THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 2 - Jean Jaurès

Présentée et soutenue par

Mehdi DJELLABI

Le 7 janvier 2021

Mesure d'interactions locales pour les nœuds d'un réseau complexe : approches théorique et pratique

---

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Bertrand JOUVE et Frédéric AMBLARD

Jury

M. Renaud LAMBIOTTE, Rapporteur
M. Christophe CRESPELLE, Rapporteur
Mme Marie-Aude AUFAURE, Examinatrice
Mme Nathalie VIALANEIX, Examinatrice
M. Bertrand JOUVE, Directeur de thèse
M. Frédéric AMBLARD, Co-directeur de thèse

Table des matières

Remerciements

Cette thèse de doctorat représente pour moi quatre années de travail. Durant cette période, j'ai beaucoup appris et beaucoup évolué tant sur le plan scientifique que personnel, et ce grâce a plusieurs personnes que je souhaite remercier.

Tout d'abord, je tiens à remercier Bertrand Jouve et Frédéric Amblard pour la confiance, l'intérêt et l'aide qu'il m'ont apporté tout au long de cette thèse. Vous avez énormément contribué à l'aboutissement de ce travail, et j'aimerais que vous sachiez toute la reconnaissance que j'en éprouve.

J'exprime également ma gratitude à Christophe Crespelle et Renaud Lambiotte pour avoir accepté de rapporter ces travaux de thèse, ainsi qu'à Marie-Aude Aufaure et Nathalie Vialaneix pour avoir accepté d'être membres de mon jury de soutenance.

Ce travail n'aurait pas été possible sans le soutien de la région Occitanie et celui de l'université fédérale de Toulouse, qui m'ont permis grâce à une allocation de recherches et diverses aides financières, de me consacrer sereinement à l'élaboration de ma thèse.

Cette épreuve aurait sans doute été moins facile à supporter sans la disponibilité et l'accueil chaleureux que m'ont témoignés les membres, permanents comme passagers, de l'IRIT de la manufacture des tabacs, avec qui j'ai passé des moments très agréables.

Il ne faudrait tout de même que j'oublie l'équilibre apporté par mes amis, Bahri et Amir avec qui j'ai partagé un toit et beaucoup plus encore, Hamza dont j'ai encombré le studio pendant les premières semaines puis la vie par la suite, Rachel, Djilali et Halima pour m'avoir relu, supporté et aidé à décompresser entre deux journées improductives.

Pour terminer, je remercie chaleureusement mes parents, mon frère, mes soeurs et mon beau-frère, qui m'ont soutenu et encouragé tout au long de mes études, moi qui avait exprimé le souhait de mettre fin à ma scolarité au bout de la toute première journée d'école.

Introduction

Le besoin qu'ont les êtres humains de comprendre, voire de reproduire certains des phénomènes qui se produisent autour d'eux remonte à la préhistoire [102]. Ainsi, les historiens des sciences s'accordent sur le fait que la technique précède la science. Bien que ce ne fut pas délibéré, on adopta très tôt une approche empirique qui a permis la fabrication d'outils nécessaires à la vie quotidienne, ainsi que la maîtrise du feu. Ceci dit, il a fallu attendre des millénaires avant que l'on prenne conscience du fait que la plupart de ces phénomènes obéissent à des lois immuables, et ce que l'on décrit aujourd'hui comme étant la science est pendant longtemps resté indissocié de la magie. Le tournant majeur dans l'histoire de la science s'est sans doute opéré lorsque le recours aux mathématiques est devenu systématique. Ainsi à la fin du XVIème siècle, et avant la découverte de la loi de l'attraction universelle des masses que l'on attribue à Isaac Newton (1687), Galilée procédait déjà à des expérimentations qui visaient à comprendre les phénomènes liés à la chute des corps. Pour cela il élabora un protocole judicieux, qui lui permit d'isoler les effets d'intérêt pour son expérimentation, en plus de pouvoir reproduire à souhait la même expérience [69]. Ensuite, à l'aide d'outils de mesure, il put rassembler assez de résultats pour pouvoir déployer l'arsenal d'outils que proposent les mathématiques.

Il serait tentant de penser qu'une approche empirique, via l'analyse des données issues d'un quelconque phénomène que l'on étudie, constitue un premier pas naturel vers la compréhension de celui-ci, voire de la mise en évidence d'une loi universelle le caractérisant (si tant est que celle-ci existe), et à plus forte raison en sachant toutes les avancées qu'ont connus les mathématiques développées autour de ce domaine.
Force est de constater que nous ne manquons nullement de données aujourd'hui. La masse générée par les populations humaines de ces dernières années ne cesse de croître, et ce d'une manière prodigieuse.

Fondamentalement, la croissance de notre monde numérique est ahurissante mais tout à fait compréhensible. Il n'y a pas si longtemps, il était admis que les données les plus importantes se trouvaient dans les bases de données d'une organisation, qu'il s'agisse de commerce électronique, de progiciel de gestion intégré, de courrier électronique, etc. Bien que ces données soient toujours importantes, ce sont aujourd'hui les données non structurées qui sont devenues l'élément vital d'une organisation, non seulement les documents de bureau traditionnels, les fichiers vidéo et audio, mais aussi les données géospatiales, le streaming, etc. En fait, on estime que plus de 200 milliards d'appareils produiront des données rien qu'en 2020, et on estime que la masse totale des données mondiales se rapprochera des 175 zettaoctet (10,000 To) en 2025, ce qui est bien étonnant quand on sait que dans les années 90, moins d'une dizaine d'organisations disposaient d'un To de données[1].

Il est aussi important de rappeler que nous sommes au début d'une nouvelle évolution du monde numérique, liée à l'avènement de l'internet des objets. De plus en plus d'objets du quotidien s'y verront assimilés, on dit alors que l'IdO est la convergence entre des objets qui peuvent communiquer électroniquement et des données qu'ils produisent ou reçoivent, avec une tendance à le faire sur Internet. Ces objets sont aussi variés que les ordinateurs, les téléphones, les réfrigérateurs, les télévisions, les voitures, les ebook, ou de tout ce qui peut être connecté [2].
Tout ces éléments pointent vers le fait que la quantité de données que l'on trouve déjà gigantesque affiche une croissance qui n'est pas prête de ralentir. Mais alors même que beaucoup d'efforts semblent se concentrer sur la mise au point de plates-formes assez puissantes pour contenir toute cette masse, nous pouvons nous demander si les outils d'analyse que l'on emploie pour extraire ou synthétiser de l'information utile à partir de ces données évoluent à des vitesses comparables. Tout semble indiquer le

[1] `https://www.aparavi.com`
[2] `https://www.lemonde.fr/blog/binaire/2019/04/17/`

contraire. Au regard de la tournure des évènements survenus récemment lors de la crise sanitaire due au SARS-CoV-2, nous avons toutes les raisons de penser que l'usage des données personnelles a montré plus d'efficacité quand il est employé à des fins de surveillance que lorsqu'on en use à des fins prédictives. Ceci ne va pas sans ajouter son lot de complications, compte tenu des législations en vigueur concernant la protection des données, en plus du risque toujours présent de contrarier les populations concernées[3].

Approche proposée

Étant donné la source inépuisable de données dont nous disposons, et leur caractère fondamentalement interconnecté, est-il possible de tirer une information pertinente à partir des données recueillies sur un seul individu, sans considération supplémentaire des informations que peuvent apporter les autres individus qui évoluent dans le même environnement. En d'autres termes, est-il possible d'analyser un système en analysant séparément ses composants ? Tout dépend du système en réalité, et de la complexité des relations entre ses individus. Nous disposons d'exemples qui montrent que dans le cas d'un système régi par des règles simples et identifiées, il n'est pas nécessaire de considérer le système dans son intégralité pour en dégager les propriétés. Nous pouvons citer l'exemple des propriétés de certains cristaux, dont les atomes sont organisés en mailles régulières, et dont la connaissance des composants élémentaires suffit souvent à la caractérisation des propriétés macroscopiques du cristal en question. Ceci n'est bien entendu pas le cas pour les systèmes dits complexes, dans lesquels les règles d'interactions sont généralement loin d'être évidentes, les rendant difficiles à appréhender. Ces systèmes peuvent se présenter sous la forme de sociétés humaines ou animales, bien qu'il soit possible grâce aux outils technologiques actuels de les avoir sous une forme plus abstraite, comme des ensembles de profils dans les plates-formes de streaming ou bien les comptes bancaires.

Les travaux de cette thèse se concentrent sur un élément spécifique de l'étude des systèmes complexes : le graphe (qu'on appellera sans distinction réseau tout au long de cette thèse), en sa qualité de support de représentation des interactions diadiques du système.
Saisir la structure de ces graphes d'interactions est un enjeu de grande importance pour la compréhension des systèmes complexes qu'ils modélisent. C'est dans cette optique que nous proposons une approche centrée sur l'analyse de la topologie des réseaux.
L'originalité de cette approche se distingue par une considération particulière : au lieu de concentrer notre analyse uniquement sur l'identification et la description des clusters de forte densité, qui certes renferment une partie importante de l'information que contient le réseau, nous choisissons d'étudier en parallèle l'ensemble des nœuds de faible densité. Ces derniers peuvent occuper diverses positions dans le réseau, et peuvent jouer un rôle très important dans sa structure globale. Or il se trouve que peu de recherches ont été effectuées dans ce sens, et parmi celles qu'on a pu identifier, les différents outils proposés sont loin d'être basés sur une approche générique.
Nous proposons donc un cadre unificateur, articulé autour du concept de la densité dans le réseau. Il faudra par conséquent revoir la définition classique de celle-ci, dont les lacunes sont bien connues et seront rappelées dans le second et le troisième chapitre. Le résumé du contenu de ce manuscrit est présenté dans les prochains paragraphes.

[3] https://www.lemonde.fr/pixels/article/2020/06/18/norvege-jugee-trop-intrusive-l-application-de-lutte-contre-le-covid-19-6043308_4408996.html

Figure 1: Réseau composé de comptes twitter ayant publié un contenu politisé, les données sont récoltées sur une période qui s'étend sur plusieurs mois précédant les élections présidentielles de 2017 en France. Les liens entre deux comptes retranscrivent les retweets. Figure tirée de [61]

Résumé du contenu

Le premier chapitre est un état de l'art de l'ensemble des outils et techniques d'analyse que nous avons identifiés dans la littérature en rapport avec notre thématique. Ces derniers sont séparés en trois classes, en fonction de l'échelle de l'information fournie par leurs résultats respectifs. On établit une analogie avec le domaine de la physique en qualifiant de "microscopique" l'échelle à laquelle l'attention est portée sur les nœuds/arêtes, en contraste avec l'échelle "macroscopique" qui regroupe quant à elle les approches considérant le graphe dans sa globalité. Les méthodes de l'échelle intermédiaire "mésoscopique" sont présentées en fin de chapitre.

Le second chapitre est le plus particulier des cinq que compte ce manuscrit. Il résume notre approche initiale de la problématique, approche que l'on a fini par laisser de côté pour des raisons pratiques. De plus, il introduit une mesure que l'on a dû tester à l'aide d'algorithmes qui ne sont pas clairement en accord avec le reste de ce qui est proposé. Il est donc considéré comme une partie autosuffisante du manuscrit, en ce sens où le lecteur n'a pas besoin de connaître les notions introduites dans l'état de l'art pour l'aborder, de même qu'à l'inverse, il n'est pas absolument nécessaire d'en connaître les détails pour aborder la suite. Les difficultés rencontrées pendant le développement de cette approche ont beaucoup influencé la réflexion qui a guidé la suite de nos travaux, il est de ce fait important de consacrer un chapitre entier à la présentation de ce contenu.

Les troisième et quatrième chapitres constituent le noyau dur de cette thèse du point de vue méthodologique, le premier des deux étant dédié à l'introduction d'une mesure que l'on classerait dans l'échelle microscopique, et le dernier à l'utilisation de celle-ci dans la mise au point d'un algorithme qui couvre les deux échelles supérieurs. On prendra soin à chaque fois de tester et de valider nos résultats sur des modèles de réseaux synthétiques.

Avant de conclure, nous présentons les résultats de notre algorithme dans un chapitre consacré à l'étude des données de réseaux réels, sur lesquels on a pu rassembler une vérité de terrain pertinente pour nos analyses. Cet algorithme pouvant avoir diverses applications, nous structurons le chapitre de sorte que chaque section se concentre sur une application différente. Les réseaux analysés sont de types relativement différents, certains statiques et d'autres dynamiques.

Chapitre 1

État de l'art

Table des matières

1.1 Résumé de l'état de l'art

Nous commençons ce manuscrit en présentant l'état de l'art des approches et outils employés dans le domaine de l'analyse structurelle des réseaux. Celui-ci commencera néanmoins par la présentation de quelques éléments de base de la théorie des graphes, qui serviront au lecteur soit comme rappels, soit comme un moyen de se familiariser avec les notations qui vont être employées tout au long de ce manuscrit.
On expliquera ensuite l'intérêt d'étudier la structure des réseaux, à travers une description succincte du rapport sous-jacent qui relie les systèmes complexes et les réseaux. On s'appuiera tout au long de cette partie sur quelques exemples d'illustration.

On enchaîne par une vue d'ensemble des outils précédemment évoqués, que l'on divise en trois parties en fonction de l'échelle de mesure de chacune. On a ainsi l'échelle microscopique, qui est l'échelle la plus fine, car elle se concentre sur les propriétés des nœuds et des liens. On présente ensuite les différents outils d'échelle macroscopique, qui ont pour particularité de résumer une information à l'échelle du graphe dans sa totalité par de simples valeurs numériques. On parlera finalement de l'échelle mésoscopique qui occupe une position intermédiaire entre les deux échelles que l'on vient de citer. Cette échelle rassemble les outils qui s'occupent de dégager une information pertinente sur une ou plusieurs sous-parties du graphe, avec la possibilité de décrire le rapport qui relie ces sous-parties entre elles.
Les éléments que contiendra chacune de ces trois descriptions ne couvrent naturellement pas la totalité des approches existantes, mais permettent au lecteur de s'approprier les concepts fondamentaux.

On consacre la dernière partie de cet état de l'art à l'introduction des concepts relatifs à un type particulier de nœuds, dont les caractéristiques sont bien identifiées dans la littérature, mais qui ne bénéficient pas d'une approche quantitative qui permettrait leur identification systématique. Nous proposons ainsi de positionner notre approche dans un cadre unificateur, qui facilite la liaison entre ces différents concepts (trous structuraux, ponts structuraux) à travers une mesure originale de la densité dans le graphe.

1.2 Éléments de la théorie des graphes

1.2.1 Définitions

Graphes simples, graphes orientés et graphes pondérés

On définit un graphe G par le couple (V, E) où V désigne l'ensemble $\{v_1, v_2, ..., v_N\}$ des nœuds du graphe et $E = \{e_1, e_2, ..., e_m\}$ celui des arêtes, qu'on appelle aussi liens du graphe. Chaque arête est représentée par un couple (v_i, v_j) de nœuds, qu'on appelle extrémités de l'arête/lien. On désigne par la taille d'un graphe, l'entier N égal au nombre de nœuds dans le graphe, et on appelle m le nombre de liens dans celui-ci.

On retrouve plusieurs types de graphes suivant les propriétés de leurs liens : d'abord on peut différencier les graphes orientés des graphes non orientés, par la distinction suivante : soient $e_1 = (v_i, v_j)$ et $e_2 = (v_j, v_i)$. On a alors $e_1 = e_2$ pour un graphe non orienté, et $e_1 \neq e_2$ pour un graphe orienté.

On peut aussi associer à chaque arête un nombre réel correspondant à son poids dans le graphe, on parle alors de graphes pondérés et on passe d'une définition des liens par un couple $e_l = (v_i, v_j)$ à une définition par un triplet $e_l = (v_i, v_j, w_l)$, où v_i et v_j sont les extrémités du lien e_l et w_l son poids. Il est bien entendu possible de combiner les propriétés des graphes pondérés à celles des graphes orientés pour obtenir des graphes pondérés et orientés, dans lesquels on aurait des liens aux directions opposées, et aux poids non égaux, ex : $e_l = (v_i, v_j, w_l) \neq e_k = (v_i, v_j, w_k)$.

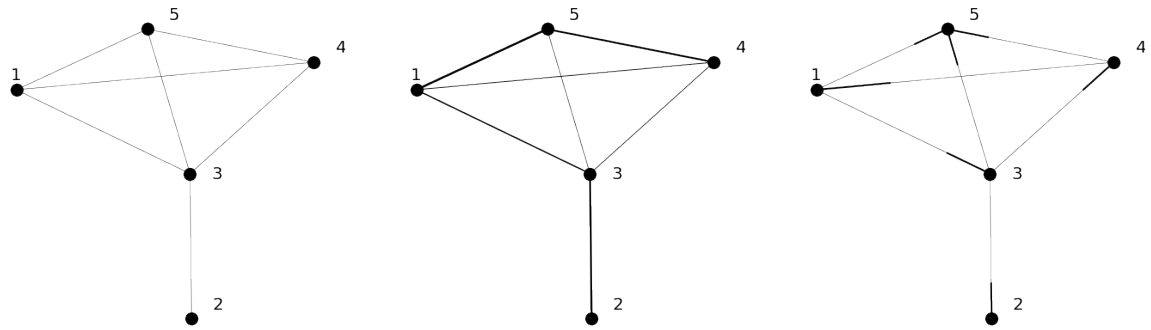
Un graphe peut se représenter sous forme d'une matrice qu'on appelle la matrice d'adjacence, celle-ci est souvent notée A pour représenter les réseaux non pondérés et W les réseaux pondérés. L'exemple le plus simple est celui d'un réseau non pondéré et non orienté, pour lequel on a $a_{i,j} = 1$ si v_i est voisin de v_j et 0 si non.
On peut noter que

- Dans le cas des réseaux non orientés on a $a_{ij} = a_{ji}$ alors que ceci n'est pas vrai pour les réseaux orientés
- $a_{ij} \in \{0, 1\}$ pour les réseaux non pondérés

- a_{ij} (ou w_{ij} suivant la notation standard) $\in \mathbb{R}$ pour les réseaux pondérés

Par exemple on donne ici les matrices d'adjacence correspondant à trois graphes de types différents, ainsi que les représentations graphiques correspondantes.

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}; \qquad B = \begin{bmatrix} 0 & 0 & 0.52 & 0.32 & 0.84 \\ 0 & 0 & 0.74 & 0 & 0 \\ 0.52 & 0.74 & 0 & 0.27 & 0.21 \\ 0.32 & 0 & 0.27 & 0 & 0.70 \\ 0.84 & 0 & 0.21 & 0.70 & 0 \end{bmatrix}; \qquad C = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



(a) (b) (c)

Figure 1.1: De gauche à droite on a représenté dans (a) un graphe simple, (b) un graphe pondéré dont l'épaisseur des liens est proportionnelle à leur poids respectifs, et (c) un graphe dirigé où la direction est représentée par une extrémité plus large dans la direction du lien.

Graphes dynamiques

Un graphe dynamique est un graphe dont l'ensemble des arêtes évolue au cours du temps. Les interactions y sont soit instantanées et on note alors chaque lien $e = (i, j, t)$, avec t l'instant auquel l'interaction entre les nœuds i et j est enregistrée, ou bien d'une durée finie auquel cas on remplace t par un intervalle et on note $e = (i, j, [t, t + T])$ avec T la durée de l'interaction.
Il existe diverses représentations pour les graphes dynamiques, la plus intuitive (mais aussi la plus coûteuse) étant celle qui consiste à associer à chaque paire de nœuds (impliquant tous les nœuds ayant fait partie du graphe à un moment donné) une fonction du temps, dont la valeur à l'instant t est égale au poids de l'arête reliant ses extrémités. Dans le cas des graphes dynamiques non pondérés, cette fonction est une succession de paliers oscillant entre 0 et 1 et dont la largeur varie. Ainsi à temps fixé, l'instantané d'un graphe dynamique est un graphe statique, comme ceux décrits plus haut. Cette représentation est équivalente à celle où à chaque instant t est associé l'ensemble $E(t)$ des liens qui existent à cet instant. On retrouve également une représentation qui regroupe toutes les interactions qui sont enregistrées sur un intervalle donné, on découpe ainsi la durée totale de l'évolution en une suite d'intervalles, et on regroupe dans chacun des intervalles l'ensemble des interactions qui y ont été enregistrées afin d'obtenir une suite de graphes statiques.

Sous-graphe et sous-graphe induit

Un sous-graphe F d'un graphe G est un graphe formé à partir d'un sous-ensemble de nœuds et d'arêtes de G.
Formellement, $F = (V_F, E_F)$ est un sous-graphe de $G = (V, E)$ si $V_F \subseteq V$ et
$E_F \subseteq \{(u, v) \in E \mid u \in V_F \wedge v \in V_F\}$. L'ensemble des nœuds du sous-graphe F est un sous-ensemble de l'ensemble des nœuds de G et l'ensemble des arêtes de F est un sous-ensemble de l'ensemble des arêtes de G ayant leur origine et leur extrémité parmi les nœuds de F.
Un sous-graphe de recouvrement est un sous-graphe qui inclut tous les sommets du graphe ; un sous-graphe induit est un sous-graphe qui inclut tous les liens dont les points d'extrémité appartiennent au sous-ensemble des nœuds.

Chaînes, cycles et connexité

On appelle chaîne, notée $\mu(s, t)$ (ou cycle si $s = t$) dans un graphe non orienté, la succession d'arêtes consécutives qui permettent de relier le nœud s au nœud t en joignant une séquence de sommets du graphe. On appelle chaîne élémentaire la chaîne qui ne passe pas deux fois par un même sommet, et chaîne simple celle qui ne passe pas deux fois par une même arête. Si en plus le graphe est non pondéré, on définit la longueur d'une chaîne par le nombre d'arêtes qui permettent de relier s à t.

Ces notions sont facilement généralisables aux réseaux orientés et pondérés, dans le cas des réseaux orientés on parle de chemins au lieu de chaînes (bien qu'il sera par la suite courant de confondre les deux termes), en autorisant uniquement les déplacements le long des arêtes sortantes.
Pour ce qui est des réseaux pondérés, on parle du poids d'une chaîne en désignant la somme des poids des arêtes qui la constituent. Il est alors facile de vérifier la possibilité d'avoir un chemin C_1 comportant un plus grand nombre d'arêtes qu'un autre chemin C_2, tout en ayant un poids inférieur à celui-ci, car passant par des arêtes de moindre poids.

On appelle chemin le plus court reliant s à t un chemin ayant la plus petite longueur, si le réseau est non pondéré, ou bien celui dont la somme des poids des arêtes est la plus petite dans le cas des réseaux pondérés.

Finalement on définit l'accessibilité entre deux nœuds (s, t) comme le fait qu'il existe un chemin reliant s à t.

On dit qu'un graphe $G = (V, E)$ non orienté est connexe si pour toute paire (s, t) avec $s \in V$ et $v \in V$, il existe une chaîne reliant s à t. La définition diffère pour les graphes orientés et les graphes dynamiques, car il se peut qu'une paire ne soit raccordable que par un chemin dans un sens précis (cas des graphes orientés), ou bien seulement sur une durée déterminée (cas des graphes dynamiques) sans que le graphe ne soit déconnecté.

1.2.2 Quelques graphes remarquables

Nous citons maintenant quelques exemples de graphes aux propriétés particulières :

Les arbres

Un arbre est un graphe non dirigé, connecté et qui ne contient aucun cycle. Il existe parmi les arbres plusieurs types remarquables, comme par exemple les arbres binaires qui sont composés d'un nœud racine, ayant tout au plus deux nœuds voisins qu'on appelle "fils", chacun de ces fils est à son tour relié au maximum à autant de nœuds (deux) dont il est le "père", et ainsi de suite. Le processus de construction d'un arbre binaire peut être limité par le nombre de niveaux le constituant. On retrouve les arbres binaires notamment comme structure pour représenter des données hiérarchiques.

Il existe bien entendu d'autres types d'arbres aux propriétés particulières, on peut rapidement citer l'exemple de l'étoile composée d'un nœud ayant N voisins, dont chacun forme un lien unique avec le nœud central, etc. Un graphe constitué de plusieurs composantes qui sont deux à deux déconnectées, et dont chacune est un arbre est appelé forêt.

Les cliques

Une clique de taille N est un graphe composé de N nœuds, dont chacun des nœuds est relié à tous les autres, c'est donc un graphe complet, dont le nombre de liens (dans le cas des graphes non dirigés) est égal à $\frac{N\cdot(N-1)}{2}$

Les graphes bipartis

Un graphe biparti $G = (V, E)$ est composé de deux sous-ensembles V_1 et V_2 disjoints de nœuds, vérifiant donc $V_1 \cap V_2 = \varnothing$ et $V_1 \cup V_2 = V$, de sorte que les nœuds du même sous-ensemble ne partagent pas de liens entre eux, tous les liens $e = (v_1, v_2) \in E$ vérifient $v_1 \in V_1$ et $v_2 \in V_2$

Les graphes aléatoires

Les graphes aléatoires sont légèrement différents de ceux décrits plus haut, car à défaut d'être distingués par des propriétés topologiques simples, comme par exemple l'absence de cycles comme c'est le cas des arbres, il sont caractérisés par des propriétés statistiques, que l'on observe sur les distributions de différentes grandeurs comme le degré, la taille de la plus grande composante connexe, etc [1].
Ces graphes sont en général le résultat d'un modèle génératif, impliquant d'une manière ou d'une autre le hasard dans son processus de génération. Il serait possible d'écrire de longues pages concernant les graphes aléatoires, les modèles à partir desquels ils sont générés et les propriétés remarquables de certains d'entre eux, mais nous nous contenterons ici de mentionner quelques exemples remarquables.

Le modèle le plus simple de graphes aléatoires consiste à fixer le nombre de nœuds N et le nombre de liens m, et ensuite de choisir aléatoirement m paires (en prenant soin de ne pas prendre deux fois la même paire pour les graphes simples) pour les connecter avec un lien. Ce modèle est aussi parfois décrit comme un tirage aléatoire de l'ensemble $\Omega_{N,m}$ contenant tous les graphes de N nœuds et m liens[92], avec pour chaque graphe G une probabilité d'être tirée égale à $P(G) = \frac{1}{|\Omega_{N,m}|}$

Il existe des modèles de graphes aléatoires qui ont des caractéristique plus spécifiques, l'un des plus répandus est celui d'Erdős–Rényi [48] qui a pour paramètres le nombre de nœuds N, et une probabilité P de relier chaque paire du graphe par un lien. Ce modèle, bien que très simple permet de dégager quelques propriétés intéressantes concernant les graphes aléatoires. Ainsi des études [49] ont établi le lien entre les paramètres (P et N) du modèle, et la taille moyenne des différentes composantes connexes du graphe généré, et en particulier l'existence ou non d'une composante géante. Ce modèle a la particularité de générer des graphes homogènes (suivant des caractéristiques qui seront décrites plus tard dans ce chapitre) car il ne distingue aucune paire de nœuds, et le choix de les relier ou à l'inverse ne pas les relier ne dépend que de P, c'est ce qu'on appelle une épreuve de Bernoulli de paramètre P pour chaque paire de nœuds du graphe.

On peut citer d'autres exemples de modèles de graphes aléatoires, qui seront détaillés plus bas dans ce chapitre, comme le modèle de Watts et Strogatz [122], ainsi que celui de l'attachement préférentiel de Barabasi-Albert [11]. Ces derniers génèrent des graphes hétérogènes, contrairement au modèle d'Erdős–Rényi, qui ont pour but de reproduire des caractéristiques bien précises des réseaux provenant du monde réel.

[1] Voir la section 1.3 pour plus de précisions sur certaines mesures de bases appliquées aux graphes

1.3 Les systèmes complexes et les réseaux.

Il existe de nombreux systèmes d'intérêt pour les scientifiques, constitués d'unités (au sens large du terme) qui interagissent les unes avec les autres de diverses façons. On peut citer l'exemple d'internet, constitué (grossièrement) de plusieurs ordinateurs connectés les uns aux autres, ou bien celui des réseaux sociaux (réels comme virtuels), qui sont constitués d'êtres vivants de même espèce et qui interagissent les uns avec les autres via des liens de connaissance. Une ville est par exemple un réseau social constitué par ses habitants et les différents liens qui les unissent : familiaux, amicaux, collègues de travail, etc. Certains domaines scientifiques se penchent sur l'étude des individus, en tenant compte d'une manière plus ou moins abstraite de l'environnement qui les entoure (la psychologie, la sociologie, la géographie, l'informatique, etc.), mais il serait aussi important de s'intéresser au système dans son intégralité, et en particulier aux motifs que forment les connexions entre les individus.
Ces systèmes sont souvent très compliqués à étudier et exigent non seulement une représentation simplifiée (sous forme de graphe par exemple) mais aussi une limite de taille (il est plus facile d'étudier le réseau social d'un petit village que celui d'un pays tout entier) pour pouvoir facilement stocker l'information qu'ils contiennent. Il est donc très important de développer des outils qui permettent la compréhension de ces réseaux, pour espérer une compréhension des systèmes qu'ils représentent, et ainsi d'améliorer ses performances (ex : le flux de données transférées sur Internet, les trafics urbains ,etc.)

