters n'est pas limité. Et en effet, dans la version par défaut, le nombre maximum de communautés est limité à 25, ce qui signifie que les méthodes $(DC)SBM$ trouvent de très grandes communautés dans les grands réseaux.
— Dans la dernière famille, enfin, on trouve principalement des petites communautés, sauf pour *RB* qui produit de très grandes communautés.

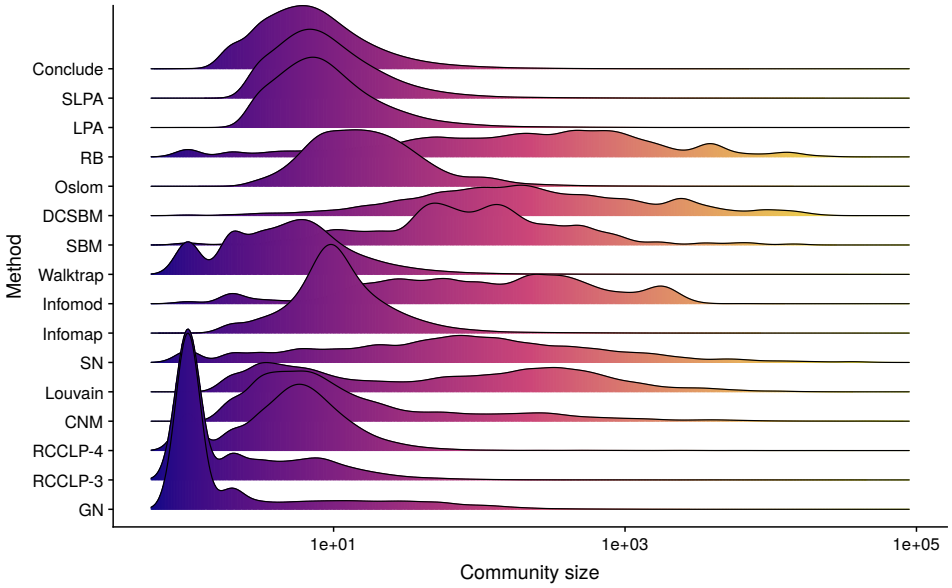


FIGURE 2.4 – Distributions de la taille des communautés, pour toutes les communautés des partitions détectées sur tous les réseaux. Les distributions sont lissées en utilisant un estimateur par noyau gaussien. La couleur du gradient est utilisée uniquement pour faciliter l'observation.

La Figure 2.5 montre le résultat du calcul de similarité de chaque couple de méthodes sur la base de notre métrique. Les méthodes peuvent être classées dans différentes classes de stratégie de partitionnement. Les séparations sont claires :

1. **Petites à très grandes** - *RB*, *DCSBM*, *SBM*, *Infomod*, *SN*, *Louvain* : Les méthodes de ce groupe découvrent des communautés dont les tailles varient dans une large gamme de spectres, de très petites à très grandes communautés. La distribution des tailles des communautés caractérisées est assez plate, ce qui signifie que toutes les tailles sont presque également prises en compte.
2. **Très petites** - *GN* and *RCCLP-3* : Ces deux méthodes identifient un grand nombre de très petites communautés, y compris des singletons, quelle que soit la taille du réseau. Par conséquent, il y a peu de diversité de taille de communautés.
3. **Petites** - les autres : Ces méthodes produisent des communautés dont la taille se rapproche d'une distribution en forme de cloche, avec des communautés de petites tailles, autour d'une dizaine de membres.

Cette caractérisation nous aide à identifier des groupes de méthodes de détection de communautés en fonction de la taille des communautés. Elle permet également

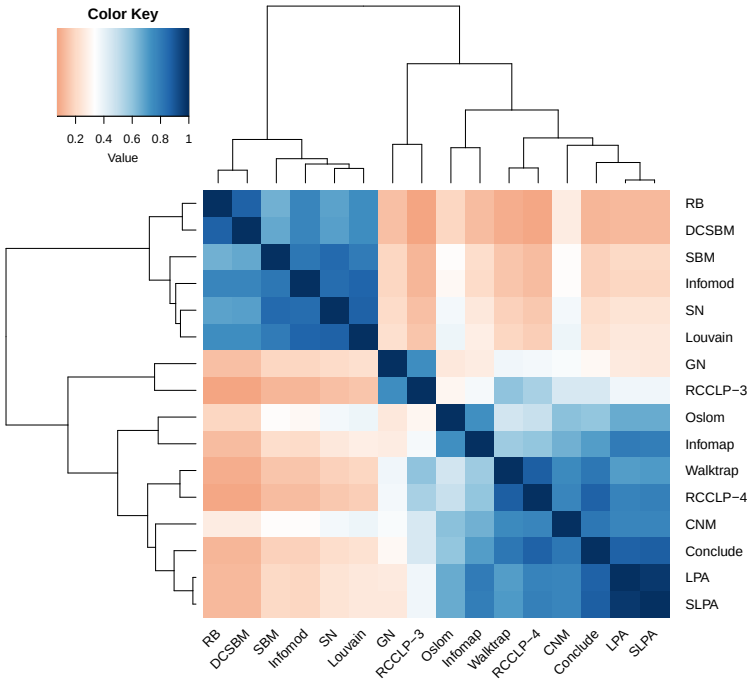


FIGURE 2.5 – Similarité des méthodes selon la taille des commmunautés produites. Deux méthodes sont proches si elles partagent une grande fraction de communautés de taille similaire, selon notre méthode (Dao, Bothorel et Lenca, 2018). Les lignes et les colonnes sont ordonnées selon une méthode de clustering hiérarchique (Joe H. Ward, 1963). Le dendogramme reflète la structure hiérarchique des clusters. Plus le bleu est foncé, et plus la similarité est grande.

d'éviter les tentatives de calcul coûteux, voire impossible, en proposant des solutions de substitution. Ainsi, en combinant avec l'analyse précédente relative au temps de calcul dans la section 2.2, on pourrait également choisir un groupe de méthodes correspondant à des critères de distinction de taille et ensuite sélectionner la méthode la plus rapide qui mène au résultat souhaité. Ou inversement, selon la taille du réseau à analyser, le critère de scalabilité pourra être prioritaire ; en cas de grand graphe, le choix se fera parmi les quatre méthodes qui passent à l'échelle. Si l'on souhaite privilégier les petites communautés, l'usage de *LPA* ou *Walktrap* est à privilégier ; *Louvain* et *SN* pourront être utilisées pour limiter le nombre de communautés à analyser qui seront alors plus grandes en termes de nombre d'individus (Table 2.5).

La taille des communautés (ou leur nombre) n'est qu'une dimension de qualité possible, bien qu'elle soit probablement l'une des informations les plus intuitives et les plus importantes lors du choix d'une méthode de clustering. Dans la partie suivante, nous mobilisons d'autres techniques qui peuvent être utilisées pour définir d'autres aspects de la similarité entre méthodes.

2.4 Stratégies de partitionnement

Dans cette section, nous nous intéressons à la manière de former les communautés. En effet, même si deux méthodes produisent des regroupements équivalents en terme de taille, il se peut que la stratégie de partitionnement diffère.

TABLE 2.5 – Classement récapitulatif des méthodes selon leur faculté à passer à l'échelle et la taille des communautés produites.

Méthode	Scalability	Size
Walktrap	high	small
LPA	high	small
Louvain	high	small to large
SN	high	small to large
Infomap	medium	small
SLPA	medium	small
CNM	medium	small
RCCLP-4	low	small
Conclude	low	small
Oslom	low	small
Infomod	low	small to large
SBM	low	small to large
DCSBM	low	small to large
RB	low	small to large
GN	low	very small
RCCLP-3	low	very small

2.4.1 Métriques de qualité et co-performance

Pour comprendre et comparer les méthodes de détection de communautés selon leur stratégie de partionnement, nous nous sommes intéressés au comportement qu'elles adoptent face aux fonctions objectifs bien connues telles que la modularité et ses variantes (D-modularité, Z-modularité...), la surprise ou encore la significance. Ces métriques sont appelées mesures de qualité, parfois "goodness metrics" ou "community scoring functions".

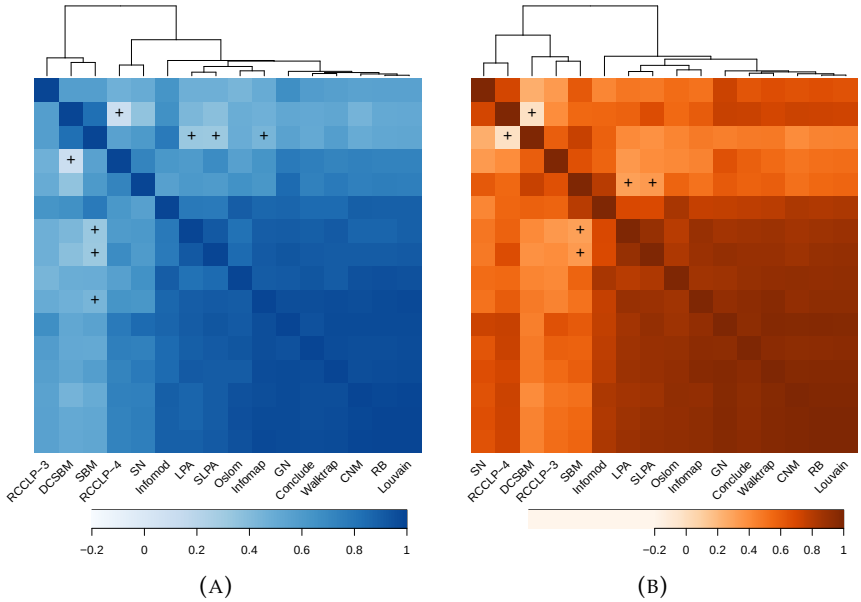
Notre approche, présentée dans (Dao, Bothorel et Lenca, 2020), se base sur l'analyse de *co-performance*. Nous définissons un *indice de co-performance* relatif à deux méthodes A et B sur un dataset $\mathcal{G}$, par leur capacité mutuelle à découvrir des structures communautaires présentant une qualité particulière Q. En d'autres termes, nous attribuons à chaque couple de méthodes un indice élevé en fonction d'une qualité Q, si la connaissance de la performance d'une méthode révèle de manière significative l'information sur la performance de l'autre. Une solution simple pour définir l'indice consiste à utiliser la corrélation de Pearson. Elle reflète en effet la covariance d'un score de qualité sur deux ensembles de partitions détectés par deux méthodes, indiquant si deux méthodes sont en accord ou non avec un critère de qualité dans le contexte donné d'un ensemble de réseaux.

Nous trouvons des indices de co-performance élevés entre les méthodes *CNM*, *Conclude*, *Oslom*, *Walktrap*, *LPA*, *SLPA* et *Infomap* dans la plupart des cas des six fonctions de qualité testées. Ces méthodes se retrouvaient également dans le cluster des méthodes associées aux communautés de petites tailles.

Cependant, il est difficile d'utiliser cet indice. Un analyste vraiment averti pourra

bien sûr chercher à optimiser via son clustering, par exemple, la modularité Erdős-Rényi plutôt que la modularité la plus connue, celle de Newman-Girvan. S'il constate que *Louvain* a de bonnes performances sur son ou ses réseau(x), il pourra être en confiance en utilisant des méthodes alternatives à fort indice de co-performance, comme par exemple *Walktrap* (Figure 2.6).

La difficulté de compréhension des mesures de qualité d'une part, et de l'indice de co-performance d'autre part, nous conduisent à proposer une autre approche, basée cette fois sur la répartition des nœuds dans les communautés.



(A) (B)

FIGURE 2.6 – Matrices de *co-performance* des 16 méthodes. Le "+" indique des p-valeurs supérieures à 0.05, donc la corrélation n'est pas avérée. (a) Newman-Girvan modularity, (b) Erdős-Rényi modularity, (c) Density modularity and (d) Z modularity.

2.4.2 Métriques de validation et constitution des clusters

Cette section est consacrée à l'utilisation de métriques de validation de clustering populaires issues de la littérature du clustering traditionnel (et également largement utilisées dans le contexte de la détection des communautés). Il s'agit de mesurer directement la similarité des partitions en utilisant leurs tables de contingence (Table 2.6). Ces mesures ne prennent pas en compte les informations structurelles des communautés comme la modularité peut le faire en comptant les arêtes inter-communautaires, mais elles utilisent uniquement le nombre commun de nœuds qui sont partagés par les paires de communautés dans deux partitions.

En comparant de manière exhaustive, sur l'ensemble de nos 108 graphes, à quelle point chaque méthode est en accord (ou pas) avec chacune des autres sur leur répartition des individus, nous aurons une sorte d'équivalence entre méthodes d'un point de vue ensembliste.

Les métriques de validation sont souvent utilisées dans le contexte de l'évaluation de la détection de communautés pour mesurer la différence entre la partition identifiée par une méthode et une partition attendue du réseau considéré (*vérité terrain* ou *groundtruth*). Plus la partition découverte est similaire à la vérité terrain, meilleure est la performance de la méthode. Cependant, dans cette étude, les métriques de validation sont exploitées comme un outil pour comparer les structures

TABLE 2.6 – Table de contingence de deux partitions P_1 et P_2 obtenues sur le même graphe.

		Partition P_2				
		$c_1^{(2)}$	$c_2^{(2)}$	$\cdots$	$c_S^{(2)}$	$\sum$
Partition P_1	$c_1^{(1)}$	n_{11}	n_{12}	$\cdots$	n_{1S}	$n_{1.}$
	$c_2^{(1)}$	n_{21}	n_{22}	$\cdots$	n_{2S}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$c_R^{(1)}$	n_{R1}	n_{R2}	$\cdots$	n_{RS}	$n_{R.}$
	$\sum$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.S}$	n

de communautés de différentes méthodes. Elles estiment ainsi la proximité empirique des différents algorithmes à travers les partitions détectées.

La Table 2.7 résume les métriques utilisées dans notre étude : Rand Index (RI) (Rand, 1971) et sa variante Adjusted Rand Index (ARI) (Hubert et Arabie, 1985), Normalized Mutual Information (NMI) (Danon et al., 2005 ; Chakraborty et al., 2017) et sa variante Adjust Mutual Information (AMI) (Vinh, Epps et Bailey, 2010).

TABLE 2.7 – Les métriques de validation utilisées pour comparer les stratégies de partionnement des méthodes de façon empirique

Mesure	**Intervalle**	**Description**
RI	$[0, 1]$	Fraction de sommets groupés et séparés en commun dans deux partitions.
ARI	$[0, 1]$	Rand Index moins sensible aux différences de taille des communautés.
NMI	$[0, 1]$	Basé sur la théorie de l'information, quantifie la quantité d'information d'une partition permettant de deviner l'autre.
AMI	$[0, 1]$	Variante robuste de NMI, moins sensible au hasard.

Le processus expérimental est le même que celui des sections précédentes. À partir des partitions détectées par les 16 méthodes sur l'ensemble du jeu de données, nous calculons les scores pour chaque paire de partitions par graphe, soit $C_2^{16} = 120$ mesures pour chacun de nos 108 graphes. La Figure 2.7 illustre les scores moyens obtenus par paire de méthodes [5].

En observant les dendogrammes de la Figure 2.7, nous notons que RI ne devrait pas être utilisée comme métrique de validation pour évaluer la performance de partitionnement. Puisque ses valeurs moyennes varient généralement sur une petite plage (0.9 à 1.0), il est difficile de différentier les partitions. D'autre part, NMI et AMI donnent des valeurs entre 0.5 et 1.0, ce qui revient à dire que globalement l'ensemble des méthodes sont plutôt en accord. Enfin, même si ARI semble amplifier les différences entre les méthodes, il n'y a pas de différence majeure dans l'évaluation de la similarité par rapport aux autres métriques.

On peut voir que les 4 métriques classent les méthodes en deux groupes principaux de manière assez similaire aux matrices de co-performance exposées dans la section précédente. Le groupe de méthodes *LPA, SLPA, Oslom, Conclude, Infomap*,

5. Quand les méthodes correspondantes sont capables de finir leur traitement en un temps raisonnable ou ne saturent pas la mémoire comme mentionné dans les expériences précédentes.

TABLE 2.8 – Classement des méthodes selon leur stratégie de partitionnement (selon la mesure *ARI*)

Méthode	**Approche**	**Cluster**
Walktrap	Dynamic process	1
GN	Edge Removal	1
CNM	Modularity optimization	1
Louvain	Modularity optimization	1
SN	Modularity optimization	1
RB	Others	1
Infomap	Dynamic process	2
Conclude	Others	2
LPA	Others	2
SLPA	Others	2
Oslom	Statistical Inference	2
Infomod	Dynamic process	3
DCSBM	Statistical Inference	3
SBM	Statistical Inference	3
RCCLP-3	Edge Removal	4
RCCLP-4	Edge Removal	4

déjà identifié dans section précédente, montre également de très fortes similitudes dans cette expérience. *LPA* et *SLPA*, en particulier, étant basées sur le mécanisme de propagation d'étiquettes, proposent des résultats presque identiques dans de nombreux cas. En outre, on peut discerner un autre groupe comprenant *Louvain, CNM, SN*, basées sur la modularité, ainsi que *GN*, *Walktrap* et *RB*, qui présentent une grande cohérence en général. Les autres méthodes présentent des scores plus faibles, mais se répartissent dans deux autres groupes : *SBM*, *DCSBM* et *Infomod*, d'une part, et les deux variantes de *RCCLP* qui se singularisent. Globalement, il semble que les méthodes ayant un même fondement théorique ont tendance à fournir des résultats assez similaires (Table 2.8).

2.5 Conclusion

Notre étude empirique, systématique, sur un ensemble varié de 108 réseaux que nous supposons représentatif des situtations réelles, nous enseigne que l'ensemble des 16 méthodes testées sont relativement interchangeables pour explorer des réseaux. Les partitionnements qu'elles offrent ne présentent pas de grandes différentes majeures en terme de stratégie de partitionnement comme le montrent les mesures de validation, puisque leurs partitions conduisent à des scores de *NMI* (ou *AMI*) plutôt élevés quelque soit la paire $((methode_i, graphe_k), (methode_j, graphe_k))$ choisie. La première conclusion est qu'un analyste peut en confiance choisir l'une ou l'autre de ces méthodes : le hasard pourra déjà être un choix en tant que tel !

Une autre conclusion est que les algorithmes de détection de communautés *"laissent leur empreinte digitale sur les communautés qu'ils renvoient"*, comme me l'a dit très justement Aaron Clauset au cours d'un de nos échanges. Autrement dit, ce qu'un algorithme trouve dans un réseau... dépend fortement de ce qu'il cherche ! Les approches

théoriques sont en effet plutôt homogènes dans nos classifications, notamment en termes de stratégie de partitionement : elles regroupent les nœuds de façon relativement similaire (même si ce constat ne soit pas systématique).

Mais cette étude nous permet d'aller un peu plus loin que ces grands principes, et nous permet d'ébaucher un guide de choix d'algorithme.

2.5.1 Vers une aide à la décision

Les fondements théoriques implémentés dans les algorithmes ne sont pas toujours simples à appréhender. Certains parmi nous seront familiers de l'optimisation de la modularité, d'autres de la théorie de l'information, quelques rares détiendront les connaissances en physique théorique pour maîtriser le modèle des verres de spin, mais globalement, je peux avancer sans risque que comprendre les algorithmes n'est qu'un critère parmi d'autres, et rarement celui qui est appliqué.

En pratique, ce sont d'avantage des critères de mise en œuvre qui sont utilisés, comme la disponibilité de l'implémentation dans un langage de programmation, facilité d'utilisation via un outil tel que Gephi...

Notre ambition est d'élargir ces critères par une cartographie simple, impliquant 3 dimensions faciles à comprendre et adopter (Figure 2.8) : la scalabilité (tenue en charge), la taille des communautés produites et une équivalence des méthodes en termes de stratégie de partionnement. Via notre calcul de similarité et notre classification, nous nous affranchissons de la compréhension fine des mécanismes implémentés tout en en tenant compte.

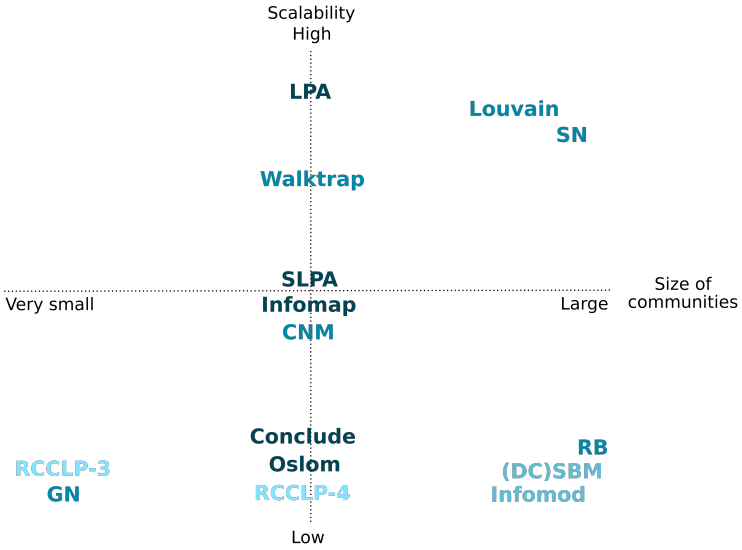


FIGURE 2.8 – Cartographie des méthodes de détection de communautés selon leur scalablité et la taille des communautés produites. La couleur correspond aux clusters de stratégie de partionnement (clusters de la Table 2.8 obtenus via la métrique de validation ARI).

Il s'agit d'un premier pas vers l'aide à la décision, permettant de sélectionner 2, 3, 4 méthodes parmi l'ensemble de la boîte à outils. Dans le chapitre suivant, nous proposons une méthode plus qualitative permettant au décideur ou l'analyste de finaliser son choix de méthodes pour n'en garder qu'une.

Un analyste peut être amené à utiliser plusieurs algorithmes avant de faire un choix final. Voici différentes possibilités de sélection.

— Pour des petits réseaux pour lesquels la complexité en temps n'est pas rédhibitoire, il pourra utiliser 3 méthodes, qui *a priori*, répartissent les nœuds de manière équivalente mais avec des niveaux de résolution/échelles différentes : *GN* plus connu sous le nom de *Edge Betweenness* (très petites communautés), *Walktrap* (petites communautés), ainsi que *Louvain* (grandes communautés).
— Pour un petit réseau toujours, on pourraît conseiller au contraire de varier les stratégies de partionnement, et de choisir 4 méthodes, l'une dans chacun des clusters, en variant les tailles de communautés : une combinaison possible est : *RCCLP-3, Walktrap, Infomap et Infomod*.
— Si l'on souhaite réaliser l'étude du point précédent exclusivement sous *igraph* en R ou Python, seuls les 2 grands clusters de méthodes sont disponibles : une combinaison possible devient alors *GN, Walktrap, Infomap et Louvain*.
— Si le réseau est grand (plus de 10 000 nœuds et 100 000 arêtes), *LPA* est tout indiqué pour offrir des petites communautés, et *Louvain* des communautés plus grandes.
— Si l'objectif est de synthétiser un grand réseau, *SN* pourraît être utilisée conjointement avec *Louvain* : elles produisent toutes les deux plutôt des grandes commmunautés, donc un nombre "raisonnable" de communautés à étudier ; mais on peut s'attendre à ce que leurs regroupements soient similaires puisqu'appartenant au même cluster et toutes deux basées sur la modularité.

2.5.2 De nombreux travaux connexes

Orman *et al.* ont publié une évaluation comparative de huit algorithmes de détection de communautés (Orman, Labatut et Cherifi, 2012), dont la plupart sont également étudiés dans ce travail. Différentes métriques de validation sont utilisées pour évaluer l'accord entre les partitions découvertes et les structures de communautés de référence (groundtruths). Comme dans notre travail, ils constatent que ces métriques (RI, ARI, NMI) *"s'accordent entre elles avec de petites différences"*, comme illustré dans la section 2.4.2. En outre, les auteurs se concentrent également sur l'analyse de nombreux aspects topologiques de la structure des communautés, notamment la transitivité, la densité, la taille des communautés, etc. Ces qualités topologiques sont ensuite utilisées pour inspecter les structures de communauté détectées par différents algorithmes. Les analyses permettent aux auteurs de conclure que ces deux approches (métriques topologiques et métriques de validation), utilisées pour évaluer les structures de communautés, sont *"complémentaires et nécessaires pour effectuer une analyse pertinente et complète des résultats de détection de communautés"*, ce que nous aborderons également dans le chapitre suivant. Ils notent également que *"l'approche traditionnelle (RI, ARI, NMI) est beaucoup plus rapide et facile à appliquer"*, et conseillent donc d'utiliser ces métriques en premier. Cependant, en pratique, les vérités terrain ne sont généralement pas disponibles. Dans le contexte où un nouvel algorithme est inventé, on utilise normalement des réseaux dont les structures de communauté sont bien connues afin de valider la méthode proposée. En réalité, comme la détection de communautés est souvent employée pour découvrir les structures de *nouveaux*

réseaux, il est donc peu probable que des structures communautaires de référence existent. Par conséquent, à partir des observations ci-dessus, nos analyses pourraient constituer un support important, en fournissant des informations supplémentaires sur la proximité entre les méthodes, tant sur l'aspect topologique que sur l'aspect partitionnel.

Agreste *et al.* évaluent également différents algorithmes de détection de communautés dans une approche empirique et comparative, notamment dans le contexte de l'analyse des données web (Agreste et al., 2017). Les auteurs constatent que la complexité temporelle est un facteur crucial dans la sélection d'un algorithme de détection de communautés et que la méthode de propagation d'étiquettes (*LPA*) présente des performances exceptionnelles en terme de scalabilité sur des graphes artificiels et réels, ce qui est également en accord avec notre analyse dans la section 2.2 qui fournit des prédictions sur le temps nécessaire pour chaque méthode, en fonction de la taille du réseau. Ils concluent également que *"L'algorithme Infomap a présenté le meilleur compromis entre la précision et les performances de calcul"* sur la base du score NMI. Une telle conclusion est valable là encore dans certains cas spécifiques où la structure de la communauté de la vérité terrain est bien connue bien sûr. A noter que des analyses supplémentaires devraient être effectuées pour déterminer si les vérités terrain correspondent à l'objectif final des algorithmes de détection de communautés (Peel, Larremore et Clauset, 2017). En effet, les informations de métadonnées des nœuds sont souvent utilisées dans la pratique comme groundtruth, alors qu'il a été constaté que les communautés issues de ces métadonnées sont parfois peu pertinentes, comme nous l'avons nous-mêmes constaté (Dao, Bothorel et Lenca, 2017b). Il convient donc d'être prudent en matière de généralisation. Les algorithmes peuvent avoir des difficultés à s'accorder avec certaines références basées sur des métadonnées, mais il ne faut pas supposer ce niveau de performance sur d'autres ensembles de données.

Ghasemian *et al.* présentent une évaluation de l'overfitting et de l'underfitting d'une quinzaine de méthodes de détection de communautés (Ghasemian, Hosseinmardi et Clauset, 2018). Les auteurs étudient le nombre de communautés détectées en pratique par de nombreuses méthodes, et le nombre maximal de clusters détectables, selon un modèle théorique, comme on l'a fait dans la section 2.3. Cette étude aide à choisir une méthode appropriée en fonction de la qualité de l'ajustement. Les méthodes de détection de communautés sont également regroupées en familles distinctes, sur la base de leurs résultats sur plus de 400 réseaux du monde réel (de manière similaire à notre analyse dans la section 2.4.2) en utilisant la métrique AMI. Les auteurs constatent également que *"ce qu'un algorithme trouve dans un réseau dépend fortement des hypothèses qu'il fait sur ce qu'il faut chercher"*, ce qui est aligné sur nos résultats à travers plusieurs analyses.

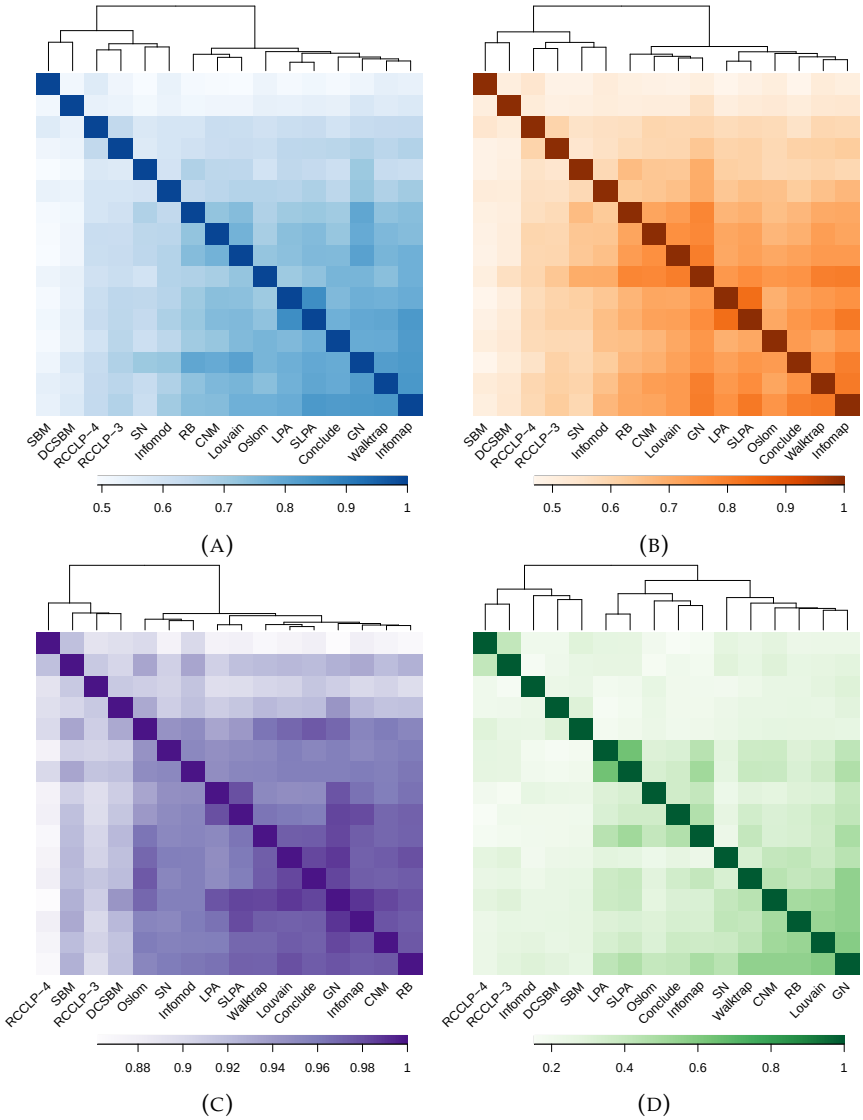
Jebabli *et al.* proposent également un cadre pour évaluer la performance des algorithmes de détection de communautés, en se basant sur les caractéristiques topologiques des graphes de communautés résultants (Jebabli et al., 2018). Dans leur article, les auteurs limitent leur attention aux réseaux avec des structures de communautés chevauchantes, alors que notre évaluation comparative concerne le clustering simple, sans chevauchement. Ils présentent et évaluent une méthodologie alternative efficace, comparée aux mesures classiques de qualité et de clustering. Comme dans notre travail, les algorithmes sont comparés dans un schéma de prise de décision. Ils fournissent différents classements des algorithmes, en fonction de différentes propriétés topologiques. Ils introduisent également une stratégie d'aide multi-critères à la décision afin de trouver le meilleur compromis entre ces différents classements. Un seul classement est produit. Une telle stratégie de prise de décision

est donc une approche très intéressante. La conversion des recommandations en un processus de décision est l'une de nos perspectives d'étude.

Cependant, Ghasemian *et al.* relativise ce genre d'étude empirique, exhaustive, qui tend à catégoriser les méthodes sur un comportement moyen basé sur leurs résultats obtenus sur un très grand nombre de réseaux très différents (sociaux, biologiques, etc). Cela cache bien évidemment des disparités. Chacune des méthodes peut potentiellement être "la meilleure" dans un contexte particulier : un type de graphe à étudier, une question de recherche, des hypothèses à vérifier, un outillage disponible, un temps de calcul à respecter, etc.

Dans le chapitre suivant, nous proposons une analyse qualitative des communautés, complémentaire à ce travail quantitatif, et nous proposons une méthodologie de choix d'une partition (d'une méthode), qui comme nous le verrons, est contextuelle à une étude donnée, encadrée par des hypothèses et une ou plusieurs questions de recherche.

FIGURE 2.7 – Similarité entre méthodes de détection de communautés quantifiée par différentes métriques de validation, basée sur les partitions découvertes sur l'ensemble du jeu de données. Les lignes et les colonnes sont ordonnées selon une méthode de clustering hiérarchique (Joe H. Ward, 1963). Dans l'ordre, le score moyen de (a) NMI, (b) AMI, (c) RI, (d) ARI.



Chapitre 3

Analyse qualitative des communautés

3.1 Introduction

Dans ce chapitre, nous abordons la problématique du choix de méthode de détection de communauté du point de vue de l'analyste, l'utilisateur final qui cherche à analyser un réseau pour répondre à une question de recherche. Nous ciblons plus spécifiquement les chercheurs en Sciences Humaines et Sociales (SHS).

Nous avons vu précédemment que les algorithmes à disposition sont nombreux et variés, et leurs fondements théoriques parfois complexes à appréhender. Et quand bien même serait-ils simples à comprendre, cela n'aide pas nécessairement à entrevoir *a priori* la *nature* des communautés détectées.

Or un analyste s'intéresse bien à la forme des communautés, à leur structure, pour en dégager des modes d'organisation, et souvent rapprocher ces organisations types d'autres critères comme l'efficacité du travail collaboratif, le caractère innovant d'une équipe, la viralité marketing, la prédiction d'achat, etc. (Mercanti-Guérin, 2010). Nous nous intéressons donc ici aux mesures *qualitatives* qui vont renseigner l'analyste sur la topologie des clusters découverts.

D'un point de vue méthodologique, utiliser la détection de communautés en SHS requiert de :

— Pouvoir décrire la topologie des communautés,
— Permettre d'explorer des topologies variées pour mettre en lumière tout type d'organisation.

En ce qui concerne le premier point, nous proposons une approche basée sur des cartes bivariées, permettant de décrire les partitions avec des métriques intuitives décrivant l'organisation interne et externe des communautés ; cette technique nous permet de décrire les communautés d'une part, mais plus généralement, appliquée

sur l'ensemble de notre jeu de données, cela nous permet également de dégager des familles de réseaux réels.

Pour adresser le deuxième point, nous proposons une méthodologie de choix de méthode permettant d'offrir le plus large éventail possible de communautés.

3.2 L'analyse de réseaux sociaux en Sciences Humaines et Sociales

Au sein de toute organisation, les collaborateurs interagissent et forment des groupes, avec leurs coutumes, leurs tâches, leurs devoirs. Ces réseaux sont pour la plupart implicites et non décrits formellement dans l'organisation. Des chercheurs, de plus en plus nombreux, notamment en Sciences de Gestion, se sont intéressés à ces réseaux informels. Il peut en coexister plusieurs : le réseau d'information reflétant l'échange d'informations, le réseau représentant l'accessibilité aux connaissances, le réseau d'échange de connaissances entre collègues qualifiés ou experts et les autres, le réseau de conseil résultant de l'entraide dans la résolution de problèmes, le réseau d'amitié, etc.

La question principale de ces études est de comprendre l'impact des pratiques informelles sur l'organisation globale, l'objectif étant de fournir des suggestions pour améliorer les processus : par exemple améliorer les flux de partage des connaissances pour augmenter la coopération ou, aligner les processus organisationnels formels sur les processus informels efficaces.

L'analyse de réseaux sociaux peut faire partie de la boîte à outils utilisée pour établir un diagnostic. (Toni et Nonino, 2010) proposent un cadre d'une telle analyse. Ils ont collecté trois réseaux par le biais de questionnaires auprès des employés d'une entreprise italienne : le réseau de connaissance, le réseau de conseil, le réseau d'accès. Dans un premier temps, pour chaque réseau, ils analysent les relations au sein et entre les départements et identifient les rôles clés. Ils utilisent des mesures de centralité du réseau et la détection des communautés. Ils identifient les rôles informels clés, à savoir : *les leaders d'opinion*, *les connecteurs centraux*, *les goulots d'étranglement*, *les experts*, *les consultants*, ou *les personnes utiles*, et les caractérisent. Par exemple, les collègues qui sont considérés comme des experts (degré de centralité le plus élevé dans le réseau de connaissances) se trouvent être les responsables de différentes unités commerciales. Mais les personnes ayant les degrés les plus faibles ne sont pas nécessairement les moins qualifiées, peut-être leurs collègues ignorent-ils leurs compétences et les considèrent-ils comme non qualifiés. Dans un deuxième temps, ils ont construit un réseau conjoint en multipliant les trois matrices d'adjacence. Ils ont alors trouvé un nouveau rôle, appelé *pilus prior* (premier lanceur) qui cumule les caractéristiques de résolution de problèmes, d'expertise et d'accessibilité. Le cadre est synthétisé dans la figure 3.1.

Une fois la structure décrite, des recommandations peuvent être faites pour formaliser les pratiques informelles et remodeler l'organisation officielle. Bien sûr, chaque entreprise, chaque contexte est différent et toute intervention doit être étudiée au cas par cas. Par exemple, si nous considérons le cas de collaborateurs qui se trouvent dans des positions isolées. Les raisons de cette situation peuvent être diverses : l'un d'entre eux peut rencontrer des difficultés à s'intégrer sur le lieu de travail ; l'élargissement de son équipe, par exemple en l'impliquant dans différents projets, peut l'aider à créer des liens avec d'autres acteurs. Ses compétences sont peut-être méconnues, et la création d'une base de données de profils de compétences permettrait de promouvoir cette personne en tant qu'*expert*, en délestant les activités d'expertise

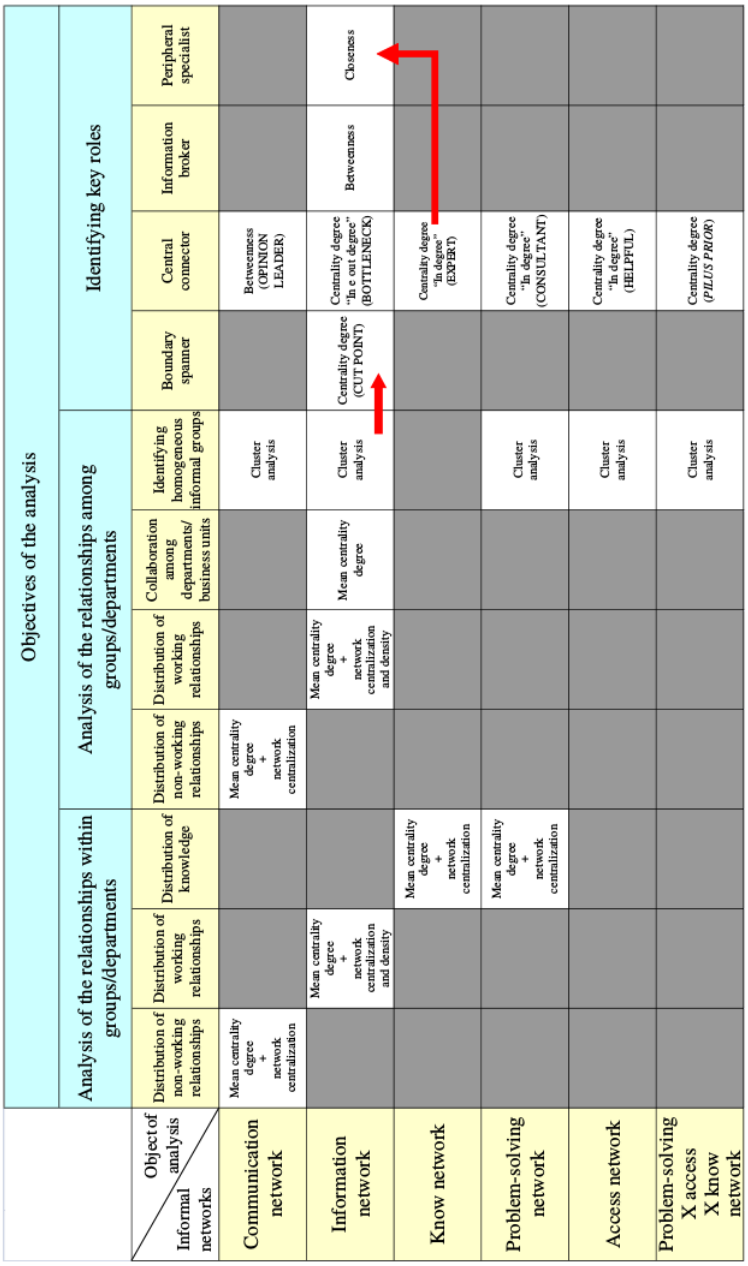


FIGURE 3.1 – Cadre d'analyse des réseaux informels et d'identification des rôles informels clés. Figure reproduite de (Toni et Nonino, 2010) avec la permission des auteurs.

d'un expert reconnu confronté à de trop nombreuses demandes (identifié comme un *goulot d'étranglement*).

Cette section ne constitue pas une étude exhaustive de la manière dont les chercheurs en sciences sociales utilisent les outils de l'analyse de réeaux sociaux. Mais la plupart des études de cas ne considèrent que les positions individuelles. Elles identifient les acteurs et qualifient leur(s) rôle(s). Selon le contexte de recherche (groupes de discussion, plateforme d'apprentissage, communautés épistémiques comme Wikipédia, communautés de marque, etc.), ils proposent des typologies *ad hoc* de types de membres, dont un exemple de catalogue a été fait par (Benamar, Balagué et Ghassany, 2017).

D'autres études se concentrent sur la perspective globale et utilisent des mesures telles que la distribution du degré, le diamètre ou la centralisation du réseau pour décrire une connectivité globale. Par exemple une configuration en étoile peut conduire à des structures fragiles, avec un manque de partage du pouvoir. Autre exemple, identifier les trous structuraux dans le réseau là où l'on s'attendrait à des relations peut refléter un manque de cohésion entre équipes (Krackhardt et Hanson, 1993). Un conseil aux managers pourrait être d'organiser des réunions périodiques inter-équipes afin de mieux partager informations et ressources.

Mais le niveau intermédiaire "méso" n'est pas si souvent considéré. Qu'apportent les communautés dans l'analyse ? Dans (Toni et Nonino, 2010) les auteurs soulèvent le problème de l'identification de groupes informels homogènes pour analyser les relations *au sein* et *entre* les groupes dans l'entreprise qu'ils étudient. Ils affirment que *"l'identification de groupes homogènes peut être très importante, surtout lorsque l'entreprise est confrontée à des changements"*. Mais ils n'utilisent pas vraiment le concept de "groupes informels homogènes" dans leur diagnostic, pas plus que celui annoncé d'"analyse en clusters". Le concept reste ambigu, et fait parfois référence à des départements, à différentes filiales/divisions (formelles) d'une entreprise, mais également, parfois, à des zones plus informelles dans un réseau connecté par un nœud de courtage ; dans ce cas, bien souvent, l'accent est mis sur le nœud et non sur les zones qu'il/elle connecte.

Pourquoi la structure méso, entre la position/le rôle de l'acteur et l'image globale, est-elle si difficile à mobiliser ? Dans ce chapitre, nous proposons des mesures et une méthodologie pour identifier et caractériser ce type de structure. Nous fournissons des outils pour décrire une communauté comme une organisation. Nous aidons à identifier quelles sont les dimensions structurelles qui pourraient se rapporter à un groupe organisé avec des règles, un but, une efficacité ou toute autre capacité d'émergence.

3.3 Les métriques de qualité de partitions

Caractériser les communautés dans un réseau permet aux analystes de discerner les différents types de structures qui le sous-tendent. Ceci dans le but de décrire et comprendre l'organisation communautaire d'un réseau, ou bien de s'appuyer sur celle-ci pour agir. Par exemple, dans un article récent, les chercheurs montrent que l'immunisation et le contrôle des épidémies sont plus efficaces lorsqu'ils exploitent les propriétés topologiques des communautés. Calculer des mesures de centralités globales sur l'ensemble du réseau est moins efficace que de détecter les hubs qui, s'ils sont infectés vont contaminer leur communauté ; si en plus ces nœuds sont des passerelles vers d'autres communautés, denses, ils peuvent propager très rapidement l'épidémie. Les auteurs tiennent compte du nombres de communautés potentiellement accessibles, leur densité et leur taille pour calculer l'influence des nœuds (Ghalmane, El Hassouni et Cherifi, 2019).

La définition d'une communauté ne faisant pas consensus, on peut imaginer de la même façon que la notion de structure communautaire varie d'un contexte à l'autre. On ne s'attend pas à ce qu'un ensemble fini de caractéristiques (le terme de "goodness metrics" est parfois employé en anglais) puisse correspondre à chaque intuition de ce qu'est une "bonne" communauté, ou une communauté "efficace". De plus le choix d'un ensemble de métriques serait controversé à moins qu'un contexte spécifique ne soit clairement défini. Parallèlement, selon le contexte, il arrive que l'on conçoive des métriques de manière *ad hoc*. Par conséquent, afin de rendre l'analyse

aussi générique que possible, nous restreignons notre liste de métriques de qualité en appliquant les critères suivants, de la plus haute à la plus basse priorité :

— Puisque nous caractérisons les communautés dans différents types de réseaux, sans contexte particulier, nous ne nous intéressons qu'aux métriques qui décrivent les communautés elles-mêmes, et non relativement à la structure globale des réseaux où elles se trouvent (comme le Cut ratio (Yang et Leskovec, 2015), la Modularity (Newman et Girvan, 2004) ou la Description Length (Rosvall et Bergstrom, 2008)) même si leur efficacité ne peut être ignorée.
— Les métriques doivent être relativement peu corrélées les unes aux autres afin d'illustrer des aspects différenciant des structures communautaires.
— Une métrique doit refléter la structure interne d'une communauté.
— Une métrique dont les concepts peuvent être représentés intuitivement et visuellement est préférable à une métrique qui reflète des idées statistiques difficiles à représenter.

Selon Yang et Leskovec, 2015, on peut distinguer les métriques de qualité ("scoring functions") qui définissent le profil interne et externe d'une communauté, et celles qui impliquent le profil de nœuds particuliers, tels que les connecteurs qui partagent des relations au sein et en dehors de la communauté.

Comme nous chercherons dans la suite à comparer des communautés issues de différentes méthodes de détection et de différents réseaux, nous privilégions les métriques relatives, telle que la fraction des arêtes/nœuds particuliers par rapport à l'ensemble des arêtes/nœuds appartenant à une communauté. Les mesures absolues, telles que le degré moyen des arêtes internes, ne permettent en effet pas de différencier les communautés, car on peut en trouver de toutes tailles dans une seule partition.

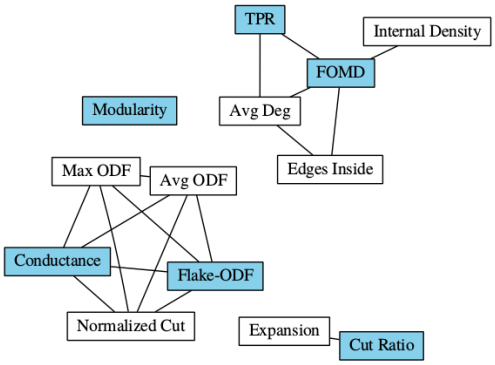


FIGURE 3.2 – 4 groupes de mesures de qualités basées sur leur corrélation (Yang et Leskovec, 2015). Les scores de ces 13 métriques ont été calculés sur chacune des 10 millions de communautés provenant de la groudtruth de 230 réseaux sociaux réels, tels que LiveJournal, Friendster, Amazon ou DBLP. Deux métriques sont connectées si elles sont corrélées. En bleu sont sélectionnées des représentants de ces clusters.

La Table 3.1 décrit quelques mesures de qualité, dont celles présentées dans la Figure 3.2. Pour un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ composé de $n = |\mathcal{V}|$ nœuds et $m = |\mathcal{E}|$ arêtes, nous considérons une communauté C de n_C nœuds comme un sous-graphe de $\mathcal{G}$ dans une partition P. Soit m_C le nombre d'arêtes internes à la communauté C, $m_C = |(i,j) \in \mathcal{E} : i \in C, j \in C|$, l_C le nombre d'arêtes externes qui connectent C à des nœuds extérieurs à C, $l_C = |(i,j) \in \mathcal{E} : i \in C, j \notin C|$. $d(i)$ est le degré de i, $d_{int}(i)$ son degré interne. Enfin, pour calculer le coefficient de clustering, nous notons Δ_C le

nombre de triangles dans la communauté C et T_C indique le nombre de cliques de tailles 3 possibles dans C.

Type	Nom	Formule	Description
Ext	Conductance	$\frac{l_C}{2m_C+l_C}$	Proportion des arêtes de C qui pointent vers l'extérieur
	MaxODF	$max_{i\in C}\frac{\lvert\{(i,j)\in l_C\rvert}{d(i)}$	Maximum Out Degree Fraction : valeur maximale de proportion de degré externe atteinte par un membre de C
	Cut Ratio	$\frac{l_C}{n_C(n-n_C)}$	Densité externe (à opposer à densité interne à la communauté)
	Expansion	$\frac{l_C}{n_C}$	Degré sortant relatif de C par rapport à sa taille
Int	FOMD	$\frac{\lvert i:i\in C,\lvert\{(i,j)\in m_C\}\rvert>d_m\rvert}{n_C}$	Fraction Over Median Degree : proportion de nœuds de C dont le degré interne est supérieur au degré médian du graphe G
	CCF	$\frac{3\Delta_C}{T_C}$	Coefficient de clustering : probabilité que deux voisins d'un membre de C soient deux membres eux-mêmes connectés.
	sc_den	$\frac{2m_C}{n_C-1}$	Scaled density : densité d'arêtes internes normalisée, relativement à la taille de C, permettant de comparer petites et grandes commmunautés
	hub_dom	$\frac{max_{i\in C}d_{int}(i)}{n_C-1}$	Hub Dominance : niveau de centralisation de la communauté C

TABLE 3.1 – Mesures de qualité décrivant différents aspects de structuration de communautés

Il existe de nombreuses métriques caractérisant la connectivité interne, ou externe, en plus de celles qui sont présentées ci-dessus. Vinh Loc Dao en présente une quinzaine dans sa thèse, Chapitre 4 (Dao, 2018). Dans (Dao, Bothorel et Lenca, 2021), nous menons une analyse empirique voisine de celle présentée par (Yang et Leskovec, 2015), sur les partitions présentées dans le Chapitre précédent issues de la quinzaine d'algorithmes de détection de communautés appliquées sur notre centaine de

réseaux.

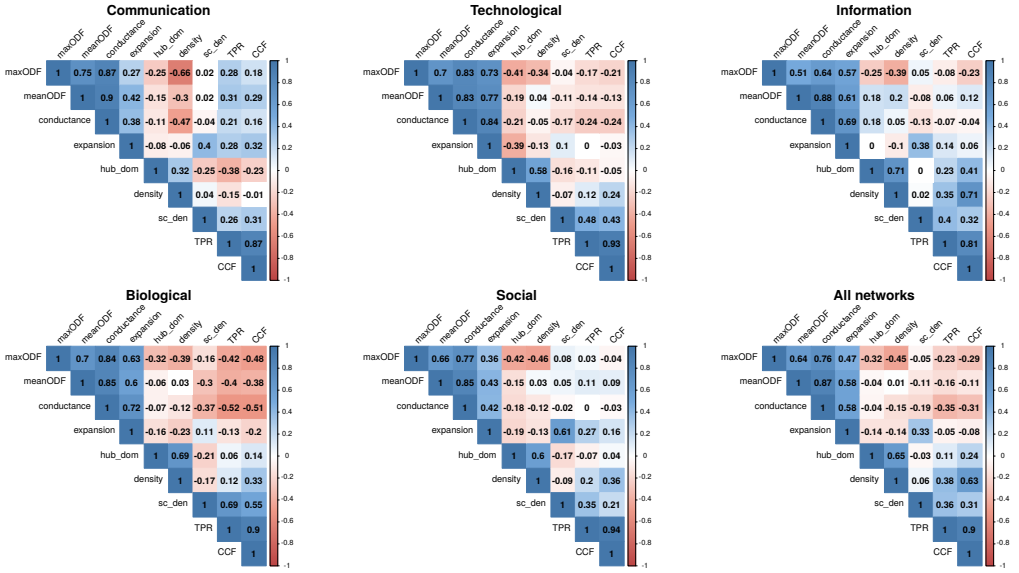


FIGURE 3.3 – Les corrélations de Pearson entre métriques communautaires. Ces corrélations sont calculées sur la base des scores des métriques mesurées sur les communautés qui contiennent au moins 3 nœuds. Les corrélations des métriques sont analysées par type de réseaux. Les métriques de qualité sont présentées dans les 6 sous-figures dans le même ordre pour une observation comparative. Les scores de corrélation dont le niveau de signification estimé est faible ($p - value > 0,01$) sont reproduits sur un fond blanc.