Nous avons classé en trois échelles différentes que nous détaillons, avec des exemples à l'appui dans la section 1.3, les outils disponibles pour l'analyse des réseaux, mais avant nous fournissons quelques exemples de systèmes complexes, dont la représentation par un réseau permet d'expliquer en partie l'émergence de comportements complexes, d'intérêts capitaux.

Les réseaux de transports

Comprendre les réseaux de transports est un enjeu important au niveau local comme au niveau global [42, 113, 41]. Localement nous pouvons penser à l'importance de la régularisation du flux, et l'impact que cela peut avoir sur la vie quotidienne des habitants d'une ville, par exemple en optimisant les liaisons entre les différents points géographiques d'une agglomération de sorte à ce que cela puisse réduire le nombre de voitures qui circulent. Au niveau global, une connaissance de la structure des réseaux de transport aérien et maritime est cruciale pour estimer l'étendue sur laquelle et la vitesse à laquelle peut se propager une épidémie à grande échelle, et ainsi avoir de meilleurs chances de la contenir.

Réseaux d'interactions de protéines

Les interactions protéine-protéine sont les contacts physiques bien spécifiques établis entre deux ou plusieurs molécules de protéines, à la suite d'événements biochimiques provoqués par des forces électrostatiques. Beaucoup sont des contacts physiques avec des associations moléculaires entre des chaînes qui se produisent dans une cellule ou un organisme vivant dans un contexte biomoléculaire spécifique [37]. Il devient alors naturel de penser à représenter un grand nombre d'interactions (complexe protéique) sous forme de graphes, dont les nœuds sont des protéines, et les liens des indicateurs sur l'interaction entre deux protéines.

L'une des questions à laquelle se proposent de répondre les biologistes concerne le lien qui existe entre la structure d'un complexe protéique, et sa fonction biologique. Les méthodes proposées par les chercheurs pour y répondre proviennent d'un grand nombre d'approches utilisant différents outils, et l'analyse de réseaux en fait partie [35, 93]. En effet c'est une approche qui s'occupe d'étudier le lien entre la structure réelle d'un complexe protéique (dans l'espace physique à 3 dimensions) et la structure de sa représentation abstraite, sous forme de réseau, dans le but d'apporter des informations supplémentaires et utiles à la compréhension de la fonction biologique de chaque complexe, en partant du réseau. Cette approche présente l'intérêt de ne dépendre que des interactions entre les nœuds du réseau (aussi appelés sous-unités protéiques), qui est plus facilement observable que l'aspect général d'un complexe, qui présente une certaine variabilité.

Les réseaux de sociétés animales

Beaucoup d'espèces animales vivent dans des sociétés de tailles plus ou moins grandes, dans lesquelles les individus interagissent les uns avec les autres. Il est ensuite possible de construire les réseaux représentant ces sociétés, en choisissant les bons critères qui expriment la formation de liens. Dans

[84] les auteurs ont étudié le même banc de dauphins près des côtes Néo-Zélandaises sur une durée de plusieurs années, en considérant comme liées les paires d'individus qui entraient plus souvent en contact que ce à quoi l'on pourrait s'attendre si cela se produisait par hasard.
D'autres équipes s'intéressent aux sociétés de macaques qui forment quant à eux leurs liens sociaux à travers l'épouillage, ce qui est reconnu comme étant une pratique qui renforce la cohésion sociale dans le groupe [17].

Les résultats de telles études permettent de mieux comprendre le comportement de ces espèces animales à travers leur vie en groupe, ainsi que la structure de leurs sociétés, à plus forte raison s'il existe des motifs récurrents pour les sociétés de même espèce.

1.4 Vue d'ensemble sur l'analyse topologique des réseaux

Nous proposons ici de donner une vue d'ensemble synthétique sur tout ce qui est fait dans le domaine de l'analyse structurelle des réseaux (on utilisera sans distinction les termes topologie et structure), à travers plusieurs outils mathématiques issus des approches les plus populaires, et en les classant en trois échelles différentes : la première concernera le niveau de précision le plus élevé qui se concentre uniquement sur les propriétés des nœuds et des liens, ensuite on décrira quelques approches qui se pratiquent à l'inverse de la première, à l'échelle de précision la plus basse, en donnant une information sur la structure globale d'un réseau, et on finira par décrire l'échelle intermédiaire, qui fournit un découpage du graphe en plusieurs sous-ensembles répondant à certaines propriétés, et de taille intermédiaire (entre le nœud et le graphe).
Nous donnons uniquement les définitions correspondant aux graphes non orientés et non dirigés, la généralisation aux différents autres cas étant souvent possible.

1.4.1 Échelle microscopique

Plusieurs des mesures qui vont être présentées ici ont un nom qui provient du domaine de la sociologie, car on y retrouve un grand intérêt pour l'analyse des réseaux sociaux. Ces noms ont par la suite continué à être utilisés dans le domaine de la science des réseaux. On divise en deux la partie consacrée à l'introduction des mesures à l'échelle microscopique, d'abord on présentera celles qui se concentrent sur les nœuds et ensuite celles qui se concentrent sur la similitude entre paires de nœuds, et qui peuvent donc être assimilées à des mesures sur les arêtes (existantes ou pas dans le graphe).

Mesures sur des nœuds

Les mesures de graphes appliquées sur les nœuds ont toutes pour but d'identifier ceux qui constituent les éléments les plus importants du réseau, mais à travers l'évolution de la théorie des graphes, la notion d'importance a pris des directions diverses, ce qui a donné lieu à plusieurs mesures que nous nous proposons de décrire ici.

Le degré

On commence par présenter la plus intuitive des mesures sur un nœud, qui est sans doute celle du degré, représentant le nombre de liens auxquels un nœud donné est rattaché, qu'on peut aussi appeler le nombre de voisins. Souvent noté d_i ou k_i (pour le nœud labelisé v_i dans le graphe), on a donc :

$$k_i = \sum_{j=1}^{N} a_{ij} \tag{1.1}$$

avec a_{ij} l'élement ij de la matrice d'adjacence. Une autre manière de noter cette somme est la suivante : $k_i = \sum_{j\sim i} a_{ij}$ avec ici la notation $i \sim j$ signifiant i voisin de j.

Le coefficient de clustering

Le coefficient de clustering d'un nœud v_i décrit la tendance de ses voisins à former des liens entre eux.

Il peut s'écrire comme :

$$C_i = \frac{|\{e_{jk} : v_j \sim v_i, v_k \sim v_i, e_{jk} \in E\}|}{k_i(k_i - 1)}. \tag{1.2}$$

Ici on note k_i le degré du nœud v_i et la notation $v_i \sim v_j$ signifie que v_i et v_j sont voisins, e_{jk} est l'arête d'extrémités v_j et v_k.

Il est possible de retrouver dans la littérature le terme de coefficient de clustering pour renvoyer vers la moyenne sur tous les nœuds du graphe de la quantité décrite par l'eq. (1.2), bien que cette mesure ne refléterait plus les propriétés locales des différents nœuds, mais plutôt une information globale sur la topologie du réseau [44].

La centralité de betweenness

Une autre question importante qui se pose concernant l'importance que tient un nœud dans le réseau est de savoir à quel point celui-ci est intermédiaire entre les nœuds du réseau, pour cela plusieurs mesures de centralité ont été développées pour essayer de capturer cette information. La centralité de betweenness [58] (aussi appelée centralité intermédiaire) mesure le degré d'implication d'un nœud dans les chemins les plus courts entre toutes les paires de nœuds du réseau de la façon suivante :

$$\text{Betweenness}(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}} \tag{1.3}$$

Avec $\sigma_{st}(u)$ le nombre de chemins les plus courts reliant la paire s et t qui passent par le nœud u (celui sur lequel la mesure est appliquée), et σ_{st} le nombre total de chemins les plus courts qui existent entre s et t.

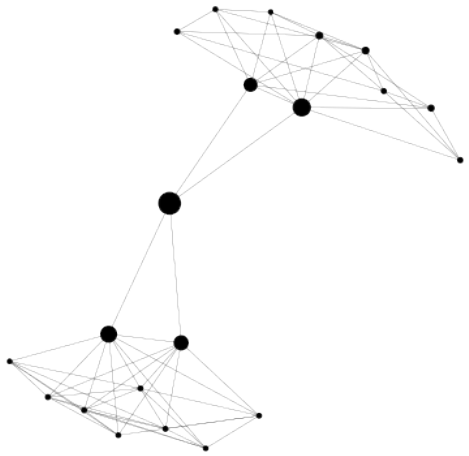


Figure 1.2: Nous montrons sur cette figure un exemple des résultats de la mesure de la centralité de betweenness sur un réseau d'une vingtaine de nœuds, la taille de chaque nœud est proportionnelle à la valeur de sa centralité de betweenness.

La centralité de proximité

La centralité de proximité (ou proximité) d'un nœud est calculée comme la somme de l'inverse de la longueur des chemins les plus courts entre le nœud considéré et tous les autres nœuds du graphe [16]. Ainsi, plus il est proche de tous les autres nœuds, plus il est central.

$$\text{Closeness}(u) = \frac{1}{\sum_v d(u,v)}. \tag{1.4}$$

Avec $d(u,v)$ désignant la longueur du chemin le plus court entre u et v

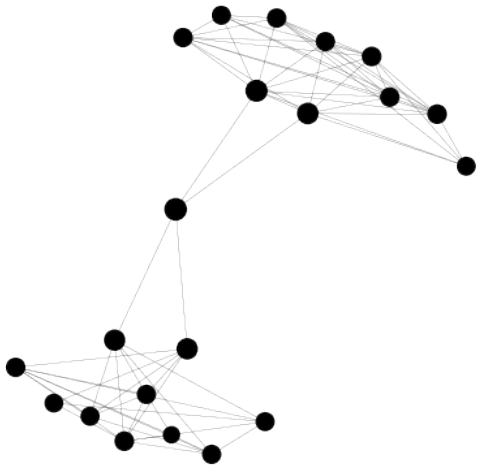


Figure 1.3: Résultats de la mesure de proximité sur le graphe précédent, la taille des nœuds est proportionnelle à leurs valeurs de la centralité de proximité

Les centralité du vecteur propre, de Katz et l'indice du PageRank

Il est souvent normal de considérer que l'importance d'un nœud au sein du réseau doit provenir de l'importance de ses voisins. Dans l'exemple du degré on attribue un "point de centralité" pour chaque voisin du nœud évalué, mais le problème avec cette mesure est qu'en pratique, deux nœuds de même degré peuvent avoir des voisinages avec des topologies radicalement différentes. Des mesures ont été développées pour distinguer ces différentes configurations, nous pouvons citer comme exemples la centralité de vecteur propre, la centralité de Katz et le score du PageRank [92, 72]. Dans ces trois cas il s'agit d'attribuer un score à chaque nœud qui prend en compte le score de ses voisins, avec des variations propres à chacune d'entre elles. La centralité de vecteur propre s'obtient de manière itérative en sommant les centralités de ses voisins :

$$\mathrm{Vect}(i) = \lambda_1^{-1} \sum_j A_{ij} \cdot \mathrm{Vect}(j) \tag{1.5}$$

Où λ_1 est la plus grande valeur propre de la matrice d'adjacence A. La valeur de la centralité de chaque nœud du réseau est en fait la valeur de la coordonnée correspondante du vecteur propre associé à la valeur propre λ_1.

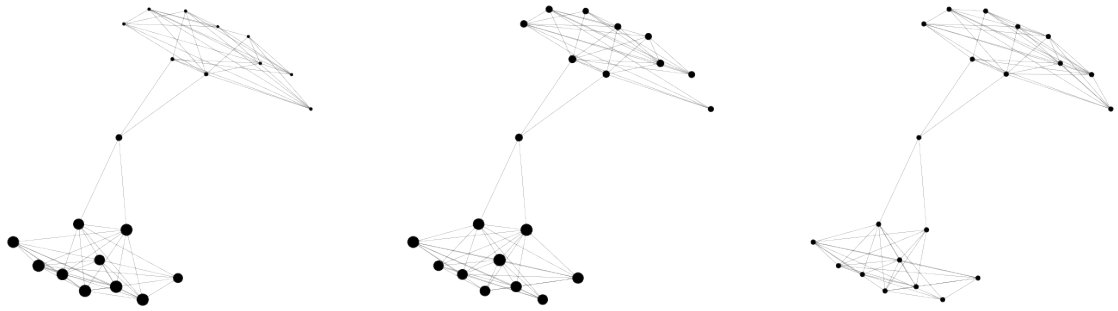
La centralité de Katz est très similaire à celle du vecteur propre, à la différence près qu'elle attribue un score initial β pour tous les nœuds du réseau :

$$\mathrm{Katz}(i) = \alpha \sum_j A_{ij} \cdot \mathrm{Katz}(j) + \beta \tag{1.6}$$

où α est une constante de normalisation, et β le score initial attribué "gratuitement" à chaque nœud du réseau. Cette modification permet d'éviter certains soucis propres à la centralité du vecteur propre. Celle-ci attribue des scores faibles à tous les nœuds qui ne sont pas dans des composantes fortement connexes, de deux ou plusieurs nœuds. La centralité de Katz a tout de même des désavantages, car elle ne différencie pas tout à fait les nœuds. Prenons l'exemple d'un nœud de forte centralité de Katz, et qui possède en même temps un degré élevé, alors tous ses voisins se retrouvent avec une grande centralité à leur tour. Prenons le cas d'une page web, vers laquelle pointe le site Google, alors suivant la centralité de Katz, cette page devrait profiter de la centralité de Google. Le problème est que Google pointe vers un très grand nombre de pages, parmi lesquelles celle dont il est question dans notre exemple. Il serait donc plus naturel de normaliser l'apport de chaque nœud par son degré pour harmoniser ce type de mesures. C'est ce qui est proposé par la mesure du PageRank :

$$\mathrm{PR}(i) = \alpha \sum_j \frac{A_{ij}}{k_j} \cdot \mathrm{PR}(j) + \beta \tag{1.7}$$

Ajoutons enfin qu'il existe des variations de cette même mesure, comme par exemple celle qui consiste à attribuer un score initial β_i différent en fonction de chaque nœud.



(a) Centralité de vecteur propre (b) Centralité de Katz (c) Pagerank

Figure 1.4: La taille des nœuds est proportionnelle aux scores des centralités correspondantes, Les paramètres choisis pour la centralité de Katz (b) et le Pagerank (c) sont $\alpha = \lambda_1^{-1}$ et $\beta = 1$, où λ_1 est la plus grande valeur propre de la matrice d'adjacence du graphe étudié.

Ces mesures ont toutes été définies dans le cadre des graphes simples non orientés, mais il est possible de passer au cas orienté et/ou pondéré, en remplaçant le degré des nœuds par les degrés sortants ou entrants, et les éléments de la matrice d'adjacence simple A_{ij} par ceux d'une matrice d'adjacence pondérée W_{ij} [43].

Indices de similarité et mesures appliquées aux arêtes

Les indices topologiques appliqués aux paires de nœuds d'un réseau mesurent leurs caractéristiques partagées, on en présente quelques unes dans cette partie.

La similarité du cosinus

Cette mesure est très répandue et est employée dans des domaines d'études très variés, allant de la sociologie à la fouille de texte [112]. Elle permet de calculer la similarité entre deux vecteurs dans un espace de dimension n muni d'un produit scalaire, en calculant l'angle qui les sépare :

$$\cos(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|} \tag{1.8}$$

où $u \cdot v$ désigne ce produit scalaire et $\|.\|$ la norme d'un vecteur.

Dans le contexte des graphes, chaque nœud v_i est représenté par le vecteur A_i de la matrice d'adjacence du graphe, que l'on peut réécrire :

$$\cos(A_i, A_j) = \frac{\sum_k A_{ik} \cdot A_{jk}}{\sqrt{k_i \cdot k_j}} = \frac{n_{ij}}{\sqrt{k_i \cdot k_j}}$$

où n_{ij} désigne le nombre de voisins communs entre le nœud i et le nœud j.

Cette mesure est bornée entre 0 si les deux nœuds ne partagent aucun voisin en commun, et 1 si les deux nœuds ont en commun l'intégralité de leur voisinage.

L'indice de Jaccard

Cette mesure tient son nom du botaniste Paul Jaccard, et est utilisée pour évaluer la similarité entre deux ensembles finis [68] :

$$J(U,V) = \frac{|U \cap V|}{|U \cup V|}. \tag{1.9}$$

Dans le contexte discuté ici, on calcule l'indice de Jaccard entre deux nœuds en mesurant la similarité décrite ci-dessus sur leurs voisinages respectifs. Nous pouvons ensuite réécrire la formule précédente en prenant en compte ces précisions. Soient v_i et v_j deux nœuds d'un graphe, notons N_i et N_j les ensembles constitués de leurs voisinages respectifs [2] et n_{ij} le nombre de voisins en commun entre les nœuds v_i et v_j:

$$J(N_i, N_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} = \frac{n_{ij}}{k_i + k_j - n_{ij}}$$

La distance Euclidienne

La distance euclidienne entre deux nœuds v_i et v_j est égale au nombre de nœuds parmi les voisinages respectifs de v_i et v_j, à être relié à l'un sans être relié à l'autre. Formellement, elle s'écrit de la manière suivante :

$$d_{ij}^2 = \sum_k (A_{ik} - A_{jk})^2. \tag{1.10}$$

On peut développer les termes de la somme pour trouver

$$d_{ij}^2 = \sum_k A_{ik}^2 + A_{jk}^2 - 2 \cdot A_{ik}A_{jk} = k_i + k_j - 2 \cdot n_{ij}.$$

On peut normaliser l'expression ainsi obtenue par sa valeur maximale, égale à la somme des degrés $k_i + k_j$, dans le cas où les deux nœuds ne partagent aucun voisin. et on obtient ainsi :

$$d_{ij}^{'2} = \frac{d_{ij}^2}{k_i + k_j} = 1 - \frac{2 \cdot n_{ij}}{k_i + k_j}.$$

L'indice de Sørensen-Dice

Cet indice [114, 39] ressemble à l'indice de Jaccard introduit plus haut. Pour deux nœuds v_i et v_j, de voisinages ouverts respectifs N_i et N_j, on a :

$$DS(N_i, N_j) = 2 \cdot \frac{|N_i \cap N_j|}{|N_i| + |N_j|} = \frac{2 \cdot n_{ij}}{k_i + k_j}. \tag{1.11}$$

Comme le coefficient de Sørensen-Dice ne satisfait pas l'inégalité triangulaire, il peut être considéré comme une version semi-métrique de l'indice de Jaccard. On note tout de même que la racine carrée de l'indice de Sørensen-Dice est quant à elle une métrique euclidienne [54]. Il reste toutefois possible d'obtenir la valeur de l'un en partant de celle de l'autre via la formule suivante :

$$J = \frac{DS}{2 - DS} \qquad \text{ou de manière équivalente :} \quad DS = \frac{2 \cdot J}{J + 1} \tag{1.12}$$

Le coefficient de clustering des liens

Il est à noter que toutes les mesures de similarité introduites précédemment peuvent être appliquées à deux nœuds, qu'ils soient ou non reliés par une arête dans le graphe. On introduit maintenant la mesure du coefficient de clustering des liens [121], qui donne un score pour chaque arête du réseau de manière analogue au coefficient défini pour les nœuds. Il compte le ratio entre le nombre de triangles dans lesquels une arête est impliquée, divisé par le nombre maximal de triangles dans lequel celle-ci peut être impliquée.

[2] On prend le voisinage ouvert, dans lequel le nœud considèré n'est pas compté comme un élément de son propre voisinage

$$C_l(v_i, v_j) = \frac{n_{ij}}{\min(k_i, k_j) - 1} \cdot A_{ij} \tag{1.13}$$

où Aij est l'élement ij de la matrice d'adjacence du graphe. Cette mesure est donc par définition nulle pour toute paire de nœuds n'étant pas reliés dans le graphe.

1.4.2 Échelle macroscopique

Les mesures introduites précédemment ont une échelle de précision qui ne permet pas d'aller au delà des considérations locales servant à les définir. Elles constituent de ce fait de bons indicateurs concernant la structure locale entourant les nœuds ou liens considérés, mais ne sont pas suffisantes pour établir une connaissance globale des caractéristiques du réseau.

Nous nous intéressons dans cette partie à un autre type de propriétés, celles qui ne peuvent être observées que si l'on considère le réseau dans son entièreté, révélant ainsi ses propriétés à l'échelle globale. Ce type d'analyse a été longuement étudié et s'est révélé très utile, car permettant parfois de trouver des similarités dans des réseaux de systèmes différents. Nous fournissons dans ce qui suit quelques unes de ces propriétés les plus remarquables.

L'effet Scale-free

Le terme scale-free (invariance d'échelle), a été pour la première fois utilisé par l'économiste italien Vilfredo Pareto, à la fin du 19ème siècle dans son célèbre "Cours d'économie politique" [96], qui remarqua qu'une minorité d'individus dans la société accaparait la majorité des revenus disponibles. Pareto fait le lien entre cette disparité et le fait que la distribution des revenus obéissait à une loi de puissance. Ce résultat fut par la suite connu sous le nom de loi des 80/20, selon laquelle à peu près 80% de la richesse mondiale profite seulement à 20% de la population.

Cette loi selon laquelle une grande part des ressources profite à une minorité, est observée dans différents secteurs : 80% des décisions sont prises durant 20% des réunions, ou encore que 80% des profits sont générés par 20% des employés. Bien entendu on observe aussi ce genre de caractéristiques sur les réseaux : 80% des citations d'articles scientifiques sont dues à 38% des chercheurs, ou encore 80% des liens dans Hollywood (un lien est créé entre deux acteurs si ils ont joué ensemble dans le même film) sont connectés à 30% d'acteurs [14, 10]. Cette observation est la signature d'une distribution en loi de puissance de la quantité étudiée.

L'étude des systèmes libres d'échelle prend ses racines dans le domaine de la physique statistique, dans lequel les distributions en loi de puissance ont été observées et largement étudiées [85, 64, 4, 118, 106]. Pour mieux comprendre la signification de ce terme, nous devons d'abord rappeler quelques notions de la théorie des probabilités.
Soit une variable aléatoire X positive ou nulle, ayant une distribution $P(X)$. On définit le moment d'ordre n par quantité suivante :

$$\langle X^n \rangle = \sum_{k=0}^{\infty} k^n \cdot P(X = k) \tag{1.14}$$

ou dans le cas où la variable aléatoire est continue:

$$\langle X^n \rangle = \int_0^{\infty} X^n \cdot P(X) dx. \tag{1.15}$$

On appelle ainsi la moyenne le moment d'ordre 1 et la déviation standard s'exprime en fonction des moments de premier et second ordre.

Pour une distribution ayant un moment d'ordre 2 à valeur finie, et une série de tirages aléatoires de la variable aléatoire X qui obéit à cette distribution, on dispose d'un indicateur sur la variabilité des valeurs de la série autour de la moyenne de X. Cet indicateur décrit de quelle façon ces valeurs s'éloignent de la moyenne, et est donné par l'intervalle

$$\langle X \rangle \pm \sigma_X$$

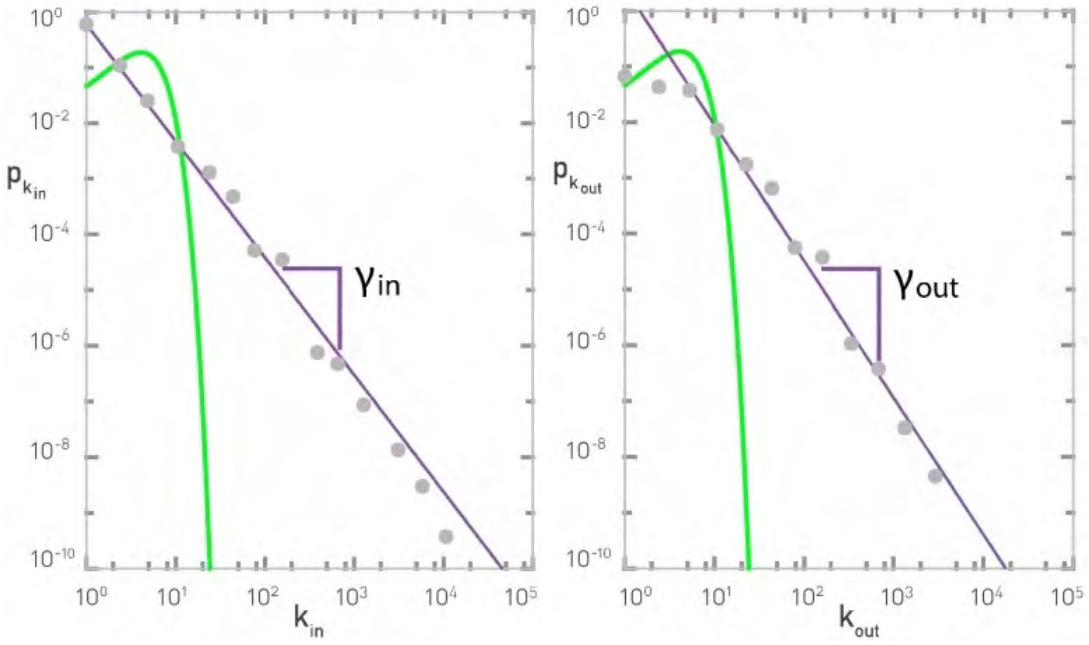
où σ_X est la déviation standard de la variable X. Nous verrons que dans le cas de certaines lois de puissance, cette propriété n'est plus valable car certaines de ces quantités ont des valeurs divergentes.

Intéressons nous maintenant à la distribution des degrés dans un réseau. Pour chaque degré k se situant entre le degré minimum k_{min} (souvent assimilé à 1) et le degré maximum k_{max} d'un réseau, nous avons une distribution empirique $P(k)$ qui décrit la probabilité de tirer au hasard un nœud de degré k parmi tous ceux qui constituent le réseau. Cette distribution peut être (pour des réseaux de taille assez grande) estimée par ses fréquences empiriques de la manière suivante :

$$p(k) = \frac{n_k}{N}$$

où n_k désigne le nombre de nœuds de degré k dans le réseau. On obtient à partir de ces données une courbe de régression représentant la distribution des degrés dans le réseau.

Sur la figure 1.5 nous pouvons apercevoir la distribution approchée des degrés rentrants et sortants, estimée sur un échantillon du réseau du web, dans lequel un lien existe de la page A vers la page B si A fait mention de B.



(a) (b)

Figure 1.5: La distribution des degrés rentrants (a) et sortants (b) d'un échantillon du web cartographié dans l'étude d'Albert et al. de 1999 [7]. La distribution est affichée en échelle logarithmique, sur laquelle une loi de puissance décrit une droite. Les points représentent les données empiriques et l'exposant de la loi est obtenu par le coefficient directeur de la droite de régression correspondante. Ici nous avons $\gamma_{in} = 2.1$ et $\gamma_{out} = 2.45$. la courbe verte correspond à la distribution attendue d'une loi de Poisson dont les paramètres seraient les valeurs moyennes des degrés rentrants et sortants. Figure tirée de [14]

En calculant l'indicateur de dispersion d'une distribution en loi de faible puissance γ, on se heurte à un problème caractéristique de ce type de distributions. En effet les quantités définies par les eq. (1.14) et eq. (1.15) sont divergentes pour $n \geq \gamma - 1$. Bien entendu ce ne sont là que des hypothèses, car en pratique le degré maximal d'un réseau quelconque est fini, mais pour de très grand réseaux celui-ci peut être relativement grand, et par conséquent la déviation standard σ_k de la distribution des degrés l'est aussi. Ceci a pour conséquence de rendre très large l'intervalle caractéristique $\langle k \rangle \pm \sigma_k$ en comparaison avec la valeur de $\langle k \rangle$, d'où le terme libre d'échelle.
Cette propriété explique les effets observées sur de tels réseaux, en effet on y retrouve un très grand nombre de nœuds de petits degrés, au milieu desquels se trouvent un petit nombre de nœuds dont le degré est très élevé. Dans l'exemple du réseau du web évoqué plus haut, on peut faire l'analogie avec les quelques sites très populaires comme Google, Facebook, etc. qui monopolisent une grande partie des liens rentrants du web, à côté desquels existent un très grand nombre de pages web qui ne sont presque jamais citées ou visitées.