Comme nous pouvons le voir sur la Figure 3.3, il existe deux groupes où les métriques sont systématiquement corrélées les unes aux autres. Le premier groupe comprend *maxODF*, *meanODF* et *conductance* qui représentent la connexion externe de la communauté avec des coefficients de corrélation très élevés (sauf pour *maxODF* et *meanODF* dans les réseaux d'information avec une relation relativement faible de $0,51$). En outre, la métrique *expansion* appartient également à ce groupe dans les réseaux technologiques, d'information et biologiques avec des scores de corrélation élevés et de manière moins marquée dans les autres types de réseaux. Le deuxième groupe se compose de *TPR* et *CCF* qui exposent des structures triadiques étroitement liées avec des scores de corrélation très élevés dans chaque catégorie de réseau. Le score de corrélation le plus faible entre *TPR* et *CCF* est de $0,81$ dans les réseaux d'information et d'environ $0,90$ dans tous les autres cas. Sans perdre la généralité, dans notre analyse, ces 2 groupes de métriques pourraient être réduits à deux métriques paragon représentatives de ces deux propriétés structurelles.

La hub dominance *hub_dom* est la seule métrique qui est tout à fait indépendante de toutes les métriques des deux groupes précédents dans chaque catégorie de réseau. Le score de corrélation le plus élevé entre *hub_dom* et ces métriques est de $0,42$ avec *maxODF* dans les réseaux sociaux, ce qui reste une corrélation relativement faible. Cette dernière, cependant, est généralement corrélée avec *density* sauf dans le cas des réseaux de communication où elles sont tout à fait orthogonales. En revanche, la densité normalisée *sc_den* présente une association hétérogène avec les autres métriques dans toutes les catégories de réseaux étudiées. Elle est proche de *CCF* et *TPR* dans les réseaux biologiques mais se rapproche de *expansion* dans les réseaux sociaux.

Metrics	Common concept
maxODF,meanODF,conductance	External activeness
expansion	External connectivity
hub_dom	Centralized connectivity
density	Internal edge density
sc_den	Average internal density
CCF, TPR	Internal triadic closure

TABLE 3.2 – Groupes de mesures de qualité qui reflètent différents aspects de la structuration des communautés. Deux métriques appartiennent à une même catégorie si elles présentent une corrélation élevée sur les ensembles de communautés de notre jeu de données. La colonne *Common concept* précise les caractéristiques structurelles communes.

Sur la base de cette analyse, les mesures de qualité des communautés ci-dessus peuvent être regroupées en 6 classes présentées dans le tableau 3.2 en fonction de leurs corrélations sur l'ensemble des communautés étudiées.

Nous proposons dans la suite de ce travail de décrire la structure d'une communauté en utilisant une combinaison de deux métriques. Une combinaison croisée de métriques issues de différents clusters peut être particulièrement riche. En particulier, les structures internes et externes des communautés ne sont généralement pas corrélées, reflétant des facettes différentes des structures communautaires.

3.4 Cartes bivariées pour caractériser l'anatomie des communautés

L'association de métriques de qualité nous permet de comparer différents réseaux selon la nature de leurs communautés, et cela de manière à ce que les informations structurelles soient exposées de manière inuitive aux analystes.

Dans la Table 3.3, nous présentons un exemple d'association de métriques qui permet de décrire des types de communautés. En prenant des paires de métriques non corrélées, comme *FOMD* et *conductance* (d'après Yang et Leskovec, 2015, Figure 3.2), nous décrivons des dimensions différentes et complémentaires de la nature organisationnelle. Lorsque la *conductance* est faible, les communautés sont assez isolées. Une *FOMD* élevée implique que la majorité des membres d'une communauté sont fortment connectés entre eux, ce qui peut signifier une bonne cohésion ou une forte activité. Un groupe qui présente une faible *FOMD* et une forte *conductance* doit être plutôt rare. En effet, une connectivité interne faible mais externe forte ne devrait pas être proposée par un algorithme de détection de communautés.

Nous verrons dans la suite que ce principe d'association de métriques nous permet d'une part de caractériser des réseaux et de les comparer, mais également, au sein d'un réseau, de comparer différentes partitions obtenues grâce à des algorithmes.

	Faible FOMD	**Forte FOMD**
Forte conductance	Communauté inactive mais très tournée vers l'extérieure (forme dégénérée ?)	Forme active et ouverte
Faible conductance	Communauté inactive et repliée sur elle-même	Communauté active et indépendante

TABLE 3.3 – Exemple de composition de mesures de qualité. Les paires associent des clusters différents tels que présentés par Yang et Leskovec, 2015

Réseau	**N**	**E**	**C**	**S**	**O**	$\bar{\mu}$	**Groundtruth**
Livejournal	4.0M	34.7M	664414	10.79	6.24	**0.95**	User-defined communities
Youtube	1.13M	3.0M	16386	7.89	2.45	**0.91**	User-defined groups
DBLP	0.32M	1.05M	13477	53.41	2.76	**0.62**	Publication venues
Amazon	0.33M	0.93M	75149	30.22	7.16	**0.58**	Product categories

TABLE 3.4 – Description des réseaux : N nombre de nœuds, E d'arêtes, C nombre de communautés groundtruth, S taille moyenne des communautés, O nombre de communautés par nœud, $\bar{\mu}$ conductance moyenne (Yang et Leskovec, 2015) des communautés. http://snap.stanford.edu/data/

3.4.1 Structures organisationnelles dans les vérités terrain

Dans ce premier travail, nous prenons des réseaux pour lesquels une vérité terrain est disponible, i.e. nous n'appliquons pas d'algorithme de détection de communautés, mais nous analysons la partition *groundtruth*.

La Table 3.4 liste les 4 réseaux que nous prenons en exemple. Si nous nous intéressons à une mesure de qualité conventionnelle, comme la conductance $\bar{\mu}$, nous pouvons décrire une partition de manière globale. Ici, la conductance indique qu'il y a plus de 90% d'arêtes dans *Livejournal* et *Youtube* qui traversent des communautés, tandis que ces chiffres sont d'environ 60% dans les réseaux *DBLP* et *Amazon*. Cette métrique est en soi informative, mais elle ne permet pas de distinguer finement les différences entre réseaux d'une part, et d'autre part, entre communautés.

En reprenant l'exemple de composition de mesures de la Table 3.3, appliquée à ces 4 réseaux bien connus, nous obtenons les cartes de chaleur de la Figure 3.4 où nous voyons clairement qu'*Amazon* et *DBLP* sont relativement similaires, comme la conductance l'indiquait. Mais nous voyons également en quoi ils sont similaires : ils contiennent des communautés variées appartenant aux 4 types attendus. Il est intéressant de noter qu'*Amazon* contient des formes de communautés "contre nature", i.e. faiblement connectées en interne et parfois connectées vers l'extérieur. Il est donc probable qu'utiliser des algorithmes sur ce réseau (et dans une moindre mesure sur *DBLP*) ne permettra pas de retrouver ces formes de regroupements "dégénérés", et qu'alors, utiliser la groundtruth pour valider ces algorithmes conduira à une évaluation mitigée sur ces 2 datasets (nous avons d'ailleurs démontré que les

partitions groundtruths ne sont pas "bien formées" selon les métriques de qualité de partitionnement (Dao, Bothorel et Lenca, 2017a)).

Livejournal et *Youtube*, également similaires du point de vue de la conductance, ne présentent quasi exclusivement que des communautés actives isolées. La faible variation de conductance (0.04 points) cache cependant la présence de communautés un peu plus isolées pour *Youtube*, communautés qu'il peut être intéressant à étudier car elles sont fermées et reflètent des groupes qui ne partagent que peu de membres.

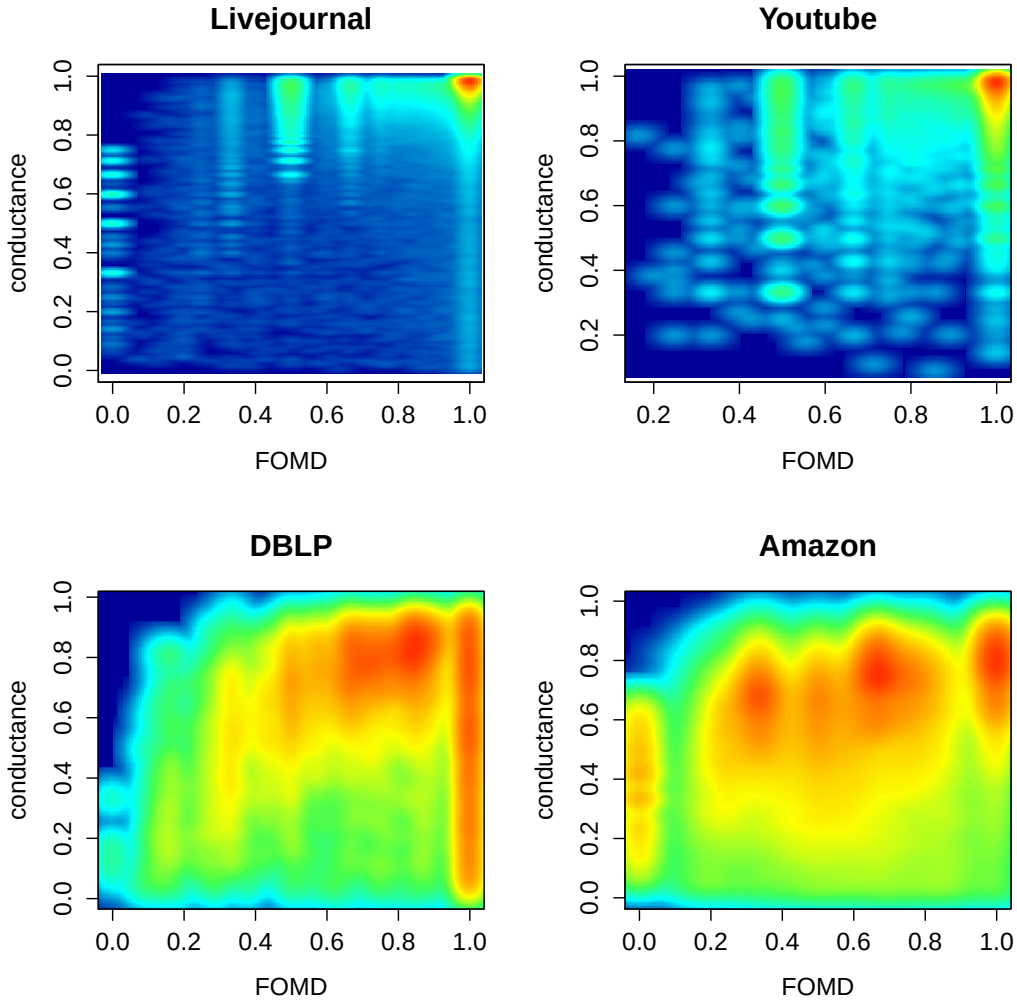


FIGURE 3.4 – Les partitions *YouTube* et *Livejournal* montrent des communautés actives qui sont très bien connectées aux autres communautés. Nous pouvons trouver différents types de communautés dans les vérités terrain d'*Amazon* et de *DBLP* : beaucoup de communautés actives avec une FOMD assez élevée, bien connectées à d'autres sujets/domaines (conductance élevée), mais nous en trouvons aussi quelques-unes plus indépendantes (conductance plus faible) qui montrent une large gamme de valeurs de FOMD, c'est-à-dire qu'elles sont plus ou moins bien inter-connectées.

Dans la même veine, nous proposons une méthodologie pour décrire les communautés par le biais de liens intra-cluster et inter-cluster (Dao, Bothorel et Lenca, 2017b). Nos résultats montrent que la composition des communautés dans les réseaux du monde réel, là encore, expose une diversité de modèles structurels d'une

part, et que les communautés présentent des variations parfois inattendues autour de la définition d'une communauté type que l'on cherche à détecter via les algorithmes, i.e. des groupes avec une grande majorité de liens internes et seulement quelques liens externes.

Nous mobilisons ici une combinaison de deux mesures basées sur *ODF* (Out Degree Fraction) que nous calculons sur l'ensemble des communautés de chaque réseau.

Le *out degree fraction* d'un noœud i de la communauté C est mesuré par le ratio entre liens externes l_C impliquant i et son degré :

$$ODF_C(i) = \frac{|\{(i,j) \in l_C|}{d(i)}$$

Pour qualifier une communauté, il est intéressant de décrire la fraction moyenne des degrés sortants de cette communauté d'une part, mais aussi comment ceux-ci sont distribués sur les nœuds. Nous calculons donc la moyenne et l'écart type des valeurs de *ODF* des nœuds d'une communauté :

$$meanODF(C) = \frac{\sum_{i \in C} ODF_C(i)}{n_C}$$

$$sdODF(C) = \left(\frac{\sum_{i \in C} [ODF_C(i) - meanODF(C)]^2}{n_C - 1} \right)^{1/2}$$

Un *meanODF* faible implique que les membres de la communauté se connectent principalement entre eux, alors qu'une valeur élevée de *meanODF* signifie que les nœuds se connectent de préférence aux nœuds d'autres communautés. Un *meanODF* faible reflète une structure (assortative en anglais) et un *meanODF* élevé (disassortative). Une valeur moyenne de *meanODF* dans ce cas-ci signifie une structure hybride de la communauté comme le montre la Figure 3.5.

L'écart-type d'une variable nous aide à comprendre la fluctuation de ses valeurs. Une faible valeur *sdODF* implique que les degrés sortants de la communauté sont distribués de façon homogène entre les nœuds. En revanche, une valeur *sdODF* élevée démontre une diversité des modèles de connexion externe des nœuds. En d'autres termes, en se basant sur la valeur *sdODF* d'une communauté, on peut déterminer s'il existe une répartition claire des rôles (Guimera et Amaral, 2005) différentiant les nœuds de la communauté ou, au contraire, si les nœuds sont équivalents dans leur fonction d'ouverture vers l'extérieur.

Après le choix de seuil pour distinguer différents niveaux des deux métriques, nous pouvons répartir les communautés dans les catégories suivantes :

— *Conventionnal communities* (S1 - faibles *meanODF* et *sdODF*) : Cette structure correspond à la définition traditionnelle de la communauté où la majorité des arêtes se situent à l'intérieur des communautés. La plupart des méthodes actuelles de méthodes de détection de communautés sont basées sur cette notion. De plus, les degrés sortants de la communauté sont répartis de manière homogène sur ses nœuds.

— *Casual communities* (S2 - *meanODF* moyenne et faible *sdODF*) : La structure modulaire n'est pas très claire dans ce type de communauté "lâche" puisqu'il n'y a pas une propension claire à la connexion interne.

— *Extrovert communities* (S3 - forte *meanODF* et faible *sdODF*) : Les membres de ces communautés sont dans l'ensemble fortement connectés vers l'exterieur.

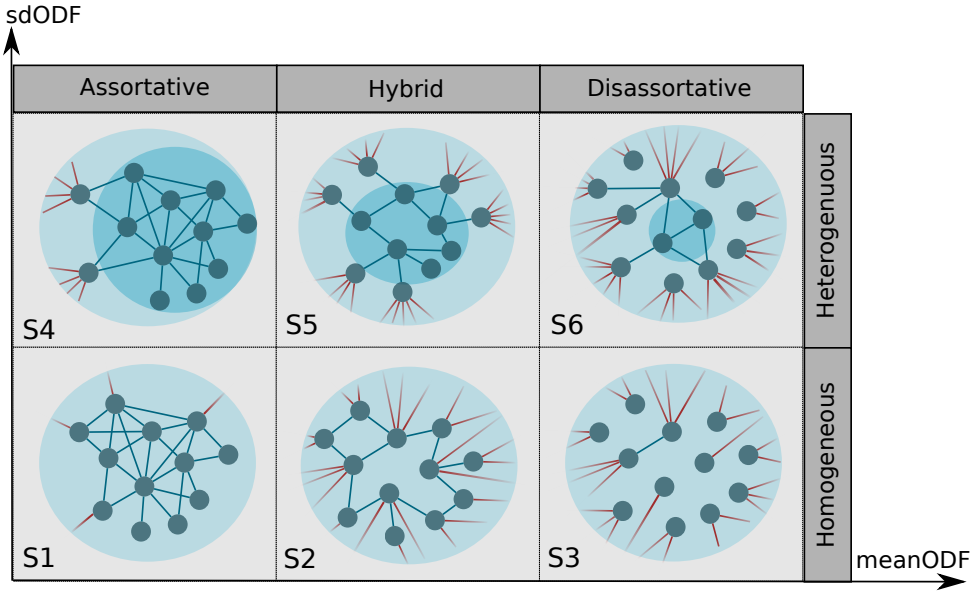


FIGURE 3.5 – Six structures communautaires représentatives mesurées par la fraction de degrés sortants (*meanODF* and *sdODF*). Les arêtes bleues représentent les connexions intra-communautaires et les rouges les relations extra-communautaires (ou inter-communautaires) . Les zones bleu foncé illustrent un arrangement core-periphery dans les structures *S*4, *S*5, *S*6.

— *Full-core communities* (S4 - faible *meanODF* and forte *sdODF*) : Ce groupe de communautés présente une similitude avec celles de structure *S*1 puisque toutes deux possèdent des connexions internes relativement denses. La seule distinction entre les structures *S*1 et *S*4 est que *S*4 contient un petit nombre de nœuds *frontière* qui attirent la plupart des liens externes. Ces connecteurs forment une zone périphérique, tandis que la majorité des membres constituent un noyau dense.
— *Half-core communities* (S5 - *meanODF* moyenne et forte *sdODF*) : Ces communautés présentent également une structure core-periphery, mais il n'y a plus de domination quantitative des nœuds du noyau sur les nœuds de la périphérie, comme c'est le cas en *S*4.
— *Seed-core communities* (S6 - fortes *meanODF* et *sdODF*) : La structure core-periphery est ici plus ou moins effacée car les nœuds frontière dominent en nombre. Cette structure présente de nombreuses similitudes avec les structures *S*3 et *S*5 et peut être considérée comme un état intermédiaire entre *S*3 et *S*5.

Nous obtenons une cartographie de ces réseaux sur la Figure 3.6 et la répartition des communautés par type de structure dans la Table 3.5.

Dans cette nouvelle expérience, nous voyons apparaître également une grande diversité d'anatomies de réseaux. Contrairement à l'intuition, les vérités terrain n'ont pas toutes le même genre d'organisation, ce qui nous permet de dire à nouveau qu'un algorithme ne pourra pas détecter de manière universelle toute vérité terrain, et nous conforte dans l'idée que notre travail exposé dans le Chapitre 3.5 est utile : à chaque jeu de données, il s'agit de trouver l'algorithme adéquat.

De plus, là encore de manière non intuitive, les structures de type core-periphery ne sont pas toujours présentes. Alors que dans *Livejournal* et *Youtube*, la majorité

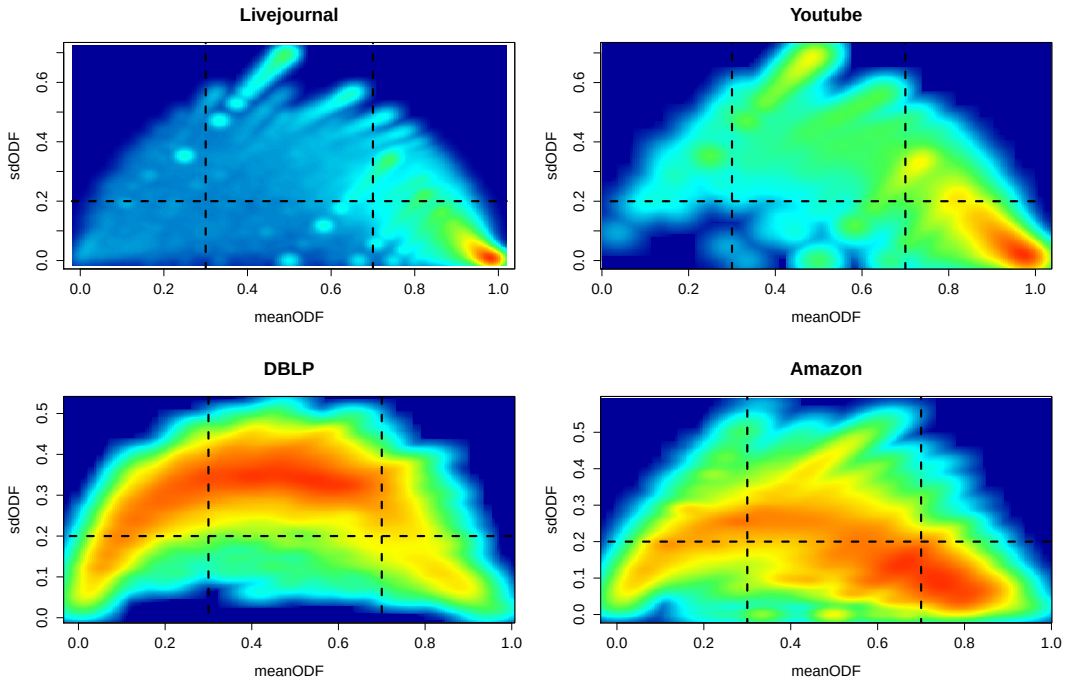


FIGURE 3.6 – Densité des communautés de la groundtruth dans un espace défini par *meanODF* et *sdODF*. Les lignes en pointillés représentent les seuils entre les structures *S*1 à *S*6 (cf. Figure 3.5) : 0.3 et 0.7 pour *meanODF* et 0.2 pour *sdODF*

Network	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Livejournal	0.29	0.74	**90.17**	0.31	3.88	4.61
Youtube	0.08	2.36	**65.36**	1.37	**17.55**	**13.28**
DBLP	6.28	2.07	4.87	**23.44**	**57.86**	5.48
Amazon	8.33	**31.13**	**23.57**	9.13	**26.63**	1.21

TABLE 3.5 – Répartition des communautés de la vérité terrain (en pourcentage) dans un espace défini par *meanODF* et *sdODF*. Seuils entre les structures *S*1 à *S*6 : 0.3 et 0.7 pour *meanODF* et 0.2 pour *sdODF*

des communautés sont concentrées dans un même type de structure, celles des réseaux *DBLP* et *Amazon* sont beaucoup plus variées. Nous constatons que la structure $S3$ occupe environ 90% et 65% des communautés dans les réseaux *Livejournal* et *Youtube* respectivement. Cela implique que la plupart des utilisateurs de ces réseaux ont généralement des amitiés en dehors de leurs communautés plutôt qu'à l'intérieur. En outre, certains groupes sont centrés autour d'un cœur actif et interne ($S5$ et $S6$ dans une moindre mesure).

Dans les réseaux *DBLP* et *Amazon*, bien qu'il y ait toujours une dominance de certaines structures, nous remarquons une répartition plus équilibrée. Dans le cas de *DBLP*, près de 60% des lieux de publication ($S5$) attirent différents types d'auteurs en termes de profil de coopération : les conférences en effet rassemblent à la fois des auteurs reconnus de la communauté scientifique, mais également des nouveaux venus, ou des chercheurs pluridisciplinaires. En même temps, environ 23.44% des conférences attirent principalement des spécialistes du domaine et très peu de chercheurs ayant des liens vers d'autres domaines ($S4$). Sur le réseau *Amazon*, la forte présence des structures $S2$ et $S3$ explique que les produits sont plus souvent co-achetés avec ceux d'autres catégories. Mais il existe également de nombreuses catégories de produits qui conduisent à des achats dans les mêmes catégories ($S1$, $S4$, $S5$). Bien sûr, sans une analyse fine des conférences ou des produits dont il est question, nous ne pouvons en dire davantage ici, mais cette démarche représente une étape intéressante pour explorer les données.

3.4.2 Les familles de réseaux

La même démarche peut être appliquée sur les partitions résultant d'algorithmes de détection de communauté. Nous avons focalisé cette fois notre étude sur l'organisation interne des communautés. Dans ce travail, nous avons utilisé l'ensemble des communautés détectées par notre quinzaine d'algorithmes sur notre centaine de réseaux, comme décrit dans le chapitre précédent (Dao, Bothorel et Lenca, 2021).

Nous avons sélectionné deux métriques internes non corrélées, *hub_dom* et *CCF*. Le coefficient de clustering (parfois appelé *transitivity*) est une métrique bien connue qui est généralement utilisée pour évaluer la structure modulaire des réseaux. Elle est basée sur le concept selon lequel les paires de nœuds ayant des voisins communs sont plus susceptibles d'être connectés (Barrat et al., 2004). La hub dominance quant à elle qualifie la centralisation de la communauté autour d'un nœud. Les arêtes internes d'une communauté peuvent en effet être distribuées de différentes manières autour de ses nœuds, soit en se concentrant autour d'un petit nombre de nœuds fortement centralisés, soit en se répartissant uniformément autour de chaque nœud. Plus la hub dominance est élevée, plus il est probable que la communauté présente une structure de type hub (Lancichinetti et al., 2010 ; Labatut et Orman, 2017).

La Table 3.6 et la Figure 3.7 illustrent les structures que ces deux métriques permettent de mettre en lumière en séparant l'espace en 4 quadrants. Les frontières entre les différentes topologies ne sont généralement pas évidentes et doivent être définies en fonction du contexte. La taille de la communauté caractéristique doit être prise en considération lors de la définition des seuils, car plus la taille de la communauté est grande, plus il est probable que les hubs et les cliques deviennent moins importants, ce qui signifie que des seuils plus bas seront à privilégier.