La distribution des degrés en loi de faible puissance est une propriété qui a été observée dans une très grande variété de réseaux réels, allant de l'exemple déjà évoqué du web [12], aux réseaux d'échanges de courriers électroniques [46], en passant par des réseaux biologiques [14]. Ceci permet d'apporter un exemple à l'idée selon laquelle il existe des régularités dans la structure des réseaux, indépendamment de leur complexité ou du système qu'ils représentent. Il est aussi intéressant de mentionner la multitude de modèles génératifs de réseaux synthétiques possédant la propriété scale-free [66, 71, 13]. Le plus connu est celui de Barabasi-Albert [6], et qui consiste à partir d'un réseau initial quelconque (qui peut ne contenir aucun nœud), puis de rajouter des nœuds en les reliant aléatoirement à ceux qui sont déjà présents dans le réseau, avec une probabilité proportionnelle à leur degré (par attachement préférentiel). Ceci induit un effet intéressant : plus le degré d'un nœud est grand, plus il a des chances de créer des liens supplémentaires, et d'augmenter encore plus son degré.

L'effet petit monde

Dans les années 1960 le psychologue américain Stanley Milgram tente une expérience qui sera plus tard connue sous le nom de l'expérience du petit monde, dans laquelle il choisit de fournir à 96 personnes choisies au hasard depuis l'annuaire téléphonique de la ville d'Omaha dans l'état du Nebraska, un courrier contenant un document semblable à un passeport et sur lequel il y avait le tampon de l'université de Harvard [119]. Il contenait aussi des instructions leur expliquant la marche à suivre pour le bon fonctionnement de l'expérience : dans le courrier est renseignée l'adresse d'un proche de Milgram vivant dans la ville de Boston, à plus d'un millier de kilomètres d'Omaha, il leur est alors demandé de faire parvenir ce courrier à la personne la plus susceptible d'être proche du destinataire et de lui demander de faire de même jusqu'à ce que le courrier finisse par atterrir chez la personne en question. Il devenait alors possible d'estimer le nombre d'intermédiaires séparant deux personnes au hasard dans les États-Unis. Les résultats de l'expérience sont les suivants : d'abord des 96 courriers envoyés, 18 ont pu atteindre leur destination finale, et ensuite le nombre d'intermédiaires estimé par l'expérience est de 5.9.

On peut émettre plusieurs remarques quant aux conditions expérimentales et aux biais auxquels elle est sujette, comme par exemple le fait que les individus qui participent soient tous de la même ville, et que la destination finale soit la même pour toutes les lettres. Par ailleurs, on peut objecter que le petit nombre de lettres qui ont finalement réussi à atteindre leur destination n'est pas assez grand pour en faire une estimation précise du véritable nombre d'intermédiaires moyen. Cette expérience a pour autant réussi à soulever la question du degré de connectivité de la société, et avec les moyens dont on dispose aujourd'hui, on a établi que l'hypothèse du petit monde [88] était bien fondée (*cf.* fig. 1.6).

Nous avons introduit plus haut le concept de diamètre d'un réseau, défini par la plus grande valeur prise par le chemin le plus court qui sépare chaque paire de nœuds possibles. Bien que cet indicateur constitue un bon moyen d'estimer la connectivité d'un réseau, il est parfois meilleur de calculer sa valeur moyenne, sur l'ensemble des chemins les plus courts séparant chaque paire de nœuds. On peut ainsi obtenir une estimation de la distance moyenne séparant deux nœuds choisis au hasard. Nous retrouvons cet effet de petit monde dans beaucoup de réseaux réels, notamment chez ceux ayant la propriété scale-free, car la présence de nœuds de grands degrés (hubs) facilite grandement la liaison entre deux nœuds, par des géodésiques (chemins les plus courts) qui traversent ces hubs.

Nous montrons ici une figure sur laquelle sont montrés les diamètres et les coefficients de clustering d'un certain nombre de réseaux réels, ainsi que ceux estimés sur des réseaux aléatoires d'Erdős–Rényi ayant le même degré moyen et la même taille que les réseaux réels auxquels ils sont comparés.

Nous pouvons tirer de cette figure des informations intéressantes : la plupart des réseaux réels de ce tableau possèdent la propriété de réseau petit monde. De plus, ils ont tous un coefficient de clustering significativement plus élevé que ceux qu'on obtient sur les réseaux aléatoires auxquels ils sont comparés. Le modèle génératif de Watts et Strogatz [122] est souvent utilisé pour produire des réseaux petit monde, plus précisément il est obtenu en recablant aléatoirement un certain pourcentage d'arêtes d'une grille régulière. Ce modèle reproduit (pour certaines valeurs du paramètre p) quelques unes des caractéristiques des réseaux réels, à savoir à la fois une faible valeur moyenne du chemin le plus court [3], et un grand coefficient de clustering (*cf.* fig. 1.7). Il est pourtant loin de reproduire toutes ces caractéristiques, en particulier la propriété scale-free.

[3]Celle-ci étant obtenue par la moyenne des chemins les plus courts calculés entre toutes les paire de nœuds du graphe

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998

Albert, R. et.al. Rev. Mod. Phy. (2002)

Figure 1.6: Tableau listant les tailles, degrés moyens ainsi que les diamètres et les coefficients de clustering réels et estimés depuis un échantillon aléatoire de nœuds, pour 17 réseaux réels. Figure tirée de [6]

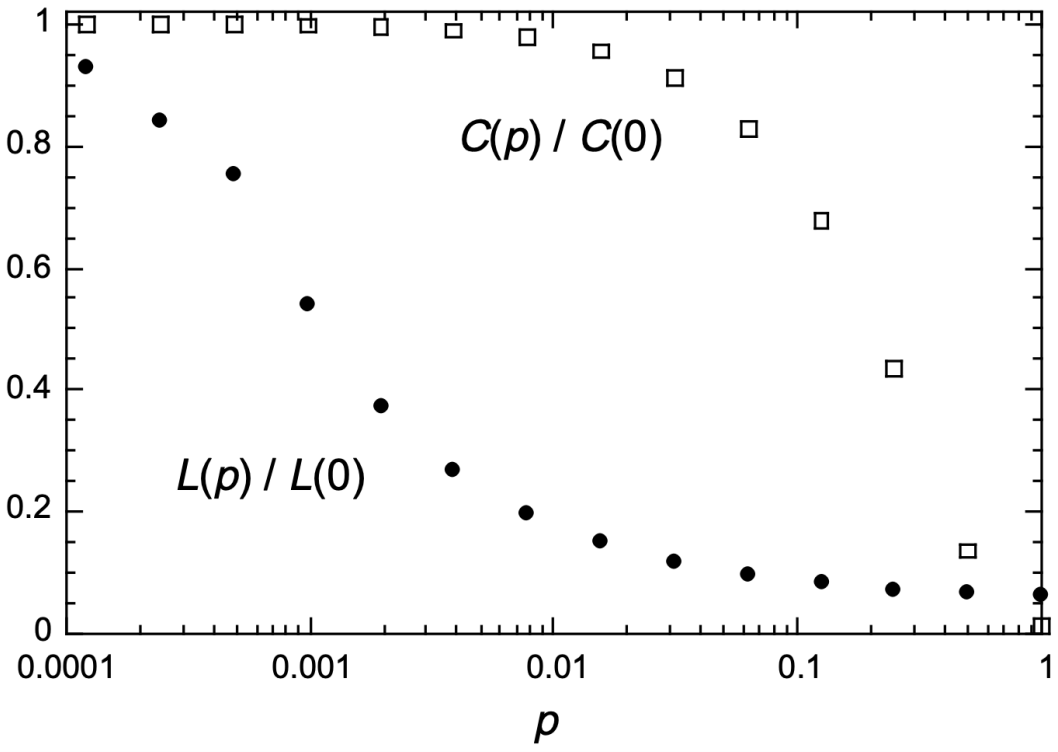


Figure 1.7: Les valeurs moyenne du coefficient de clustering et du chemin le plus court (tous les deux normalisés), en fonction du paramètre p de recâblage aléatoire. Nous pouvons voir sur cette figure que la moyenne du chemin le plus court diminue plus rapidement que le coefficient de clustering. Figure tirée de [122]

Assortativité et disassortativité

Nous disposons jusqu'à présent de trois caractéristiques globales que l'on retrouve dans les réseaux réels, souvent toutes les trois en même temps, qui sont une distribution des degrés en loi de faible puissance

(effet scale-free), une faible valeur moyenne du chemin le plus court et un coefficient de clustering élevé. Nous poursuivons par la notion de mixage assortatif, qui quantifie la préférence qu'ont les nœuds d'un réseau à former des liens avec des voisins qui leur sont similaires d'une manière ou d'une autre. On peut penser à une multitude de mesures de similarité (comme celles introduites dans la section 1.3.1), mais celle à laquelle nous allons nous intéresser ici concerne le degré des nœuds, afin de mettre en évidence les corrélations entre nœuds de degrés similaires, c'est à dire leur tendance à être connectés à des voisins ayant des degrés du même ordre de grandeur. Si cette corrélation existe et est positive, on dit qu'un réseau est assortatif, si au contraire elle est négative, on dira plutôt que le réseau est disassortatif (c'est à dire que les nœuds de degrés similaires ont délibérément tendance à s'éviter). On dit que le réseau est neutre si aucune corrélation n'est observée.

L'assortativité est mesurée par un indicateur qu'on appelle le coefficient de corrélation de Pearson [98], qu'on applique aux paires de nœuds qui sont reliés par une arête. Son expression est donnée par la formule suivante [92] :

$$r = \frac{\sum_{jk} j \cdot k \cdot (e_{jk} - q_j q_k)}{\sigma_q^2} \tag{1.16}$$

où q_k est la distribution des degrés restant, c'est à dire le nombre d'arêtes rattachées au nœud, autres que celle qui relie la paire. e_{jk} fait référence à la distribution de probabilité commune des degrés restants des deux sommets. Cette quantité est symétrique sur un réseau non dirigé et vérifie les normalisations :

$$\sum_{jk} e_{jk} = 1 \text{ ; et } \sum_{j} e_{jk} = q_k = \frac{(k+1)p_{k+1}}{\sum_{j \geq 1} j p_j}$$

où p_k est la distribution des degrés du réseau. Pour plus de détail sur les calculs, nous référons le lecteur au papier de [89] dont nous montrons quelques coefficients d'assortativité sur certains réseaux provenant du monde réel, ainsi que deux modèles de réseaux synthétiques :

	network	n	r
real-world networks	physics coauthorship[a]	52 909	0.363
	biology coauthorship[a]	1 520 251	0.127
	mathematics coauthorship[b]	253 339	0.120
	film actor collaborations[c]	449 913	0.208
	company directors[d]	7 673	0.276
	Internet[e]	10 697	−0.189
	World-Wide Web[f]	269 504	−0.065
	protein interactions[g]	2 115	−0.156
	neural network[h]	307	−0.163
	food web[i]	92	−0.276
models	random graph[u]		0
	Callaway *et al.*[v]		$\delta/(1+2\delta)$
	Barabási and Albert[w]		0

Figure 1.8: Coefficients d'assortativités et tailles respectives calculés sur 13 réseaux réels, tableau tiré de [89]

Nous remarquons là aussi que les réseaux réels présentent les deux tendances, l'assortativité dans certains cas, comme nous pouvons l'observer sur les réseaux de collaborations entre mathématiciens ou entre physiciens, et que d'un autre côté il existe de la dissortativité entre les pages internet. Les modèles synthétiques d'Erdős–Rényi et Barabasi-Albert sont quant à eux neutres, de coefficient nul.

L'effet rich club

On sait qu'un réseau ayant une valeur positive de son coefficient d'assortativité des degrés a parmi ses propriétés la suivante : ses nœuds de degrés élevés tendent à former des liens entre eux. On peut alors aller plus loin dans la compréhension des structures formées par ces liens. Pour cela les chercheurs dans [124] ont formalisé l'idée sous forme d'un filtre, d'abord en classant les nœuds par degrés croissants, ensuite en mesurant la proportion de liens que se partagent entre eux ces individus de grands degrés.

Le filtre devient par moment tellement sélectif qu'on se retrouve avec un petit nombre d'individus dans le réseau, qui porte le nom de club huppé, ou "rich club" en anglais[4].

Formellement, on quantifie l'effet du rich club dans un réseau en utilisant deux mesures appelées le paramètre de richesse $\phi(k)$ pour la première et le coefficient du rich club $R(k)$ pour la seconde. Ces deux mesures sont détaillées ci-dessous.

$$\phi(k) = \frac{2 \cdot E_{>k}}{N_{>k} \cdot (N_{>k} - 1)} \tag{1.17}$$

où $E_{>k}$ désigne le nombre d'arêtes parmi les $N_{>k}$ nœuds ayant un degré supérieur à k. Ainsi, $\phi(k)$ est la densité en liens du sous-réseau induit par l'ensemble des nœuds ayant des degrés supérieurs à l'entier k (k joue ici le rôle de filtre). Un réseau qui possède la propriété du rich club doit de ce fait avoir une courbe $\phi(k)$ d'allure croissante.

Par ailleurs, les nœuds de forts degrés ont une plus forte probabilité à être reliés entre eux que les nœuds de faibles degrés. Pour limiter ce biais, Colizza et al [34] proposent une normalisation du coefficient du rich club.

$$R(k) = \frac{\phi(k)}{\phi_{null}(k)} \tag{1.18}$$

où $\phi_{null}(k)$ est le résultat attendu de la même analyse sur un modèle nul, composé d'une version aléatoire (mais conservant certaines propriétés) du réseau étudié. Une telle mesure permet d'évaluer à quel point les nœuds de degrés supérieurs ou égaux à k sont liés entre eux, car elle quantifie l'effet évoqué plus haut sur un modèle de comparaison, dans lequel chaque nœud à le même degré que dans le réseau étudié, mais dont les liens sont recâblés de manière aléatoire. On reviendra sur cette propriété du rich club dans le chapitre 4, car elle constitue le point central de l'un des algorithmes développés durant cette thèse. Nous montrons maintenant l'évolution du coefficient $R(k)$ sur certains réseaux, tirés de [124].

On remarque sur la figure 1.9 que l'effet du rich club est bien distinguable sur le réseau des collaborations scientifiques, le réseau des transports aériens, ainsi que celui généré à travers le modèle de Barabasi-Albert. Cet effet est caractérisé par la présence d'un pique du coefficient $R(k)$ d'une valeur supérieure à 1, situé autour d'une grande valeur de k. À l'inverse nous observons sur les figures représentant les réseaux d'interactions entre protéines, et le réseau des pages internet qu'il existe un minimum inférieur à 1 du coefficient $R(k)$ pour les grandes valeurs de k, ce qui indique que les nœuds de grands degrés sont moins liés entre eux que ce que l'on obtient si ces derniers avaient formé leurs liens aléatoirement. Ces résultats sont en accord avec ceux de la figure 1.8, sur laquelle on observe que les réseaux de collaborations scientifiques étaient assortatifs, et qu'à l'inverse les réseaux d'interactions entre protéines et d'internet (dont le coefficient du rich club est négatif) étaient disassortatifs. Il est important de préciser qu'on parle ici de tendances globales, propres à chaque type de réseaux, car les deux résultats montrés sur la figure 1.9 et ceux de la figure 1.8 proviennent de jeux de données différents.

1.4.3 Échelle mésoscopique

En physique, le terme mésoscopique est employé pour indiquer une échelle d'observation intermédiaire, se situant entre le microscopique (comme par exemple la physique quantique, physique statistique, etc.) et l'échelle macroscopique (exemple : mécanique des objets solides, thermodynamique). Nous l'utilisons ici pour introduire les outils d'analyse qui se situent entre l'échelle du nœud et/ou du lien comme entité unique à analyser, et celle qui exprime des grandeurs qui couvrent le réseau dans son entièreté. Il s'agit donc d'une échelle qui traite des ensembles dont la taille se situe entre l'unité et l'ensemble du graphe. Elle est l'objet d'un certain nombre d'approches et d'algorithmes qui, pour un réseau en entrée, fournissent un découpage de celui-ci en sous-ensembles vérifiant certaines caractéristiques structurelles. Ces caractéristiques peuvent être, comme nous le verrons, soit basées sur des définitions précises, soit de nature plus subjective.

Certains de ces algorithmes ont été largement étudiés durant les dernières années et on ne pourra pas tous les présenter. Nous nous limitons, comme nous l'avons fait jusqu'ici, aux plus répandus parmi les différents types d'approches identifiées dans cette catégorie.

[4] On emploiera la nomination rich club par la suite

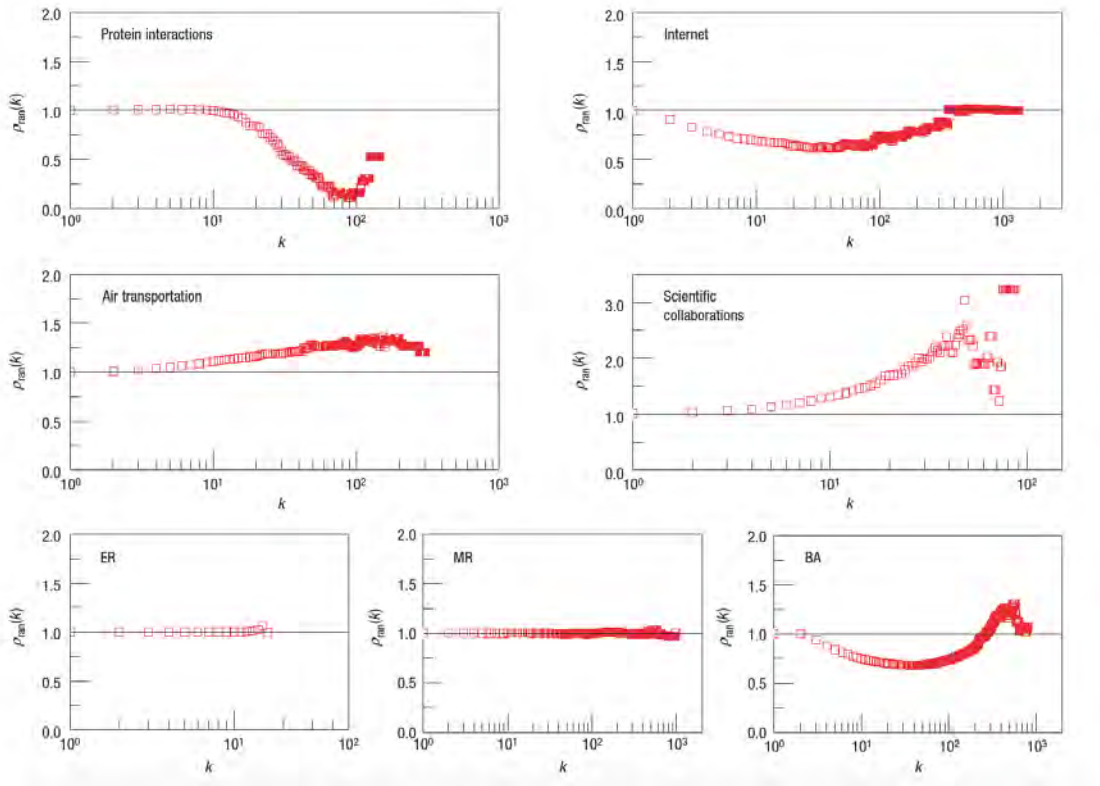


Figure 1.9: Coefficient du rich club en fonction du degré k pour 4 réseaux réels (partie haute), et 3 réseaux synthétiques (rangée du bas). Le coefficient est noté $\rho_{ran}(k)$ sur cette figure, alors que nous l'avons noté $R(k)$ sur eq. (1.18). Les courbes présentant des valeurs supérieures à 1 démontrent que les graphes correspondants présentent un effet rich club, alors que celles qui sont de valeurs inférieures ou égales à 1 n'en ont pas, ou manifestent l'effet inverse (dissortativité). Figure tirée de [34]

La décomposition en $k-$cores

La notion de noyau (core en anglais) a été introduite par Seidman [107] en 1983, dans une étude visant à apporter une nouvelle approche à la mesure de la cohésion dans les réseaux, le degré présentant quelques lacunes en la matière. Soit $G = \{V, E\}$ un graphe non orienté et non pondéré tel que $|V| = N$ et $|E| = m$. Soit $V_F \subset V$ un sous-ensemble de nœuds, on note F le sous-graphe induit par V_F.

Le $k-$core, ou noyau d'ordre k, est défini comme le sous-graphe induit par $V_F \subset V$, tel que chaque nœud de V_F est au moins de degré k dans F : $\forall v \in V_F : k_v \geq k$. On associe chaque nœud du graphe au nombre k du noyau d'ordre le plus élevé auquel il appartient. Cette définition est facilement généralisable aux graphes dirigés, en restreignant les définitions précédentes aux degrés entrants, et sortants, obtenant ainsi des noyaux dans ces deux directions. En revanche il n'y a pas à ce jour de méthode faisant l'unanimité, et qui généralise le concept de k-core aux graphes pondérés, car cette notion repose principalement sur le degré.

On note que les noyaux sont en relation d'inclusion, *i.e.* $k-\text{core} \subset (k+1)-\text{core}$, et qu'ils n'induisent pas nécessairement de sous-réseaux connexes. On retrouve dans la littérature une définition relative à celle des $k-$cores, notée $k-$shell (signifiant la couche d'ordre k) qui est définie par l'ensemble de nœuds dont le nombre de core maximal est k, qui en d'autres mots désigne les nœuds appartenant au $k-$core mais pas au $(k+1)-$core.

Un algorithme simple permet de retrouver de manière itérative les différents $k-$cores de la façon suivante : on commence par classer tous les nœuds de degré 0 (s'il y en a) dans le $0-$core, ensuite on supprime de façon itérative tous les nœuds de degré 1 jusqu'à ce que le réseau obtenu ne contienne plus de nœuds de degré 1. L'ensemble des nœuds ainsi supprimés est par définition le 1-shell, et le réseau composé par les nœuds qui n'ont pas été supprimés est le $2-$core. On reproduit ensuite ce même processus, en se concentrant sur les nœuds de degré 2, et ainsi de suite pour chaque entier k, jusqu'à ce

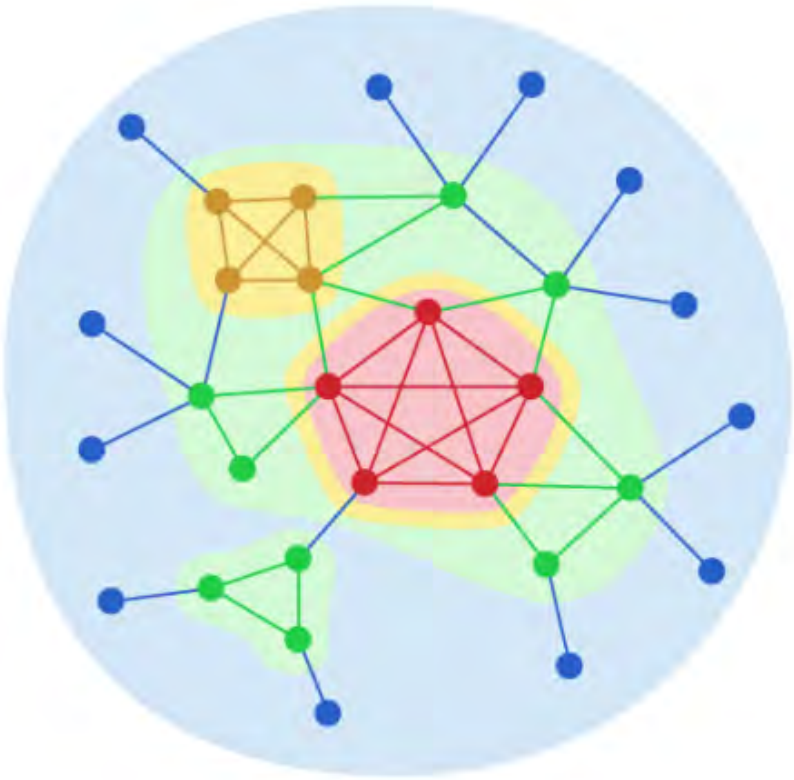


Figure 1.10: Décomposition d'un réseau en $k-$cores : les nœuds en rouge représentent la $4-$shell, ceux en jaune la $3-$shell, ceux en vert la $2-$shell et la $1-$shell en bleue. Chaque $k-$core inclut tous les nœuds de l'ensemble des $s-$shell avec s un entier supérieur ou égal à k. Figure tirée de [111]

que le graphe ne contienne plus aucun nœud.

Le partitionnement en communautés

On vient de montrer qu'il existait un algorithme permettant de décomposer le graphe en plusieurs sous-ensembles imbriqués les uns dans les autres, suivant le degré de chaque nœud. Nous allons dans cette partie décrire une famille d'algorithmes correspondant à un autre type de décomposition, basée sur le contraste entre cohésion locale et globale des nœuds, et dont le résultat est en général (dans une version simplifiée du problème mais ceci n'est pas nécessairement le cas) une partition dans laquelle chaque nœud appartient à un unique sous-ensemble. Chaque ensemble est caractérisé par le fait d'être constitué de nœuds qui ont la propriété d'être plus reliés entre eux qu'ils ne le sont au reste du réseau, et que l'on appelle communautés. Ces communautés sont très utiles à la compréhension de la structure globale du réseau, car elles séparent celui-ci en différents sous-ensembles, nous permettant de mieux connaître la façon dont les nœuds sont reliés du point de vue structurel, en plus de fournir une échelle intermédiaire de représentation, à travers la manière dont ces communautés sont reliées les unes aux autres.

Il existe aujourd'hui un grand nombre d'algorithmes dont l'objectif est de fournir une partition du réseau en communautés [56], et bien que les chercheurs s'y soient penchés relativement tôt [74, 15, 117], cette approche n'a connu son véritable essor que depuis l'introduction par Newman de la modularité [91], une mesure permettant d'évaluer la qualité d'une partition donnée, et par conséquent de comparer différentes partitions, puis d'en choisir la meilleure. Nous ne pouvons pas présenter ici chacune des méthodes existantes pour la détection de communautés (nous référons le lecteur à [56] pour une lecture plus approfondie sur le sujet) mais nous allons donner une brève description de certains des algorithmes les plus répandus.

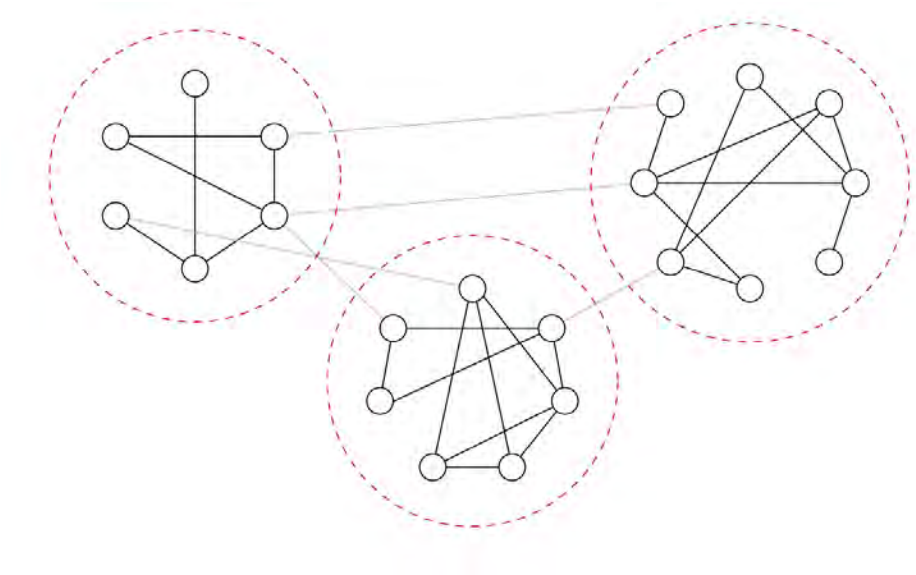


Figure 1.11: Exemple d'un graphe comportant 3 communautés, chacune étant entourée d'un cercle en pointillés rouges, présentant la propriété d'avoir plus de liens intra-communautaires que de liens inter-communautaires. Figure tirée de [90]

La modularité et l'algorithme de Louvain

Soit $C = \{c_1, c_2, ..., c_l\}$ une partition, où c_i est une communauté composée d'un certain nombre de nœuds. La modularité est définie alors comme :

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i \cdot k_j}{2m}) \cdot \delta(c_i, c_j) \tag{1.19}$$

où m est le nombre de liens dans le réseau, c_i désigne le sous-ensemble de nœuds appelé communauté (mais aussi parfois clusters, ou modules) et $\delta(m, l)$ est le symbole de Kronecker, qui est égal à 1 si $m = l$ et 0 sinon. Ici le rapport $\frac{k_i \cdot k_j}{2m}$ désigne la probabilité que deux nœuds i et j de degrés respectifs k_i et k_j forment un lien dans un réseau aléatoire de même distribution de degré que le réseau original.