Il est parfois difficile de nommer les types d'organisation trouvés dans nos cartes bivariées (cf. nos 6 structures dans l'espace *meanODF* et *maxODF* de la Figure 3.5). Dans cette étude, nous avons cherché à rapprocher les organisations internes de modèles bien connus de la littérature des réseaux complexes :

Transitivity	Hub dominance	Topology
Low	Low	String-based
High	Low	Grid-based
Low	High	Star-based
High	High	Clique-based

TABLE 3.6 – Quatre topologies distinctes caractérisées par la *Transitivity* (*CCF*) et la hub dominance (*hub_dom*).

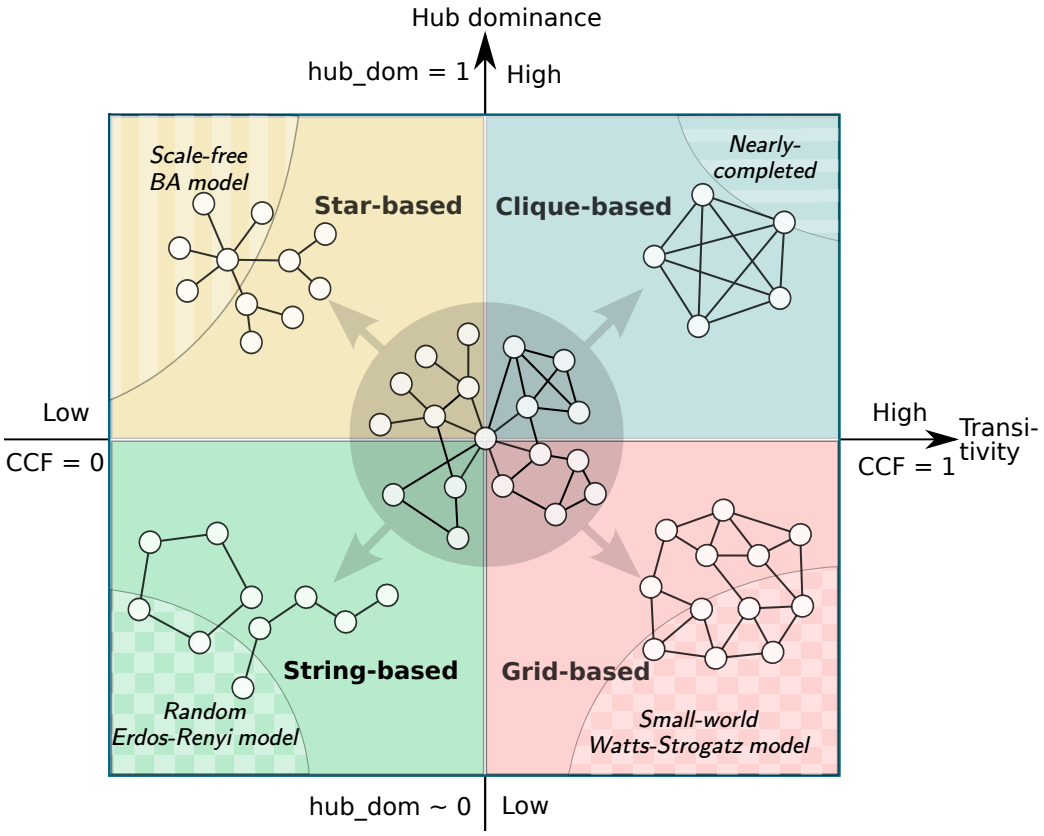


FIGURE 3.7 – Une catégorisation de la structure interne des communautés selon deux dimensions de propriétés structurelles : la hub dominance et la transitivité représentées par *hub_dom* et *CCF* respectivement. Quatre communautés topologiques représentatives sont illustrées en fonction de leurs scores correspondants.

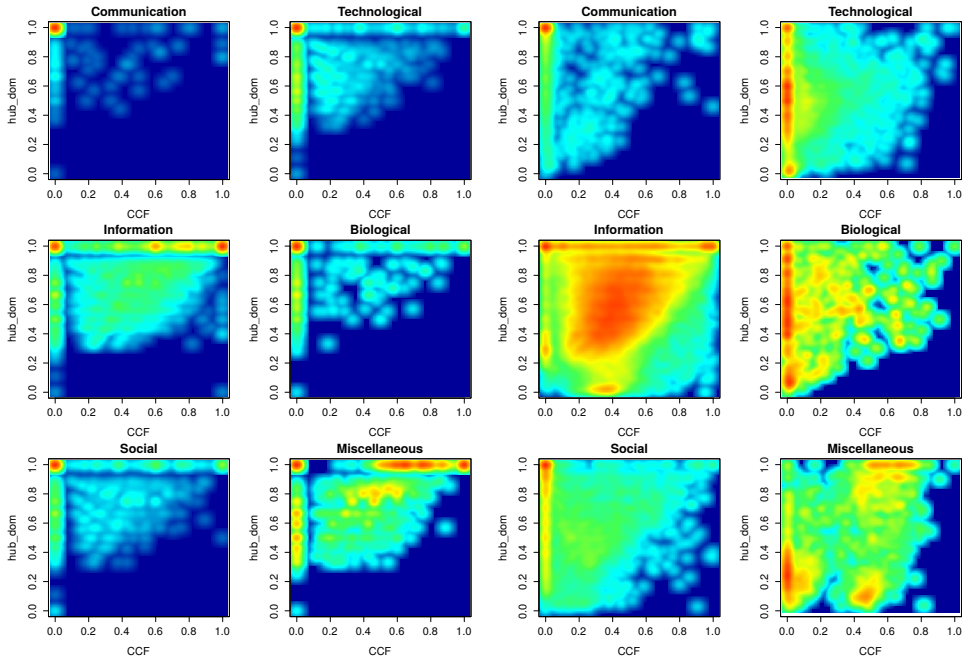
— Le modèle *Erdős-Rényi* (Erdős et Rényi, 1959) est l'un des premiers modèles proposés pour décrire la génération de *graphes aléatoires*. Dans ce modèle, deux paramètres sont nécessaires pour générer un graphe, à savoir un nombre fixe de sommets n et une probabilité de connexion p entre deux sommets arbitraires (ou encore le nombre d'arêtes m). Chaque paire de sommets est ensuite connectée indépendamment des autres paires avec la probabilité p, qui reflète la propriété aléatoire du graphe résultant. Si nous fixons les paramètres n et p du modèle de telle sorte que le modèle crée un graphe aléatoire dont le degré moyen se rapproche des réseaux réels : $\langle k \rangle = p(n-1) = c > 1$, où c est une

constante et $c \ll n$, le graphe aura presque certainement une grande composante connexe contenant une grande partie des sommets et de très petites composantes de moins de $\mathcal{O}(\log(n))$ sommets. Cette configuration produit des sommets qui ont tous environ $c > 1$ connexions. Ici, nous faisons référence aux graphes aléatoires créés par cette configuration. Puisqu'un réseau aléatoire est construit à partir d'un mécanisme stochastique homogène, il n'y a normalement pas de hubs ni de cliques, ce qui signifie des valeurs de transitivité et de hub dominance faibles (Figure 3.7 en bas à gauche). Un graphe aléatoire typique construit avec une petite valeur de p aura une topologie de type chaîne (string en anglais). Dans un régime extrême, lorsque la probabilité de connexion p approche 1, le graphe aléatoire associé devient *presque complet* lorsque le degré moyen $< k >$ approche $n - 1$, ce qui signifie que chaque sommet est connecté avec presque tous les autres sommets comme illustré dans la Figure 3.7 en haut à droite, correspondant à des fortes valeurs de hub dominance et CCF.

— Le modèle *Watts-Strogatz* produit des réseaux ayant la propriété *small-world*, ce qui signifie que toute paire de nœuds peut être connectée par un petit nombre de nœuds intermédiaires et que la distance géodésique moyenne croît proportionnellement au logarithme du nombre de nœuds n du réseau : $L \propto \log(n)$. Le modèle est construit pour caractériser l'observation selon laquelle de nombreux réseaux du monde réel présentent cette propriété de connectivité à faible longueur de chemin et forment des groupes denses proches de treillis réguliers, ce qui implique une forte présence de fermetures triadiques (Watts et Strogatz, 1998). À partir d'un anneau avec n nœuds et k arêtes par nœud, chaque arête est redistribuée aléatoirement avec une probabilité $0 < p < 1$. Les auteurs constatent qu'une petite valeur de p réduit considérablement la longueur du chemin caractéristique d'un réseau où les nœuds ne sont initialement que connectés localement. Cela peut s'expliquer par le fait que les arêtes reconnectées créent des raccourcis entre les zones éloignées du réseau et réduisent donc considérablement le diamètre. Un réseau small-world typique peut être décrit à l'aide d'une valeur intermédiaire de p, de sorte que la distance entre deux nœuds arbitraires est très faible, mais le coefficient de clustering reste élevé puisque la perturbation aléatoire n'est pas assez forte pour briser les structures locales des nœuds dans l'anneau du réseau. De plus, la forme de la distribution des degrés dans le réseau est assez similaire à celle d'un graphe aléatoire où chaque nœud a environ k de voisins et où il n'y a normalement pas de phénomène de hub dominance. La topologie d'un réseau small-world typique est relativement homogène et ressemble à une grille, comme illustrée dans la Figure 3.7 en bas à droite, avec un score de CCF élevé et un score *hub_dom* faible.

— Le modèle *Barabási-Albert (BA)* (Barabási et Albert, 1999) est né de la découverte de la distribution hétérogène des degrés dans de nombreux réseaux du monde réel. Plus précisément, la connectivité des sommets suit une *distribution en loi de puissance*, ce qui signifie que la probabilité qu'un sommet se connecte à k voisins dans son réseau est égale à $p(k) = Ck^{-\alpha}$ où la constante C est fixée par une exigence de normalisation et α est le coefficient de loi de puissance. Ce coefficient varie entre 2 et 3. Les réseaux possédant cette caractéristique statistique sont appelés *scale-free* par Barabási *et al.* pour souligner la propriété d'invariance d'échelle. Cette caractéristique est expliquée par les auteurs comme étant la conséquence de deux mécanismes principaux :

premièrement, les réseaux s'étendent progressivement en attirant de nouveaux nœuds vers les nœuds existants ; deuxièmement, ces nouveaux nœuds ont tendance à s'attacher préférentiellement aux sommets qui sont déjà bien connectés. C'est pourquoi ce modèle est souvent appelé *modèle d'attachement préférentiel*, ce qui implique que plus un sommet est connecté, plus il a de chances de recevoir de nouvelles arêtes ("richer nodes get richer"). Ce mécanisme ainsi produit nativement des hubs, et donc les valeurs de hub dominance sont généralement élevées. D'autre part, les coefficients de clustering associés sont généralement faibles et se dégradent rapidement en fonction de la taille du réseau (Klemm et Eguíluz, 2002 ; Fronczak, Fronczak et Hołyst, 2003), ce qui implique une faible transitivité. Par conséquent, les réseaux *scale-free* typiques ont une structure proche de celle des topologies en étoile, et se situent dans le coin supérieur gauche de la Figure 3.7.



(A) Petites communautés (B) Grandes communautés

FIGURE 3.8 – Distributions des communautés structurelles détectées sur différentes catégories de réseaux dans un espace à deux dimensions caractérisé par la transitivité (*CCF*) et la dominance du hub (*hub_dom*). Les petites communautés contiennent moins de 10 membres. De gauche à droite, de haut en bas : (a) Communication, (b) Technologique, (c) Information, (d) Biologique, (e) Social, (f) Divers : réseaux d'énergie, réseaux écologiques, réseaux artificiels, etc.

Puisqu'il a été remarqué que certaines caractéristiques structurelles pourraient différer entre les petites communautés appelées *micro-communautés* et les grandes communautés appelées *macro-communautés* (Lancichinetti et al., 2010), nous procédons à leur analyse séparée. La Figure 3.8a montre les distributions des petites communautés de 10 nœuds ou moins dans nos 6 groupes de réseaux différents (les réseaux de communication, technologiques, d'information, biologiques, sociaux et divers). Les distributions homologues pour les grandes communautés de plus de 10 nœuds sont représentées dans la Figure 3.8b.

À première vue, nous remarquons qu'il existe une plus grande diversité de structures dans les grandes communautés. Cela s'explique par le fait qu'il y a beaucoup

plus de possibilités de connecter des nœuds dans une grande communauté que dans une petite. Par conséquent, les structures des grandes communautés sont plus différentiantes et en même temps plus complexes. Plus précisément, la plupart des petites communautés se trouvent autour de deux axes où $CCF = 0$ ou $hub\_dom = 1$, surtout à leur point de croisement où $CCF = 0$ et $hub\_dom = 1$. Cela signifie que les structures en étoile et dominées par un hub sont très bien représentatives des petites communautés de chaque catégorie de réseau. En revanche, la structure en grille est totalement absente, ce qui est assez prévisible puisqu'il faut un grand nombre de nœuds pour qu'une grille se forme. De plus, la longue traîne de la distribution des degrés que l'on retrouve dans de nombreux réseaux du monde réel rend l'établissement de grilles moins probable.

Dans les réseaux d'information et divers, les communautés sont beaucoup plus riches en structures par rapport aux autres catégories, et cela pour tout type de communautés. Concrètement, outre les groupes en étoile, il existe également de nombreuses communautés de type clique et des structures mixtes, car les valeurs des coefficients de clustering dans ces groupes s'étendent sur toute la gamme. Il en va de même pour les valeurs de hub dominance de hub qui varient entre 0, 4 à 1 à petite échelle et de 0 à 1 à grande échelle.

Bien qu'il existe des différences de topologies entre les diverses catégories de réseaux, il n'est pas très évident de les distinguer à l'aide de la représentation proposée. Nous avons procédé à une inspection détaillée pour les grandes communautés (Dao, Bothorel et Lenca, 2021). En guise d'exemple, nous montrons ici le résultat de cette étude empririque pour les macro-communautés des réseaux d'informations dans la Figure 3.9.

Les communautés d'information contiennent des sous-réseaux dans les réseaux de citations, les réseaux de collaboration scientifique, les réseaux de moteurs de recherche, les réseaux de recommandation, etc. Globalement, les communautés d'information se distinguent des autres catégories par leur forte transitivité. Ainsi, les cliques sont très bien représentées. De plus, de nombreuses communautés d'information peuvent être considérées comme des mélanges des différentes topologies de base (en étoile, en chaîne, en clique et en grille), comme la communauté de collaboration du réseau Arxiv Condensed Matter illustrée sur la Figure 3.9(h). La présence de concentrateurs dans les réseaux d'information est toujours élevée, mais ils ne sont plus les seuls éléments qui relient les différents membres des réseaux. Par conséquent, les réseaux d'information sont beaucoup plus denses et mieux connectés que d'autres types de sous-réseaux de la même taille. Il s'agit probablement de la caractéristique de connectivité la plus représentative des réseaux d'information. Des résultats similaires liés aux structures denses et de cliques ont également été trouvés par (Lancichinetti et al., 2010). Les Figures 3.9(a-h) décrivent d'autres communautés représentatives. Alors que la structure de la Figure 3.9(d) ressemble à une topologie en étoile avec une séquence de connexions périphérie-périphérie, celle de la Figure 3.9(e) issue la collaboration Arxiv High Energy Physics, qui ressemble davantage à un réseau complet avec quelques nœuds mal connectés. Les Figures 3.9(c,g) illustrant des systèmes de recommandation et de web révèlent une structure mixte où les hubs peuvent être bien reconnus et la présence de cliques est également remarquable en même temps.

La diversité organisationnelle constatée dans les réseaux d'information peut s'expliquer par la manière dont nous définissons cette catégorie. En effet, un système de recommandation commercial s'avère être très différent des citations web ou d'un réseau de collaboration, même s'ils sont tous considérés comme des systèmes d'information dans la communauté des sciences des réseaux. Une étude approfondie,

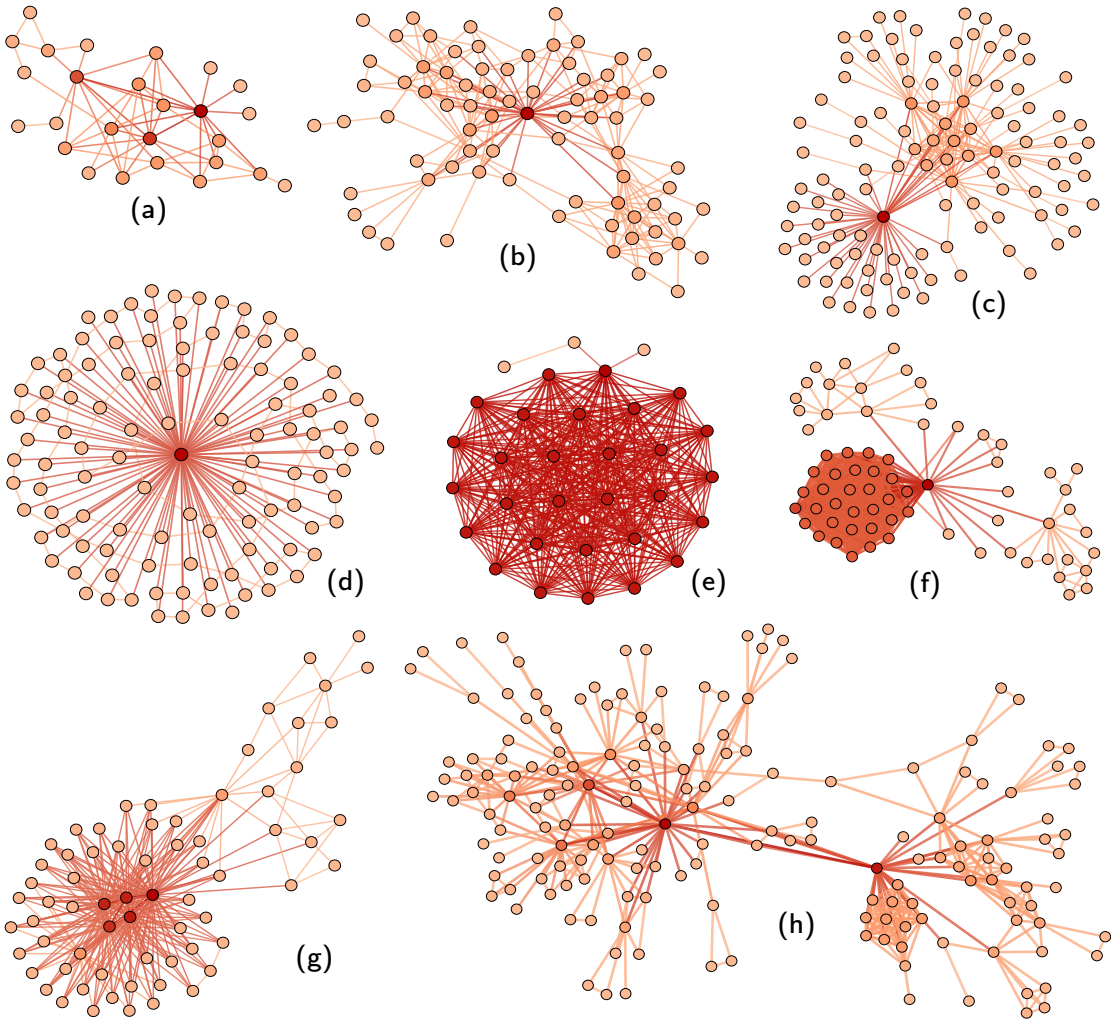


FIGURE 3.9 – Quelques topologies représentatives détectées dans les *Réseaux d'Information* avec leurs scores respectifs (CCF, hub_dom). Les hub sont plus foncés que les nœuds périphériques. (a,b,g) Groupes de produits Amazon recommandés - $(0.40, 0.52)$, $(0.33, 0.45)$ et $(0.24, 0.76)$ respectivement ; (c) Un groupe issu d'un système éducatif en ligne - $(0.30, 0.43)$; (d) Un groupe de sites web sur l'Indochine en 2004 - $(0.05, 0.98)$; (e-f) Une communauté de collaboration sur Arxiv High Energy Physics - $(0.99, 0.97)$ and $(0.95, 0.99)$; (h) Une communauté de collaboration sur Arxiv Condensed Matter network - $(0.44, 0.36)$.

au cas par cas, serait donc intéressante à mener afin de comprendre les mécanismes que reflète l'organisation mésoscopique. C'est précisément l'ambition de nos travaux, proposer des outils d'exploration pour disséquer ces phénomènes complexes qui régulent les interactions au sein d'un réseau.

3.5 Conclusion

L'objectif de ce travail est de pouvoir caractériser une partition par des critères organisationnels et ainsi contribuer à outiller les chercheurs en Sciences Humaines et Sociales, et plus généralement tout chercheur et décideur qui souhaite explorer et comprendre finement un réseau, au niveau mesoscopique (communautés), mais aussi intra-communautaires. En apportant une description des structures communautaires, nous avons montré comment comparer des réseaux, ou plus généralement émettre des propriétés de familles de réseaux réels. Ainsi, des chercheurs citent nos travaux, et utilisent nos résultats pour fournir un modèle de génération de réseaux qui contrôle la hub dominance et le clustering coefficient (Yamaguchi et al., 2020).

Cette description structurelle a un intérêt en soi, nous venons de le constater, mais qu'en est-il de l'intégrer à notre un processus de choix d'algorithmes ?

Dans le Chapitre 3.5 précédent, nous avons ébauché une aide à la décision de choix de méthodes de détection de communautés. Pour mémoire, ce choix repose sur la disponibilité des méthodes bien sûr, mais ensuite sur leur complexité en temps de calcul, la taille visée des communautés que l'on recherche et une équivalence des méthodes en termes de stratégie de partitionnement. Un classement des méthodes est proposé en relation avec ces critères.

A travers l'analyse quantitative empirique exposée dans ce chapitre, nous avons désormais des outils complémentaires. Nous pouvons désormais proposer des critères *qualitatifs* relatifs à la nature même des communautés produites par les méthodes, caractérisant leur organisation interne et/ou externe.

Autant la cartographie de méthodes du Chapitre est-elle générique, et comme on l'a vu, elle permet de pré-sélectionner rapidement quelques méthodes parmi les 16, autant le choix final d'une méthode/partition parmi cette pré-sélection ne peut se faire que sur la base d'un choix orienté. Pourquoi préférer telles tailles de communautés par exemple ?

Selon nous, et en accord avec (Smith et al., 2020), la "meilleure" méthode dépend du contexte, de la question de recherche, c'est-à-dire de la manière dont les communautés seront utilisées. Cette décision ne peut donc se prendre que de manière contextuelle, pour un réseau donné, lorsqu'il s'agit de répondre à une problématique donnée, et dans le but d'explorer plutôt telles ou telles caractéristiques que d'autres.

Ainsi, nous proposons aux praticiens une méthodologie :

— **L'étape 1 - niveau partition** vise le choix d'un ensemble de partitions. Comme nous n'avons pas de connaissances a priori sur les communautés de notre réseau, nous pouvons ici sélectionner des méthodes produisant des partitions les plus différentes possibles, offrant alors une amplitude de diversité pour l'exploration des données à venir ; nous pouvons aussi au contraire, faire le choix de réduire les possibles, et de sélectionner des méthodes qui font consensus. Concernant le choix à proprement parler :
  — Soit nous avons mobilisé tous les algorithmes à disposition, et dans ce cas, il s'agit de comparer les partitions produites : avec des métriques de validation comme vu en Section 2.4, telles que NMI, AMI, etc,; grâce aux distributions de tailles de communautés. L'idée est de reproduire nos expérimentations du Chapitre précédent sur le réseau courant, et ainsi d'obtenir des clusters de méthodes dans le contexte particulier de l'étude.
  — Soit nous nous basons sur l'étude exhaustive empirique réalisée dans le Chapitre 3.5, et nous en utilisons les résultats, à savoir la cartographie de la Figure 2.8 proposant l'équivalence de méthodes. Il s'agit alors de pré-sélectionner des méthodes, puis de calculer les partitions avec les méthodes sélectionnées.

— **L'étape 2 - niveau communauté** a pour but de caractériser les communautés avec des cartes bivariées basées sur les mesures qualitatives. Comme il existe un nombre indécent de combinaison, cela n'aurait pas de sens de toutes les réaliser. Au contraire, c'est à ce stade que nous préconisons de sélectionner la ou les combinaisons qui sont *pertinentes pour le problème à traiter sur le réseau donné*. On voudra privilégier tantôt la nature des organisations internes (avec la transitivité et la hub dominance par exemple), tantôt l'ouverture des communautés et les relations inter-groupes (*ODF*, *FOMD*, etc).

— Enfin, **l'étape 3 - question métier** consiste à sélectionner une partition (et ses communautés) de façon à mener l'analyse de manière approfondie. Ici encore, c'est bien la question métier que l'on cherche à résoudre qui va guider le choix final.

Cette méthodologie est générique, et laisse encore beaucoup de questions ouvertes, notamment pour les choix stratégiques à chaque étape : privilégier la diversité de partitions ou bien le consensus ? Quels couples de métriques sont révélateurs de quelles types de topologie ? Le graal serait de fournir un arbre de décision, ou une table du type de la Figure 3.1 proposée par (Toni et Nonino, 2010), par exemple, avec un choix exhaustif et explicite des situations à étudier et les outils correspondants à mobiliser.

En attendant, dans le Chapitre 4, à suivre, nous montrons comment nous avons instancié cette méthodologie de sélection de méthodes. A travers l'étude de cas d'Ulule, une plate-forme participative de financement, et une question de recherche spécifique — quelles organisations communautaires mènent à une collecte de fonds réussie —, nous montrons comment sélectionner la méthode la plus pertinente.

Chapitre 4

Le cas d'étude Ulule

Des travaux tendent à montrer qu'il n'existe pas de "meilleur" algorithme de manière intrinsèque. En moyenne, les méthodes ont les mêmes performances (Peel, Larremore et Clauset, 2017), et de manière qualitative et complémentaire, nous avons montré quelles produisent des partitions assez équivalentes (Chapitre 3.5). Mais nous avons également montré que, bien qu'il n'existe pas de "meilleur" algorithme, les méthodes sont libres de se spécialiser pour détecter certains types de structures (tout en en ignorant d'autres). Cela signifie également qu'il n'y a pas de définition "correcte" de la structure d'une communauté... seulement différentes perspectives.

Aussi la diversité des méthodes de détection de communautés (définition d'une communauté, fonction objectif, partition obtenue) est-elle un atout pour détecter différentes structures dans les réseaux étudiés ? Nous avons en effet contribué à dresser une cartographie des communautés et des réseaux bien connus issus du monde réel et provenant de différents domaines dans le Chapitre 3, comme ont pu le faire d'autres études, par exemple Peel, Larremore et Clauset, 2017.