Cette métrique est intéressante dans la mesure où elle accorde des valeurs positives élevées à des partitions dans lesquelles les nœuds partagent plus de liens internes (reliant les nœuds de la même communauté), que de liens externes (ou intercommunautaires), tout en comparant les résultats obtenus aux valeurs que l'on obtiendrait si le réseau étudié était d'une structure aléatoire, à travers le recâblage arbitraire des arêtes du réseau (ce qui, comme nous l'avons déjà mentionné, permet de préserver le degré de chaque nœud). La modularité peut aussi prendre des valeurs négatives, si par exemple la partition évaluée ne contient aucune arête reliant 2 nœuds du même module, tout en ayant au moins une arête entre deux nœuds de modules différents. Sa valeur tend vers zéro pour une partition aléatoire sur un réseau dans lequel les liens sont aussi aléatoirement formés.

On peut maintenant introduire l'algorithme de Louvain qui calcule une partition en optimisant la modularité de manière gloutonne, l'idée est simple : commencer par assigner une communauté différente à chaque nœud dans le réseau, et ensuite fusionner les communautés qui induisent la plus grande hausse de modularité (après avoir testé les manières possibles de les fusionner) et répéter ainsi jusqu'à atteindre la valeur maximale de la modularité, en s'assurant que le moindre changement ultérieur baisse sa valeur [18]. On répète ensuite cette opération sur un nouveau réseau obtenu en fusionnant les nœuds d'une même communauté (les nœuds de la même communauté sont maintenant représentés par le même nœud dans le nouveau réseau et les liens entre les nœuds de la même communauté deviennent des boucles), et en pondérant chaque lien du nouveau réseau avec un poids égal au nombre de liens entre les deux communautés correspondantes.

La méthode de Louvain a été largement étudiée, et il existe beaucoup de variantes qui corrigent certains de ses désavantages. Parmi ces désavantages, on peut citer le fait qu'elle soit gloutonne et s'arrête au premier maximum de modularité rencontré, ce qui n'assure pas que celui-ci soit un maximum

global. Une autre caractéristiques importante est que la méthode est non déterministe, en effet l'ordre dans lequel on parcourt les modules durant l'optimisation peut changer le résultat final de manière remarquable, de même que l'ajout d'un faible nombre de nœuds au réseau. Finalement, on peut aussi citer le fait qu'elle ait du mal à détecter les communautés de petites tailles. Il existe d'autres commentaires sur les limites de cette méthode, nous renvoyons le lecteur à [87, 26, 9] pour plus de détails.

Un des points importants qui n'est cependant pas pris en compte par les algorithmes basés sur l'optimisation de la modularité, est la possibilité d'un regroupement entre communautés. En effet le fait que chaque nœud du réseau soit contenu dans une seule et unique communauté constitue une contrainte assez restrictive. Par exemple, dans les réseaux sociaux, un individu peut appartenir à la fois à la communauté de son cercle familial, mais aussi à celle constituée par son entourage professionnel. Cet exemple en est un parmi d'autres, et en principe on ne devrait imposer aucune contrainte sur la nature des communautés qui constituent le réseau sans connaître au préalable la structure de celui-ci. Ainsi de nombreuses études se sont penchées sur la question en apportant chacune des contributions différentes [31], nous en citons quelques unes ici.

Méthodes basées sur les communautés de liens

Ce type de méthode se concentre sur le regroupement de liens. En maximisant une fonction objectif définie au préalable, on calcule un ensemble L de communautés $l_i, i \in [1, N_L]$ constituées de liens, on extrait ensuite de chacune une communauté de nœuds c_i constituée de tous les nœuds qui sont à l'extrémité d'au moins un des liens dans l_i. Dans Ahn et al [3]. Les communautés de liens sont calculées en rassemblant dans le même cluster les arêtes les plus similaires, en quantifiant celle-ci par la mesure suivante : pour chaque paire de liens $((i,k),(j,k))$ ayant comme extrémité commune un nœud k , on calcule l'indice de Jaccard défini dans la partie 1.3.1, entre les voisinages $Nb(i)$ et $Nb(j)$ des nœuds i et j.

$$S((i,k),(j,k)) = \mathrm{J}(Nb(i), Nb(j))$$

Il reste ensuite à choisir la partition optimale, pour cela on interrompt le processus d'agglomération en maximisant la fonction objectif suivante :

$$D = \frac{2}{m} \sum_l m_l \cdot \frac{m_l - (n_l - 1)}{(n_l - 2)(n_l - 1)} \tag{1.20}$$

avec m le nombre de liens, m_l et n_l sont respectivement le nombre de liens et de nœuds dans le cluster l.

Méthode de percolation de cliques

Cette approche se base sur l'idée qu'une communauté étant dense en connexions, elle doit être constituée de plusieurs petites cliques de tailles différentes et qui partagent un certain nombre de nœuds en commun [95]. On choisit alors un entier k représentant la taille la plus petite possible pour une communauté, de sorte à ce que tout nœud n'appartenant pas à une clique de taille au moins égale à k ne fera partie d'aucune communauté. On définit ensuite une communauté, comme l'union de toutes les k-cliques que l'on peut atteindre en passant de l'une à l'autre par une série de k-cliques adjacentes.

Cette méthode implique le choix empirique de l'entier k, mais est connue pour son efficacité sur les réseaux de terrains, sur lesquels il est d'usage de choisir un entier k supérieur ou égal à 3.

La méthode OSLOM

La méthode OSLOM est l'une des rares approches qui s'appuient sur les propriétés statistiques des clusters, tout en ayant une propriété qui la différencie d'un grand nombre d'algorithmes : certains nœuds peuvent ne faire partie d'aucune communauté et sont considérés comme du bruit autour des clusters auxquels ils sont reliés (on retrouve aussi cette propriété dans la méthode de la percolation de cliques, pour des valeurs assez élevées de k, mais ce résultat est la conséquence systématique du choix de k, alors que pour OSLOM chaque situation est traitée à part). OSLOM est initialisé avec le résultat d'un

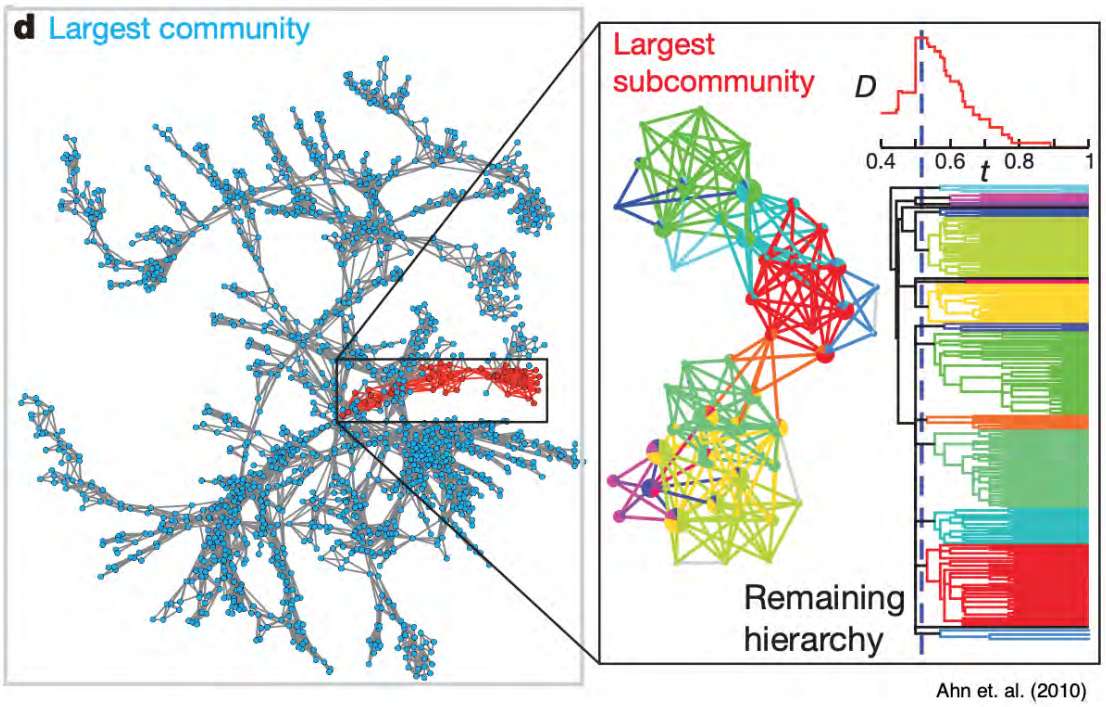


Figure 1.12: Sur la partie gauche est représentée l'échantillon d'un réseau de téléphonie mobile, avec en rouge les nœuds de la plus grande communauté détectée par la méthode décrite dans cette partie. À droite de la figure, on peut voir que cette communauté est obtenue en coupant le dendrogramme au maximum de la fonction objectif, et qu'elle est elle-même constituée de communautés de plus petites tailles. Figure tirée de [3]

algorithme classique de partitionnement de sommets (par exemple Louvain). Cette partition constitue l'ensemble des communautés de départ qu'on appelle graines, on calcule ensuite une probabilité pour chaque nœud d'être connecté à la communauté qui lui est assignée, en prenant en considération le nombre de nœuds que celle-ci contient, ainsi que le degré du nœud. On peut ainsi estimer le nombre d'arêtes qui relient le nœud à la communauté dans le cas d'un modèle nul aléatoire, finalement on choisit de le garder si ce nombre est plus grand que ce qui est attendu dans le cas aléatoire, ou au cas contraire, de l'enlever de la communauté. Le processus est répété N fois pour chaque graine, afin d'éviter les minimums locaux, et passe aléatoirement d'une graine à l'autre jusqu'à l'obtention d'un résultat stable.

Nous ne pouvons bien entendu pas donner une description détaillée de chacun des algorithmes de détection de communautés.

En plus des méthodes décrites ci-dessus, il existe d'autres types d'approches comme par exemple celles basées sur la marche aléatoire [99], les méthodes de compression d'information [105], les approches spectrales [27, 82], la méthode de propagation de labels [100] et bien d'autres.

La propriété qu'il est important de retenir de la description (sommaire) que l'on vient de donner, est qu'il n'y a aucun consensus sur la définition formelle d'une communauté. Au lieu de cela, toutes les approches s'appuient sur des considérations empiriques qui servent ensuite à construire l'algorithme correspondant. Par exemple l'idée que les nœuds d'une communauté sont supposés avoir plus de cohésion entre eux qu'ils n'en ont avec le reste du réseau a servi à introduire la mesure de la modularité, qui est elle-même optimisée afin de donner la meilleure partition compte tenu de cette considération, mais il est fondamental de rappeler qu'en l'absence d'une réelle définition de ce qu'est une communauté, on ne peut que choisir parmi les approches disponibles celle qui est la plus appropriée aux données étudiées.

Les nœuds sans communautés

Dans la plupart des algorithmes décrits précédemment, l'entrée de l'algorithme est un graphe $G = (V, E)$ et la sortie une partition $C = \{c_1, c_2, ..., c_l\}$ qui assigne une communauté à chaque nœud du réseau, de sorte à ce que l'union de tous les sous-ensembles recouvre l'ensemble des nœuds du graphe.

$$\bigcup_{i=1}^{l} c_i = V. \tag{1.21}$$

Il existe toutefois quelques exceptions, comme nous l'avons vu pour les algorithmes de percolation de cliques (aussi noté CPM pour "clique percolation method") et celui d'OSLOM, dans lesquels il est possible que certains nœuds demeurent sans communauté à la sortie. En effet il semble normal de se poser des questions sur ce type de nœuds, car à priori imposer la contrainte dictée par eq. (1.19) est encore une fois une restriction à l'étude de la structure d'un réseau, qui peut contenir plus de richesse que ce que peut capturer un algorithme de détection de communautés. On distingue alors deux cas de figure, le premier concerne les situations triviales (qui constituent la majorité des cas rencontrés à la sortie d'algorithmes de type OSLOM ou CPM), comme par exemple celui où le nœud identifié sans communauté est une feuille (nœud de degré égal à 1) dont le voisin fait partie d'un cluster, ou bien pour le cas de l'algorithme CPM, un nœud qui n'est inclus dans aucune clique de taille supérieure ou égale à k, qui est l'entier représentant la taille de la graine. Ces nœuds-là peuvent bien entendu constituer un ensemble qui contient des informations importantes sur la structure du réseau, mais ils sont faciles à retrouver, et dans la majorité des cas, on n'a pas besoin de connaître les résultats de l'algorithme au préalable pour les identifier. On s'intéresse plus particulièrement au deuxième cas de figure, qui concerne les nœuds qui n'ont pas d'appartenance claire à un quelconque cluster dense, sans qu'il soit pour autant possible de les identifier de façon triviale. Il n'existe aujourd'hui aucun algorithme qui s'occupe exclusivement de retrouver ce type de nœuds, dont on définit l'ensemble comme la partie non dense du réseau. La section 1.4 est consacrée à l'introduction des différents concepts relatifs à ce type de nœuds, et qui ont participé à l'élaboration de l'algorithme qui permet de les calculer.

1.5 La partie non dense

Considérons pour commencer le fait qu'il puisse y avoir dans certains réseaux un sous-ensemble de nœuds, qu'on reconnaît grâce à une caractéristique particulière : sa faible densité en connectivité. Ces nœuds feraient alors soit partie de la couche périphérique du réseau [116], de par le fait que leurs uniques voisins soient inclus dans des clusters de forte densité, alors qu'eux-mêmes n'en font pas partie (à cause de leur faible nombre de voisins), ou bien de manière moins évidente, sont positionnés à des endroits stratégiques, par exemple en qualité d'intermédiaires entre deux communautés, sans pour autant faire partie de l'une ou de l'autre.

On peut alors se poser quelques questions, concernant l'impact de ces nœuds-là sur la topologie du réseau, ces questions constituent le sujet qui a occupé nos recherches tout au long de cette thèse, et on essayera d'apporter des éléments de réponse tout au long du manuscrit. Mais avant, nous commençons par donner dans ce qui suit un sommaire des travaux qui se rapprochent le plus de notre problématique.

1.5.1 Trous structuraux de Burt

Le terme de trous structuraux a été évoqué et conceptualisé par le sociologue américain Ronald Stuart Burt en 1992 dans un article où il évoque l'importance de la bonne gestion de son réseau personnel, afin de bénéficier des meilleurs avantages dans un milieu compétitif, où les relations de chaque individu jouent un rôle important pour la réussite [23].

Burt définit un trou structurel comme la séparation qui existe entre les contacts non redondants. Ici le terme séparation fait allusion à l'absence de liens, les contacts sont l'équivalent pour nous des nœuds dans un réseau (en particulier les voisins dans un réseau égo-centré), et la non redondance entre deux nœuds exprime le fait que les deux mènent à des cercles sociaux qui ne se recoupent pas : par exemple, deux individus qui travaillent dans la même entreprise sont considérés comme redondants car faisant partie du même cercle.

Il n'existe pas d'expression formelle permettant d'identifier les trous structuraux dans un réseau si celui-ci en contient. Au lieu de ça, Burt fournit des indicateurs empiriques (qui peuvent quant à eux

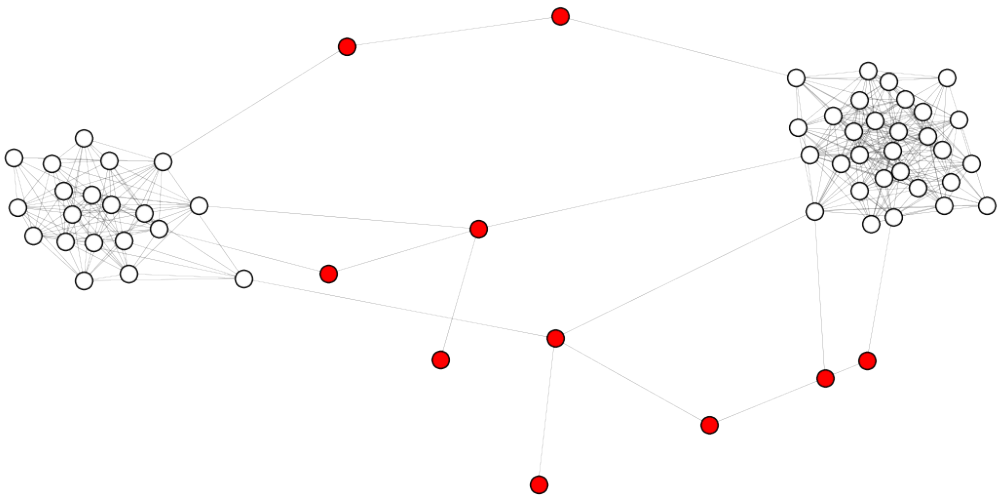


Figure 1.13: Exemple d'un réseau comprenant deux clusters (dont les nœuds sont représentés en blanc) avec une forte densité de liens internes dans chacun d'eux, et un ensemble de nœuds faiblement reliés entre eux comme au reste du réseau, faisant office de partie non dense, représentée en rouge sur la figure.

être formellement exprimés), notamment en distinguant différents types de redondance. La redondance par cohésion, si le voisinage d'un nœud est densément connecté (comme c'est le cas par exemple pour une clique) alors on parle de redondance par cohésion, si en revanche les voisins ne sont pas liés entre eux mais qu'ils sont liés à des individus du même groupe, on parle alors de redondance par équivalence structurelle.

Ainsi Burt estime que dans les deux cas illustrés sur la figure 1.14, les nœuds A, B et C sont redondants (par cohésion dans (a) et par équivalence structurelle dans (b)) car menant dans un cas comme dans l'autre à la même source d'informations. Ensuite il introduit une autre notion importante pour sa théorie des trous structuraux, celle de l'efficacité : en maximisant les contacts non redondants, on maximise les trous structuraux dans son propre voisinage.

Sur la figure 1.15 nous pouvons voir d'un côté (fig. 1.15a) un réseau ego-centré non efficace, car le nœud central (en noir) crée des liens avec plusieurs nœuds du même cluster, alors que de l'autre côté (fig. 1.15b), l'efficacité est maximale car le nœud central ne partage qu'un seul lien avec un nœud de chaque cluster, ayant accès aux mêmes sources d'informations, tout en maximisant les trous structuraux.

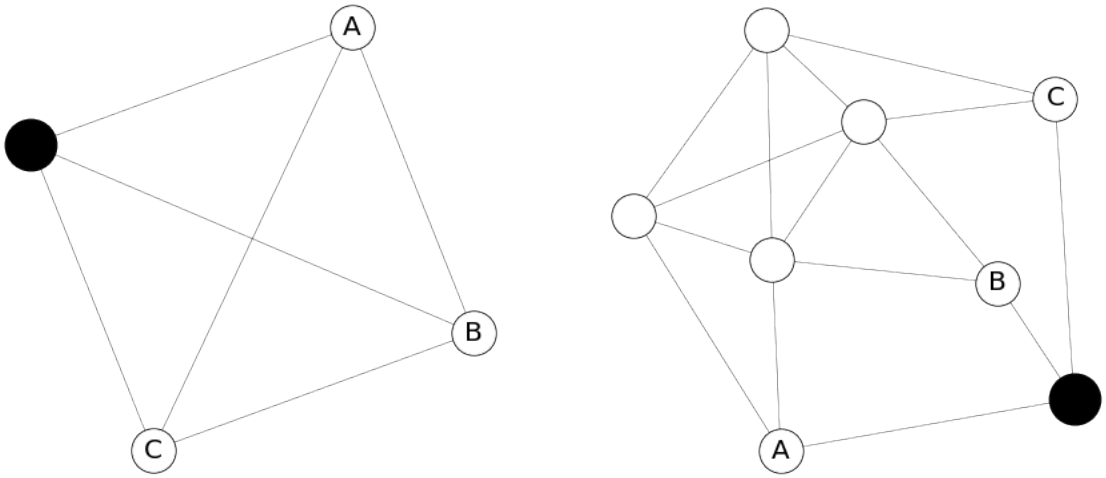
On constate que l'étude de Burt n'est pas dédiée à l'analyse structurelle des réseaux, mais plutôt orientée vers la stratégie que doit adopter un agent dans un milieu compétitif, afin d'y occuper une position avantageuse, en ce sens où l'accent est explicitement mis sur la façon de gérer ses contacts, et non à l'élaboration d'un outil analytique permettant d'identifier systématiquement les trous structuraux. On y retrouve cependant certaines des idées fondamentales en rapport avec notre problématique, notamment dans le fait de mettre en évidence l'existence d'un ensemble de nœuds jouant un rôle important dans le réseau, sans pour autant bénéficier d'une grande densité en connectivité.

1.5.2 Ponts structuraux

On retrouve dans la littérature diverses références [23, 22, 123, 32, 51] qui sont liées à notre problématique, bien que là encore, aucune des ces études ne soit exclusivement dédiée aux parties non denses, mais plutôt à des problématiques spécifiques, dans lesquelles les nœuds des parties non denses (qui portent des noms différents d'une référence à l'autre) jouent un rôle important.

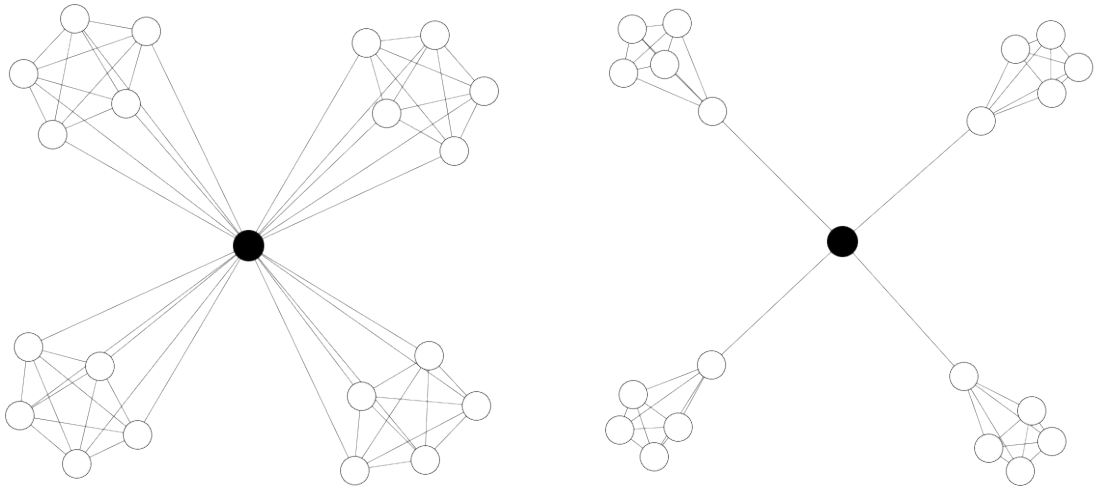
Par exemple dans [55] l'auteur définit par point d'articulation, aussi appelé pont, le nœud dont la suppression augmente le nombre de composantes connexes dans le graphe, comme le montre l'exemple de la figure 1.16.

On remarquera que l'exemple du nœud central de la figure 1.15 vérifie aussi cette définition, ce qui montre le lien qui existe entre les points d'articulation et les trous structuraux, mais les trous structuraux de la figure 1.15 ne sont qu'un cas particulier des points d'articulation, car dans la définition



(a) (b)

Figure 1.14: Le nœud représenté ici en noir est sur (a) redondant par cohésion aux nœuds A, B et C car tout le monde est connecté à tout le monde, et sur (b) redondant à A, B et C par équivalence structurelle, car ses trois voisins donnent accès au même ensemble de nœuds par la suite.



(a) (b)

Figure 1.15: (a) Réseau ego-centré non efficace, car le voisinage du nœud central est localement redondant par cohésion. (b) Cas analogue avec le maximum d'efficacité, car le nombre de voisins est minimal (4 au lieu de 20), alors que la source d'information accessible demeure la même.

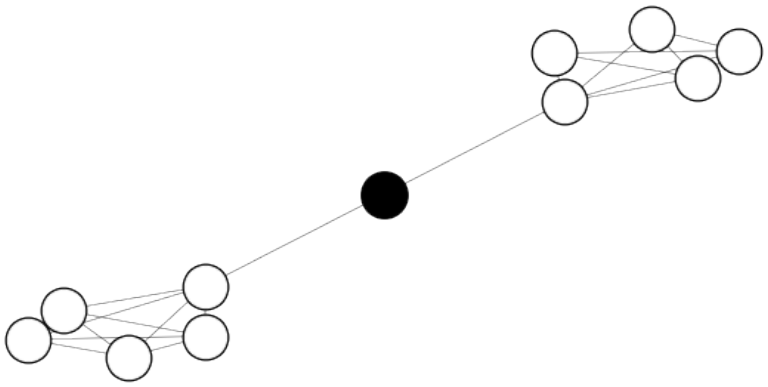


Figure 1.16: Le nœud central est un exemple de pont, la transition de l'un à l'autre des deux clusters passe nécessairement par lui, et sa suppression résulterait en la division du réseau en deux composantes connexes.

de ces derniers on n'impose aucune condition sur la structure des composantes connexes obtenues après la suppression, qui peuvent être composées de clusters, d'un nœud unique, ou toute autre structure imaginable. Les auteurs de [55] proposent un algorithme permettant la détection des points d'articulation, dans le cas des réseaux orientés, ce qui change relativement les choses car la définition d'un point d'articulation devient reliée à la notion de composante fortement connexe et non à la connexité simple.

Dans le travail de [52, 51] on retrouve aussi quelques points de ressemblance avec les parties non denses. L'auteur y identifie un ensemble de nœuds qu'il appelle "goulots d'étranglement" structurels ("structural bottlenecks" en anglais) comme un petit ensemble de nœuds S tel que G/S est constitué d'au moins deux grandes composantes connexes. Le reste de l'étude se concentre sur une classification des réseaux basée sur les propriétés spectrales, dont l'une des classes est celle des réseaux homogènes, sans "bottlenecks". On retrouve aussi cette appellation dans les études des flux d'information sur les réseaux [19], mais elle désigne de façon plus générale les endroits du réseau où le flux est le plus saturé.

1.6 La densité comme cadre unificateur

Nous venons de voir qu'il existe des points de similitudes entre plusieurs recherches menées dans des secteurs relativement éloignés (sociologie, analyse structurelle de réseaux, optimisation de flux) et que l'on relie qualitativement par la propriété de non densité, d'où l'intérêt d'une telle approche.

Nous pouvons constater en parcourant la littérature relative à l'analyse de la topologie des réseaux, que la notion de densité y est relativement peu développée, d'autant plus que celle-ci est généralement utilisée comme indicateur à l'échelle macroscopique, à travers le rapport entre le nombre de liens du réseau et le nombre de paires de nœuds qu'il contient. L'équivalent de cette mesure à l'échelle des nœuds (échelle microscopique) est la centralité du degré, dont la valeur moyenne calculée sur l'ensemble des nœuds est effectivement égale au rapport que l'on vient d'évoquer.
Nous pouvons cependant nous demander si la définition de la densité à l'échelle des nœuds, telle qu'on la connaît aujourd'hui, prend en compte tous les aspects relatifs à une telle notion, et dans le cas contraire, quelles seraient les propriétés principales que devrait couvrir une mesure microscopique de la densité ?

Cette question constitue le point de départ de cette thèse, dont l'objectif est de décrire les différentes étapes d'une étude quantitative de la notion de densité, qui puisse être compatible avec les différentes échelles décrites plus haut, dans le but de créer un cadre permettant d'unifier les approches telles que celles énoncées dans la section 1.4.

Ensuite, nous nous tournerons vers la description détaillée de notre algorithme, qui prend en entrée un graphe non pondéré et non dirigé, et qui à l'aide de cette nouvelle définition de la densité fournit en sortie un découpage du graphe en deux classes : la partie dense et la partie non dense. Chacune de ces deux classes est à son tour composée de plusieurs sous-ensembles de nœuds, répondant à des critères topologiques que l'on décrira dans un chapitre dédié. Pour finir, nous étudierons chacune de ces deux

parties à travers plusieurs applications, en portant l'accent sur les nœuds de la partie non dense qui sont susceptibles de jouer un rôle important dans la structure des réseaux correspondants, et qui n'ont jusqu'ici bénéficié que de peu d'attention.

Pour conclure cet état de l'art, nous positionnons les travaux de cette thèse parmi l'ensemble des outils méthodologiques dédiés à l'analyse topologique des réseaux qui, à travers un découpage hiérarchique basé sur notre propre définition de la densité, fournit un point de vue original et une meilleure compréhension de la structure des réseaux.

Chapitre 2

Mesure stochastique : la densité spatiale

Table des matières

2.1 Introduction

Lors de la phase de réflexion consacrée à la mise au point d'une nouvelle mesure de la densité, nous nous sommes heurtés à plusieurs questions élémentaires, notamment en ce qui concerne les critères qui doivent être pris en compte par une telle mesure. Le choix de ces critères n'étant jamais universel, nous avons d'abord cherché à contourner le problème, en nous tournant vers une approche assez classique en analyse des données, qui est celle de plonger le graphe (ou les données étudiées en général) dans un espace euclidien à faible dimension [30, 36, 63], afin de pouvoir se rapporter à la mesure spatiale de la densité (au sens euclidien du terme), sous l'hypothèse que la densité dans le graphe, quelle que soit sa forme la plus adaptée, soit une propriété émergente, si la fonction décrivant le plongement en question le permet.
Ainsi nous avons étudié les différentes façons possibles de plonger un graphe dans un espace euclidien, et nous avons finalement opté pour les algorithmes de positionnement à l'aide d'une fonction qui modélise les liens comme une force attractive entre les nœuds, et l'absence de liens comme une force répulsive. Cette approche a l'avantage de bien fonctionner à faibles dimensions, ce qui est plutôt encourageant car il est préférable d'éviter les espaces à grandes dimensions, dans lesquels la notion de densité devient de moins en moins évidente à mesure que la dimension augmente.

Ce chapitre occupe une place particulière au sein du manuscrit, car il fait appel à des concepts qui n'ont pas été décrits dans l'état de l'art, en raison de la différence entre les thèmes abordés. De plus, ses résultats ne sont pas nécessaires aux développements qui suivent ce chapitre, bien qu'ils aient fortement influencé les réflexions ultérieures. Il constitue donc une partie indépendante et autosuffisante.

Nous déroulons le chapitre en deux phases, la première est consacrée à l'introduction et la description rapide des concepts et outils qui nous sont nécessaires pour la suite, et la seconde portera sur le développement et l'expérimentation de la mesure de densité qu'on y développe.
Ainsi nous commençons par introduire quelques concepts élémentaires de l'analyse topologique des données (aussi connu sous l'abréviation TDA), et les principaux algorithmes de clustering basés sur la densité. Ensuite nous décrivons le cheminement qui mène à notre mesure de la densité, que nous évaluons à travers ses résultats sur les algorithmes de clustering précédemment introduits. Nous décrivons finalement l'algorithme de plongement du graphe dans un espace euclidien et finissons sur un examen statistique des résultats, basé sur l'analyse topologique des données.

2.2 L'analyse topologique des données (TDA)

Parmi le grand nombre de méthodes qui sont aujourd'hui employées pour analyser les données, beaucoup se concentrent sur un objectif particulier, défini au préalable. Nous pouvons par exemple citer le cas de la régression linéaire, dont le but est de fournir une représentation des données sous la forme d'un modèle linéaire, dans lequel chaque valeur obtenue en sortie est une combinaison linéaire des valeurs en entrée. On peut bien entendu tomber sur des cas où cette représentation linéaire n'est pas valable, auquel cas il faut avoir recours à d'autres méthodes, parmi lesquelles les méthodes de clustering (qui seront mieux détaillées par la suite) et dont le but est de regrouper les données par paquets, répondant à des critères de similarité prédéfinis. On peut remarquer que dans les deux exemples qui viennent d'être cités, l'information à extraire en sortie est prédéfinie et réduit donc le champ de vision à celui qu'elle propose. C'est sur cet aspect que l'analyse topologique des données représente une approche moins contraignante, car comme nous le verrons, elle s'intéresse en particulier à la forme d'un nuage de points, sans aucun *a priori* [1].

La TDA est représentée par un ensemble de méthodes qui recherchent une forme (au sens topologique) dans les données [29, 28], portée par les récentes avancées en topologie informatique [24, 126], qui ont rendu possible le calcul d'invariants topologiques sur des données, ces dernières ne répondant pas nécessairement à la condition de continuité, et pouvant être représentées par un nuage fini (et potentiellement bruité) de points. Mais alors, on peut se demander par quel moyen peut-on rendre intelligible un jeu de données dont on connaît la topologie, et c'est à cette question en particulier que la TDA tente d'apporter des réponses, en se basant sur une phrase bien connue de ses utilisateurs : les données ont une forme, et la forme une signification. Il est cependant important de souligner que la forme dont il est question n'est pas le produit de l'approche par la TDA, mais provient plutôt d'une

[1] La transformation qui permet de plonger un jeu de données dans un espace métrique ne fait cependant pas partie du champs d'opération de la TDA.

transformation effectuée au préalable, permettant de plonger les données dans un espace de dimension d, souvent à l'aide de mesures de similarités, et que c'est sur cette représentation que la TDA est appliquée. Pour le moment nous ne nous préoccupons pas de ces transformations, et on se contente de fournir une description introductive de certains éléments (ceux que nous utilisons dans ce chapitre) de l'analyse topologique des données.

Imaginons donc que l'on dispose d'un jeu de données représenté par un nuage de N points, dans un espace de dimension d, avec N et d deux entiers, et dont on voudrait étudier la forme. Le premier point important à souligner est que l'objet étudié est constitué d'un ensemble fini de points, sur lequel il est encore difficile d'identifier une quelconque structure. La première étape consiste donc à donner à ce nuage de points une représentation plus pratique, sur laquelle il est possible d'effectuer des opérations mathématiques, afin d'en tirer une information pertinente, et c'est par ce point que nous proposons de commencer notre description de la TDA.

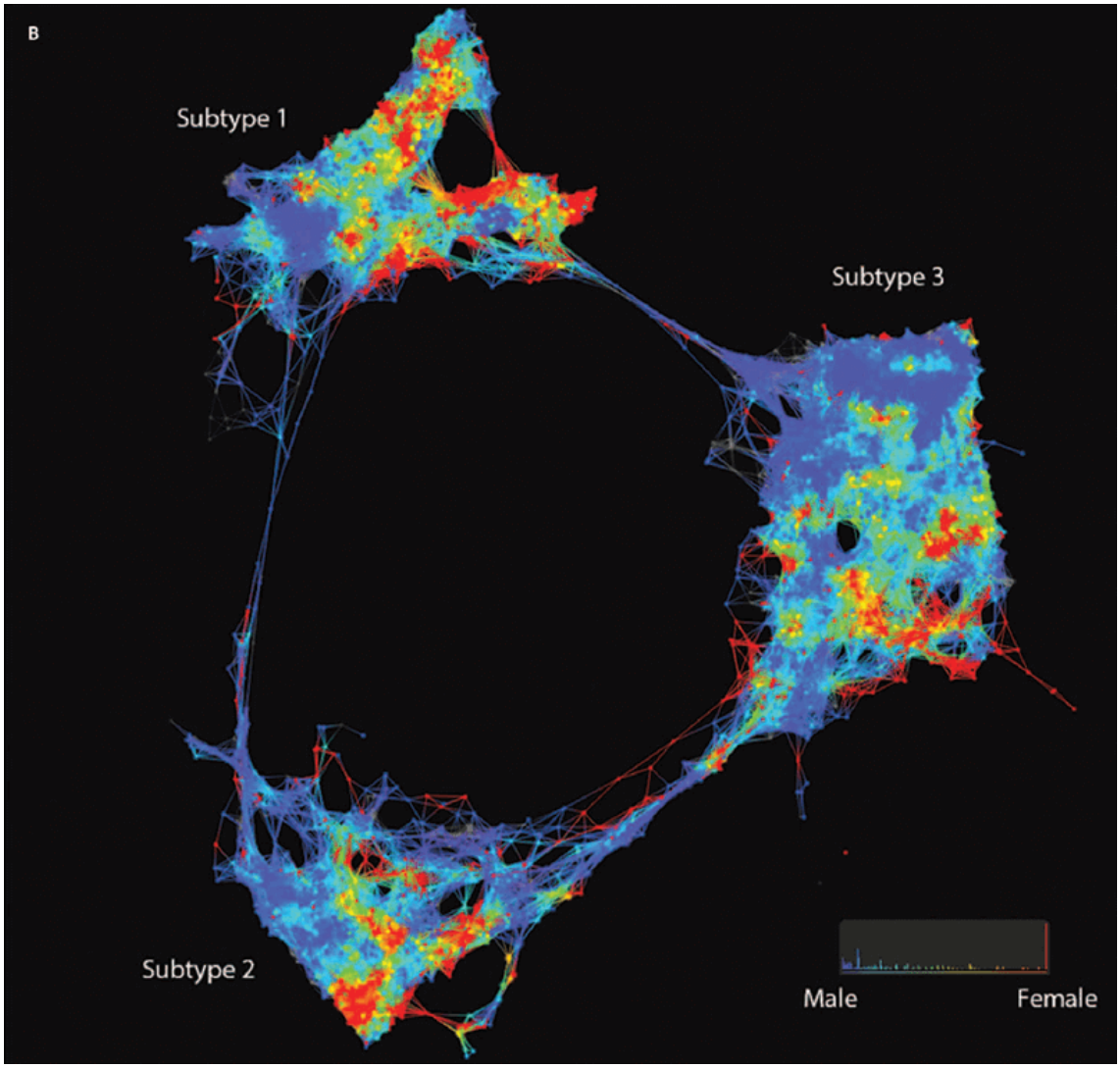


Figure 2.1: Réseau patient-patient pour 2551 malades atteints de diabète. Chaque nœud peut représenter un seul ou plusieurs patients, ayant une similarité significative dans leurs caractéristiques cliniques. Un lien relie les groupes qui ont des patients en commun. Ce réseau a été issu de l'algorithme du Mapper, très employé en TDA, à partir des données cliniques des patients. Figure tirée de [80]. La figure représentée ici est le résultat final d'un certain nombre de transformations effectuées sur les données de départ, le développement dans sa totalité nécessite d'aller plus loin dans les détails techniques de la TDA, et c'est ce que nous proposons de faire dans ce qui va suivre.

2.2.1 Du nuage de points au complexe simplicial

Parmi tous les schémas possibles que propose la TDA, il y en a un qui se distingue en particulier par son efficacité à résumer l'information contenue dans une grande masse de données, tout en explorant plusieurs échelles successives : Le diagramme de persistance. Mais avant d'en arriver à ce point, il faut passer par une représentation intermédiaire, qu'on obtient après le filtrage du nuage de points en entrée, pour en faire un complexe simplicial. Nous commençons donc par introduire les différentes notions liées aux complexes simpliciaux, avant de se concentrer sur la méthode principale de l'homologie persistante, et du diagramme de persistance.

Simplexe

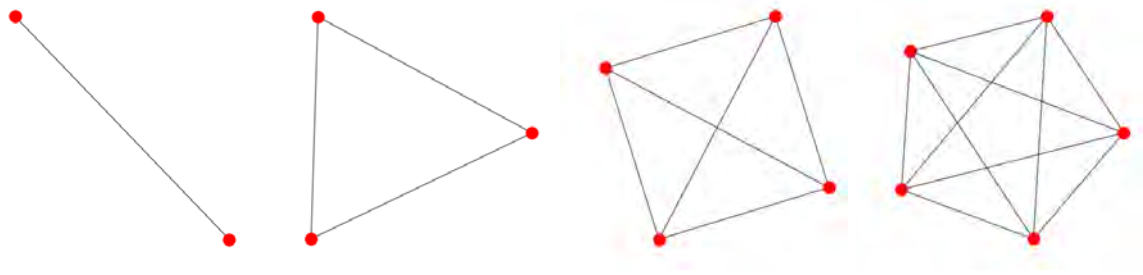
Un k-simplexe ou simplexe de dimension k dans $\mathbf{R}^n$ est défini à l'aide de $k+1$ points linéairement indépendants dans $\mathbf{R}^n$. Soit $\{v_0, ..., v_k\}$ cet ensemble. Le k-simplexe noté $[v_0, ..., v_k]$ est l'espace topologique donné par l'ensemble

$$\{\sum_{i=0}^{k} t_i v_i | \sum_{i=0}^{k} t_i = 1; t_i \geq 0\}$$

avec la topologie induite par la métrique euclidienne [76]. Les nombres t_i sont les coordonnées du point $x = \sum_i t_i v_i \in [v_0, ..., v_k]$. Les 0-simplexe sont les points v_i. Un 1-simplexe est un ensemble de la forme

$$\{t_0 v_0 + t_1 v_1 | t_1 + t_2 = 1; t_{0,1} \geq 0\} = \{t_0 v_0 + (1 - t_0) v_1\}$$

qui est un segment ou un bord avec des points d'extrémité v_0 et v_1. De la même façon, le 2-simplexe est un triangle avec les sommets v_0, v_1 et v_2 en plus des trois arêtes $v_0 + v_1$, $v_1 + v_2$ et $v_0 + v_2$.



(a) le 1-simplex (b) le 2-simplex (c) le 3-simplex (d) le 4-simplex

Figure 2.2: Exemples de plusieurs k-simplexe $k \in \{1, 2, 3, 4\}$, chacun d'entre eux étant représenté (dans l'espace à deux dimensions) sur cette figure par le graphe complet de $k-1$ nœuds.

Complexe simplicial

Soit X un ensemble fini. Un complexe simplicial K sur X est un ensemble de sous-ensembles de X de telle sorte que:

- $\{x\} \in K$ pour tout $x \in X$
- Si $\sigma \in K$ et $\tau \subset \sigma$ alors $\tau \in K$

Un élément $\sigma \in K$ est appelé un simplexe de dimension $|\sigma| - 1$ où $|\sigma|$ indique le nombre d'éléments dans l'ensemble σ . Nous utilisons le symbole K_i pour désigner l'ensemble des simplexes dans K de dimension i, par exemple $K_0 = X$.

Choisissons un ordre sur l'ensemble X. Pour n'importe quel n-simplexe $\sigma = \{x_0 < x_1 < ... < x_n\} \in K_n$ et $0 \leq i \leq n$, on note $d_i\sigma$ le $(n-1)$-simplexe dans K obtenu à partir de σ en supprimant x_i. Un complexe simplicial K sur un ensemble fini X peut être réalisé comme un espace topologique comme suit : choisissons tout d'abord un ensemble de points linéairement indépendants $\{v_x\}_{x \in X}$ dans $\mathbf{R}^n$, alors nous définissons la réalisation de K comme l'espace topologique donné par le sous-ensemble du simplexe $[v_x]_{x \in X}$ composé des points $\sum_{x \in X} t_x v_x$ tels que $\{x \in X | t_x \neq 0\}$ est un simplexe dans K.

2.2.2 Filtrage des données

Les points dans un espace métrique peuvent être considérés comme des 0-simplexe, mais on peut aussi construire des complexes simpliciaux de rangs plus élevés à partir de ces points, à travers des méthodes de filtrage dont nous présentons ici deux parmi les plus utilisées : celle de Vietoris-Rips et celle de Čech [120].

Complexe de Vietoris-Rips

Soit $\epsilon > 0$ un nombre réel. On définit $VR(X;\epsilon)$ comme étant le complexe simplicial sur l'ensemble X composé des sous-ensembles $\sigma \subset X$ où $d(x,y) < \epsilon$ pour tout $x, y \in \sigma$.
Notez que $VR(X;\epsilon) \subseteq VR(X;\epsilon')$ pour $\epsilon \leq \epsilon'$.

En pratique pour obtenir un complexe de Vietoris-Rips à partir de nos données, il suffit de tracer une boule de rayon ϵ dans l'espace euclidien de dimension d dans lequel elles sont plongées, et ce autour de chaque point. On relie ensuite par une arête chaque paire de points dont les boules se chevauchent. Un n-simplexe est obtenu si il existe un sous-ensemble σ de $n+1$ points dans X tel que chaque deux de points dans σ sont à distance inférieure ou égale à ϵ. Un exemple d'un tel complexe est montré en bas à droite de la figure 2.3, les différents n-simplexes y sont représentés par différentes couleurs.

Complexe de Čech

Soit $\epsilon > 0$ un nombre réel. On définit $C(X;\epsilon)$ comme étant un complexe simplicial sur l'ensemble X composé des sous-ensembles $\sigma \subset X$ pour lesquels il existe un $y \in X$ tel que $d(x,y) < \epsilon$ pour tout $x \in \sigma$. Là encore, on retrouve la propriété $C(X;\epsilon) \subseteq C(X;\epsilon')$ si $\epsilon \leq \epsilon'$. Contrairement aux complexes de Vietoris-Rips, il ne suffit plus que les points d'un sous-ensemble σ soient deux à deux à distance inférieure ou égale à ϵ, il faut aussi qu'il y ait au moins un point à distance inférieure ou égale à ϵ de tous les points de σ, en pratique cela se traduit pour un $n-$simplexe par l'existence d'au moins un point où se chevauchent toutes les boules centrées autour des $n+1$ points de σ.
Un exemple d'un tel complexe est montré en bas à gauche de la figure 2.3, les différents n-simplexes y sont représentés par différentes couleurs. La différence que l'on vient de décrire peut être vérifiée sur le triangle en bas à gauche des deux complexes représentés sur la figure 2.3. Celui-ci est considéré comme un $2-$simplexe dans le complexe de Vietoris-Rips mais pas dans celui de Čech.

À noter que les deux complexes définis précédemment sont étroitement liés, puisque le complexe de Čech est un sous-complexe du complexe de Vietoris-Rips, bien que le complexe de Čech soit plus coûteux en calcul que le complexe Vietoris-Rips, car il nécessite de vérifier les intersections d'ordre supérieur des boules dans le complexe.

Nous pouvons à présent analyser un nuage de points par l'intermédiaire de ces complexes, à l'aide d'outils mathématiques que nous décrirons dans la section suivante. Pour cela l'approche la plus répandue consiste à construire plusieurs complexes d'échelles différentes (différentes valeurs de ϵ), et d'identifier les caractéristiques les plus résilientes aux changements d'échelle. Nous donnons dans les deux prochaines parties une vue d'ensemble sur les caractéristiques qu'on peut extraire à partir des complexes simpliciaux, ainsi qu'une description de la méthode qui évalue leur persistance.

2.2.3 Opérateur bord, groupes d'homologie, et nombres de Betti

Un concept clé dans la définition de l'homologie sur les complexes simpliciaux est la notion d'orientation d'un simplexe. Par définition, l'orientation d'un k-simplexe est donnée par un ordre de sommets, qu'on note $(v_0, ..., v_k)$, avec la règle que deux ordres définissent la même orientation si et seulement si ils diffèrent d'une permutation paire. Ainsi, chaque simplexe a exactement deux orientations, et changer l'ordre de deux sommets change une orientation vers l'orientation opposée. Par exemple, choisir l'orientation d'un 1-simplexe revient à choisir l'une des deux directions possibles, et choisir l'orientation d'un 2-simplexe revient à choisir la convention du sens antihoraire.

Soit S un complexe simplicial. On appelle k-chaîne simpliciale la somme :

$$\sum_{i=1}^{N} n_i \sigma_i \tag{2.1}$$

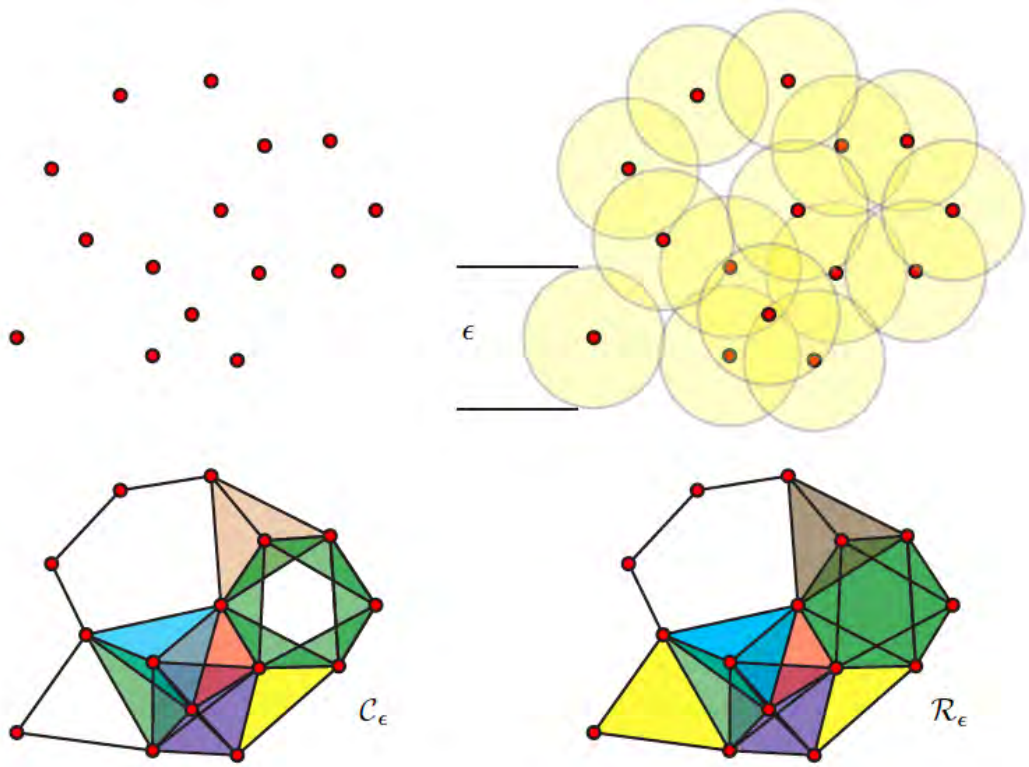


Figure 2.3: Nous avons en haut à gauche un ensemble fixe de points, on construit à partir de cet ensemble un complexe de Čech (en bas à gauche), ainsi qu'un complexe de Vietoris-Rips (en bas à droite), basés sur un paramètre de proximité ϵ (en haut à droite). [60]

où chaque n_i est un entier et σ_i est un k-simplexe orienté . Dans cette définition, chaque simplexe orienté est égal au négatif du simplexe ayant l'orientation opposée.
Par exemple $(v_i, v_j) = -(v_j, v_i)$ [2]. Le groupe de k-chaînes sur S est noté C_k. Il s'agit d'un groupe abélien libre qui a une correspondance élément par élément avec l'ensemble des k-simplexes dans S. On a ainsi un isomorphisme entre l'espace des k-chaînes de simplexes et l'espace des k-chaînes de simplexes orientés modulo les relations

$$[\phi(v_0), \phi(v_1), ..., \phi(v_k)] = (-1)^{signe(\phi)}[v_0, v_1, ..., v_k] \tag{2.2}$$

où ϕ désigne une permutation quelconque de $v_0, ..., v_k$.

Soit $\sigma = [v_0, ..., v_k]$ un k-simplexe orienté, considéré comme un élément de la base de C_k. L'opérateur de bord

$$\partial_k : C_k \to C_{k-1} \tag{2.3}$$

est l'homomorphisme défini par :

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, ..., \widehat{v_i}, ..., v_k] \tag{2.4}$$

avec $[v_0, ..., \widehat{v_i}, ..., v_k]$ le simplexe orienté représentant la $i-ème$ face de σ qu'on obtient en supprimant le $i-ème$ 0-simplexe (un nœud, si on s'en tient à l'analogie évoquée plus haut). Notons que l'opérateur bord dépend de l'orientation par défaut de chaque simplexe. Une propriété fondamentale de cet opérateur est que le bord d'un bord est nul, ce qui se traduit par $\partial_k(\partial_{k+1}(C_{k+1})) = 0$.

Notons L_k le noyau de l'opérateur bord

$$L_k = \mathrm{Ker}(\partial_k) \tag{2.5}$$

[2] Nous référons le lecteur vers [127] pour des définitions plus complètes

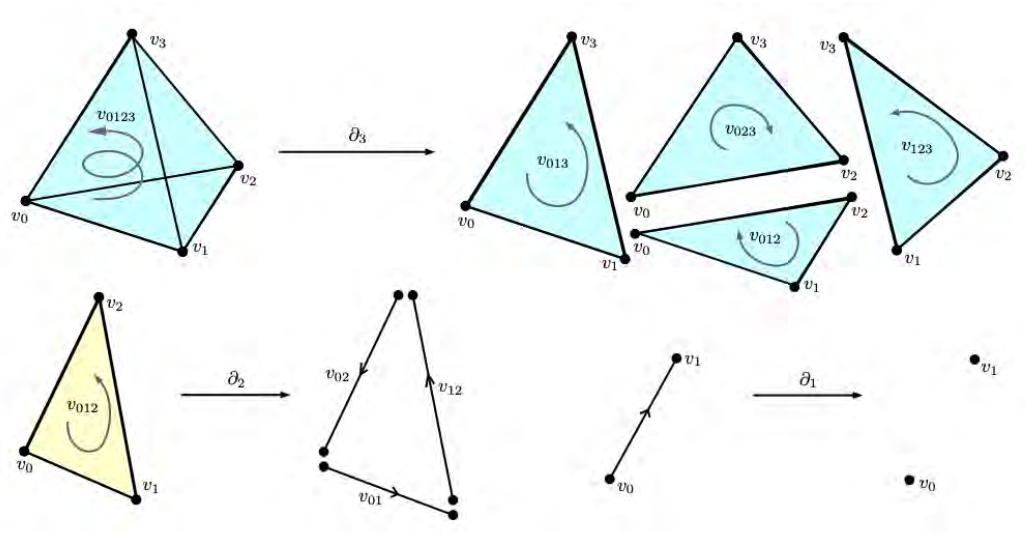


Figure 2.4: Passage d'un 4-simplexe à deux 0-simplexe par applications successives de l'opérateur bord. Le tétraèdre a pour bord 4 triangles, qui ont chacun pour bords trois arêtes, puis deux points. L'orientation au sein de chaque simplexe est représentée par une flèche. figure tirée de [108]

composé des éléments qu'on appellera cycles dans C_k, et notons B_k l'image de l'opérateur ∂_{k+1} :

$$B_k = \mathrm{Im}(\partial_{k+1}) \tag{2.6}$$

où B_k est constitué de bords. Il est facile de vérifier que $B_k \subset L_k$. On définit alors le k-ème groupe d'homologie H_k du complexe simplicial S comme étant le groupe abélien quotient

$$H_k = L_k / B_k \tag{2.7}$$

ce qui veut dire que si le résultat de la différence entre deux éléments de L_k est un élément de B_k, alors ces deux éléments font partie de la même classe d'équivalence

$$[l] = \{m \in L_k | m + l \in B_k\}. \tag{2.8}$$

Il s'ensuit que le groupe d'homologie $H_k(S)$ est non nul quand il y a exactement k-cycles en S qui ne sont pas des bords. Dans un sens, cela signifie qu'il y a des trous k-dimensionnels dans le complexe. Le rang du k-ème groupe d'homologie, le nombre défini par :

$$\beta_k = \mathrm{rang}(H_k(S)) \tag{2.9}$$

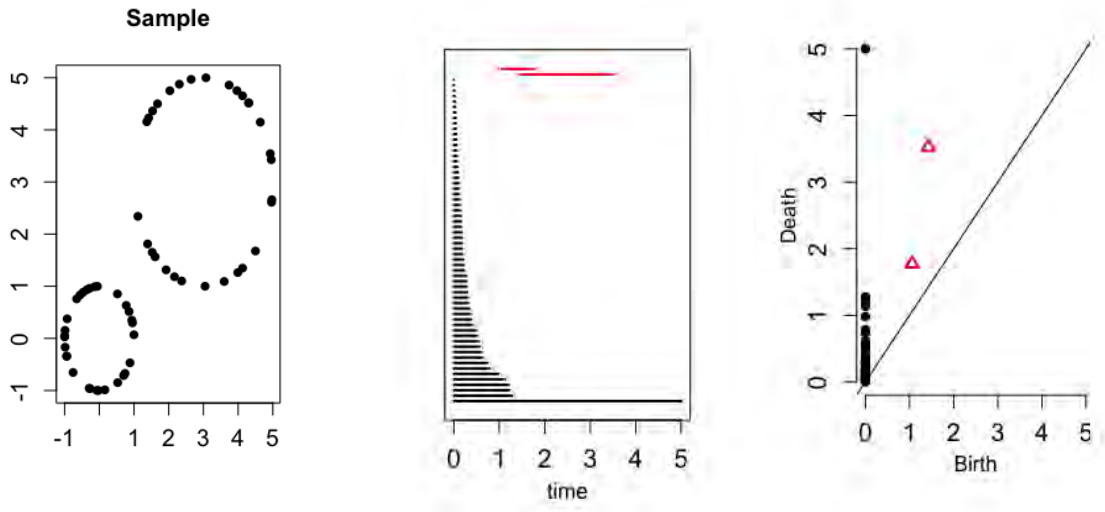
est appelé le k-ème nombre de Betti de S. Il donne accès au nombre de cavités k-dimensionnelles dans S.