Dans ce chapitre, nous montrons comment tirer partie des particularités des méthodes pour explorer un réseau particulier. Nous déroulons pour ce faire la méthodologie de sélection de méthodes de détection de communautés présentée dans le Chapitre 3, à travers un cas d'étude en économie du numérique. L'objectif y est d'étudier les pratiques sociales de la plateforme de financement participatif Ulule, et d'identifier comment et lesquelles de ces pratiques influencent la réussite de financement des projets.

Cette étude a donné lieu à des publications en SHS (Lyubareva et al., 2019 ; Lyubareva et al., 2020 ; Bothorel, Brisson et Lyubareva, 2022), mais également en Informatique (Bothorel, Brisson et Lyubareva, 2021) où nous avons décrit la méthodologie que nous retraçons ici de manière synthétique.

4.1 Problématique et hypothèses

L'étude des communautés en ligne occupent une place importante dans les recherches socio-économiques depuis l'arrivée du web social. Cela est lié au rôle des technologies numériques (Rheingold, 2000), qui, grâce à leurs fonctionnalités de communication, invitent les utilisateurs à construire des formes de dialogue et d'interaction en réseau du type "many-to-many", (par opposition à "one-to-one", comme le téléphone, ou "few-to-many", comme l'imprimerie ou la radio).

Dans ce type d'étude en Sciences Humaines et Sociales, les communautés étudiées sont délimitées et connues *a priori*. Il s'agit des abonnés à un service en ligne, des participants à un forum de discussion, les contributeurs inscrits sur une plateforme. Par exemple, dans le contexte des plateformes de crowdfunding, de nombreuses études se concentrent sur les interactions directement observables entre les participants des projets individuels et montrent leur rôle en relation avec le succès des campagnes de collecte de fonds (Kuppuswamy et Bayus, 2018 ; Agrawal, Catalini, Goldfarb et al., 2010 ; Zheng et al., 2014).

Cependant, peu d'études empiriques s'intéressent aux structures relationnelles qui sont *non explicitement* énoncées (Inbar et Barzilay, 2014). Il est pourtant intéressant de se demander si les cercles relationnels peuvent aller au-delà des projets individuels, élargissant le capital social initial des porteurs de projets, pour former un réseau social inter-projets au niveau de la plateforme.

Et si un tel réseau existe, la participation des membres à ce *réseau inter-projets* garantit-elle des taux de réussite plus élevés des campagnes de crowdfunding ? Cette question a de fortes implications managériales et économiques pour les plateformes : une telle mise en réseau pourrait-elle conduire à la formation d'un noyau dur avec une participation intense ? Les plateformes doivent-elles développer davantage les mécanismes de mise en réseau inter-projets ? Comment les porteurs de projets doivent-ils s'entourer pour se donner toutes les chances de réussite ?

Dans notre étude, nous faisons l'hypothèse qu'un tel réseau inter-projets existe et que des communautés *non directement observables* de ce réseau sont un atout pour la plateforme Ulule en termes de dynamique sociale. Cependant, comme dans de nombreuses études exploratoires, nous ne disposons d'aucune information préliminaire sur leur nombre ou leurs structures. Notre objectif est 1) de découvrir ces communautés algorithmiquement, 2) de décrire leurs organisations internes, et 3) d'explorer, in fine, s'il existe une relation entre leurs structures et le succès des campagnes de collecte de fonds qu'elles portent.

La question se pose donc de *choisir* un algorithme de détection de communautés approprié, adapté à notre contexte particulier.

Nous allons pré-sélectionner d'abord 3 méthodes candidates, puis qualifierons leurs partitions résultantes par des caractéristiques topologiques. Ensuite, nous les comparerons pour finalement sélectionner la plus pertinente par rapport à notre question de recherche en introduisant des indicateurs métiers liés au financement participatif.

4.2 Jeu de données Ulule

Depuis 2010, Ulule est devenu un des premiers sites européens de financement participatif avec plus de deux millions de membres, 24000 projets financés et un taux de succès de 63% en 2018. Les projets publiés sur la plateforme s'inscrivent dans des catégories thématiques variées comme la vidéo, la musique, l'art, l'éducation,

la technologie, etc. 90 jours est la période maximale durant laquelle la collecte de fonds peut avoir lieu, et les dons peuvent commencer à partir de 5 euros. La plateforme accepte les projets, quel que soit le statut du porteur : particulier, organisation marchande ou association.

Les données que nous analysons représentent les cinq premières années du fonctionnement de Ulule, de janvier 2010 jusqu'à mars 2016. Après nettoyage, les données incluent 19 544 projets dont 11 900 ont été financés avec succès et 7 644 ont échoué. Ces projets ont réuni 876 758 contributeurs, qui ont versé au total 47,75 millions d'euros.

4.2.1 Le graphe de contributeurs actifs

La plateforme Ulule ne permettant pas à ses utilisateurs de mettre en évidence leurs liens d'amitié ou d'intérêt avec les autres utilisateurs, nous avons donc fait le choix d'utiliser les seules traces d'interaction à notre disposition : les contributions des utilisateurs aux mêmes projets. Nous construisons ainsi un graphe de co-contributions qui va nous permettre de vérifier si les co-contributions sont aléatoires ou, si, comme nous en faisons l'hypothèse, il existe une dynamique communautaire qui influe sur le succès des projets.

Nous définissons ainsi un graphe de co-contributions non-orienté, dans lequel chaque arête signifie que les utilisateurs ont contribué à 3 mêmes projets. En choisissant de ne créer des arêtes qu'entre deux individus ayant co-contribués à 3 mêmes projets nous éliminons les co-contributions fortuites pour ne conserver que celles qui seraient les plus susceptibles de mettre en évidence une interaction inter-projet entre les individus.

Le graphe ainsi obtenu possède de nombreuses composantes connexes. Beaucoup sont de très petites tailles et ne présentent donc aucun intérêt lors de la phase d'analyse. Nous nous focalisons donc sur la plus grande, qui contient 2081 nœuds et 4749 arêtes.

L'existence même de ce graphe met en évidence le fait qu'il existe bel et bien un réseau social propre à la plate-forme Ulule, un réseau transverse à ses différents projets. La densité du graphe est faible (0,002), avec une majorité de noeuds à faible coefficient de clustering. La moyenne des coefficients de clustering est en effet de 0,26, mais un nombre significatif (25%) de noeuds sont impliqués dans des cliques où tous leurs voisins sont eux-mêmes connectés entre eux, ce qui signifie qu'ils ont eux aussi co-contribué à au moins 3 projets communs. Concernant le degré, nous retrouvons une distribution des degrés en loi de puissance, classique dans les réseaux sociaux en ligne (propriété *scale-free* des réseaux complexes réels). Le degré moyen est de 4,56 et seuls plus de 25% ont un degré supérieur à 4, le degré maximal étant de 199. Il existe donc des *Ululers* qui co-financent des projets avec beaucoup de contributeurs différents (24 d'entre eux ont plus de 50 voisins dans le graphe). La longueur moyenne des plus courts chemins est de 3,97, avec un diamètre de 13 (longueur du plus court chemin le plus long), faisant de ce graphe un réseau *petit monde*, signature, en plus de la propriété *scale-free*, de l'existence d'un réseau social classique (Barabási, 2002).

Si l'on s'intéresse à la thématique des projets, nous constatons que la répartition des catégories, selon que les projets sont financés ou non par les membres du graphe, révèle des informations intéressantes. L'amélioration du taux de réussite des campagnes des projets *appartenant* au graphe est observée pour toutes les catégories (28,6% d'amélioration globale), en particulier pour les catégories thématiques *jeux*, *bande dessinée*, *technologie* et *édition*. Ceci n'est pas surprenant car ces domaines ont

Variable	Définition
degree	Degré : nombre d'individus ayant co-contribués à au moins 3 même projets.
clustering coefficient	Coefficient de clustering local : mesure à quel point le voisinage d'un nœud est connecté. Plus il est grand plus le voisinage tend vers une clique, i.e. tous les voisins sont eux-mêmes connectés entre eux.
betweenness	Centralité d'intermédiarité : nombre de fois où le nœud figure sur le plus court chemin entre deux nœuds du graphe.
closeness	Centralité de proximité : définit à quel point un nœud est central (c'est à dire qu'il a la distance la plus faible à tous les autres noeuds).

TABLE 4.1 – Mesures de centralité des contributeurs du graphe social

naturellement une forte composante sociale dans la production et la consommation de biens (Lyubareva et al., 2019). Ces résultats préliminaires confirment notre intérêt pour l'étude des communautés d'Ululers actifs.

4.2.2 Les contributeurs actifs

Afin de caractériser les *Ululers* faisant partie du réseau social de la plate-forme et susceptibles de représenter un cercle original des financeurs, nous combinons attributs relationnels (Table 4.1) et socio-économiques (Table 4.2). En plus des caractéristiques habituellement mobilisées dans la littérature pour décrire les contributeurs du crowdfunding, nous créons des variables supplémentaires liées (i) à un niveau de spécialisation thématique d'un contributeur et de ses voisins dans le graphe ; et (ii) aux délais d'arrivée dans les projets du contributeur et de ses voisins dans le graphe. Ces informations nous permettent d'analyser la proximité cognitive et comportementale entre les contributeurs et de modéliser le principe d'homophilie.

Nous constatons également que les membres de notre graphe social varient beaucoup en termes d'activité sociale (degré des nœuds, coefficient de clustering et centralités) mais aussi de comportement de contribution (nombre de projets financés, montant moyen des contributions, taux de spécialisation qui quantifie la variété des catégories thématiques abordées (catégorie des projets) et leurs similarités avec les nœuds voisins). Une analyse des correspondances multiples sur ces attributs, suivie d'une classification ascendante hiérarchique, conduit à 5 clusters de contributeurs actifs d'Ulule : les *Précurseurs* et les *Suiveurs*, les *Spécialistes* et *Spécialistes collaboratifs*, ainsi que les *Sponsors*. Une version détaillée est disponible (Lyubareva et al., 2020), mais pour des raisons de lisibilité, nous n'en faisons qu'un résumé dans ce document.

Selon leur arrivée dans les projets, nous trouvons les *Précurseurs* (ils sont 538) et les *Suiveurs* (653). Ces derniers se caractérisent par un intérêt pour des projets d'envergure très importante. Ces deux profils sont les plus nombreux. Nous trouvons aussi les *Spécialistes*, très nombreux également (504), qui ont la particularité de se focaliser sur certaines catégories thématiques. Les *Sponsors*, très peu nombreux (18), quant à eux, sont ceux qui ont une activité sociale forte en termes du positionnement central et du nombre de liens sociaux dans le graphe (forte centralité d'intermédiarité et de degré), et qui participent à un nombre important de différents

Attributs économiques	
project goal	Moyenne en euros de l'objectif des projets financés.
contrib. amount	Montant moyen des contributions.
projects contributed	Nombre de projets financés.
specialization rate	Taux de spécialisation thématique : ratio de projets sur la thématique la plus financée par rapport au nombre total de projets financés.
contribution time	Avancement médian des projets lors de la contribution.
precursor rate	Proportion des projets financés pour lesquels l'avancement médian des projets financés est inférieur à celui de ses voisins.
Attributs d'homophilie	
neighbor specialisation rate	Moyenne du taux de spécialisation thématique de l'ensemble des voisins du contributeur.
neighbors contribution time	Moyenne de l'avancement médian des projets lors de la contribution des voisins.
neighbors similarity	Corrélation de rangs dans le classement des catégories thématiques financées par le contributeur et celles financées par son voisinage.

TABLE 4.2 – Attributs socio-économiques et comportementaux d'un contributeur du graphe social

projets d'Ulule. Dans la même lignée, nous trouvons les *Spécialistes collaboratifs* (368), qui se distinguent par un fort coefficient de clustering, dont l'activité sociale se traduit donc par la cohésion des liens et la solidarité ; ils contribuent en moyenne à un nombre de projets plus modeste, mais avec des montants plus importants que les autres profils. Ce résultat met en lumière le lien entre le volume de contributions et l'implication sociale, mais également que ce lien peut prendre des formes variées.

4.3 Etape 1 : Pré-sélection de 3 méthodes

Notre analyse porte sur la structure interne des communautés. En effet, nous cherchons quels types d'organisations de contributeurs mènent au succès des projets : quels en sont les contributeurs, leurs thématiques, mais aussi leur interactions, c'est-à-dire comment ils co-contribuent.

Pour faire apparaître les structures, nous cherchons à détecter des communautés que nous pourrons observer à différents niveaux d'échelle, car nous faisons l'hypothèse ici que petites, moyennes et grandes communautés ne feront pas apparaître les mêmes topologies. Pour ce faire, nous allons privilégier 3 méthodes "équivalentes" d'un point de vue répartition des nœuds (métrique ARI) mais qui proposent différentes tailles de communautés. Nous nous appuyons sur notre étude exhaustive des $108x16$ partitions synthétisée sur notre Figure 2.8 du Chapitre 3.5 que nous reproduisons ici pour faciliter la lecture (Figure 4.1). Comme on peut le voir, *Edge betweenness (GN)*, *Louvain* et *Walktrap* appartiennent au même cluster de stratégie de partitionnement et répondent à notre critère multi-échelle. De plus, ces trois méthodes étant

bien connues, elles facilement mobilisables car largement documentées et accessibles via les librairies R ou Python.

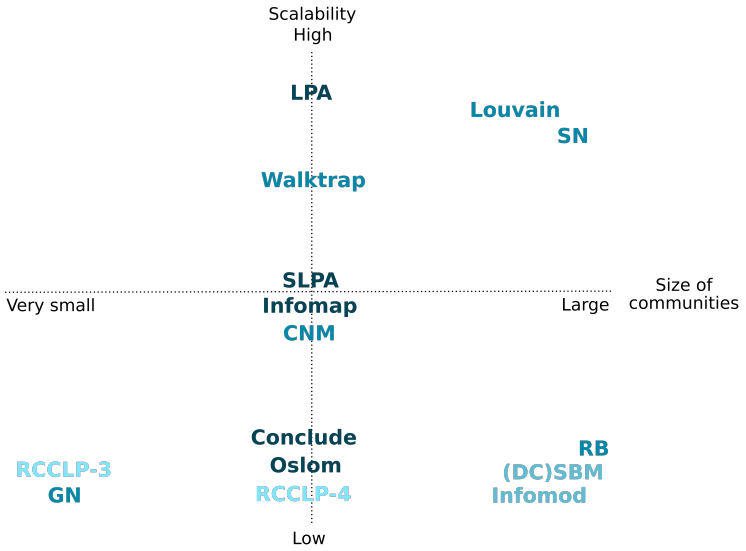
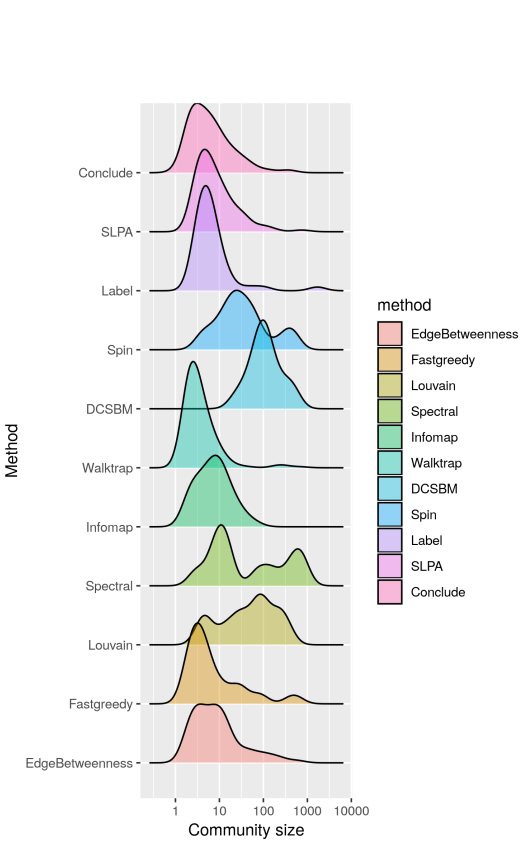


FIGURE 4.1 – Cartographie des méthodes de détection de communautés selon leur scalablité et la taille des communautés produites. La couleur correspond aux clusters de stratégie de partionnement (clusters de la Table 2.8 obtenus via la métrique de validation ARI).

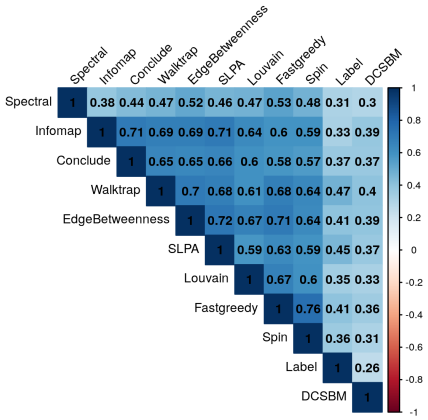
Nous avons calculé également ces indicateurs sur notre partition Ulule, mais comme nous allons le voir, les résultats sont tout à fait cohérents avec ceux de notre expérimentation exhaustive. La figure 4.2b montre comment les partitions sont étonnamment similaires du point de vue de la NMI où tous les scores sont positifs, allant de 0,26 à 0,76, avec une grande majorité supérieure à 0,5. Huit d'entre elles affichent des scores supérieurs à 0,6, avec un groupe très consensuel composé de quatre méthodes : *Edge Betweenness*, *SLPA*, *Fast greedy* et *Walktrap* avec des scores supérieurs à $0,7$. *Louvain* est également proche de *Edge Betweenness* et *Walktrap* avec des scores supérieurs à $0,6$. Il existe trois méthodes légèrement différentes : *Spectral*, *Label Propagation* et *Spin Glass* qui produisent des partitions plus spécifiques, qui sont toutes différentes les unes des autres. Si cette différence peut s'expliquer facilement par le fait que *Spectral* et *Spin Glass* mettent en œuvre des mécanismes intrinsèquement distincts, les résultats de *Label Propagation*, qui est une variante de *SLPA*, sont assez surprenants.

Concernant la taille des communautés, nous retrouvons également les résultats de notre étude exhaustive. *Louvain* génère des groupes d'assez grande taille en général. Les deux autres méthodes, tout en étant complémentaires, permettent de révéler des plus petits groupes, créant des partitions plus fines, surtout la méthode *Walktrap* (Figures 4.2c et 4.2a).

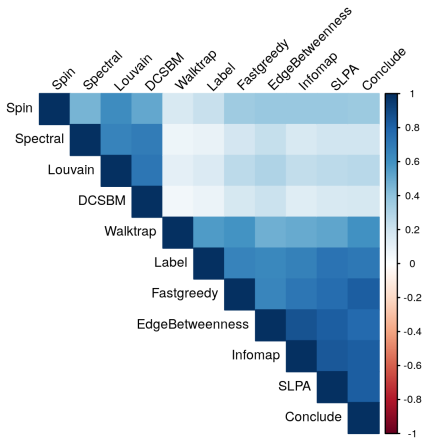
Le Tableau 4.3 confirme que l'algorithme de *Louvain* partitionne le réseau en moins de communautés que les deux autres méthodes. Cependant, leur grande taille en nombre de membres influe relativement peu sur le degré, dont la moyenne est de $3,63$ pour des communautés qui rassemblent 90 membres en moyenne contre $2,29$ pour 12 membres en moyenne pour *Walktrap*. La centralité de proximité moyenne et la densité moyenne restent stables quelque soit la taille des communautés et l'algorithme utilisé.



(A) Distribution de taille



(B) NMI



(C) Taille

FIGURE 4.2 – Comparaison des partitions Ulule issues de 11 algorithmes : Distribution des tailles, et similarité du point de vue de la stratégie de partitionnement NMI et de la distribution de taille

4.4 Etape 2 : Qualification de la structure interne des communautés

Afin de caractériser les schémas organisationnels au sein des communautés, ce qui est notre objectif, nous proposons l'utilisation de mesures structurelles appliquées aux communautés. Comme nous l'avons vu, il en existe un nombre non négligeable : densité des liens internes, centralité moyenne des nœuds, degré moyen, etc, que nous pouvons combiner dans des cartes bivariées. Par exemple, plonger les communautés dans l'espace meanODF x stdODF permet d'explorer différentes situations concernant l'ouverture des communautés et la coopération entre ces groupes d'Ululers. Cependant, la hub dominance et la transitivité sont particulièrement pertinentes lorsque l'on considère les organisations internes reflétant la coopération, car leur combinaison conduit aux modèles bien connus décrits sur la figure 4.3 :

— *Hub dominance* : les arêtes internes d'une communauté peuvent être distribuées de différentes manières autour de ses nœuds, soit en se concentrant autour de quelques nœuds fortement centralisés, soit en étant distribuées uniformément sur les nœuds. La métrique de la dominance des nœuds identifie

Méthode	Nb communautés.	Nb. membres	Degré	Coeff. clustering	Centralité intermédiarité	Centralité proximité
Louvain	22	90,48	3,63	0,26	2 871,45	0,24
Edge Bet.	71	28,90	2,96	0,20	2 234,84	0,23
Walktrap	167	12,39	2,29	0,15	1 695,03	0,22

TABLE 4.3 – Communautés générées par trois méthodes de l'état de l'art. Les indicateurs caractérisant les nœuds (4 dernièrs colonnes) sont calculés pour chaque membre de chaque communauté, puis une moyenne est calculée pour chaque communauté. La synthèse montre ici les moyennes des indicateurs pour toutes les communautés par méthode.

la propriété de centralisation. Plus cette métrique est élevée pour une communauté, plus il est probable qu'elle ait une structure de type hub. La dominance de hub peut être considérée comme une version normalisée de la centralité des degrés. Une forte dominance concentrée sur quelque nœuds donne lieu à des modèles en étoile, comme le montre la figure 4.3.

— *Transitivité* : la transitivité reflète la probabilité que les sommets adjacents d'un sommet soient connectés. Cette métrique est généralement employée pour caractériser les structures en treillis (grilles) ou des cliques dans les réseaux (Figure 4.3). Une transitivité élevée associée à des sphères d'intérêt (ou d'autres attributs) similaires entre les individus indique souvent l'existence d'une homophilie sociale.

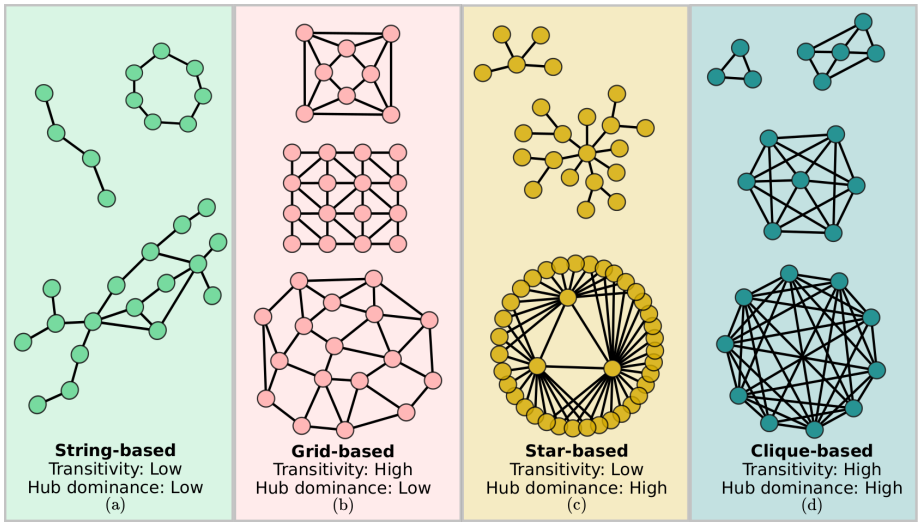


FIGURE 4.3 – Quatre topologies représentatives projetées dans l'espace *(Hub dominance, Transitivity)* : (a) String-based, (b) Grid-based, (c) Star-based, (d) Clique-based.

Sur la Figure 4.4, nous pouvons remarquer que les grandes communautés sont concentrées dans la même zone. Les méthodes responsables de ces grandes communautés (par exemple *Louvain*) produisent en effet a priori très peu de modèles structurels différents. Avec une faible hub dominance et une faible transitivité, la plupart des communautés peuvent être considérées comme des structures "string-based" (Figure 4.3a) qui ne reflètent pas vraiment de coopération entre les individus.

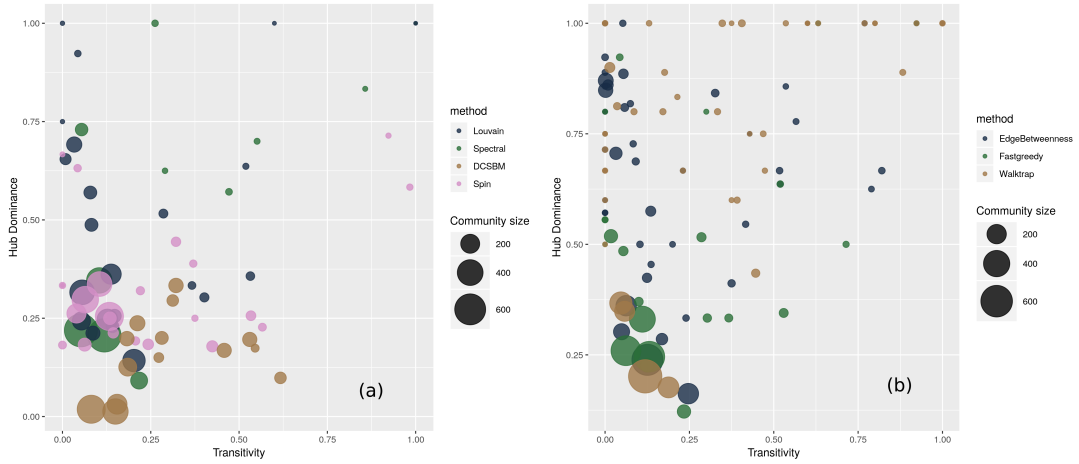


FIGURE 4.4 – Les communautés Ulule plongées dans l'espace *(Hub dominance, Transitivity)*. Les partitions de méthodes produisant des regroupements de grande taille à gauche, et petites taille à droite.

Cependant, cette réalité cache probablement des zones plus denses. Afin de détecter si la coopération existe ou non au sein des grandes communautés, nous pourrions faire un zoom pour en extraire les sous-zones denses, c'est-à-dire appliquer à nouveau une détection de communauté à chaque communauté, puis projeter les nouvelles petites communautés dans notre carte bivariée.