2.2.4 Homologie et persistance

On a maintenant plusieurs outils pour résumer l'information contenue dans un nuage de points d'un espace métrique (ici on se limitera au cas des espaces euclidiens), à un ensemble de nombres entiers qui reflètent leur topologie, en passant par la construction d'un complexe simplicial moyennant le choix d'un paramètre d'échelle ϵ fixé.

Pour simplifier on établit une analogie entre le complexe simplicial et un graphe : pour chaque valeur de ϵ, on construit un complexe simplicial qui est éventuellement scindé en plusieurs composantes connexes. Le nombre de ces composantes correspond au nombre β_0 de Betti.
Chacune de ces composantes est ensuite analysée séparément à la recherche de cavités de dimension supérieure. Si pour une composante donnée, il existe un certain nombre de cycles indépendants de dimension 1, le nombre total de cycles ainsi détectés sur l'ensemble des composantes constitue le nombre

β_1 de Betti, et ainsi de suite pour les dimensions supérieures.
Ainsi, dans le cas d'un nuage de points tridimensionnels, si le choix du paramètre d'échelle aboutit à un complexe simplicial ayant pour nombres de Betti ($\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$) on peut en déduire que ce complexe est homomorphe à une sphère. On peut ensuite se poser la question suivante : est-ce que cette structure qui est apparue pour une valeur arbitraire de ϵ, est significative par rapport au reste du complexe, ou est-ce juste un effet relatif à l'échelle de construction choisie ? Pour éviter de se confronter à ce genre de questions, auxquelles il est souvent difficile d'apporter une réponse claire, il existe une approche plus générale. Cette approche consiste à reproduire l'analyse qu'on vient de décrire à différentes échelles. Les structures les plus significatives ne sont pas toutes celles qui sont apparues à une échelle donnée, mais plutôt celles qui sont les plus résilientes au changement d'échelle. En d'autres mots, on construit une série de complexes répondant à des paramètres d'échelles croissants (on part d'une petite valeur de ϵ qui génère un complexe simplicial composé de 0-simplexes pour finir avec un N-simplexe obtenu pour une grande valeur de ϵ, N étant le nombre de points dans le nuage). On calcule les nombres de Betti pour chaque complexe, et à chaque échelle. On s'intéresse en particulier aux cycles qu'on continue à observer malgré l'augmentation du paramètre d'échelle (cycles persistants), qui sont porteurs d'informations pertinentes, mettant en évidence les formes les plus importantes dans le complexe.



(a) Un nuage de points échantillonné sur deux cercles

(b) Le diagramme de persistance

(c) Le diagramme apparition/disparition

Figure 2.5: Du nuage de points (a) au diagramme apparition/disparition (c), les différentes étapes de la TDA. Sur (b) le diagramme se lit en effectuant une coupe verticale à une échelle donnée (ici notée "time" au lieu de ϵ). Le nombre d'intersections entre cette droite verticale et les barres donne accès aux nombres de Betti du complexe. Les barres noires représentent le nombre β_0 de Betti et les barres rouges le nombre β_1. Par exemple pour time= 0 on dispose d'un grand nombre d'intersections (60 en tout) avec les barres noires, ce qui se traduit par autant de composantes connexes dans le complexe, chacune étant un point du complexe d'échelle nulle. Pour time= 1.5 on a $\beta_0 = 1$ et $\beta_1 = 2$, ce qui indique une composante connexe qui contient deux cavités distinctes. La figure (c) donne pour chacune des barres les coordonnées qui correspondent à son instant d'apparition et de disparition. Ces schémas sont générés à l'aide du package R-TDA [53]

La figure 2.5 montre l'exemple d'un diagramme de persistance sur un nuage de 60 points, qui a été échantillonné à partir de deux cercles (30 points pour chaque cercle) de rayons différents et de centres distincts (*cf.* fig. 2.5a). Nous pouvons voir sur la 2.5b la représentation en code barre de l'homologie persistante, avec sur l'axe des abscisses l'échelle de filtrage des données (ici le filtrage se fait via un complexe de Vietoris-Rips). En ordonnées nous pouvons observer plusieurs barres, chacune correspond à un intervalle du type $\epsilon \in [t_1, t_2]$, qui est celui durant lequel l'échelle de filtrage aboutit à complexe

simplicial dans lequel la structure correspondante continue d'exister. En pratique, chacune de ces barres représente un cycle de dimension k, ou de manière équivalente le bord d'une cavité de dimension $k+1$. Il est important de remarquer que dans cet exemple, nous nous limitons aux cas $k \in \{0, 1\}$ avec la couleur noire pour $k = 0$ et la couleur rouge pour $k = 1$ (il n'aurait d'ailleurs pas été nécessaire d'aller au delà, car le nuage de points étudié est échantillonné sur le plan, on n'y trouvera par conséquent aucune cavité de dimension 3). Les cavités recherchées correspondent donc au nombre de composantes connexes pour le cas $k = 0$, représentées sur la figure par des barres noires, ainsi que les cavités bidimensionnelles (qui représentent ici l'intérieur des cercles) que l'on désigne par des barres rouges sur cette figure. Nous pouvons donc observer que les barres noires indiquent que le complexe part de plusieurs composantes connexes (autant de composantes que de points), que celles-ci fusionnent à mesure que le paramètre ϵ grandit, jusqu'à ne plus former qu'une seule composante connexe à la fin. En ce qui concerne les barres rouges, on peut remarquer l'apparition de deux cavités bidimensionnelles distinctes, à des échelles non nulles (qui reflètent la taille des cavités correspondantes), et qui finissent elles aussi par fusionner une fois le paramètre ϵ assez grand, avant de disparaître complètement quand l'échelle dépasse une certaine valeur, d'environ 3.5 sur la figure 2.5b.

La figure 2.5c est une autre représentation de l'homologie persistante. Cette fois-ci on associe un point à chaque barre, en traçant l'échelle à laquelle la structure correspondante disparaît, en fonction de l'échelle à laquelle elle apparaît. Dans ce diagramme, appelé diagramme apparition/disparition (aussi appelé parfois diagramme mort/naissance), plus un point est situé loin au dessus de la droite $y = x$, plus la structure correspondante est persistante. Nous pouvons facilement voir dépasser les deux points en rouge qui correspondent à la persistance de deux cavités bi-dimensionnelles, ainsi qu'un point noir pour la composante connexe correspondante. Il est donc possible de déduire, juste en observant ce diagramme, que le nuage de points duquel il est issu est composé deux cavités bi-dimensionnelles, qui représentent l'intérieur des deux cercles de la figure 2.5a.

2.3 Aperçu des algorithmes de clustering basés sur la densité

Il existe deux sortes de clustering : la classification hiérarchique et le partitionnement. Le premier fournit une décomposition hiérarchique d'un ensemble de données D, dans laquelle chaque cluster est divisé à son tour en plusieurs sous-clusters et ce jusqu'à obtenir une partition dans laquelle chaque cluster est composé d'une unité. On obtient ainsi une représentation en dendrogramme de cette décomposition, et on choisit de retenir la meilleure partition. Ceci s'obtient en coupant le dendrogramme à l'endroit qui (par exemple) maximise une mesure objectif de cette partition, établie au préalable.

Le deuxième type est celui du partitionnement, qui consiste à construire à partir d'un jeu de données D, k sous-ensembles dont l'union recouvre l'ensemble D de départ. Là encore on peut distinguer deux types d'approches, la première est celle dans laquelle le nombre de clusters k est prédéfini (par exemple le k-means), et la seconde, est celle dans laquelle on ne connaît pas le nombre de clusters au préalable. C'est vers cette seconde approche qu'on va se tourner ici, en présentant certains des algorithmes les plus populaires, ceux basés sur la densité.

2.3.1 DBSCAN

L'algorithme DBSCAN (Density Based Spatial Clustering of Application with Noise) constitue la pierre angulaire de l'édifice qu'est devenue la famille des algorithmes de clustering basés sur la densité. Il est par conséquent celui avec le plus d'imperfections, car il a ouvert le champ à une multitude d'autres algorithmes qui s'en sont inspirés pour en améliorer les résultats. Il est basé sur une idée
fondamentale : pour chaque point p d'un cluster donné, le voisinage de ce point doit au moins contenir un nombre k de points voisins. Ce voisinage désigne l'ensemble de points qui sont à distance inférieure ou égale à d du point évalué, relativement à la mesure de distance choisie au préalable.

Pour plus de précisions nous donnons les définitions présentées dans l'article [50], soit D un nuage de points dans un espace métrique, et d une distance on définit alors:

Le voisinage

Le voisinage d'un point p relativement à ϵ est défini comme l'ensemble de points noté N, contenant tous les points à distance inférieure ou égale à ϵ de p.

$$N_\epsilon(p) = \{q \in D | d(p,q) \leq \epsilon\} \tag{2.10}$$

Directement densité-atteignable

Un point p est directement densité-atteignable depuis q relativement à ϵ et M, si

- $p \in N_\epsilon(q)$
- $|N_\epsilon(q)| \geq M$

À noter que cette propriété n'est pas symétrique, ceci est typiquement le cas pour les points p se situant aux bords des clusters, qui sont directement densité-atteignables depuis des points q qui se situent au coeur, sans que l'inverse ne soit vrai.

Densité-atteignable

Un point p est densité-atteignable depuis q relativement à ϵ et M, s'il existe une série de points $p_1, p_2, ..., p_n$ telle que $p_n = p$ et $p_1 = q$ et pour tout i on dispose de la propriété p_{i+1} est directement densité-atteignable depuis p_i. Cette propriété est transitive, au sens où si r est densité-atteignable depuis q qui est lui-même densité-atteignable depuis p alors r est densité-atteignable depuis p. En revanche, cette propriété étant une extension de la propriété de densité-atteignabilité directe, elle n'est par conséquent pas symétrique.

Densité-Connecté

Un point p est densité-connecté au point q relativement à ϵ et M, s'il existe un point o tel que p et q sont tous les deux densité-atteignables depuis o. Cette propriété est quant à elle symétrique.

On définit finalement un cluster C, relativement à ϵ et à M, comme un sous-ensemble non vide de D qui satisfait les propriétés suivantes :

- $\forall p, q$: si $q \in C$ et p est densité-atteignable depuis q, relativement à ϵ et M, alors $p \in C$
- $\forall p, q \in C$: p est densité-connecté à q, relativement à ϵ et M.

Il s'ensuit que les points considérés comme du bruit, sont tous ceux qui sont dans D sans être dans aucun des clusters calculés.

Lemme 1
Soit p un point de D, et $|N_\epsilon(p)| \geq M$. Alors l'ensemble :
$\{o | o \in D$ et o est densité-atteignable depuis p relativement à ϵ et M $\}$
constitue un cluster relativement à ϵ et M.

Lemme 2
Soit C un cluster relativement à ϵ et M, et P un point quelconque appartenant à C ayant pour propriété $N_\epsilon(p) \geq M$: Alors C est défini par
$\{o | o \in D$ et o est densité-atteignable depuis p relativement à ϵ et M $\}$

De ces deux lemmes on retient qu'un cluster est assimilable à un point p quelconque appartenant à son noyau, *i.e* $|N_\epsilon(p)| \geq M$. Il suffit alors de choisir les bons paramètres ϵ_i et M_i pour chaque cluster C_i, ainsi qu'un point appartenant à son noyau, pour le recouvrir intégralement. En pratique il est difficile d'avoir accès à ces informations *a priori*, et on procède autrement, en définissant un ϵ et un M globaux pour tous les clusters.

La recherche des points densité-atteignables s'effectue par la collecte itérative de points directement densité-atteignables. DBSCAN vérifie le voisinage de chaque point de D. Si le voisinage $N_\epsilon(p)$ du point p contient plus que M points, un nouveau cluster C contenant les objets de $N_\epsilon(p)$ est créé. Ensuite les voisinages de tous les points q dans C qui n'ont pas encore été traités sont évalués. Si $N_\epsilon(q)$ contient plus que M points, les voisins de q qui ne sont pas déjà contenus dans C sont ajoutés au cluster et leur voisinage est évalué lors de l'étape suivante. Cette procédure est répétée jusqu'à ce qu'aucun nouveau point ne puisse être rajouté à C, et un nouveau cluster sera dès lors recherché. Plus de détails sur le choix des paramètres ϵ et M, ainsi que sur l'algorithme DBSCAN sont disponibles sur l'article [50].

En plus de ne pas nécessiter la connaissance du nombre de clusters à retrouver, et d'intégrer la notion de bruit dans ses résultats, DBSCAN peut trouver des clusters aux formes arbitraires. Il peut même trouver un cluster complètement entouré par (mais non connecté à) un cluster différent. Il possède cependant quelques inconvénients, on peut citer par exemple le fait qu'il ne soit pas entièrement déterministe : les points du bord accessibles à partir de plus d'un cluster peuvent faire partie de l'un ou l'autre, selon l'ordre dans lequel les points sont explorés. Mais son principal désavantage réside dans le fait que DBSCAN ne sait pas bien traiter les jeux de données contenant de grandes différences de densités, puisque il n'existe pas de moyen efficace de choisir les bons paramètres M et ϵ pour chacun des clusters. Pour celà il existe une correction de l'algorithme DBSCAN qui ajoute une composante hiérarchique : OPTICS.

2.3.2 OPTICS

Les auteurs dans [8] introduisent deux notions supplémentaires afin de pallier l'incapacité identifiée chez DBSCAN, à retrouver des clusters de densités différentes, la distance d'accessibilité , et la distance noyau (respectivement reachability distance et core-distance en anglais).

Distance noyau

Soit p un point de l'ensemble D de points, dans l'espace métrique muni de la distance d. On appelle $M(p)$ la distance qui sépare le point p de son $M-$ème voisin le plus proche. On définit alors la distance noyau de p, relativement à ϵ et à M, notée $d^{noy}_{\epsilon,M}(p)$ comme :

$$d^{noy}_{\epsilon,M}(p) = \begin{cases} \text{Indéfinie si } |N_\epsilon(p)| \leq M \\ M(p) \text{ sinon} \end{cases}$$

Distance d'accessibilité

La distance d'accessibilité d'un point p en partant d'un point o, relativement à ϵ et M, est définie comme:

$$d^{att}_{\epsilon,M}(o,p) = \begin{cases} \text{Indéfinie si } |N_\epsilon(o)| \leq M \\ max(d^{noy}_{\epsilon,M}(o), d(o,p)) \text{ sinon} \end{cases}$$

où $d(o,p)$ est la mesure de distance définie dans l'espace métrique dans lequel sont plongées nos données.

L'algorithme est très similaire à DBSCAN, en ce sens où il va parcourir le nuage de points et étendre les points en clusters si ceux-ci n'ont pas déjà été explorés. L'innovation que OPTICS apporte est qu'il explore les points dans un ordre pertinent, pour ensuite créer des clusters à des échelles qui prennent en compte la densité locale de chaque point. Pour ce faire, les points sont (linéairement) ordonnés de telle sorte que ceux qui sont proches spatialement se retrouvent proches dans l'ordre d'exploration. Une fois cet ordre établi, on trace un graphe d'accessibilité (reachability plot) dans lequel on représente la distance d'accessibilité des différents points dans l'ordre établi par OPTICS. Comme les points appartenant à un cluster ont une faible distance d'accessibilité à leurs voisins les plus proches, les clusters apparaissent comme des vallées dans le graphique d'atteignabilité. Plus la vallée est profonde, plus le cluster est dense, il est possible ensuite de choisir un seuil au-delà duquel les points seront considérés comme du bruit. La figure 2.6 représente un exemple d'application de cet algorithme sur un nuage de points, avec les résultats du clustering par OPTICS ainsi que le graphique d'accessibilité correspondant.

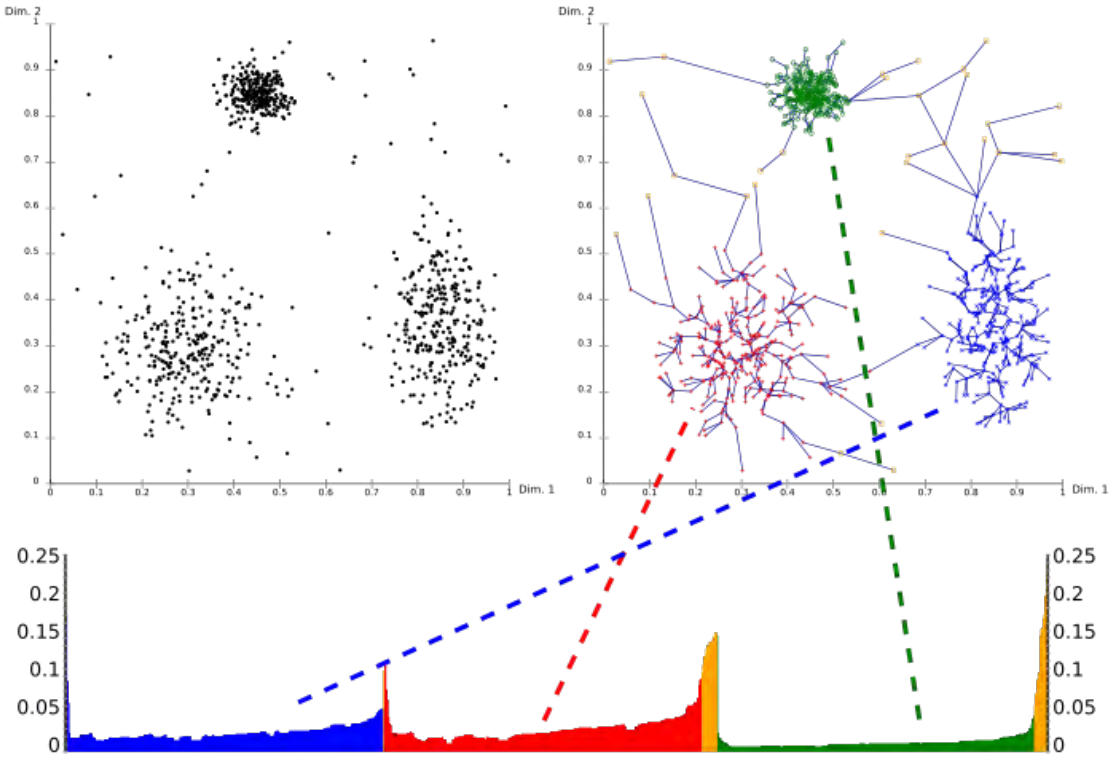


Figure 2.6: Graphique d'accessibilité d'un nuage de points contenant trois clusters. En haut à gauche nous avons le nuage de points, à droite les clusters découverts par OPTICS et en bas le graphe d'accessibilité. Figure générée à l'aide du logiciel de traitement de données ELKI [1].

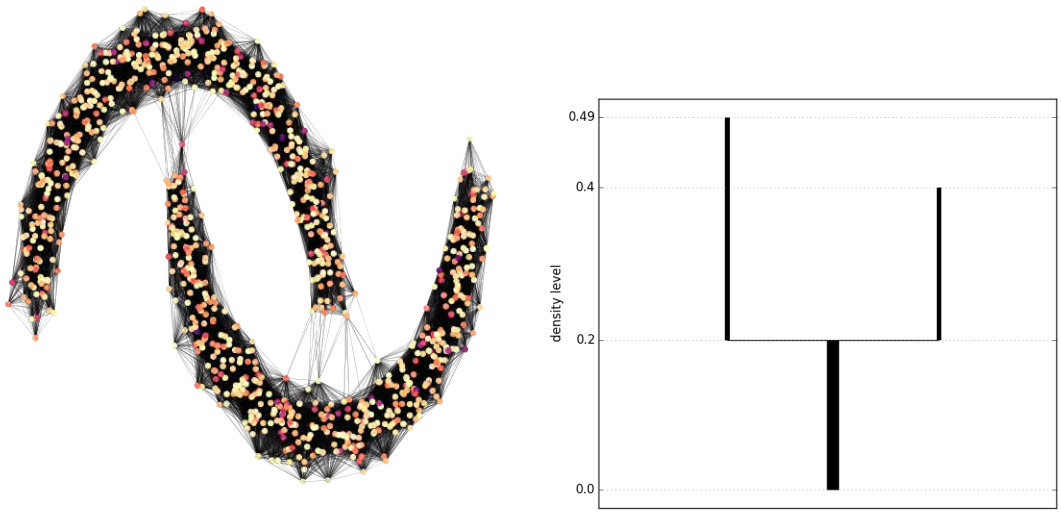
2.3.3 Algorithmes non inspirés de DBSCAN

Nous consacrons cette dernière partie à la mention de quelques-uns des algorithmes se basant sur la densité pour le clustering. Le premier est celui proposé dans l'article [73] qui calcule la densité spatiale de chaque point, sans contrainte particulière sur le choix de la mesure de densité. Dans cet article, les auteurs choisissent par défaut la densité obtenue en fixant un nombre k entier. On calcule ensuite la distance $d_k(p)$ séparant chaque point p de son k-ème voisin le plus proche, et on obtient la densité

$$\rho_k(p) = \frac{k}{2 \cdot \pi \cdot d_k(p)^2}. \tag{2.11}$$

On procède ensuite à la construction d'un graphe de voisins les plus proches, en reliant par une arête chaque point à ses k plus proches voisins. Une fois ce graphe obtenu, l'algorithme LST (pour "Level Set Trees") procède en supprimant tour à tour les nœuds ayant une densité plus faible qu'un seuil s qu'on fait varier de 0 à la densité maximale de l'ensemble des points. On garde une trace sur le nombre de composantes connexes dans le graphe qu'on obtient après chaque suppression, celles-ci pouvant être scindées à leur tour en plusieurs petites composantes. On obtient ainsi un arbre qu'on peut choisir de couper à la valeur adéquate de s, pour obtenir nos clusters qui ne sont rien d'autre que les composantes connexes du graphe, les nœuds supprimés étant le bruit. La figure 2.7 montre le résultat d'un tel algorithme sur un nuage de points composé de deux croissants de lune.

Un autre algorithme décrit dans [101] procède en recherchant d'abord le centre de chaque cluster, pour ensuite retrouver l'ensemble des points le constituant. L'algorithme repose sur les hypothèses suivantes : d'abord les centres des clusters sont entourés (spatialement) de voisins ayant une densité locale plus faible, ensuite ces centres doivent être à une distance relativement grande de tous les points ayant une densité locale plus élevée. Pour chaque point p, on calcule deux grandeurs : premièrement la



(a) (b)

Figure 2.7: (a) Le graphe des plus proches voisins d'un nuage de points généré sous la forme de deux croissants de lunes contenant 500 points chacun, avec $k = 10$ pour la construction du graphe. La couleur de chaque nœud est proportionnelle à la valeur de sa densité. (b) Le résultat de l'algorithme LST. Selon la valeur du seuil, on obtient soit un seul cluster si $0 \leq s \leq 0.2$ soit deux clusters si $0.2 \leq s \leq 0.4$. Figure générée à l'aide de la bibliothèque DeBaCl de Python [73].

densité locale ρ_p, obtenue en fixant un rayon r et en calculant le nombre de points dont la distance à p est inférieure à r

$$\rho_p = \sum_o \chi(d(o,p) - r) \tag{2.12}$$

avec

$$\chi(x) = \begin{cases} 1 \text{ si } x < 0 \\ 0 \text{ sinon} \end{cases}$$

et deuxièmement sa distance d_p par rapport au plus proche point de densité supérieure.

$$d_p = \min(\{d(q,p) | \rho_p < \rho_q\}).$$

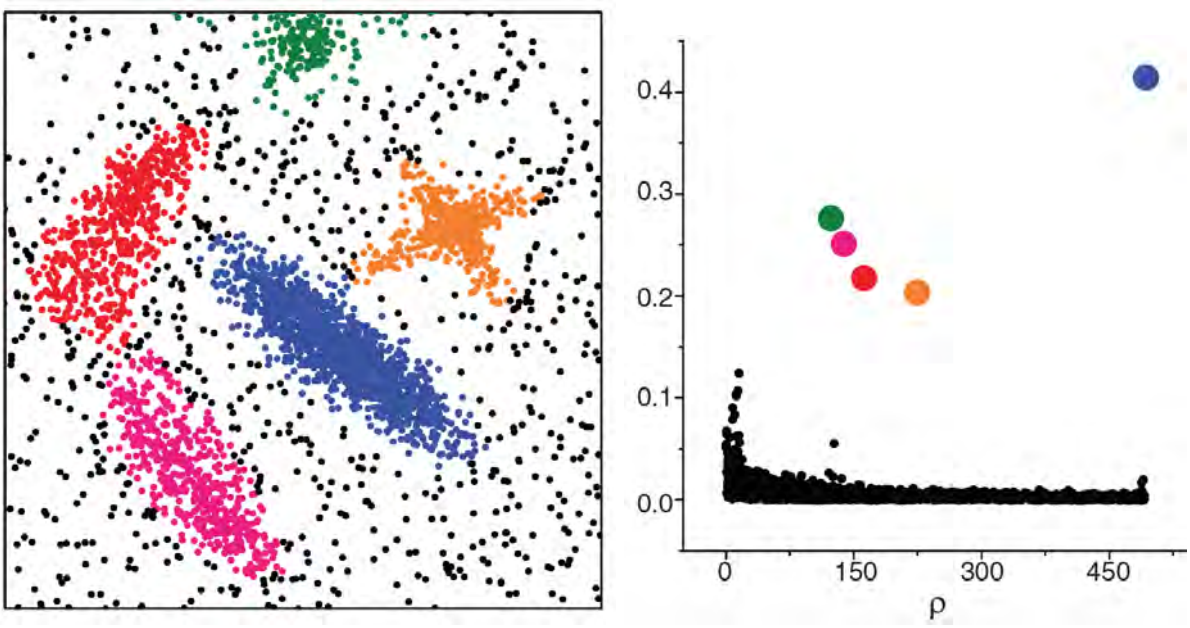
On trace ensuite les points (ρ_p, d_p) pour chaque point p, et on choisit le seuil adéquat (ce choix est empirique) pour obtenir les points représentant les centres. Ces derniers doivent être le plus éloignés possible du reste du nuage de points (*cf.* fig. 2.8b). Une fois les centres identifiés, chaque point est ensuite assigné au cluster représenté par son centre le plus proche. Ensuite pour chaque cluster c_i, on calcule le bord ∂_{c_i} comme étant l'ensemble des points qui se trouvent plus proches de n'importe quel point q appartenant à un cluster c_j différent de c_i (celui où se trouve p), qu'à son propre centre qu'on note ici c_p. Ceci se traduit par

$$q \in \partial_{c_i} \text{ s'il existe } p \in c_j \neq c_i \text{ ; tel que } d(q,p) < d(q,c_p)$$

On calcule ensuite la plus grande valeur de densité dans l'ensemble de points qui constituent le bord d'un cluster que l'on note $\rho_b^i = \max(\{\rho_q | q \in \partial_{c_i}\})$. On considère finalement comme du bruit tous les points du cluster c_i ayant une densité inférieure à ρ_b^i. On répète le calcul pour tous les clusters pour finalement obtenir l'ensemble des points qui constituent le bruit dans le nuage de points.

2.4 Densité spatiale

Nous avons vu jusqu'à présent deux façons différentes d'estimer la densité, à travers les mesures présentes dans eq. (2.11) et eq. (2.12). La première consiste à fixer un entier k pour calculer une densité



(a) (b)

Figure 2.8: (a) Les clusters résultants de l'algorithme décrit dans [101] et (b) les centres correspondants.

proportionnelle à l'inverse du carré du rayons r_k, séparant le point de son $k-$ème voisin le plus proche, la seconde consiste à fixer un rayon r pour tous les points, et de calculer une densité proportionnelle au nombre de points se trouvant à l'intérieur du cercle centré sur le point évalué, et de rayon r. Il est utile de rappeler qu'il existe dans la littérature d'autres mesures de densité, telles que celles présentées dans [21], et qui n'apparaissent pas dans ce chapitre.

Nous allons dans cette section introduire une nouvelle mesure de densité, dont on décrira l'intérêt et les caractéristiques. On testera ensuite cette mesure en l'incorporant dans DBSCAN, et en comparant ses résultats à ceux de la version classique de ce dernier, ainsi qu'à ceux de l'algorithme OPTICS, sur un nuage de points synthétiquement généré.

2.4.1 La mesure

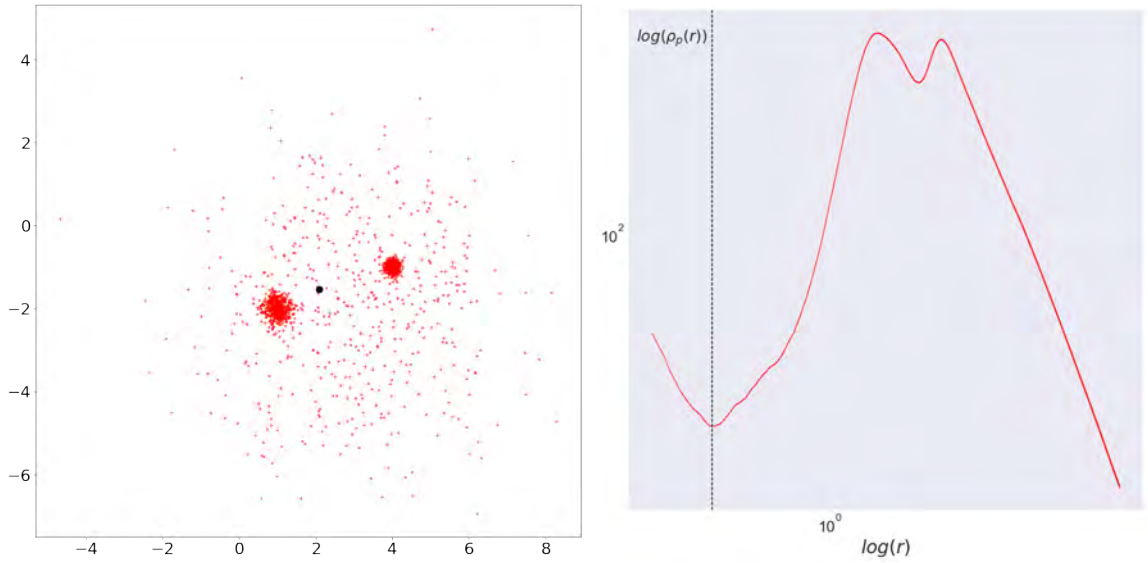
Soit D un nuage de points dans un espace métrique de dimension d, et p un point de ce nuage. Nous pouvons alors tracer la fonction $\rho_p(r)$ qui à ce point p associe pour chaque valeur de r la valeur obtenue en divisant le nombre de points $N_r(p)$ (*cf.* eq. (2.10)) par le volume V_d de la sphère d-dimensionnelle de rayon r. Par exemple nous avons $V_2 = \pi r^2$; $V_3 = \frac{4\pi}{3} r^3$,etc.
Pour simplifier, nous prendrons $d = 2$ dans les exemples à venir.

$$\rho_p(r) = \frac{N_r(p)}{V_d(r)} \tag{2.13}$$

Nous montrons sur la figure 2.10 le résultat de cette fonction sur un point sélectionné arbitrairement depuis un nuage de points.

Nous pouvons voir sur la figure que la densité du point p examiné suit une évolution difficilement prévisible. On peut cependant facilement interpréter les deux derniers maxima, dus à l'intégration progressive des points présents dans les deux clusters, à mesure que le rayon augmente. L'évolution qui précède cet évènement reste cependant difficile à deviner. La mesure décrite par eq. (2.12) est obtenue en décidant arbitrairement de la valeur du rayon seuil r_s, qui est la même pour tous les points. Ce choix est difficile à justifier, d'autant plus qu'il est possible de voir sur la figure 2.10 que certains choix de r_s peuvent être plus pertinents que d'autres, compte tenu des extrema locaux de la fonction $\rho_p(r)$.

L'algorithme que nous allons introduire consiste à calculer un rayon seuil $r_s(p)$ différent pour chaque point p. Ce rayon correspond à l'antécédent minimal dont l'image est un extremum local de la fonction



(a) (b)

Figure 2.9: Sur (a) nous avons un nuage de points dont celui examiné, représenté en noir, avec une taille supérieure à celle du reste des points, et sur (b) nous avons l'évolution de la densité de ce dernier, en fonction du rayon r, sur une échelle logarithmique. Nous indiquons par une droite verticale la valeur du premier extremum local de la fonction $\rho_p(r)$. La courbe est lissée par une fenêtre glissante d'une largeur égale au millième de la largeur totale de l'intervalle d'évaluation.

$\rho_p(r)$.

$$r_s(p) = \min(\{r_e | \rho_p(r_e - h) \leq \rho_p(r_e) \leq \rho_p(r_e + h) \text{ ou } \rho_p(r_e - h) \geq \rho_p(r_e) \geq \rho_p(r_e + h)\}) \tag{2.14}$$

où h est le pas d'échantillonage [3]. En effet nous ne pouvons pas écrire que r_s est le minimum de l'ensemble des points qui annulent la fonction dérivée de ρ, car cette fonction n'est pas dérivable en tout point. Nous nous contentons donc de la description donnée plus haut. Des détails sur l'échantillonnage sont donnés à la fin de cette section.

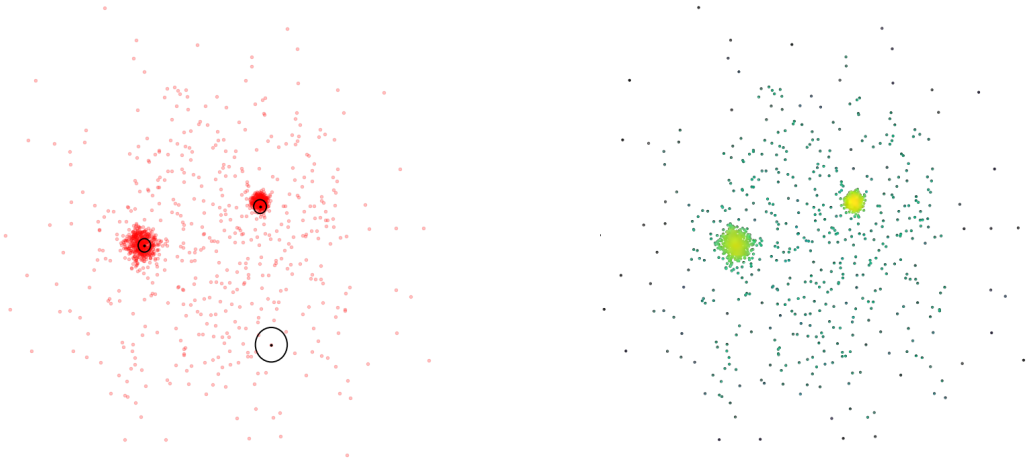
Nous justifions ce choix par le fait qu'un extremum local indique les limites d'une zone au-delà de laquelle un changement de régime est enregistré, par exemple si un point se trouve à l'intérieur d'un cluster dense et homogène, alors au moment où le rayon dépasse les limites de ce cluster, on enregistre une baisse de densité, ce comportement est aussi vrai pour les points se trouvant à l'extérieur des clusters, qui voient leur densité baisser jusqu'à ce que leur rayon encercle des points d'un cluster à proximité. Pour finir nous calculons la densité de chaque point p comme $\rho_p(r_s(p))$

Nous montrons sur la figure 2.10 différents rayons seuils calculés sur un nuage de points, ainsi que la densité calculée à partir de ces rayons.

Il est important de noter la différence entre cette mesure et celle décrite dans eq. (2.11). Cette dernière attribue certes différents rayons à chaque point, mais elle dépend d'un paramètre k, qu'on choisit empiriquement, et souvent de manière arbitraire. À l'inverse la mesure décrite par eq. (2.14) n'a pour seul paramètre que le nombre N de points échantillonnés, en commençant à $r = 0$, puis en augmentant r par pas de $h = \frac{r_{max}}{N}$ jusqu'à $r = r_{max}$, qui désigne le rayon de valeur maximale à laquelle on évalue la densité. Ce paramètre a l'avantage d'être intuitif, car il reflète le niveau de précision du calcul.

Il doit cependant satisfaire quelques conditions : d'un côté N doit être grand afin que la précision du calcul soit suffisamment bonne, de l'autre N doit être relativement petit devant l'inverse de la distance minimale séparant deux points dans le nuage. Si tel est le cas, la fonction $\rho_p(r)$ connaît un extremum local à chaque fois qu'un nouveau point est rajouté à l'intérieur du cercle de rayon r. Formellement, ceci s'écrit:

[3] La valeur de $r_s(p)$ est sensible à la largeur de la fenêtre de lissage. Il faut donc s'assurer que celle-ci soit petite devant l'intervalle total de la fonction, comme c'est le cas sur la figure 2.9b



(a) (b)

Figure 2.10: (a) Un nuage de points avec les rayons seuils calculés pour trois points différents, (b) la densité obtenue après le calcul de tous les rayons seuils pour tous les points. La couleur de chaque point correspond à sa densité, et va du bleu (faible densité) au jaune (forte densité).

$$1 \ll N \ll \frac{1}{\min(\{d(i,j)|i,j \in D\})}.$$

Ici on considère que les coordonnées de chaque point sont normalisées, de sorte que la distance maximale séparant deux points soit égale à 1.

2.4.2 Application à l'algorithme DBSCAN

Nous présentons dans cette partie un algorithme qui s'appuie sur la mesure de densité présentée dans eq. (2.14) pour générer les paramètres ϵ et M (*cf.* 3.1) de l'algorithme DBSCAN. Ces paramètres n'ont pas les mêmes valeurs pour tous les points dans l'application que nous allons montrer, et sont obtenus automatiquement.

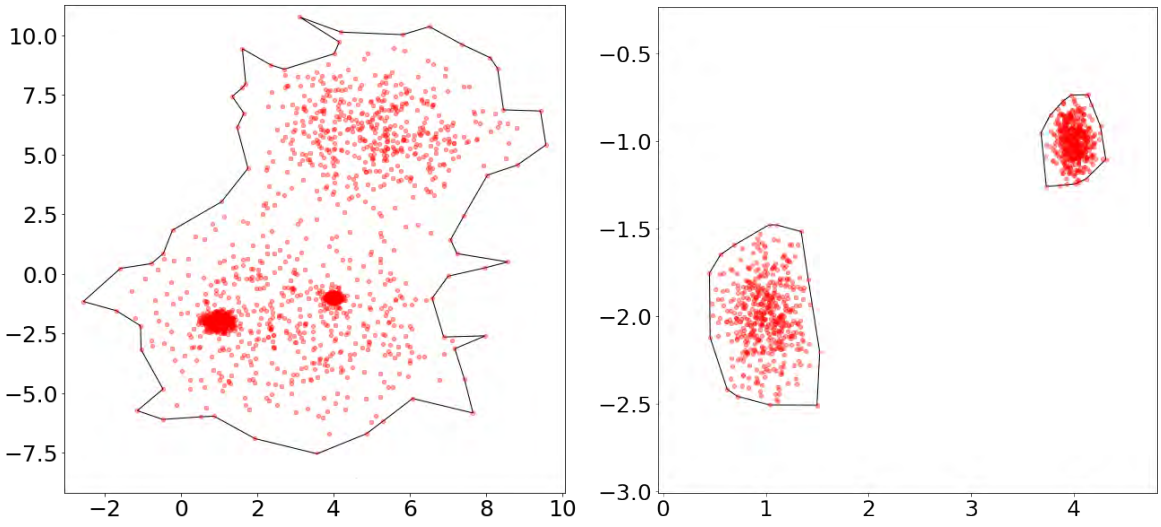
Nous comparerons ensuite les résultats d'un tel clustering avec ceux obtenus par l'algorithme DBSCAN classique, ainsi que ceux d'OPTICS. L'algorithme se divise en trois étapes, la première consiste à évaluer la densité moyenne dans différentes régions du nuage de points, à l'aide de l'enveloppe concave [47] calculée sur les données, la seconde à estimer la densité de chaque point à l'aide de eq. (2.14), et la dernière à utiliser ces deux informations afin d'attribuer les paramètres ϵ et M à chacun des points des données. Une fois les paramètres de chaque point générés, on suit les mêmes étapes que dans l'algorithme DBSCAN pour calculer les clusters.

Densité moyenne dans un nuage de points.

Pour estimer les paramètres ϵ et M de chaque point, nous allons avoir besoin de connaître la densité moyenne dans le nuage de points. Celle-ci s'obtient en divisant le nombre total de points N_{pts} par la surface S_0 qu'occupent ces derniers. On peut alors se demander quelle peut être cette surface. Le plus simple serait par exemple le rectangle (ou hyperrectangle dans le cas des dimensions supérieures à 2) délimité par les coordonnées maximales et minimales dans chacune des dimensions. Cette estimation n'est cependant pas précise dans le cas où le nuage de points est constitué par plusieurs clusters séparés par des régions vides, que l'on inclut dans la surface S_0.

Il existe cependant des algorithmes qui permettent de délimiter plus efficacement les frontières d'un nuage de points [47], en calculant son enveloppe concave. Ils sont généralement basés sur la triangulation de Delaunay [38] comme le modèle de forme alpha [5]. Ce dernier associe à chaque valeur réelle et positive

d'un paramètre α une délimitation sous forme d'une enveloppe concave. Un exemple d'application de cet algorithme est représenté sur la figure 2.11 ci-dessous :



(a) (b)

Figure 2.11: (a) La forme alpha d'un nuage de points composé de plusieurs clusters de densité différentes, dont deux plongés dans du bruit. Le résultat est une composante unique. En revanche sur (b) nous avons une forme alpha qui se divise en deux composantes, car les clusters sont éloignés et ne sont pas plongés dans du bruit.

Une fois cette surface délimitée, il ne reste plus qu'à calculer son aire S_0, et ensuite estimer la densité moyenne :

$$\rho_0 = \frac{N_{pts}}{S_0}. \tag{2.15}$$

À noter qu'il est possible que cette enveloppe soit composée de plusieurs composantes, si les clusters sont trop éloignés les uns des autres, comme on peut le voir sur la figure 2.11b. Dans ce cas là, nous faisons le choix de calculer une densité globale pour chacune de ces composantes, que l'on notera $\rho_0^{(i)}$, pour i allant de 1 jusqu'au nombre total des composantes disponibles. Cette densité est obtenue par le rapport entre le nombre de points dans la composante et son aire.

Choix des paramètres et résultats

À l'aide des résultats obtenus par eq. (2.14) et eq. (2.15), nous définissons pour chaque point p d'un jeu de données D (qui consiste en un nuage de points) les paramètres

$$\epsilon_p = r_s(p)$$

ainsi que le nombre minimal de points :

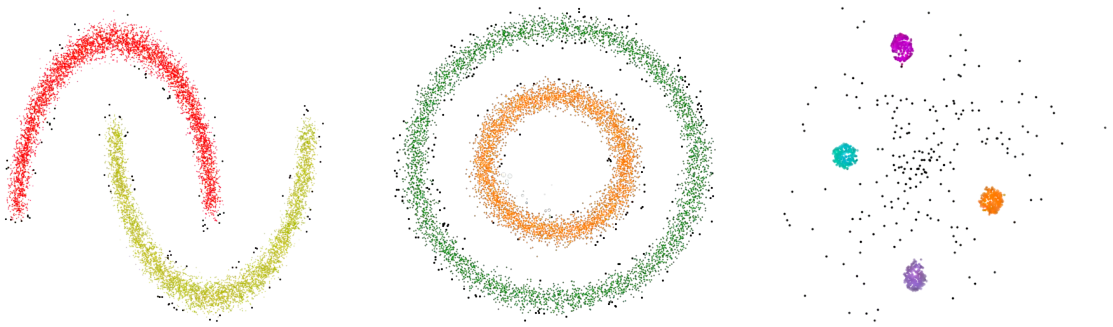
$$M_p = \frac{\rho_0 \cdot N_{\epsilon_p}(p)}{\rho_p(r_s(p))}.$$

D'un côté nous avons le paramètre ϵ_p, qui n'est autre que le rayon défini par eq. (2.14). De l'autre, on considère que le nombre minimal de points que doit contenir un cercle de rayon ϵ_p centré autour du point p, est celui qui rend sa densité $\rho_p(r_s(p))$ au moins égale à la valeur moyenne ρ_0. Ceci implique que tous les points dont la densité est inférieure à la densité moyenne ρ_0 seront considérés comme du bruit.

Ainsi pour chaque point on obtient une estimation différente du rayon ϵ et du nombre minimum de points. Il est important de préciser que le choix du paramètre α pour le calcul de la forme concave du nuage de points est important, car il permet en fonction des résultats de faire une séparation préliminaire des clusters, étant donné que chaque composante de l'enveloppe concave est traitée séparément des

autres. Des valeurs élevées de α entraînent une enveloppe de plus en plus resserrée autour des points, ce qui peut la séparer en plusieurs composantes distinctes.

Nous montrons, dans ce qui va suivre, les résultats d'un tel choix de paramètres d'abord sur des cas simples, ou les clusters sont de densités homogènes, ensuite sur un cas plus compliqué où les densités varient d'un cluster à l'autre.



(a) (b) (c)

Figure 2.12: Les résultats de l'algorithme DBSCAN avec les choix de paramètres décrits dans cette section, sur des données où les points des différents clusters sont de densité homogène. Les points classés comme du bruit sont ici représentés en couleur noire.

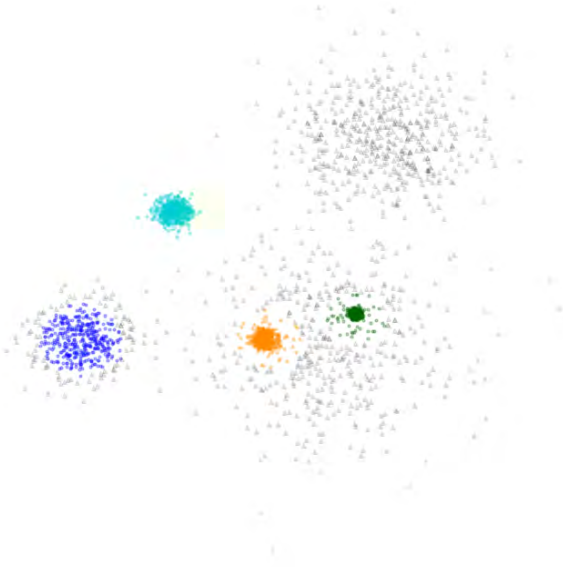
Nous pouvons voir sur la figure 2.13 que ce choix permet de retrouver tous les clusters disponibles, en adoptant le choix de paramètres décrit plus haut (*cf.* fig. 2.13c), au contraire du cas où les paramètres sont choisis arbitrairement(*cf.* fig. 2.13a). En revanche la comparaison entre la figure 2.13b et la figure 2.13c montre que OPTICS gère d'une manière plus douce les bordures des clusters. En effet en examinant en détail la figure 2.13c on se rend compte que le nombre de points qui se situent sur la périphérie des clusters et qui sont considérés comme du bruit est beaucoup plus important que sur la figure 2.13b. Ceci est dû au fait qu'ayant adopté cette nouvelle densité et les choix de paramètres correspondants, tous les points dont la densité est inférieure à la moyenne sont considérés comme du bruit, les points qui se situent sur la périphérie des clusters étant plus susceptibles de satisfaire cette condition.
Nous en concluons que la densité obtenue à l'aide de eq. (2.14) accentue la différence de valeurs entre les points qui se trouvent au coeur des clusters et ceux de la périphérie.

2.5 La densité dans le graphe et la densité spatiale

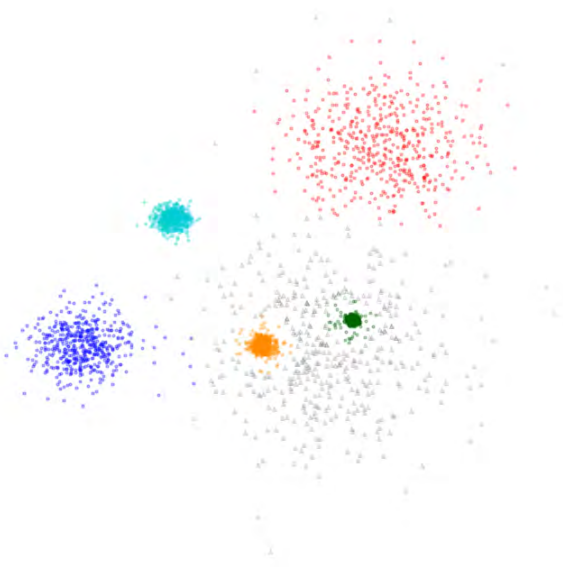
Dans le domaine de l'analyse des données, la visualisation est la représentation graphique de ces données. Il s'agit de produire des images qui reflètent les relations entre les données. Ceci est réalisé par l'utilisation d'une correspondance systématique entre les aspects imagés, et les valeurs des données lors de la création de la visualisation. Cette cartographie établit la façon dont les valeurs des données seront représentées visuellement, en déterminant comment et dans quelle mesure la propriété d'un indicateur graphique, comme sa taille ou sa couleur, changera pour refléter un changement de valeurs dans les données.

Pour transmettre l'information de façon claire et efficace, plusieurs outils de visualisation des données sont disponibles, comme par exemple les graphiques statistiques, les graphiques d'information et bien d'autres outils. Les données numériques peuvent être encodées à l'aide de points, de lignes ou de barres pour communiquer visuellement un message quantitatif. Une visualisation efficace aide les utilisateurs à analyser et à raisonner au sujet des données. Il rend les données complexes plus accessibles, compréhensibles et utilisables.

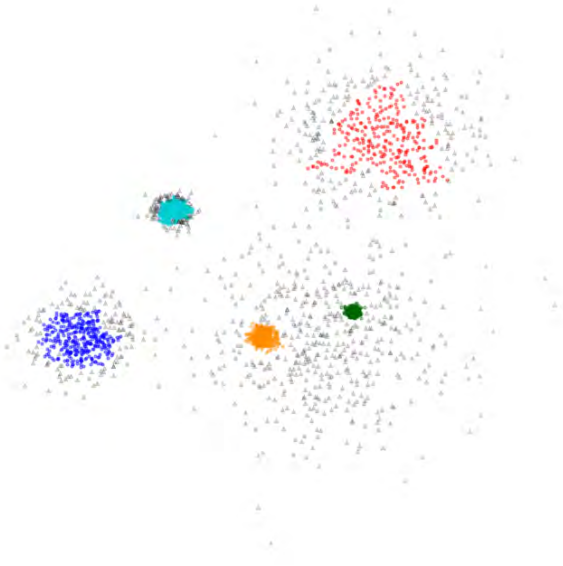
Parmi les multiples techniques de visualisation, certaines consistent à explorer les similarités présentes dans les données, pour leur attribuer une position dans un espace de faible dimension m (généralement $m = 2$ ou $m = 3$) pour ensuite les représenter comme un nuage de points, permettant ainsi leur visualisation. On peut évoquer le principe de positionnement multidimensionnel MDS, qui consiste à représenter des données d'un espace de dimension N, dans un nouvel espace $p < N$, à l'aide d'une matrice de similarité D, tout en conservant au mieux les proximités entre les points. Ce principe est



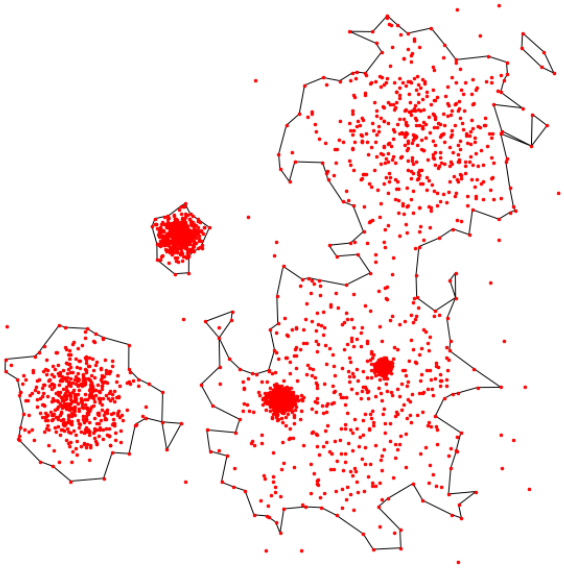
(a) DBSCAN avec des paramètres arbitraires



(b) OPTICS



(c) DBSCAN avec le choix de paramètres décrit dans cette section



(d) l'enveloppe alpha du nuage de points

Figure 2.13: (a) Les choix de paramètres sont $\epsilon = 0.5$ $M = 10$, (b) nous exécutons OPTICS avec le même paramètre $M = 10$, et dans (c) nous avons pris une précision de calculs de $N = 10000$ points. (d) Montre le résultat de l'enveloppe alpha calculée via un paramètre $\alpha = 2$ qui sépare en trois composantes distinctes l'enveloppe concave du nuage de points, les points en dehors de l'enveloppe sont automatiquement considérés comme du bruit, celui-ci étant représenté dans les trois autres figures par l'ensemble des points de couleur grise.

applicable sur des réseaux, et le choix de la mesure de similarité à employer dépend des propriétés que l'on souhaite mettre en évidence. .

Nous choisissons de nous concentrer ici sur des approches basées sur une analogie avec le domaine de la physique [20]. Ces méthodes intuitives ont l'avantage d'être faciles à programmer et donnent de bons résultats sur de petits graphes ou des graphes de tailles moyennes. En général, ces méthodes consistent à sélectionner les deux aspects suivants :

- D'abord un modèle, représentant le système physique avec lequel l'analogie est faite, comprenant notamment les interactions entre ses éléments.

- Un algorithme qui se charge de calculer la configuration optimale recherchée.

Un des exemples les plus connus de ce type d'approches est celui dans lequel chaque nœud du graphe est considéré comme un anneau solide, relié par des ressorts (de longueur au repos nulle, et qui peuvent s'étirer à l'infini si nécessaire) aux nœuds avec lesquels il partage une arête, tout en exerçant une force répulsive sur tous les autres nœuds du graphe. Ceci a pour effet de rapprocher les nœuds qui appartiennent aux mêmes clusters, et en même temps d'éloigner les clusters les uns des autres.

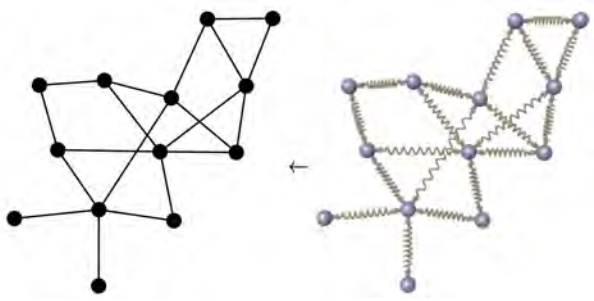


Figure 2.14: Représentation du modèle décrit plus haut [20]

Le positionnement qui est fournit par ce type d'algorithme est à la base de notre première tentative de mise au point d'une mesure de densité à l'échelle des nœuds, celle-ci se base sur l'hypothèse suivante : plus deux nœuds partagent de voisins en communs, plus ceux-ci se trouveront rapprochés spatialement dans le positionnement, grâce à la force attractive, faisant que les clusters dans le graphe forment aussi des clusters dans l'espace géométrique de positionnement. Inversement, moins il existe de liens entre les nœuds de deux clusters distincts, plus ces clusters seront éloignés dans l'espace de positionnement, par l'effet de la force de répulsion. On associe alors la densité d'un nœud dans le graphe à la densité de celui-ci dans l'espace de positionnement.

Dans ce qui suit nous donnons une description plus détaillée du modèle, de l'algorithme d'optimisation, ainsi que quelques résultats de tests que nous avons effectués sur des réseaux réels et synthétiques.

2.5.1 Les algorithmes de type attraction/répulsion, et notre version personnalisée

Soit $G = (V, E)$ un graphe de taille $|V| = N$, non dirigé et non pondéré. On note $p = (p_v)_{v \in V}$ avec $p_v = (x_v, y_v)$ les coordonnées du nœud i dans le plan à deux dimensions, muni d'une norme. Nous nous limitons dans toute la suite de ce chapitre au cas d'un espace de positionnement à deux dimensions. On note $||p_v - p_u||$ la distance entre les points p_u et p_v. On note aussi $\overrightarrow{p_u p_v}$ le vecteur unitaire de coordonnées $\frac{p_u - p_v}{||p_v - p_u||}$.