Car en effet, à l'inverse, *Edge Betweenness*, *Walktrap* et *Fastgreedy* qui produisent des communautés de petite et moyenne taille semblent générer divers types d'organisations directement observables. Les points sont répartis dans 3 des 4 zones de notre carte bivariée de la Figure 4.4 (panneau de droite). Nous trouvons surtout beaucoup de groupes dans la partie supérieure de la carte. Cela signifie que leurs membres sont organisés autour des hubs, mais de deux manières différentes. Lorsque la transitivité est élevée, nous trouvons des organisations basées sur des cliques (Figure 4.3d), où les Ululers contribuent (presque) tous les uns avec les autres à des projets communs. Lorsque la transitivité est faible, au contraire, les organisations imitent les structures en étoile, avec une très forte centralisation (Figure 4.3c). Les Ululers de ces groupes sont moins impliqués dans la coopération horizontale, mais semblent suivre des influenceurs (Ululers avec des degrés élevés) qui concentrent les projets communs avec beaucoup de contributeurs peu connectés.

Avec ces nouvelles perspectives, *Edge Betweenness* et *Walktrap* semblent être de très bonnes candidates : (i) elles appartiennent au groupe consensuel des méthodes précédemment présenté ; et (ii) elles offrent diverses formes d'organisation interne. *Louvain*, sans offrir de structures topologiques diversifiées, démontre cependant des propriétés intéressantes de forte hub dominance dans ses grandes communautés. Pour cette raison, et par curiosité, nous proposons de conserver *Louvain* pour explorer plus avant ses communautés.

4.5 Etape 3 : Apport de connaissances métier

Même si d'ores et déjà nous pensons que *Louvain* ne nous apportera pas la richesse topologique de *Edge Betweenness* et *Walktrap*, nous allons conduire une analyse complémentaire des 3 partitions conformément au principe d'alignement métier de (Smith et al., 2020).

4.5.1 Sélection finale de la partition

Pour rappel, nous souhaitons évaluer si les membres du réseau social de la plateforme Ulule forment des sous-groupes qui vont avoir une influence positive sur le succès des projets. Le cas échéant, nous cherchons à connaître la configuration de ces groupes, en particulier, quels types de membres est-il intéressant d'associer pour donner toutes ses chances à une campagne de financement participatif.

Les communautés sont ici définies selon trois types de variables :

— les profils des contributeurs qui les constituent (cf. Section 4.2.2),
— leurs caractéristiques organisationnelles : coefficient de clustering et poids des liens entre les noeuds,
— leurs caractéristiques métier en lien avec la problématique : volume de financement, nombre d'interactions via un système de commentaires mis à disposition, spécialisation thématique et taux de succès des projets.

Pour classer les communautés de nos partitions, les méthodes de clustering suivantes ont été utilisées : (i) une analyse en composantes principales (2 dimensions) suivie d'un clustering hiérarchique ascendant (distance euclidienne, méthode Ward de minimisation de la variance), (ii) la méthode K-means et (iii) un arbre de décision. Nous identifierons ainsi différents types de communautés, et nous pourrons procéder à une analyse comparative du succès de ces communautés vis à vis des campagnes de levée de fonds.

L'arbre de décision produit des clusters de communautés (Familles de communautés dans la Figure 4.5) presque identiques à la méthode K-means (complétude = 0.964 ; Adjusted Rand Index : ARI = 0.854). La proximité avec les clusters produits par l'ACP n'est pas évidente lorsque l'on considère ces mesures (complétude = 0.513 ; ARI = 0.398), mais néanmoins, la même variété de formes communautaires peut être observée. Elles sont au nombre de 3 :

— Famille 1 (avec les Sponsors) : très grandes communautés équilibrées composées de tous les profils. Elles ont la particularité d'avoir attiré les *Sponsors* et beaucoup de *Followers*.
— Famille 2 (Spécialistes) : communautés très nettement dominées par les *Spécialistes*, qu'ils soient collaboratifs ou non.
— Famille 3 (Followers, Precursors) : petites voire micro communautés plutôt dominées par les *Precursors* et dans une moindre mesure les *Followers*.

Quelles que soient la méthode de clustering et la méthode de détection des communautés, les trois familles de communautés susmentionnées sont clairement détectées. Le seul élément distinctif est la proportion de chaque famille de communautés. *Edge Betweenness* propose les clusters les plus équilibrés (Tableau 4.4), tandis que *Louvain* produit principalement des éléments des Familles 1 et 2, et sans surprise, la plupart des 167 (petites) communautés produites par Walktrap sont de la Famille 3.

Sur cette base, nous choisissons la partition *Edge Betweenness*, composée de 72 communautés de tailles diverses, avec différents types d'organisation (en étoile, treillis, formes plus lâches, quelques pseudo-cliques) et présentant des exemples substantiels de chaque Famille. La distribution exacte des profils de ses Ululers est montrée dans la Figure 4.5.

4.5.2 Caractère collaboratif des Familles de communautés

Les trois familles de communautés présentent des différences cruciales en termes de collaboration. Sur les plateformes de crowdfunding, elle peut prendre différentes formes : partage de projets au sein d'une communauté (poids des liens dans le réseau social Ulule), cohésion des membres d'une communauté autour des mêmes

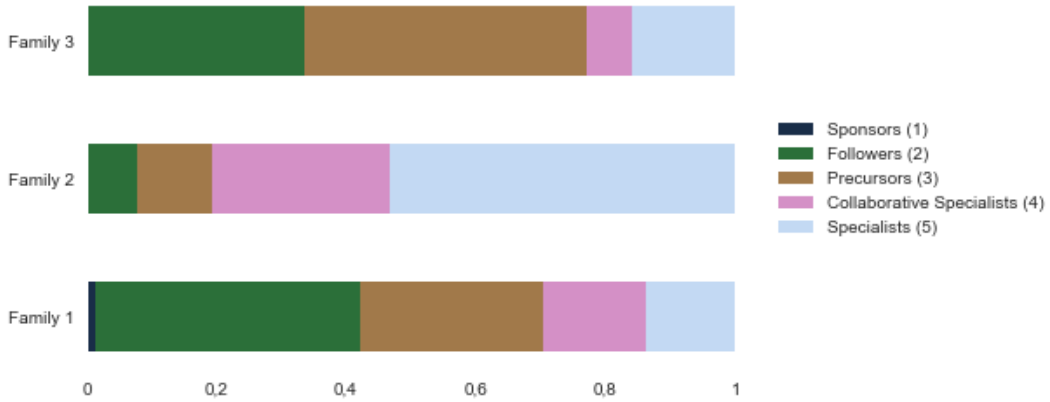


FIGURE 4.5 – Typologie des communautés (*Edge Betweenness*, arbre de décision).

Famille	Nombre communautés	Nombre membres moyen par communauté	Contribution moyenne d'un membre	Nombre projets par membre	Objectif moyen des projets	Nombre projets moyen par communauté
1	8	161,5	41,4	14,4	8 050	469,5
2	28	19,1	47,7	13,3	8 501	124,8
3	36	7,0	41,3	13,7	9 589	42,5

TABLE 4.4 – Profil de taille et de contribution des communautés par famille.

projets (transitivité, coefficient de clustering), ou communication via un système de feedback (commentaires). Comme présenté dans le tableau 4.5, chaque famille de communautés dans le cas de la plateforme Ulule privilégie une de ces formes.

La combinaison de deux aspects de la collaboration — le poids des liens dans le graphe et le coefficient de clustering — peut nous renseigner sur l'organisation de la communauté. Par exemple, dans la Famille 1, le fait que le nombre moyen de projets partagés soit très élevé mais que les membres soient peu connectés entre eux (faible coefficient de clustering) met en évidence une organisation très centralisée autour de quelques acteurs centraux. On retrouve ici les communautés à faible transitivité mais pas nécessairement à forte hub dominance. Ces grandes communautés se trouvent en effet au centre du réseau, et les Sponsors y jouent certes le rôle de hub, mais les structures en étoile très marquées autour d'eux sont locales, au niveau

Famille	Taille moyenne	Nombre projets	Projets partagés	Nombre commentaires par projet	Coefficiant de clustering
1	161,5	1 642	4,3	73,2	0,25
2	19,1	781	3,7	137,3	0,32
3	7,0	547	3,4	176,0	0,15

TABLE 4.5 – Taille des communautés et leur caractère collaboratif

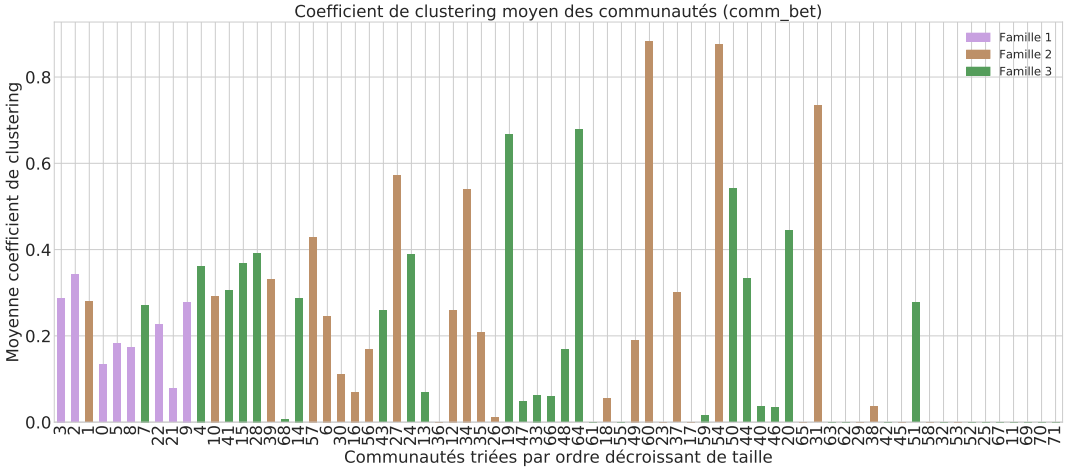


FIGURE 4.6 – Coefficient de clustering moyen des communautés. Les communautés sont triées par ordre décroissant de taille, celles contenant le plus de membres se trouvant donc sur la gauche.

du cœur des communautés ; ces communautés sont grandes, plus d'une centaine de membres, et ces hubs centraux sont entourés d'une périphérie large, peu dense, qui fait baisser le score de hub dominance. Les contributeurs liés aux membres centraux n'interagissent pas forcement entre eux (ce qui explique, au moins en partie, le faible nombre de commentaires dans ces projets). Par conséquent, le fonctionnement des communautés dans ce cas est étroitement lié à l'activité des membres centraux les plus actifs. Ces communautés attirent toujours beaucoup de gens, grâce à leur ouverture thématique, probablement aussi grâce à une meilleure visibilité sur la plateforme Ulule (mise en avant éditoriale des projets populaires) et, très certainement grâce à d'autres propriétés des projets non observables dans notre étude (par exemple communication hors plateforme).

Les communautés dans la Famille 2 regroupent les membres qui partagent les mêmes centres d'intérêt. Ces groupes thématiques, fortement connectés et solidaires, prennent collectivement leurs décisions de financement des projets et contribuent en moyenne avec des montants plus importants que les autres communautés. Cependant leurs choix thématiques sont plus restreints que dans la Famille 1. En affichant explicitement l'intérêt vers un thème spécifique, ces communautés attirent des financeurs intéressés par ce sujet, mais leur population est *a priori* plus faible que dans la Famille 1 (orientée grand public). Le nombre de commentaires, tout en étant relativement élevé, est plus faible que dans la Famille 3. Cela peut être lié à l'existence d'autres plates-formes de communication, en plus de la plate-forme Ulule, utilisées par les membres de ces communautés pour interagir. Une observation intéressante concerne le fait que les communautés spécialisées, basées sur le principe de l'homophilie, émergent principalement dans les catégories thématiques *Games*, *Charities*, *Comics*, *Films and video* et *Publishing*. En revanche, les catégories thématiques comme *Art*, *Crafts and Food*, *Fashion*, *Music* ou *Technology* attirent principalement les communautés "grand public" non-spécialisées.

Finalement, dans la Famille 3, les communautés se structurent également autour de certains thèmes, sans pour autant que leurs membres partagent tous les mêmes centres d'intérêt ou soient intéressés par un thème spécifique. La cohésion de ces communautés, en moyenne assez faible, varie en réalité au cas par cas (Figure 4.6). Le nombre de projets partagés dans cette Famille est moins élevé que dans les autres

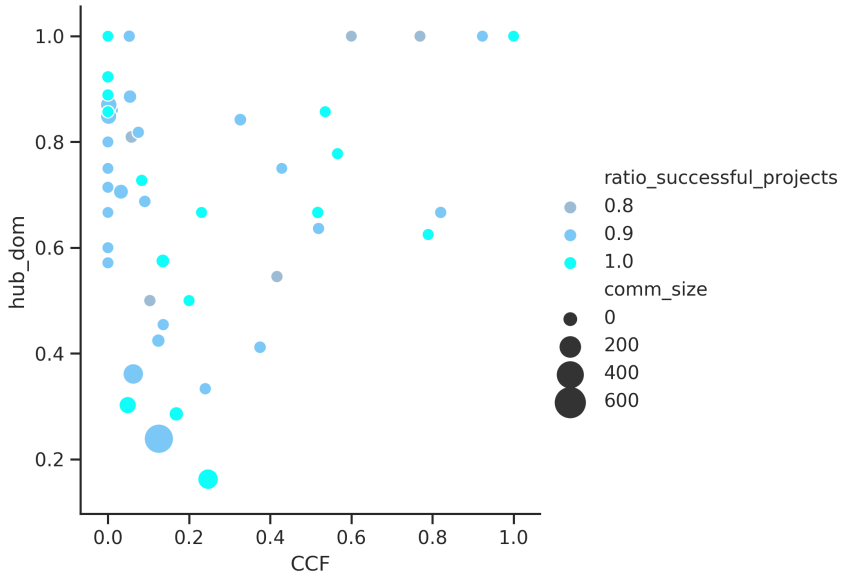


FIGURE 4.7 – Les communautés découvertes par *Edge Betweenness* plongées dans l'espace *(Hub dominance, Transitivity)*. La couleur représente le taux de succès des projets.

Familles. En revanche, les membres de ces communautés communiquent beaucoup via le système de commentaires, probablement pour compenser le manque d'autres formes d'interaction. Par ailleurs, la faible ouverture thématique combinée avec l'absence de spécialisation claire n'attirent pas beaucoup de monde dans ces communautés.

4.6 Quelles communautés pour quel succès ?

Une fois notre partition sélectionnée, nous pouvons procéder à une analyse comparative du succès de ces communautés vis à vis des campagnes de levée de fonds.

Les meilleures communautés, qui portent des projets tous financés, proviennent des Familles 2 et 3 (Table 4.6). Elles sont de petites tailles comprenant de 3 à 13 membres. Celles-ci sont assez spécialisées et financent des projets portant en moyenne sur 4 thématiques différentes. Les projets sont très variés en objectif à atteindre (tous les types de montants sont représentés), mais moins de 25% des projets ont une envergure plus importante que la moyenne, on a donc plutôt des projets modestes. Cependant, les communautés qui réussissent le moins appartiennent aussi aux Familles 2 et 3. Lorsque l'on regarde la Figure 4.7, il n'y a pas de zones clairement marquées par le taux de succès. Peut-être précisément parce que finalement tous les types de communautés trouvées et analysées dans le graphe mènent à bien leurs campagnes ? Nous remarquons cependant que les communautés les moins performantes se situent sur la moitié gauche. En effet, et la Table 4.6 le confirme, ce qui différencie clairement les communautés à très fort taux de réussite, c'est le coefficient du clustering et le nombre de commentaires, qui sont significativement plus élevés. En d'autres termes, la spécialisation thématique et le principe fondateur d'homophilie qui caractérisent la Famille 2 et partiellement la Famille 3, ne garantissent pas *en soi* le succès des projets du crowdfunding. Ils doivent s'articuler avec une implication sociale forte et une cohésion de ses membres pour donner lieu à des performances économiques marquantes des campagnes de levée de fonds.

	Nb projets partagés	Nb commentaires par projet	Objectif projets	Nb de membres	Nb de thèmes	Degré	Centralité d'intermédiarité	Coefficient de clustering
Reussite 100% (13 communautés)								
mean	11	517	8 826	6	4	2,84	2 126	0,39
std	6	1 120	6 132	3	3	1,06	1 356	0,35
min	3	2	2 630	3	1	1,67	777	0
25%	6	21	4 408	3	1	2,00	1 419	0
50%	9	60	7 946	4	3	2,50	1 559	0,43
75%	15	123	10 807	6	7	3,25	2 078	0,67
max	22	3 840	25 019	13	9	4,85	5 869	0,88
Reussite < 85% (9 communautés)								
mean	11	234	9 126	4	4	2,06	1 796	0,09
std	10	534	6 271	2	3	0,85	721	0,17
min	5	15	2 202	2	2	1,50	693	0
25%	6	29	6 305	3	3	1,67	1 039	0
50%	6	35	7 006	3	3	1,80	2 078	0
75%	12	88	11 189	5	3	2,00	2 078	0,06
max	35	1 653	23 916	8	11	4,25	2 970	0,44
Global (72 communautés)								
mean	58	393	10 832	29	6	2,96	2 235	0,20
std	122	686	5 799	78	4	1,18	1 092	0,22
min	3	2	2 202	2	1	1,50	693	0
25%	8	33	6 894	3	3	2,00	1 558	0
50%	17	78	9 053	8	5	2,93	2 077	0,15
75%	36	381	14 915	13	8	3,63	2 728	0,30
max	789	3840	27 555	579	15	6,46	5 869	0,88

TABLE 4.6 – Communautés à fort taux de réussite vs. communautés à plus faible taux de réussite. 13 communautés ont un taux de réussite de 100%, tous les projets ont atteint leur objectif (filtrage des communautés de plus de 2 membres). 9 communautés ont un taux de succès inférieur à 85%, la communauté la moins performante ayant vu pourtant 75% de ses projets projets réussir. Enfin, nous rappelons ici (sous-table Global) les différents indicateurs pour l'ensemble des 72 communautés afin de faciliter la comparaison. Les indicateurs sont arrondis pour plus de lisibilité.

Ainsi, contrairement à certains travaux existants, nos résultats mettent en évidence que le principe d'homophilie et de proximité thématique n'est donc pas le seul déterminant de formation des communautés en ligne et de leur succès. La diversité thématique peut être aussi un moteur de développement des communautés en ligne (le rôle de la diversité et les externalités du réseau dans les communautés en ligne a été aussi démontré par (Wang et Kraut, 2012)). En revanche, la spécialisation thématique, qui se trouve à l'origine du principe d'homophilie dans les communautés en ligne, peut donner lieu à un niveau de cohésion et de solidarité particulièrement élevés au sein des communautés et garantir, dans ce cas, la performance marquante des projets soutenus. Nos résultats mettent aussi en évidence que dans certaines thématiques, comme *Games*, *Comics*, *Vidéo*, *Publishing* ou *Charities*, ces communautés spécialisées et solidaires ont plus de chances de se développer. Ce résultat est en lien avec les travaux existants sur le rôle particulièrement important des communautés dans la production et/ou la consommation des biens de ces secteurs culturels (Auray et Georges, 2012 ; Proulx, 2017 ; Cohendet, Grandadam et Simon, 2008 ; Asur et Huberman, 2010 ; Throsby, 2001 ; Pélissier et Chaudy, 2009).

4.7 Conclusion

La détection de communautés permet d'identifier des groupes très divers dans un réseau social. Ce Chapitre instancie notre méthodologie permettant de choisir un algorithme de détection de communauté pertinent, en relation avec une problématique métier, parmi une douzaine d'algorithmes bien connus, en apportant un éclairage différentiant sur les formes de coopération non directement observables sur une plateforme de crowdfunding.

Le choix d'une méthode particulière n'est pas un choix facile ou neutre. Comme nous le montrons ici, selon les méthodes de partition, les praticiens obtiennent un éventail de types de communautés différents. Décider d'une méthode peut impacter fortement les résultats finaux de leur analyse. Ce travail prouve qu'une manière précise de choisir une méthode appropriée est une tâche complexe. En particulier dans le contexte d'études exploratoires, il est nécessaire de combiner une série de techniques, par exemple, dans notre cas, les similarités des partitions, les critères qualitatifs et les indicateurs structurels.

En accord avec Smith et al., 2020, cette étude confirme que le choix d'une méthode est déterminé par le contexte et la problématique de la recherche. Des techniques supplémentaires, des données et des indicateurs spécifiques permettent de réduire le champ des options disponibles. L'alignement avec la question de recherche des praticiens joue un rôle crucial dans le choix final. Dans le cadre de l'étude de cas présentée, le choix de la méthode *Edge Betweenness* résulte de l'analyse des caractéristiques socio-économiques et de l'exploration de la distribution des profils des *Ululers*. Ainsi, nous avons identifié 3 Familles de communautés sur la plateforme ; leurs caractéristiques distinctives, à savoir l'organisation, le nombre de participants, l'intensité de la collaboration, la spécialisation thématique et la performance dans les campagnes de collecte de fonds. En fonction du contexte et des données disponibles, différents indicateurs socio-économiques peuvent être mobilisés pour obtenir la classification des communautés et une série d'autres questions orientées business peuvent être abordées : par exemple, la distribution précise des formes de chaînes, d'étoiles ou de cliques dans les familles de communautés, le cycle de vie et la dynamique d'évolution des communautés et bien d'autres.

D'un point de vue méthodologique, le choix des métriques utilisées dans les cartes bivariées restent une question ouverte. Si la *transitivité* et la *hub dominance* paraissent appropriées pour caractériser les formes de collaboration (large gamme d'organisations plus ou moins centralisées, plus ou moins horizontales), il n'en demeure pas moins que cette technique à elle-seule ne permet pas de faire émerger des structurations plus complexes. Par exemple, nous l'avons vu, les communautés moyennes à grandes montrent des signes de centralisation locale mais reste globalement proches de structures de types chaînes, peu dense en arêtes. Choisir une méthode telle que *Walktrap* permet de réduire la taille des communautés, mais par la même, nous prive de ces communautés core-periphery qui peut peuvent avoir un sens, en l'occurrence dans notre cas, les Familles 1 et 2 n'étaient pas représentées. Il y a bien un niveau d'échelle à gérer. Cela suppose-t'il d'inventer de nouvelles métriques ? De mener jusqu'au bout conjointement l'analyse sur des partitions consensuelles différentes, ce qui revient à ne pas choisir une méthode, mais plusieurs ? Explorer d'autres techniques de combinaison de métriques qualitatives ?

Thomas Sychterz, COO North America & CMO at Ulule, published a comment to our study (Lyubareva et al., 2020) advertised on LinkedIn in June 2020 :

« Funny timing ! So we just launched our very first online Pitch Pitch today, supercharged by the new group of Ulule superbackers (The Jury of Ululers).

Literally at the same time, I see an amazing 20 page deep-dive analysis (completely independent !) published by Inna Lyuberava, Cécile Bothorel, Laurent Brisson and Romain Billot in Revue Française de Gestion regarding the power of, specifically, the Ulule community of over 3 Million members. They've identified 5 types of backers in the Ulule community, with specific types of characteristics and profiles.

Long story short : there's a real network effect created by superbackers and a significant potential to leverage the input of Specialist Backers, get insight from Precursor backers, create interesting content for Follower Backers and get the Collaborative backers involved in projects in more than just monetary ways.

More than ever, we need help entrepreneurs and creators as much as possible to get funding, exposure and reach for their good ideas, projects and businesses. Not only does the Ulule staff offer unparalleled support and individualized coaching, free of charge, but now we're working on getting the torque of our superbackers to help push projects even further. »

Conclusion

Tout au long du manuscrit, j'ai déjà donné quelques conclusions relatives aux différents travaux présentés et j'ai discuté de certaines questions et perspectives immédiates qu'ils soulèvent. En guise de conclusion générale, je voudrais décrire quelques axes de recherche ouverts auxquels je crois particulièrement pour le domaine des réseaux complexes.

Graphes temporels et dynamique communautaire

Dans les travaux présentés dans ce document, les réseaux sont statiques, ou s'ils ne le sont pas, nous en avons ignoré la composante temporelle. Pourtant analyser la dynamique communautaire s'avère pertinente pour étudier les phénomènes sociaux et en comprendre les mécanismes sous-jacents. Il s'agit de l'un des grands enjeux de la décennie : modéliser la dynamique des réseaux pour améliorer la performance des tâches de prédiction, de classification. Dans mon cas, c'est une approche plus descriptive qui m'intéresse : comprendre la dynamique, et en particulier celle des communautés.

Si nous prenons l'exemple d'un forum de discussion, le réseau temporel modélisant une discussion ou un ensemble de discussions est une séquence d'arêtes datées, i.e. de tuples (t, u, v). Pour analyser la dynamique communautaire, la technique couramment utilisée est de transformer les flux de liens en séquences de graphes statiques, appelés "snapshots". Le temps est découpé en périodes, et les interactions de chaque période sont agrégées pour constituer un graphe statique. Sur chacun des graphes statiques, il s'agit d'appliquer un algorithme de détection de communautés. L'objectif est ensuite d'apparier les communautés entre partitions consécutives, de façon à reconstituer des communautés évolutives. Celles-ci, appelées parfois communautés dynamiques, sont ainsi décrites par un durée et des flucatuations de membres au fil du temps, d'un snapshot à l'autre.

De nombreux problèmes restent à ce jour non résolus. Le premier est l'échantillonnage du temps. Certains réseaux, comme ceux qui modélisent les collaborations, par exemple la co-rédaction de publications scientifiques, ont une échelle de temps naturelle (chaque article scientifique a une année de publication). Pour autant, est-ce que la taille de la fenêtre temporelle pertinente est l'année ? Si nous élargissons à plusieurs années, nous pourrions observer des phénomènes plus lents, des tendances à rapprocher des disciplines par exemple, ou au contraire à éclater une discipline en sous-thèmes. Pour d'autres cas, un forum de discussion par exemple, nous n'avons aucune intuition quant à la période adéquate à étudier. Les phénomènes qui régissent la modification des communautés sont-ils hebdomadaires, mensuels, annuels ? Sont-ils homogènes en termes de durée d'une communauté à l'autre ? Et pour une communauté évolutive donnée, a-t'elle une rythmique régulière ou bien la fenêtre temporelle doit-elle être réajustée en permanence au fil du temps ? A ma

connaissance, très peu de travaux abordent cette question dans le domaine des réseaux complexes, et tout au plus la signalent.

Pourtant le domaine des séries temporelles proposent des solutions pour la segmentation et la compression de l'information, notamment via l'identification de points importants (PIP pour perceptually important points) dans une courbe, technique largement utilisée en finance (Fu, 2011). La segmentation pourrait passer par la détection d'événements dans le flux d'interactions. Mais alors quels types d'événements ? Des événements liés à la volumétrie des interactions, comme un pic d'interactions ? On se ramènerait alors au problème bien adressé de la segmentation de série temporelle. Mais qu'en est-il d'événements à la fois structurels et temporels ? Existe-t'il une mesure d'activité, ou une combinaison de mesures, reflétant une modification significative de la structure modulaire ?

La détection d'anomalies adresse en partie ce problème, en détectant des *change points* au niveau d'un nœud, d'une arête ou d'un sous-graphe. Par point de changement, on entend ici une situation différente par rapport à une situation de référence. Une anomalie peut avoir plusieurs formes : une arête qui ne devrait pas exister entre deux noeuds, un nœud diffèrent des autres nœuds, un sous-graphe plus dense que le reste du graphe, ou encore un changement de voisinage. Mais ces méthodes n'exploitent pas toute la richesse de l'analyse de séries temporelles (changement de phases, saisonnalité) ni la richesse de l'analyse de la topologie (voisinage étendu, motifs complexes, communautés).

Le deuxième problème lié à la détection de communautés évolutives via des séries de graphes statiques est l'instabilité de communautés entre pas de temps consécutifs. En effet, les algorithmes statiques utilisés sont très souvent non déterministes, et peuvent sur le même graphe révéler des partitions différentes ; lorsque le graphe évolue, même très peu, cela ne fait qu'accroître le risque d'incohérences, et cela même dans le cas de méthodes non stochastiques. Il est très difficile de savoir si l'instabilité des partitions est due à la modification de graphe de l'instant $t + 1$ par rapport à l'instant t, ou si cela résulte de l'algorithme, voire même de la segmentation, qui si elle trop fine, peut générer des graphes très différents entre 2 pas de temps, en ne mettant en avant que des artefacts furtifs par exemple.

L'enjeu est ici pourtant de découvrir des partitions stables pour en retracer une évolution. Des mécanismes de lissage temporel existent et sont couramment utilisés, mais le choix de la technique a des impacts forts sur le résultat (Rossetti et Cazabet, 2018). Privilégie-t'on une stratégie globale (par l'optimisation d'une métrique sur toutes les partitions temporelles par exemple) ou une stratégie locale (recalcul des communautés impactées par un changement entre t et $t + 1$) ? Cela fait écho avec la problématique du choix d'algorithme statique lui-même, comme nous l'avons abordé dans le Chapitre 3. Le choix de l'algorithme en contexte de série temporelle de graphes reste à traiter, mais s'ajoute à cela, entre autres le choix des snapshots et des mécanismes de lissage.

Le troisième problème est lié à la reconstitution des communautés évolutives. Il s'agit du mécanisme d'appariement. Comment faire le *matching* des communautés statiques détectées à t et $t + 1$ comme étant deux stades de l'évolution de la même communauté ? Quid des fusions, scissions, qui peuvent se produire ? Là encore, il existe des solutions ensemblistes (des *matching metrics*) qui permettent de calculer des scores de similarités entre ensembles. L'impact du choix de ces métriques est très peu exploré à l'heure actuelle.

Le quatrième problème est lié à l'évaluation. Quelles métriques considérer ? Nous

n'avons pour l'heure pas de dataset avec une vérité terrain de communautés évolutives. L'une des approches pour pallier ce problème, est de générer des données synthétiques étiquetées. Nous travaillons sur ce point et allons délivrer dans les mois qui viennent un benchmark de communautés évolutives. L'évaluation se fait sur trois niveaux : 1) niveau *macro* où la groundtruth livre les communautés évolutives et les événements de naissance, fusion, scission, etc. selon la méthode ICEM (Mohammadmosaferi et Naderi, 2020) ; l'évaluation revient à comparer ces événements ; 2) niveau *meso*, celui des partitions de chaque snapshot, où les métriques classiques de comparaison de partitions (ARI par exemple) permettent de mesurer l'écart d'affectation des membres à des communautés statiques par rapport à la groundtruth ; 3) niveau *micro* où nous vérifions si les membres sont affectés aux bonnes communautés statiques, mais également, en terme de dynamique, s'ils transitent avec les nœuds attendus aux snapshots suivants.

Revisiter le formalisme et la notion de communauté

Le processus de détection de communautés évolutives sur une série de graphes temporels en lui-même est discutable. Il réside en effet sur le *principe d'agrégation*, qui consiste à former le graphe de tous les liens qui se produisent pendant une fenêtre temporelle donnée, et cela pour l'ensemble des fenêtres temporelles qui couvrent l'ensemble de la période d'étude. Certes, la motivation est de retomber sur un formalisme bien connu et outillé – le graphe statique —, mais nous l'avons vu, cette manipulation soulève autant de problèmes qu'elle n'en résout.

L'une des voies à explorer est d'adopter le formalisme du *link stream* ou flux d'arêtes. Dans ce cadre, l'idée est de traiter chaque nouvelle interaction au fil de l'eau, et donc de s'affranchir du découpage en snapshots. L'équipe Complex Networks du LIP6 développe de nombreux outils en lien avec ce formalisme, et redéfinit la notion de densité et de clique par exemple. Étendre ces travaux à la notion de communauté reste un problème ouvert.

Le passage aux flux d'interactions offre des perspectives de passage à l'échelle, tant en termes de volumétrie que de temps de calcul. En maintenant un modèle dynamique mis à jour de manière incrémentale, des travaux commencent à voir le jour pour maintenir à la volée des représentations latentes des nœuds (dits *node embeddings*), par exemple pour détecter des anomalies dans les échanges centrés autour de chacun des nœuds. La problématique de mise à jour incrémentale de clustering dynamique reste elle-aussi à explorer, à des fins de monitoring par exemple.

A noter que le formalisme du *link stream* offre une richesse d'outils du côté des séries temporelles. Si l'on considère la série des tuples (t, e), il devient possible d'analyser l'évolution de chaque arête, et avec des outils de type ARIMA (et bien d'autres) de trouver des régularités d'évolution, des périodicités ou encore des anomalies. Des travaux comme ceux initiés par Bautista et Latapy vont plus loin : ils montrent comment décomposer un graphe temporel selon des séries temporelles de sous-structures, offrant ainsi un moyen d'étudier l'évolution de cliques, de triangles ou tout autres groupes de nœuds (Bautista et Latapy, 2022). Le choix du dictionnaire des sous-structures n'est cependant pas immédiat et relève d'un choix *a priori*. J'entrevois ici un champ de recherches en *dictionary learning* : comment définir un catalogue de sous-graphes qui soit représentatifs d'un graphe temporel ? Existe-t'il un tel dictionnaire, qui, à la manière des décompositions spectrales, serait la signature d'un link stream ? Si oui, existe-t'il une méthode pour les trouver ?

Plus généralement, on peut se poser la question de ce qu'est une communauté. Comme on l'a vu notamment dans le Chapitre 3.5, la définition dans les graphes statiques ne fait pas consensus, et au final dépend de la fonction objectif utilisée par les méthodes de partitionnement, ou de la méthode de clustering basée sur des embeddings. Dans notre étude concernant les vérités terrain de communautés (Dao, Bothorel et Lenca, 2017a), nous avons montré qu'il n'y avait pas adéquation entre ces groupes "naturels" issus des métadonnées, étiquetés, et la structure du graphe, et avons conclu, comme une boutade que les vérités terrain sont mal formées. Mais si nous prenons le problème dans l'autre sens, il se peut que les groundtruths basées sur les métadonnées ne soient pas bien représentés par la densité, ou du moins pas par cette seule mesure. D'autres caractéristiques, comme par exemple les corrélations de degrés, la densité des boucles (la densité des triangles ou autres structures plus lâches), peuvent jouer un rôle (Hric, Darst et Fortunato, 2014). *Hric et al.* avancent que le meilleur pari serait d'effectuer une étude détaillée des propriétés topologiques de ces communautés terrain, et de rester ouvert au fait que des composantes non-topologiques interviennent probablement dans la formation d'un groupe. C'est ici que les travaux sur les graphes avec attributs peuvent se révéler particulièrement intéressants, car leur pouvoir expressif permettrait de matérialiser ce qui confère un sentiment d'appartenance à une communauté, si bien-sûr on arrive à comprendre cette notion.

Dans le cadre dynamique, le concept de communauté est encore moins bien défini. On pourrait par exemple y voir des *groupes de nœuds qui interagissent en moyenne plus fréquemment que d'autres*. Ou... faire d'autres hypothèses. Mais c'est probablement en étudiant des plateformes aux "communautés" bien délimitées que nous pourrions en comprendre la nature. Des équipes de chercheurs à Oslo se concentrent par exemple sur la plateforme Reddit qui s'avère être une mine de données encore facilement accessible de nos jours. Les internautes s'y regroupent autour de centres d'intérêt. En considérant les sujets comme des communautés, le premier constat est *a priori* contre nature : une communauté, lorsqu'elle évolue, renouvelle ses membres ; et le noyau, loin d'être stable lui-même, se renouvelle également beaucoup. CE turnover existe dans tout groupe en réalité : les communautés de joueurs en ligne, le personnel d'un hôpital, un club sportif, une équipe pédagogique, etc.

Une étude exhaustive de telles communautés bien délimitées serait intéressante à mener. Le graal serait d'y trouver des lois décrivant la dynamique communautaire, à la manière des études qui ont permis de révéler le principe de l'*attachement préférentiel*, ou la nature *scale-free* et *petit monde* des grands graphes réels à la fin des années 90. Une alternative, plus réaliste, est de s'appuyer sur les compétences de nos collègues chercheurs en Sciences Humaines et Sociales.

4.8 Computational Social Science

Etudier des terrains bien délimités, théoriser les comportements, proposer des modèles sont le lot quotidien des chercheurs en Sciences Humaines et Sociales. J'ai eu la chance d'expérimenter depuis une dizaine d'années, plusieurs collaborations avec des chercheurs de différentes disciplines, en marketing, en économie, ou encore management, à l'occasion de la Chaire Réseaux Sociaux (Section 1.4 du Chapitre 1), du projet ANR PIL 2018-2022 (analyse du phénomène de chambres d'écho sur YouTube, (Bothorel, Brisson et Lyubareva, 2022)), de l'étude de la plate-forme de financement participatif Ulule (Chapitre 3) ou encore du phénomène de promotion dans Wikipédia (Picot Clemente, Bothorel et Jullien, 2015).

Comme nous l'avons vu en détail avec le travail réalisé sur la plateforme Ulule, à travers une recherche appliquée, se posent des questions techniques mais surtout scientifiques. Ces questions sont profitables à toutes les disciplines impliquées. Chaque cas d'étude est l'occasion d'expliciter des concepts (par exemple, qu'est-ce qu'une communauté ?), des modèles, de définir de nouvelles méthodologies d'analyse, mais également de lever des verrous scientifiques. Une bonne partie, si ce n'est la totalité des pistes de travaux que j'évoque plus haut dans cette conclusion, émane de nos réflexions communes pour analyser Ulule, YouTube ou Wikipédia. Réciproquement, grâce à nos problématiques, méthodes, algorithmes et mesures, les Sciences Humaines et Sociales peuvent affiner leurs modèles, en proposer de nouveaux, mais aussi les évaluer. Comme dit Gilles Dowek dans l'émission "La Conversation Scientifique" [1], *« Les réseaux d'ordinateurs sont aujourd'hui aux Sciences Humaines et Sociales ce que la lunette de Galilée fut à l'astronomie »*.

Mon implication dans ces recherches pluridisciplinaires s'est concrétisée tout d'abord dans le partage de vues scientifiques autour d'une même problématique et une compréhension mutuelle des méthodes et objets d'étude. Ensuite, ces collaborations rapprochées ont permis de travailler *ensemble*, de proposer de nouvelles problématiques et les méthodologies idoines. Enfin, nous convergeons vers le développement d'une boite à outils de détection et d'analyse des communautés et plus récemment de leur dynamique. L'idée est d'intégrer et enrichir ces outils au sein d'une plateforme et d'ouvrir cette plateforme à la communauté scientifique.

Mon ambition est de consolider ces recherches, et pourquoi pas de monter une équipe pluridisciplinaire *Computational Social Science*. Dans la même veine que le MediaLab de SciencePo ou encore l'observatoire OSoMe (prononcer awe•some) de l'université de l'Indiana, une telle équipe pourrait :

— Mener des projets pluridisciplinires combinant Data & Network Science et Sciences Humaines et Sociales s'inscrivant dans le long-terme,
— Constituer un vecteur de formation (initiale et continue),
— Développer des partenariats avec des entreprises et des administrations,
— Constituer une référence pour le débat public sur des thématiques telles que la presse, la transition numérique et les modèles organisationnels et de gouvernance.

Il s'agit ici de fédérer les chercheurs en Sciences Humaines Sociales d'IMT Atlantique (dans un premier temps) mais aussi d'augmenter les forces vives en informatique et analyse de réseaux complexes, de sorte à constituer une équipe de recherche dédiée à l'analyse de dynamiques communautaires.

1. Gilles Dowek, chercheur à l'Inria, et enseignant à l'Ecole Normale Supérieure de Paris-Saclay était l'invité de l'Episode "Homo sapiens devient-il homo informaticus ?", Emission "La Conversation Scientifique" par Etienne Klein, diffusée le Samedi 15 septembre 2018 sur France Culture, https://www.franceculture.fr/emissions/la-conversation-scientifique/homo-sapiens-devient-il-homo-informaticus

Bibliographie

Agrawal, Ajay, Christian Catalini, Avi Goldfarb et al. (2010). *Entrepreneurial finance and the flat-world hypothesis : Evidence from crowd-funding entrepreneurs in the arts*. Rapp. tech.

Agreste, Santa, Pasquale De Meo, Giacomo Fiumara, Giuseppe Piccione, Sebastiano Piccolo, Domenico Rosaci, Giuseppe M. L. Sarne et Athanasios V. Vasilakos (2017). « An Empirical Comparison of Algorithms to Find Communities in Directed Graphs and Their Application in Web Data Analytics ». In : *IEEE Transactions on Big Data* 3.3, p. 289-306. DOI : 10.1109/tbdata.2016.2631512. URL : https://doi.org/10.1109/tbdata.2016.2631512.

Albert, R., H. Jeong et A.-L. Barabási (sept. 1999). « Internet : Diameter of the World-Wide Web ». In : *Nature* 401, p. 130-131. DOI : 10.1038/43601.

Aldrich, Howard et Diane Herker (1977). « Boundary Spanning Roles and Organization Structure ». English. In : *The Academy of Management Review* 2.2, pp. 217-230. ISSN : 03637425. URL : http://www.jstor.org/stable/257905.

Ames, Brendan P. W. (2013). « Guaranteed clustering and biclustering via semidefinite programming ». In : *Mathematical Programming* 147.1-2, p. 429-465. DOI : 10.1007/s10107-013-0729-x. URL : https://doi.org/10.1007/s10107-013-0729-x.

Asur, Sitaram et Bernardo A Huberman (2010). « Predicting the future with social media ». In : *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, p. 492-499.

Auray, Nicolas et Fanny Georges (2012). « Les productions audiovisuelles des joueurs de jeux vidéo ». In : *Réseaux* 5, p. 145-173.

Barabási, Albert-László (2002). *Linked : The New Science of Networks*. Cambridge, MA, Perseus Publishing.

Barabási, Albert-László et Réka Albert (1999). « Emergence of Scaling in Random Networks ». In : *Science* 286.5439, p. 509-512. ISSN : 0036-8075. DOI : 10.1126/science.286.5439.509. URL : http://science.sciencemag.org/content/286/5439/509.

Barrat, A., M. Barthélemy, R. Pastor-Satorras et A. Vespignani (2004). « The architecture of complex weighted networks ». In : *Proceedings of the National Academy of Sciences* 101.11, p. 3747-3752. ISSN : 0027-8424. DOI : 10.1073/pnas.0400087101. eprint : http://www.pnas.org/content/101/11/3747.full.pdf. URL : http://www.pnas.org/content/101/11/3747.

Bautista, Esteban et Matthieu Latapy (2022). *A Frequency-Structure Approach for Link Stream Analysis*. arXiv : 2212.03804 [eess.SP].

Beaudouin, Valérie, Tomas Legon et Dominique Pasquier (2016). *«Moi je lui donne 5/5» : Paradoxes de la critique amateur en ligne*. Presses des Mines via OpenEdition.

Bellman, R.E. (1961). *Adaptive control processes : a guided tour*. Rand Corporation Research studies. Princeton University Press. URL : http://books.google.fr/books?id=POAmAAAAMAAJ.

Benamar, Lamya, Christine Balagué et Mohamad Ghassany (2017). « The identification and influence of social roles in a social media product community ». In : *Journal of Computer-Mediated Communication* 22.6, p. 337-362.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte et Etienne Lefebvre (2008). « Fast unfolding of communities in large networks ». In : *Journal of Statistical Mechanics : Theory and Experiment* 2008.10, P10008 (12pp). ISSN : 1742-5468. DOI : 10.1088/1742-5468/2008/10/P10008. URL : http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Bohlin, Ludvig, Daniel Edler, Andrea Lancichinetti et Martin Rosvall (2014). « Community Detection and Visualization of Networks with the Map Equation Framework ». In : *Measuring Scholarly Impact*. Springer International Publishing, p. 3-34. DOI : 10.1007/978-3-319-10377-8_1. URL : https://doi.org/10.1007/978-3-319-10377-8_1.

Bothorel, Cécile, Laurent Brisson et Inna Lyubareva (juin 2021). « How to Choose Community Detection Methods in Complex Networks ». In : *Series Computational Social Science*. Methods and applications in social networks analysis. Evidence from Collaborative, Governance, Historical and Mobility Networks. ISBN 9788835124603, pp 16-36. URL : http://ojs.francoangeli.it/_omp/index.php/oa/catalog/book/682.

Bothorel, Cécile, Laurent Brisson et Inna Lyubareva (2022). « Plateformes en ligne et analyse des dynamiques communautaires ». In : *Diversité des approches méthodologiques en sciences sociales*. ISTE. URL : https://hal.science/hal-03755833.

Bothorel, Cécile, Juan David Cruz Gomez, Magnani Matteo et Barbora Micenkova (sept. 2015). « Clustering attributed graphs : models, measures and methods ». In : *Network Science* 3.03, p. 408 -444. DOI : 10.1017/nws.2015.9. URL : https://hal.archives-ouvertes.fr/hal-01257833.

Bourqui, R., D. Auber et P. Mary (2007). « How to Draw Clustered-Weighted Graphs using a Multilevel Force-Directed Graph Drawing Algorithm ». In : *Information Visualization, 2007. IV '07. 11th International Conference*, p. 757 -764. DOI : 10.1109/IV.2007.65.

Burt, Ronald S (1992). *Structural holes*. Harvard university press.

Cao, Jinxin, Di Jin, Liang Yang et Jianwu Dang (2018). « Incorporating network structure with node contents for community detection on large networks using deep learning ». In : *Neurocomputing*. ISSN : 09252312. DOI : 10.1016/j.neucom.2018.01.065. URL : http://linkinghub.elsevier.com/retrieve/pii/S0925231218300985.

Cao, Qiang, Michael Sirivianos, Xiaowei Yang et Tiago Pregueiro (2012). « Aiding the detection of fake accounts in large scale social online services ». In : *9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, p. 197-210.

Chakraborty, Tanmoy, Ayushi Dalmia, Animesh Mukherjee et Niloy Ganguly (août 2017). « Metrics for Community Analysis : A Survey ». In : *ACM Comput. Surv.* 50.4, p. 1-37. ISSN : 0360-0300. DOI : 10.1145/3091106. URL : http://doi.acm.org/10.1145/3091106.

Chen, Qi et Lingwei Wei (2019). « Overlapping Community Detection of Complex Network : A Survey ». In : *2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*. IEEE, p. 513-516.

Chunaev, Petr (2020). « Community detection in node-attributed social networks : a survey ». In : *Computer Science Review* 37, p. 100286.

Clauset, Aaron, M. E. J. Newman et Cristopher Moore (déc. 2004). « Finding community structure in very large networks ». In : *Physical Review E* 70.6. DOI : 10.1103/physreve.70.066111. URL : https://doi.org/10.1103/physreve.70.066111.

Cohendet, Patrick, David Grandadam et Laurent Simon (2008). « Réseaux, communautés et projets dans les processus créatifs ». In : *Management international* 13.1, p. 29.

Combe, David, Christine Largeron, Elöd Egyed-Zsigmond et Mathias Géry (2012). « Combining relations and text in scientific network clustering ». In : *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, p. 1248-1253.

Coscia, Michele, Fosca Giannotti et Dino Pedreschi (sept. 2011). « A classification for community discovery methods in complex networks ». In : *Statistical Analysis and Data Mining* 4.5, p. 512-546. DOI : 10.1002/sam.10133. URL : https://doi.org/10.1002/sam.10133.

Creusefond, Jean (fév. 2017). « Caractériser et détecter les communautés dans les réseaux sociaux ». Theses. Normandie Université. URL : https://tel.archives-ouvertes.fr/tel-01497593.

Cross, Rob et Andrew Parker (2004). *The hidden power of social networks : Understanding how work really gets done in organizations*. Sous la dir. d'Harvard Business School Press. Harvard Business School Press.

Cruz Gomez, Juan David et Cécile Bothorel (août 2013). « Information integration for detecting communities in attributed graphs ». In : *CASoN 2013 : 5th IEEE International Conference on Computational Aspects of Social Networks*. Fargo, United States. URL : https://hal.archives-ouvertes.fr/hal-00857229.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (oct. 2010). « Détection de communautés dans les réseaux socio-sémantiques par point de vue ». In : *Journée thématique : fouille de grands graphes*. Toulouse, France. URL : https://hal.archives-ouvertes.fr/hal-00540871.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (oct. 2011a). « Entropy based community detection in augmented social networks ». In : *International Conference on Computational Aspects of Social Networks*. Salamanca, Spain, p. 163-168. URL : https://hal.archives-ouvertes.fr/hal-00640722.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (jan. 2011b). « Identification et visualisation des partitions de réseaux sociaux à l'aide de points de vue sémantiques ». In : *AVEC 2011 : 9e atelier visualisation et extraction des connaissances - EGC 2011 : 11e conférence internationale francophone sur l'extraction et la gestion des connaissances*. Brest, France, p. 25-36. URL : https://hal.archives-ouvertes.fr/hal-00767746.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (mars 2011c). « Semantic clustering of social networks using points of view ». In : *CORIA : conférence en recherche d'information et applications 2011*. Avignon, France. URL : https://hal.archives-ouvertes.fr/hal-00609291.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (oct. 2012). « Détection et visualisation des communautés dans les réseaux sociaux augmentés ». In : *Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle* 26.4, p. 369 -392. DOI : 10.3199/RIA.26.369-392. URL : https://hal.archives-ouvertes.fr/hal-00739426.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (déc. 2013a). « Community detection and visualization in social networks : integrating structural and

semantic information ». In : *ACM Transactions on Intelligent Systems and Technology* 5.1, p. 11-26. DOI : 10.1145/2542182.2542193. URL : https://hal.archives-ouvertes.fr/hal-00763931.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (août 2013b). « Integrating heterogeneous information within a social network for detecting communities ». In : *ASONAM 2013 : the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Niagara Falls, Canada. URL : https://hal.archives-ouvertes.fr/hal-00857225.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (mai 2013c). « Layout Algorithm for Clustered Graphs to Analyze Community Interactions in Social Networks ». In : *INSNA 2013 : XXXIII Sunbelt Social Networks Conference of the International Network for Social Network Analysis*. Hamburg, Germany, p. 2. URL : https://hal.archives-ouvertes.fr/hal-00780515.

Cruz Gomez, Juan David, Cécile Bothorel et François Poulet (2014). « Analyse intégrée des réseaux sociaux pour la détection et la visualisation de communautés ». In : *Revue des Sciences et Technologies de l'Information - Série TSI : Technique et Science Informatiques* 33.4, p. 399-427. URL : https://hal.archives-ouvertes.fr/hal-00937849.

Cui, P., X. Wang, J. Pei et W. Zhu (2019). « A Survey on Network Embedding ». In : *IEEE Transactions on Knowledge and Data Engineering* 31.5, p. 833-852. DOI : 10.1109/TKDE.2018.2849727.

Danon, Leon, Albert Díaz-Guilera, Jordi Duch et Alex Arenas (sept. 2005). « Comparing community structure identification ». In : *Journal of Statistical Mechanics : Theory and Experiment* 2005.09, P09008-P09008. DOI : 10.1088/1742-5468/2005/09/p09008. URL : https://doi.org/10.1088/1742-5468/2005/09/p09008.

Dao, Vinh-Loc (déc. 2018). « Characterizing community detection algorithms and detected modules in large-scale complex networks ». Theses. Ecole nationale supérieure Mines-Télécom Atlantique. URL : https://tel.archives-ouvertes.fr/tel-02121358.

Dao, Vinh Loc, Cécile Bothorel et Philippe Lenca (2017a). « Community detection methods can discover better structural clusters than ground-truth communities ». In : *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM 17*. ACM Press. DOI : 10.1145/3110025.3110053. URL : https://doi.org/10.1145/3110025.3110053.

Dao, Vinh-Loc, Cécile Bothorel et Philippe Lenca (2017b). « Community structures evaluation in complex networks : A descriptive approach ». In : *NetSci-X 2017 : International School and Conference on Network Science*. 3rd International Winter School and Conference on Network Science (NetSci-X 2017). Tel Aviv, Israel, p. 11-19. DOI : 10.1007/978-3-319-55471-6_2. URL : https://hal.archives-ouvertes.fr/hal-01513246.

Dao, Vinh-Loc, Cécile Bothorel et Philippe Lenca (déc. 2018). « Estimating the similarity of community detection methods based on cluster size distribution ». In : *Complex Networks 2018, The 7th International Conference on Complex Networks and Their Applications*. T. 812. Studies in Computational Intelligence. Cambridge, United Kingdom : Springer, p. 183-194. URL : https://hal.archives-ouvertes.fr/hal-01911077.

Dao, Vinh Loc, Cécile Bothorel et Philippe Lenca (mars 2020). « Community structure : A comparative evaluation of community detection methods ». In : *Network Science* 8.1, p. 1-41. DOI : 10.1017/nws.2019.59. URL : https://hal.archives-ouvertes.fr/hal-02459508.

Dao, Vinh-Loc, Cécile Bothorel et Philippe Lenca (déc. 2021). « An empirical characterization of community structures in complex networks using a bivariate map of quality metrics ». In : *Social Network Analysis and Mining* 11.1. 18 pages, 12 figures, 41 reference items, p. 37. DOI : 10.1007/s13278-021-00743-1. URL : https://hal-imt-atlantique.archives-ouvertes.fr/hal-01809064.

Eades, Peter et Qing-Wen Feng (1997). « Multilevel visualization of clustered graphs ». In : *Graph Drawing*. Sous la dir. de Stephen North. T. 1190. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, p. 101-112. URL : http://dx.doi.org/10.1007/3-540-62495-3-41.

Enright, Anton J, Stijn Van Dongen et Christos A Ouzounis (2002). « An efficient algorithm for large-scale detection of protein families ». In : *Nucleic acids research* 30.7, p. 1575-1584.

Erdős, P. et A. Rényi (1959). « On random graphs, I ». In : *Publicationes Mathematicae (Debrecen)* 6, p. 290-297. URL : http://www.renyi.hu/\~{}p_erdos/Erdos.html\#1959-11.

Falih, Issam, Nistor Grozavu, Rushed Kanawati et Younès Bennani (2018). « Community detection in attributed network ». In : *Companion Proceedings of the The Web Conference 2018*, p. 1299-1306.

Forsé, Michel (2008). « Définir et analyser les réseaux sociaux ». In : *Informations sociales* 3, p. 10-19.

Fortunato, S. et M. Barthelemy (déc. 2006). « Resolution limit in community detection ». In : *Proceedings of the National Academy of Sciences* 104.1, p. 36-41. DOI : 10.1073/pnas.0605965104. URL : https://doi.org/10.1073/pnas.0605965104.

Fortunato, Santo (fév. 2010). « Community detection in graphs ». In : *Physics Reports* 486.3-5, p. 75-174. DOI : 10.1016/j.physrep.2009.11.002. URL : https://doi.org/10.1016/j.physrep.2009.11.002.

Fortunato, Santo et Darko Hric (nov. 2016). « Community detection in networks : A user guide ». In : *Physics Reports* 659, p. 1-44. DOI : 10.1016/j.physrep.2016.09.002. URL : https://doi.org/10.1016/j.physrep.2016.09.002.

Fronczak, Agata, Piotr Fronczak et Janusz A. Hołyst (oct. 2003). « Mean-field theory for clustering coefficients in Barabási-Albert networks ». In : *Physical Review E* 68.4. DOI : 10.1103/physreve.68.046126. URL : https://doi.org/10.1103/physreve.68.046126.

Fu, Tak chung (2011). « A review on time series data mining ». In : *Engineering Applications of Artificial Intelligence* 24.1, p. 164-181. ISSN : 0952-1976. DOI : https://doi.org/10.1016/j.engappai.2010.09.007. URL : https://www.sciencedirect.com/science/article/pii/S0952197610001727.

Ghalmane, Zakariya, Mohammed El Hassouni et Hocine Cherifi (2019). « Immunization of networks with non-overlapping community structure ». In : *Social Network Analysis and Mining* 9.1, p. 45.

Ghasemian, A., H. Hosseinmardi et A. Clauset (fév. 2018). « Evaluating Overfit and Underfit in Models of Network Community Structure ». In : *ArXiv e-prints*. arXiv : 1802.10582 [stat.ML].

Giacomo, Emilio. di, W. Didimo, L. Grilli et G. Liotta (2007). « Graph Visualization Techniques for Web Clustering Engines ». In : *Visualization and Computer Graphics, IEEE Transactions on* 13.2, p. 294 -304. ISSN : 1077-2626. DOI : 10.1109/TVCG.2007.40.

Girvan, M. et M. E. J. Newman (juin 2002). « Community structure in social and biological networks ». In : *Proceedings of the National Academy of Sciences* 99.12, p. 7821-7826. DOI : 10.1073/pnas.122653799. URL : https://doi.org/10.1073/pnas.122653799.

Granovetter, Mark S (1973). « The strength of weak ties ». In : *American journal of sociology* 78.6, p. 1360-1380.

Grover, Aditya et Jure Leskovec (2016). « Node2vec : Scalable Feature Learning for Networks ». In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA : Association for Computing Machinery, 855–864. ISBN : 9781450342322. DOI : 10.1145/2939672.2939754. URL : https://doi.org/10.1145/2939672.2939754.

Guimera, Roger et Luis A Nunes Amaral (2005). « Cartography of complex networks : modules and universal roles ». In : *J. Stat. Mech.-Theory Exp.*, art. no. P02001. DOI : 10.1088/1742-5468/2005/02/P02001.

Helme-Guizon, Agnès, Fanny Magnoni et al. (2013). « Les marques sont mes amies sur Facebook : vers une typologie de fans basée sur la relation à la marque et le sentiment d'appartenance ». In : *Revue Française du marketing* 243.3, p. 5.

Hogan, Bernard (2011). *NameGenWeb*. Facebook Application. URL : http://apps.facebook.com/namegenweb/.

Hric, Darko, Richard K. Darst et Santo Fortunato (déc. 2014). « Community detection in networks : Structural communities versus ground truth ». In : *Physical Review E* 90.6. DOI : 10.1103/physreve.90.062805. URL : https://doi.org/10.1103/physreve.90.062805.

Hu, Weihua, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta et Jure Leskovec (2020). « Open graph benchmark : Datasets for machine learning on graphs ». In : *arXiv preprint arXiv :2005.00687*.

Hubert, Lawrence et Phipps Arabie (déc. 1985). « Comparing partitions ». In : *Journal of Classification* 2.1, p. 193-218. DOI : 10.1007/bf01908075. URL : https://doi.org/10.1007/bf01908075.

Inbar, Yael et Ohad Barzilay (2014). « Community impact on crowdfunding performance ». In.

Ingram, Stephen, Tamara Munzner et Marc Olano (2009). « Glimmer : Multilevel MDS on the GPU ». In : *IEEE Transactions on Visualization and Computer Graphics* 15, p. 249-261. ISSN : 1077-2626. DOI : http://doi.ieeecomputersociety.org/10.1109/TVCG.2008.85.

Jebabli, Malek, Hocine Cherifi, Chantal Cherifi et Atef Hamouda (2018). « Community detection algorithm evaluation with ground-truth data ». In : *Physica A : Statistical Mechanics and its Applications* 492, p. 651-706. ISSN : 0378-4371. DOI : https://doi.org/10.1016/j.physa.2017.10.018. URL : http://www.sciencedirect.com/science/article/pii/S0378437117310282.

Jerome, Kunegis (2013). « The Koblenz Network Collection ». In : *Proceedings Conference on World Wide Web Companion*, p. 1343-1350. URL : http://konect.uni-koblenz.de.

Jin, Di, Zhizhi Yu, Pengfei Jiao, Shirui Pan, Philip S. Yu et Weixiong Zhang (2021). *A Survey of Community Detection Approaches : From Statistical Modeling to Deep Learning*. arXiv : 2101.01669 [cs.SI].

Joe H. Ward, Jr. (1963). « Hierarchical Grouping to Optimize an Objective Function ». In : *Journal of the American Statistical Association* 58.301, p. 236-244. DOI : 10.1080/01621459.1963.10500845.

Kane, Gerald C. et Sam Ransbotham (2012). « Collaborative Development in Wikipedia ». In : *CoRR* abs/1204.3352. arXiv : 1204.3352. URL : http://arxiv.org/abs/1204.3352.

Kane, Gerald C et Sam Ransbotham (2016). « Research note—content and collaboration : an affiliation network approach to information quality in online peer production communities ». In : *Information Systems Research* 27.2, p. 424-439.

Kang, Zhao, Zhanyu Liu, Shirui Pan et Ling Tian (2022). « Fine-grained attributed graph clustering ». In : *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, p. 370-378.

Kipf, Thomas N et Max Welling (2016). « Variational graph auto-encoders ». In : *arXiv preprint arXiv :1611.07308*.

Klemm, Konstantin et Víctor M. Eguíluz (mai 2002). « Growing scale-free networks with small-world behavior ». In : *Physical Review E* 65.5. DOI : 10.1103/physreve.65.057102. URL : https://doi.org/10.1103/physreve.65.057102.

Kohonen, Teuvo (1997). *Self-Organizing Maps*. Sous la dir. de Teuvo Kohonen. Secaucus, NJ, USA : Springer-Verlag New York, Inc. ISBN : 3-540-62017-6.

Krackhardt, David et Jeffrey R Hanson (1993). « Informal networks ». In : *Harvard business review* 71.4, p. 104-111.

Kuppuswamy, Venkat et Barry L Bayus (2018). « Crowdfunding creative ideas : The dynamics of project backers ». In : *The Economics of Crowdfunding*. Springer, p. 151-182.

Labatut, V. et G. K. Orman (2017). « Community Structure Characterization ». In : *Encyclopedia of Social Network Analysis and Mining*. Springer New York, p. 1-13. DOI : 10.1007/978-1-4614-7163-9_110151-1. URL : https://doi.org/10.1007/978-1-4614-7163-9_110151-1.

Lambiotte, R. (2010). « Multi-scale modularity in complex networks ». In : *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, p. 546-553.

Lancichinetti, Andrea, Mikko Kivelä, Jari Saramäki et Santo Fortunato (août 2010). « Characterizing the Community Structure of Complex Networks ». In : *PLoS ONE* 5.8. Sous la dir. d'Olaf Sporns, e11976. DOI : 10.1371/journal.pone.0011976. URL : https://doi.org/10.1371/journal.pone.0011976.

Lancichinetti, Andrea, Filippo Radicchi, José J. Ramasco et Santo Fortunato (avr. 2011). « Finding Statistically Significant Communities in Networks ». In : *PLoS ONE* 6.4. Sous la dir. d'Eshel Ben-Jacob, e18961. DOI : 10.1371/journal.pone.0018961. URL : https://doi.org/10.1371/journal.pone.0018961.

Largillier, Thomas, Guillaume Peyronnet et Sylvain Peyronnet (2010). « SpotRank : a robust voting system for social news websites ». In : *Proceedings of the 4th workshop on Information credibility*, p. 59-66.

Leskovec, Jure et Andrej Krevl (juin 2014). *SNAP Datasets : Stanford Large Network Dataset Collection*. http ://snap.stanford.edu/data.

Leskovec, Jure, Kevin J. Lang, Anirban Dasgupta et Michael W. Mahoney (2008). « Statistical properties of community structure in large social and information networks ». In : *Proceeding of the 17th international conference on World Wide Web - WWW 08*. DOI : 10.1145/1367497.1367591. URL : https://doi.org/10.1145/1367497.1367591.

Leydesdorff, Loet et Inga Ivanova (2020). « The measurement of "interdisciplinarity" and "synergy" in scientific and extra-scientific collaborations ». In : *Journal of the Association for Information Science and Technology*.

Lyubareva, Inna, Laurent Brisson, Cécile Bothorel et Romain Billot (juin 2019). « Crowdfunding platform and inter-project social network : example of Ulule ». In : *SASE 2019 : Fathomless Futures : Algorithmic and Imagined*. New York City, United States. URL : https://hal.archives-ouvertes.fr/hal-02505858.

Lyubareva, Inna, Laurent Brisson, Cécile Bothorel et Romain Billot (juin 2020). « Une plateforme de crowdfunding et son réseau social ». In : *Revue Française de Gestion*. Les mutations de l'accompagnement entrepreneurial 1.286, p. 135-151. DOI : 10.

3166/rfg.2019.00402. URL : https://hal.archives-ouvertes.fr/hal-02880083.

Meo, Pasquale De, Emilio Ferrara, Giacomo Fiumara et Alessandro Provetti (fév. 2014). « Mixing local and global information for community detection in large networks ». In : *Journal of Computer and System Sciences* 80.1, p. 72-87. DOI : 10.1016/j.jcss.2013.03.012. URL : https://doi.org/10.1016/j.jcss.2013.03.012.

Mercanti-Guérin, Maria (2010). « Analyse des réseaux sociaux et communautés en ligne : quelles applications en marketing ? » In : *Management & Avenir* 2, p. 132-153.

Milgram, Stanley (1967). « The small world problem ». In : *Psychology today* 2.1, p. 60-67.

Mohammadmosaferi, Kaveh Kadkhoda et Hassan Naderi (2020). « Evolution of communities in dynamic social networks : An efficient map-based approach ». In : *Expert Systems with Applications* 147, p. 113221.

Moreno, Jacob Levy (1934). « Who shall survive ? : A new approach to the problem of human interrelations. » In.

Nerurkar, Pranav, Madhav Chandane et Sunil Bhirud (2019). « A Comparative Analysis of Community Detection Algorithms on Social Networks ». In : *Computational Intelligence : Theories, Applications and Future Directions - Volume I*. Sous la dir. de Nishchal K. Verma et A. K. Ghosh. Singapore : Springer Singapore, p. 287-298. ISBN : 978-981-13-1132-1.

Newman, M. E. et M. Girvan (fév. 2004). « Finding and evaluating community structure in networks ». In : *Phys. Rev. E* 69.2, 026113, p. 026113. DOI : 10.1103/PhysRevE.69.026113. eprint : cond-mat/0308217.

Newman, M. E. J. (2006). « Finding community structure in networks using the eigenvectors of matrices ». In : *Phys. Rev. E* 74 (3), p. 036104. DOI : 10.1103/PhysRevE.74.036104. URL : https://link.aps.org/doi/10.1103/PhysRevE.74.036104.

Olteanu, Madalina, Nathalie Villa-Vialaneix et Christine Cierco-Ayrolles (2013). « Multiple kernel self-organizing maps ». In : *21. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013)*.

Orman, Günce Keziban, Vincent Labatut et Hocine Cherifi (août 2012). « Comparative evaluation of community detection algorithms : a topological approach ». In : *Journal of Statistical Mechanics : Theory and Experiment* 2012.08, P08001. DOI : 10.1088/1742-5468/2012/08/p08001. URL : https://doi.org/10.1088/1742-5468/2012/08/p08001.

Page, Lawrence, Sergey Brin, Rajeev Motwani et Terry Winograd (1999). *The PageRank citation ranking : Bringing order to the web.* Rapp. tech. Stanford InfoLab.

Pan, Shirui, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao et Chengqi Zhang (2018). « Adversarially regularized graph autoencoder for graph embedding ». In : *arXiv preprint arXiv :1802.04407*.

Papadopoulos, Symeon, Yiannis Kompatsiaris, Athena Vakali et Ploutarchos Spyridonos (juin 2011). « Community detection in Social Media ». In : *Data Mining and Knowledge Discovery* 24.3, p. 515-554. DOI : 10.1007/s10618-011-0224-z. URL : https://doi.org/10.1007/s10618-011-0224-z.

Peel, Leto, Daniel B. Larremore et Aaron Clauset (mai 2017). « The ground truth about metadata and community detection in networks ». In : *Science Advances* 3.5, e1602548. DOI : 10.1126/sciadv.1602548. URL : https://doi.org/10.1126/sciadv.1602548.

Pélissier, Nicolas et Serge Chaudy (2009). « Le journalisme participatif et citoyen sur Internet : un populisme dans l'air du temps ? » In : *Quaderni. Communication, technologies, pouvoir* 70, p. 89-102.

Perozzi, Bryan, Rami Al-Rfou et Steven Skiena (2014). « Deepwalk : Online learning of social representations ». In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 701-710.

Picot Clemente, Romain, Cécile Bothorel et Nicolas Jullien (août 2015). « Contribution, Social networking, and the Request for Adminship process in Wikipedia ». In : *OpenSym 2015 : 11th International Symposium on Open Collaboration*. San Francisco, United States : ACM. URL : https://hal.archives-ouvertes.fr/hal-01192597.

Pons, Pascal et Matthieu Latapy (2005). « Computing Communities in Large Networks Using Random Walks ». In : *Computer and Information Sciences - ISCIS 2005*. Sous la dir. de pInar Yolum, Tunga Güngör, Fikret Gürgen et Can Özturan. Springer Berlin Heidelberg, p. 284-293. ISBN : 978-3-540-32085-2.

Pons, Pascal et Matthieu Latapy (2011). « Post-processing hierarchical community structures : Quality improvements and multi-scale view ». In : *Theoretical Computer Science* 412.8-10, p. 892-900. DOI : 10.1016/j.tcs.2010.11.041. URL : https://doi.org/10.1016/j.tcs.2010.11.041.

Porter, M. A., J.-P. Onnela et P. J. Mucha (fév. 2009). « Communities in Networks ». In : *Notices of the American Mathematical Society* 9.

Proulx, Serge (2017). « L'injonction à participer au monde numérique ». In : *Communiquer. Revue de communication sociale et publique* 20, p. 15-27.

Radicchi, F., C. Castellano, F. Cecconi, V. Loreto et D. Parisi (fév. 2004). « Defining and identifying communities in networks ». In : *Proceedings of the National Academy of Sciences* 101.9, p. 2658-2663. DOI : 10.1073/pnas.0400054101. URL : https://doi.org/10.1073/pnas.0400054101.

Raghavan, Usha Nandini, Réka Albert et Soundar Kumara (sept. 2007). « Near linear time algorithm to detect community structures in large-scale networks ». In : *Physical Review E* 76.3. DOI : 10.1103/physreve.76.036106. URL : https://doi.org/10.1103/physreve.76.036106.

Rand, William M. (déc. 1971). « Objective Criteria for the Evaluation of Clustering Methods ». In : *Journal of the American Statistical Association* 66.336, p. 846-850. DOI : 10.1080/01621459.1971.10482356. URL : https://doi.org/10.1080/01621459.1971.10482356.

Reichardt, Jörg et Stefan Bornholdt (juill. 2006). « Statistical mechanics of community detection ». In : *Physical Review E* 74.1. DOI : 10.1103/physreve.74.016110. URL : https://doi.org/10.1103/physreve.74.016110.

Rheingold, Howard (2000). *The virtual community : Homesteading on the electronic frontier*. MIT press.

Richardson, Lizzie (2015). « Performing the sharing economy ». In : *Geoforum* 67, p. 121 -129. ISSN : 0016-7185. DOI : https://doi.org/10.1016/j.geoforum.2015.11.004. URL : http://www.sciencedirect.com/science/article/pii/S001671851530141X.

Riolo, Maria A., George T. Cantwell, Gesine Reinert et M. E. J. Newman (sept. 2017). « Efficient method for estimating the number of communities in a network ». In : *Physical Review E* 96.3. DOI : 10.1103/physreve.96.032310. URL : https://doi.org/10.1103/physreve.96.032310.

Rossetti, Giulio et Rémy Cazabet (2018). « Community discovery in dynamic networks : a survey ». In : *ACM Computing Surveys (CSUR)* 51.2, p. 1-37.

Rossi, Ryan A. et Nesreen K. Ahmed (2015). « The Network Data Repository with Interactive Graph Analytics and Visualization ». In : *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. URL : http://networkrepository.com.

Rosvall, M., D. Axelsson et C. T. Bergstrom (nov. 2009). « The map equation ». In : *European Physical Journal Special Topics* 178, p. 13-23. DOI : 10.1140/epjst/e2010-01179-1. arXiv : 0906.1405.

Rosvall, M. et C. T. Bergstrom (avr. 2007). « An information-theoretic framework for resolving community structure in complex networks ». In : *Proceedings of the National Academy of Sciences* 104.18, p. 7327-7331. DOI : 10.1073/pnas.0611034104. URL : https://doi.org/10.1073/pnas.0611034104.

Rosvall, Martin et Carl T. Bergstrom (2008). « Maps of random walks on complex networks reveal community structure ». In : *Proceedings of the National Academy of Sciences* 105.4, p. 1118-1123. ISSN : 0027-8424. DOI : 10.1073/pnas.0706851105. URL : http://www.pnas.org/content/105/4/1118.

Rusch, T Konstantin, Michael M Bronstein et Siddhartha Mishra (2023). « A Survey on Oversmoothing in Graph Neural Networks ». In : *arXiv preprint arXiv :2303.10993*.

Santamaría, Rodrigo et Roberto Therón (2008). « Overlapping Clustered Graphs : Co-authorship Networks Visualization ». In : *Smart Graphics*. Sous la dir. d'Andreas Butz, Brian Fisher, Antonio Krüger, Patrick Olivier et Marc Christie. T. 5166. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, p. 190-199. URL : http://dx.doi.org/10.1007/978-3-540-85412-8-17.

Schaub, Michael T., Jean-Charles Delvenne, Martin Rosvall et Renaud Lambiotte (fév. 2017). « The many facets of community detection in complex networks ». In : *Applied Network Science* 2.1. DOI : 10.1007/s41109-017-0023-6. URL : https://doi.org/10.1007/s41109-017-0023-6.

Shen, Xiaobo, Shirui Pan, Weiwei Liu, Yew-Soon Ong et Quan-Sen Sun (2018). « Discrete network embedding ». In : *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, p. 3549-3555.

Shi, Dan, Lei Zhu, Yikun Li, Jingjing Li et Xiushan Nie (2019). « Robust Structured Graph Clustering ». In : *IEEE Transactions on Neural Networks and Learning Systems*.

Smith, Natalie R, Paul N Zivich, Leah M Frerichs, James Moody et Allison E Aiello (2020). « A Guide for Choosing Community Detection Algorithms in Social Network Studies : The Question Alignment Approach ». In : *American Journal of Preventive Medicine* 59.4, p. 597-605.

Su, Xing et al. (2022). « A Comprehensive Survey on Community Detection With Deep Learning ». In : *IEEE Transactions on Neural Networks and Learning Systems*, p. 1-21. DOI : 10.1109/TNNLS.2021.3137396.

Tamassia, Roberto (1987). « On embedding a graph in the grid with the minimum number of bends ». In : *SIAM J. Comput.* 16 (3), p. 421-444. ISSN : 0097-5397. DOI : http://dx.doi.org/10.1137/0216030. URL : http://dx.doi.org/10.1137/0216030.

Throsby, David (2001). *Economics and culture*. Cambridge university press.

Toni, Alberto F. de et Fabio Nonino (2010). « The key roles in the informal organization : a network analysis perspective ». In : *The Learning Organization* 17.1, p. 86-103. DOI : 10.1108/09696471011008260. URL : https://doi.org/10.1108/09696471011008260.

Traud, Amanda L, Peter J Mucha et Mason A Porter (2012). « Social structure of facebook networks ». In : *Physica A : Statistical Mechanics and its Applications* 391.16, p. 4165-4180.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio et Yoshua Bengio (2017). « Graph attention networks ». In : *arXiv preprint arXiv :1710.10903*.

Vinh, Nguyen Xuan, Julien Epps et James Bailey (déc. 2010). « Information Theoretic Measures for Clusterings Comparison : Variants, Properties, Normalization and Correction for Chance ». In : *J. Mach. Learn. Res.* 11, p. 2837-2854. ISSN : 1532-4435. URL : http://dl.acm.org/citation.cfm?id=1756006.1953024.

Wang, Chun, Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang et Chengqi Zhang (2019). « Attributed graph clustering : A deep attentional embedding approach ». In : *arXiv preprint arXiv :1906.06532*.

Wang, Chun, Shirui Pan, Guodong Long, Xingquan Zhu et Jing Jiang (2017). « Mgae : Marginalized graph autoencoder for graph clustering ». In : *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, p. 889-898.

Wang, Yi-Chia et Robert Kraut (2012). « Twitter and the development of an audience : those who stay on topic thrive ! » In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, p. 1515-1518.

Wasserman, Stanley et Katherine Faust (1994). *Social Network Analysis : Methods and Applications*. T. 8. Structural analysis in the social sciences, 8 1. Cambridge University Press. Chap. 6, p. 825. ISBN : 0521387078. DOI : 10.2307/2077235. URL : http://www.amazon.com/dp/0521387078.

Watts, Duncan J. et Steven H. Strogatz (juin 1998). « Collective dynamics of 'small-world' networks ». In : *Nature* 393.6684, p. 440-442. DOI : 10.1038/30918. URL : https://doi.org/10.1038/30918.

Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang et S Yu Philip (2020). « A comprehensive survey on graph neural networks ». In : *IEEE transactions on neural networks and learning systems* 32.1, p. 4-24.

Xie, Jierui et Boleslaw K. Szymanski (2012). « Towards Linear Time Overlapping Community Detection in Social Networks ». In : *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, p. 25-36. DOI : 10.1007/978-3-642-30220-6_3. URL : https://doi.org/10.1007/978-3-642-30220-6_3.

Yamaguchi, Hiroto, Yuya Ogawa, Seiji Maekawa, Yuya Sasaki et Makoto Onizuka (2020). « Controlling Internal Structure of Communities on Graph Generator ». In : *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

Yang, Chao, Robert Harkreader, Jialong Zhang, Seungwon Shin et Guofei Gu (2012). « Analyzing spammers' social networks for fun and profit : a case study of cyber criminal ecosystem on twitter ». In : *Proceedings of the 21st international conference on World Wide Web*, p. 71-80.

Yang, Hong, Shirui Pan, Peng Zhang, Ling Chen, Defu Lian et Chengqi Zhang (2018). « Binarized attributed network embedding ». In : *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, p. 1476-1481.

Yang, Jaewon et Jure Leskovec (2015). « Defining and evaluating network communities based on ground-truth ». In : *Knowledge and Information Systems* 42.1, p. 181-213.

Zheng, Haichao, Dahui Li, Jing Wu et Yun Xu (2014). « The role of multidimensional social capital in crowdfunding : A comparative study in China and US ». In : *Information & Management* 51.4, p. 488-496.

Zhou, Kuangqi, Yanfei Dong, Wee Sun Lee, Bryan Hooi, Huan Xu et Jiashi Feng (2020). « Effective training strategies for deep graph neural networks ». In : *arXiv preprint arXiv :2006.07107*.