Le modèle de Fruchterman et Reingold [59] est l'un des plus utilisés parmi les algorithmes de ce type. Dans ce modèle la force d'attraction entre deux nœuds est la suivante :

$$f_a(u, v) = \frac{||p_v - p_u||^2}{k} \cdot a_{u,v} \cdot \overrightarrow{p_u p_v}. \tag{2.16}$$

Cette force n'est prise en compte que lorsque les deux nœuds considérés sont reliés par une arête $(u, v) \in E$, d'où le terme $a_{u,v}$. La force de répulsion s'écrit quant à elle :

$$f_r(u, v) = \frac{-k^2}{||p_v - p_u||} \cdot \overrightarrow{p_u p_v} \tag{2.17}$$

Ce terme en revanche existe entre chaque paire de nœuds (u, v), indépendamment de leur situation dans le graphe.

Ici k est une constante positive, qui représente un paramètre d'échelle. En effet, il est facile de vérifier que la somme des deux forces s'annule quand $||p_v - p_u|| = k$. En général, on fixe $k = C \cdot \sqrt{\frac{S}{N}}$ où S est la surface de l'espace fini de positionnement (rappelons qu'on choisit de normaliser les coordonnées pour que la distance maximale entre deux points soit égale à 1), N le nombre de nœuds, et C une constante positive. Ainsi, la valeur de k fournit la distance caractéristique qui sépare deux nœuds voisins.
À noter que le modèle de Fruchterman et Reingold ne repose pas sur des forces de type gravitation et

ressorts. En effet, la force attractive varie en fonction du carré de la distance, alors que la force répulsive varie comme l'inverse de celle-ci.

On connaît plusieurs façons de calculer le positionnement optimal des nœuds, l'une des plus intuitives [20] consiste à calculer l'énergie totale d'un système obéissant à de telles forces, et de dire que le positionnement optimal est celui qui minimise une telle énergie. On peut commencer par calculer l'expression de l'énergie du système, que l'on peut facilement obtenir en calculant la primitive de la somme des deux forces décrites plus haut :

$$E = \sum_{u,v} -\frac{1}{3k}||p_v - p_u||^3 \cdot a_{u,v} + k^2 \cdot log(||p_v - p_u||) \tag{2.18}$$

Ce qui en terme de coordonnées (x_v, y_v) de chaque nœud v s'écrit :

$$E(x_1, y_1, x_2, y_2, ..., x_N, y_N) = \sum_{u,v} -\frac{1}{3k}((x_u - x_v)^2 + (y_u - y_v)^2)^{\frac{3}{2}} \cdot a_{u,v} + k^2 \cdot log((x_u - x_v)^2 + (y_u - y_v)^2)^{\frac{1}{2}})$$

On a ainsi une fonction de $2N$ ($d \cdot N$ pour une dimension d quelconque) variables à minimiser, N étant le nombre de nœuds dans le réseau.

Une autre méthode utilisée dans [45, 59] consiste à calculer itérativement, et pour chaque nœud, son déplacement dans le sens des forces qu'il subit, tout en limitant le déplacement maximal à une valeur finie. Ce procédé est répété un certain (grand) nombre de fois, et en ajoutant une modulation qui atténue progressivement la longueur du déplacement, on aboutit alors à des résultats stables pour des graphes de taille moyenne, de l'ordre de quelques milliers à quelques dizaines de milliers de nœuds.

2.5.2 Le modèle de Fruchterman et Reingold revisité

Nous généralisons ici le modèle de Fruchterman et Reingold. En gardant les caractéristiques principales de celui-ci, cette version est celle de l'algorithme de positionnement utilisé par la suite pour mesurer la densité. Soient les forces attractives et répulsives suivantes :

$$f_a(u, v) = ||p_v - p_u||^\nu \cdot k_{u,v}^n \cdot a_{u,v} \cdot \overrightarrow{p_u p_v} \tag{2.19}$$

$$f_r(u, v) = \frac{-k_{u,v}^m}{||p_v - p_u||^\mu} \cdot \overrightarrow{p_u p_v} \tag{2.20}$$

où $k_{u,v}$ est un paramètre qui change d'une paire à l'autre :

$$k_{u,v} = \sqrt{\frac{S}{N}} \cdot (1 - J(u, v)) \tag{2.21}$$

Ici S est la surface totale, N le nombre de nœuds, et $J(u, v)$ l'indice de Jaccard entre les voisinages des deux nœuds u et v (rapport entre les tailles de l'intersection et de l'union des voisinages). Dans le cas classique défini précédemment, le paramètre k est le même pour toutes les paires de nœuds, ici il dépend de leur similarité : plus ils ont des voisinages similaires, plus ils sont proches l'un de l'autre dans le positionnement (sous réserve qu'ils partagent un lien, sinon ils ne subissent que la force répulsive qui les éloigne l'un de l'autre).

Ensuite nous avons les paramètres m, n, μ et ν qui interviennent dans le modèle, avec pour seule restriction la positivité des paramètres μ et ν, de sorte que f_a soit attractive et f_r répulsive. Si nous fixons comme contrainte supplémentaire[4] $f_a(k_{u,v}) + f_b(k_{u,v}) = 0$ pour toutes les paires de nœuds qui partagent une arête, nous obtenons :

$$k_{u,v}^\nu \cdot k_{u,v}^n - k_{u,v}^{-m} \cdot k_{u,v}^\mu = 0$$

Ce qui n'est vérifié que si $k_{u,v} = 0$ ou plus généralement si :

$$\nu + n = m - \mu \tag{2.22}$$

[4]Cette contrainte est justifiée par le fait que la distance d'équilibre entre deux nœuds reliés par une arête est égale à $k_{u,v}$ et que la somme des forces à cette distance doit être nulle.

Par exemple les forces décrites par le modèle de la section 2.5.1 sont celles qui correspondent au cas $m = 2$, $\mu = 1$, $n = -1$, $\nu = 2$

Nous pouvons ainsi modifier les paramètres pour mettre en évidence des propriétés différentes. On distingue cependant trois configurations caractéristiques : celle ou $\nu = \mu$, qui correspond au cas où la force de répulsion et la force d'attraction sont de même exposant, les deux autres cas sont soit celui où l'exposant de la force d'attraction est le plus fort, où à l'inverse celui où l'exposant de la force de répulsion plus grand. Les paramètres m et n servent en revanche à agir sur l'ordre de grandeur de ces forces. Nous donnons sur la figure 2.15 quelques exemples qui correspondent à différents choix de paramètres, sur un réseau réel échantillonné à partir de l'égo-graphe d'un utilisateur de Facebook. Les données relatives à ce graphe sont disponibles sur la plate-forme de Stanford d'analyse des réseaux (SNAP) [79].
Nous remarquons sur cette figure que ces choix ne modifient pas radicalement la forme générale du positionnement, car on observe bien la formation de clusters sur les trois figures. À une échelle plus fine, nous pouvons cependant remarquer quelques différences. En particulier l'effet que les différents rapports de forces ont sur le positionnement des nœuds de faible densité spatiale. Ces derniers sont moins dispersés autour des clusters quand $\nu \geq \mu$ (*cf.* fig. 2.15b et fig. 2.15d), que quand $\nu < \mu$ (*cf.* fig. 2.15f). Ceci est dû à une plus forte répulsion dans ce derniers cas. Nous faisons finalement le choix de garder cette dernière configuration pour l'estimation de la densité, dont nous montrons le résultat sur le même réseau présenté précédemment, sur la figure 2.16. On note aussi que la configuration représentée sur les figures 2.15e et 2.15f décrit des forces réalistes, avec une force d'attraction qui s'écrit comme la force de rappel d'un ressort, et une force de répulsion qui décroît comme la force de gravitation.

Il reste cependant important de rappeler que le calcul du positionnement repose sur l'optimisation d'une fonction objectif non linéaire, à $N \cdot d$ variables (d étant la dimension de l'espace de positionnement). Pour de grandes valeurs de N, trouver le minimum global d'une telle fonction est théoriquement possible [75], mais avec un temps de calcul qui croit de façon exponentielle avec la taille des données. Nous nous contentons ainsi de solutions approchées, obtenues depuis des minima locaux de la fonction objectif, en des temps raisonnables. La conséquence de ceci est que les positionnements obtenus par différentes optimisations sont très rarement identiques. Nous consacrons la prochaine section à l'étude de cette variabilité.

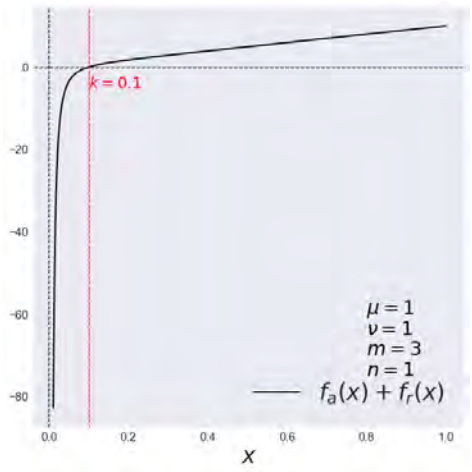
2.6 Analyse statistique de la variabilité des résultats

Nous allons à présent soumettre l'approche que l'on vient de décrire à des tests, en l'appliquant à des nuages de points issus du positionnement calculé d'abord sur des graphes synthétiques, puis sur des réseaux réels. Pour cela nous calculons d'abord pour chaque réseau $N_{pos} = 1000$ positionnements différents, ayant tous les mêmes paramètres : $m = 2, n = -1, \mu = 2, \nu = 1$.

Pour chacun de ces graphes, nous calculons ensuite la densité moyenne de chaque nœud, sur les N_{pos} positionnements différents, obtenue grâce à la mesure décrite précédemment par les eq. (2.13) et eq. (2.14), ainsi que sa déviation standard. Ensuite nous appliquons les outils de l'analyse topologique des données sur les nuages de points des différents positionnements, et en calculant l'homologie persistante de chacun des N_{pos} nuages de points. Ces derniers étant calculés dans le plan (de dimension $d = 2$), nous nous concentrons en particulier sur la distribution des instants d'apparition, ainsi que celle des durées de vie des cycles 1-dimensionnels [5]. Nous rappelons que nous pouvons obtenir des informations sur la forme des données en se concentrant sur la présence de ces cycles et de leur persistance.

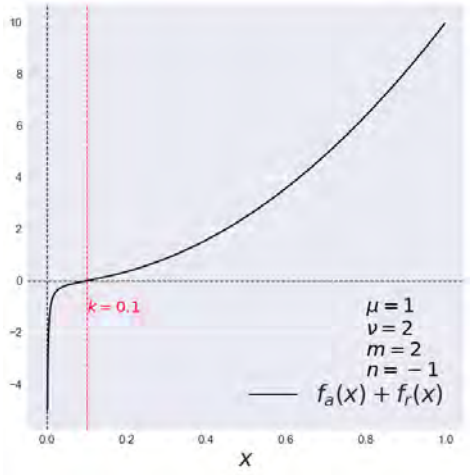
Les données sur lesquelles sont calculées ces quantités contiennent d'abord trois réseaux synthétiques : un modèle d'Erdös-Renyi et un modèle de Barabasi-Albert de 500 nœuds chacun. Le troisième modèle est celui qu'on appelle treeCom, qui a été construit à partir d'un arbre aléatoire composé de 100 nœuds, représentant la partie de faible densité du modèle. Cet arbre est relié à quatre clusters de 100 nœuds chacun, représentant quant à eux la partie de forte densité. Une description plus détaillée du modèle treeCom, prenant en compte les paramètres structurels est donnée dans le prochain chapitre de cette thèse. Pour le moment, nous ne nous concentrons que sur la forme de son positionnement, que l'on peut voir sur la figure 2.17b. Nous avons aussi trois réseaux réels, le premier est un réseau électrique, dont les données sont disponibles sur le lien [104], le second un réseau social fictif, issu de la saga "Le trône de fer" que l'on peut trouver sur le lien `https://networkofthrones.wordpress.com/` (et qui sera analysé en détails dans le dernier chapitre) et enfin un réseau de co-citations d'articles scientifiques [104].

[5] Représentant les trous dans le plan, et leurs tailles respectives



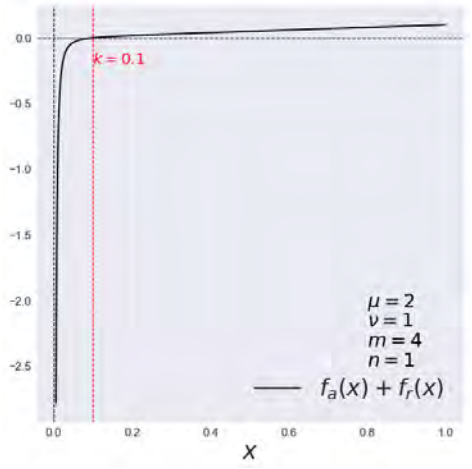
(a)

(b)

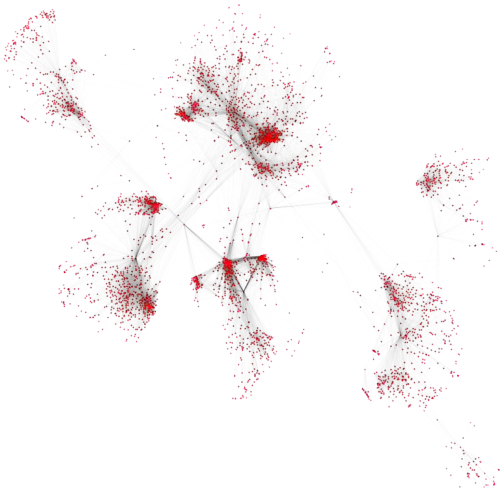


(c)

(d)



(e)



(f)

Figure 2.15: Les trois configurations possibles : (a),(b) $\mu = \nu$; (c),(d) $\mu < \nu$; (e),(f) $\mu > \nu$

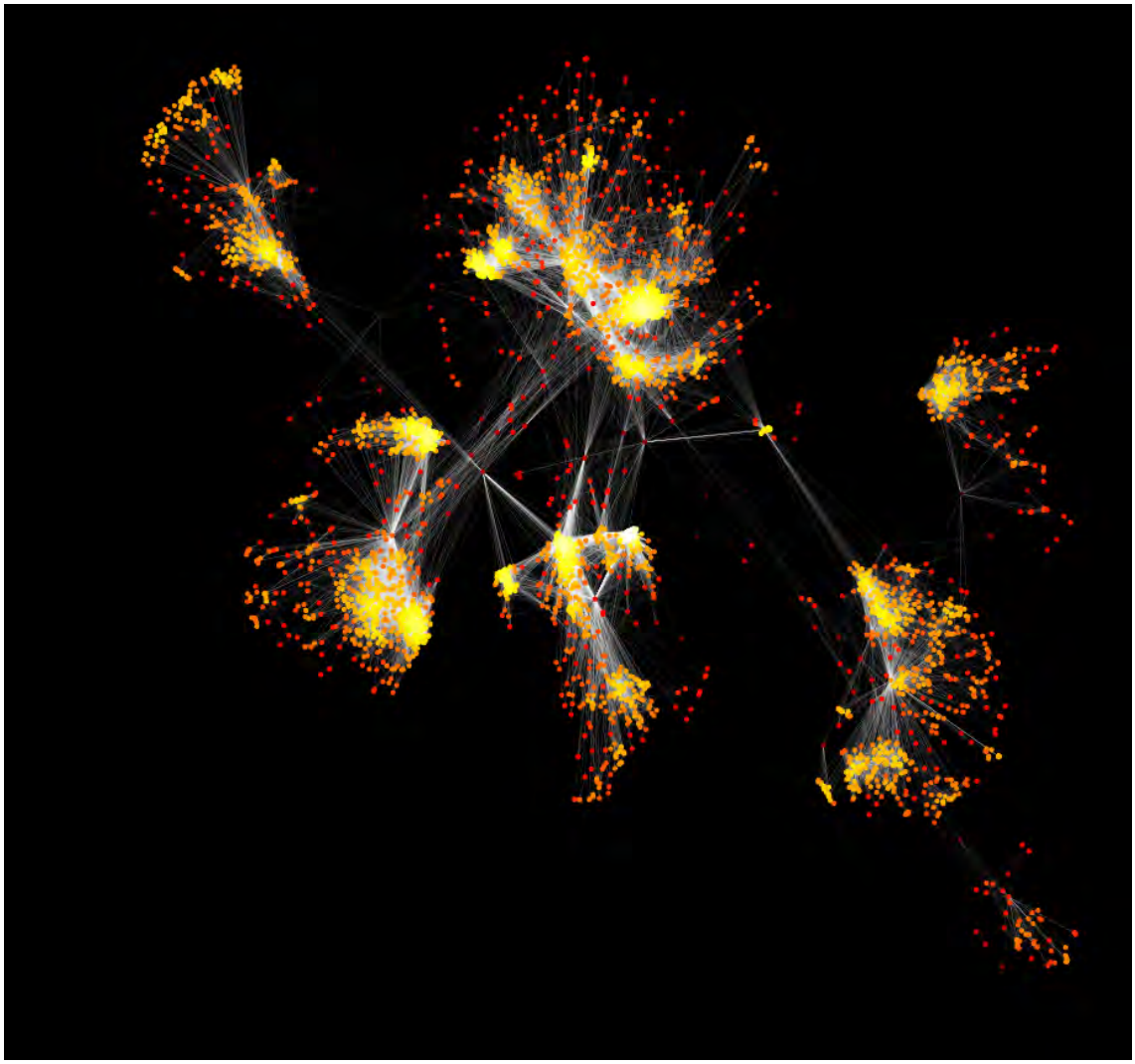


Figure 2.16: Résultats de la mesure de densité décrite dans la section 2.4 à travers eq. (2.14) sur un réseau réel, correspondant à un échantillon du graphe de Facebook, extrait de la plate-forme de Stanford d'analyse des réseaux [79]. Les niveaux de couleurs représentent la valeur de la densité, allant du jaune pour les hautes valeurs de densité, jusqu'au rouge pour les faibles valeurs.

Nous représentons par la suite un positionnement calculé sur chacun de ces réseaux, avec la densité de chaque nœud représentée sur la figure 2.17 par un code de couleurs, la densité moyenne et son écart-type pour chacun de ces réseaux sur la figure 2.18, et enfin les distributions estimées des instants de naissance et des durées de vie, obtenues depuis l'analyse de l'homologie persistante des données issues des positionnements sur la figure 2.19.

Nous appelons t le paramètre d'échelle dans l'analyse de l'homologie persistante (nous l'avons appelé ϵ dans la section 2.2, mais il est aussi courant de le noter t pour une analogie avec la variable temps), $B(t)$ la distribution des instants de naissance et $L(t)$ celle des durées de vies. Ces notations font référence aux mots anglais "Birth" et "Lifetime", que l'on retrouve souvent dans la littérature. On estime ces dernières par la méthode de Parzen-Rosenblatt [97], aussi appelée méthode de l'estimation par noyau, qui à partir d'un échantillon statistique, d'une largeur de bande, et d'une fonction noyau, estime la densité de probabilité de la variable aléatoire représentée par l'échantillon. Pour les densités affichées sur la figure 2.19 le noyau employé pour le calcul est un noyau Gaussien :

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Nous rappelons que pour un échantillon $X = \{x_1, x_2, ..., x_n\}$ de variables aléatoires indépendantes et

identiquement distribuées (on peut supposer que c'est le cas de nos observations), la densité estimée par la méthode de Parzen-Rosenblat est donnée par :

$$\hat{f}_h(x) = \frac{1}{N \cdot h} \sum_{i=1}^{N} k(\frac{x - x_i}{h})$$

La largeur de bande choisie lors de nos calculs est donnée par $h = \frac{\max(X) - \min(X)}{1000}$ où X est l'échantillon d'observations.

Les résultats sur la figure 2.19 montrent deux types de comportements différents :

- D'un côté les nuages de points qui forment des cycles qui naissent à l'instant $t = 0$ et qui vivent une certaine durée distribuée autour d'une valeur positive de t, comme c'est le cas pour les positionnements des réseaux d'Erdös-Renyi, et Barabási-Albert, représentés par la figure 2.19d et 2.19f. Ce résultat signifie que les cycles qui naissent à l'instant $t = 0$ se mettent tour à tour à fusionner les uns avec les autres, jusqu'à ce que t soit assez grand pour qu'il n'y en ait plus aucun. Ce type de profil indique des nuages de points distribués de façon plus ou moins homogène sur le plan, comme nous pouvons l'apercevoir sur les positionnements représentés sur la figure 2.17d et la figure 2.17f. D'une manière plus générale, l'absence de pic dans la distribution des instants de naissance $B(t)$, hormis celui qu'on observe à $t = 0$, indique l'absence de creux de tailles significatives dans les positionnements issus des modèles de Barabási-Albert et Erdös-Renyi.

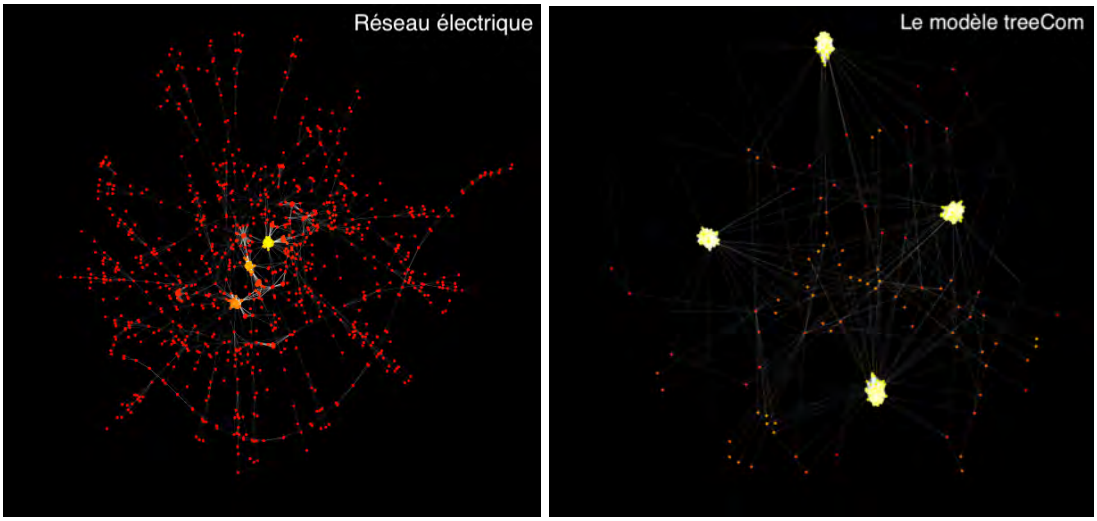
- L'autre comportement observé est celui où on relève l'existence de plusieurs modes dans les distributions des instants de naissance $B(t)$, (*cf.* fig. 2.19a,b,c,e). Le fait qu'il y ait un pic d'instant de naissance, autour d'une valeur de t supérieure à zéro, signifie qu'il y a un creux dans le plan qui n'a pas pu être relevé pour les faibles valeurs de t, si en plus celui-ci coexiste avec un autre pic autour d'une valeur proche de $t = 0$, alors on peut en déduire l'existence de différentes échelles de "vides" dans le nuage de points étudié. De manière générale, plus il y a de modes dans la distribution des instants de naissance, et plus ces modes sont éloignés les uns des autres, plus ces différentes échelles de vide sont distinctes. Ceci équivaut à dire qu'il existe plusieurs échelles de densité dans le nuage de points étudié.

Pour ce qui est de la durée de vie de ces cycles 1-dimensionnels (qu'on a appelé vides), on rappelle que plus celle-ci est longue, plus la composante correspondante est persistante. Ainsi on remarque sur la figure 2.19a et la figure 2.19b que les durées de vie des cycles aux naissances les plus tardives (courbes en vert) sont légèrement plus longues que celles des cycles qui sont nés à des échelles intermédiaires (courbes en rouge). Ceci s'explique intuitivement par le fait que plus l'instant de naissance t d'un cycle est tardif, plus le vide qu'il contient est grand, et par conséquent plus il est persistant à l'augmentation de t. Ceci n'est pas toujours vrai pour les cycles dont la naissance est à $t = 0$ (courbes en bleu), car certaines composantes résultant de la fusion entre plusieurs cycles dont la durée de vie a été courte, peuvent persister durant un intervalle relativement long (comme c'est le cas sur la figure 2.5b).
Ainsi nous observons deux modes avec des durées de vie relativement élevées sur la figure 2.19a, qui correspondent aux composantes apparues à $t = 0$. Ce réseau a pour particularité d'être composé de plusieurs clusters d'une densité spatiale très élevée par rapport au reste des points (*cf.* fig. 2.17a, fig. 2.18a). Il en résulte une forte force répulsive localement concentrée autour de ces points, qui aboutit à un positionnement comme celui affiché sur la figure 2.17a, sur lequel on observe des clusters de points entourés par une zone vide. Cette dernière est à l'origine des modes qui correspondent à de longues durées de vie, observées sur la courbe bleue de la figure 2.19a . Cet effet n'est en revanche pas retrouvé dans la distribution des durées de vies issue du positionnement du modèle treeCom, bien que sa distribution des instants de naissance soit similaire à celle issue du réseau électrique, comme le montrent les figures 2.19a et 2.19b.
Pour le modèle treeCom nous remarquons que les durées de vie les plus longues sont celles des composantes qui sont apparues tardivement. Ceci est dû au fait que la densité moyenne des nœuds qui se situent en dehors des clusters, est significativement plus grande dans le réseau électrique en comparaison avec le treeCom. À titre d'exemple, la densité moyenne calculée sur l'ensemble de nœuds ayant une densité moyenne inférieure à 10% du maximum de la densité moyenne est 23.54 fois supérieure dans le cas du réseau électrique par rapport au treeCom. Par conséquent on observe plus de persistance dans les cycles de naissances tardives de ce dernier.

Les deux autres réseaux à considérer sont celui des co-citations (*cf.* fig. 2.19c et fig. 2.19e), et le réseau social fictif. Dans les deux cas, on observe deux modes dans la distribution des instants de naissance, le premier autour de $t_0 = 0$ et le second autour de $t_1 > 0$. Chacun de ces modes possède à son tour un seul mode dans la distribution des durées de vie qui lui correspond. Cette configuration indique l'existence de deux échelles différentes de vide dans les nuages de points desquels les données correspondantes sont issues, et donc une hétérogénéité dans les densités de ces derniers.

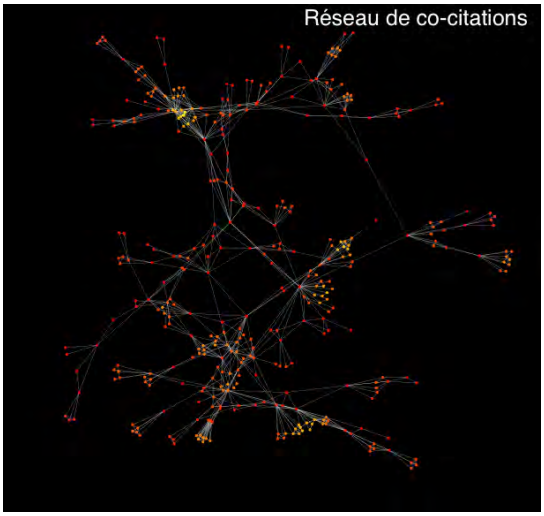
Pour finir nous montrons sur la figure 2.20 les nuages de points représentant les coefficients de variations[6] pour chacun des nœuds des réseaux étudiés, obtenus sur les 1000 positionnements calculés. Nous pouvons apercevoir que ce rapport connaît une décroissance significativement plus importante dans le cas du réseau électrique et du modèle treeCom, ce qui nous permet de supposer l'existence d'un lien entre la faible variabilité des résultats de notre algorithme, caractérisée par un faible coefficient de variations, et le nombre d'échelles distinctes de densité, caractérisé quant à lui par le nombre de pics distincts dans la distribution des instants de naissance. Cette propriété est aussi observée, bien que d'une manière moins évidente, en comparant les résultats des réseaux de co-citations, et le réseau social fictif, avec ceux des modèles d'Erdös-Renyi et de Barabási-Albert. Ces derniers n'ayant qu'un seul mode dans leurs distributions respectives des instants de naissance, on voit que leur coefficient de variations connaît une décroissance atteignant une valeur proche de 0.33 dans le cas du modèle de Barabási-Albert, tandis que ce rapport atteint environ la valeur de 0.22 dans le cas du réseau social fictif et celui des co-citations, qui possèdent quant à eux deux modes distincts dans leurs distributions respectives des instants de naissance.

[6]Le coefficient de variation est une mesure relative de la dispersion des données autour de la moyenne. Il se calcule comme le rapport entre l'écart-type et la moyenne

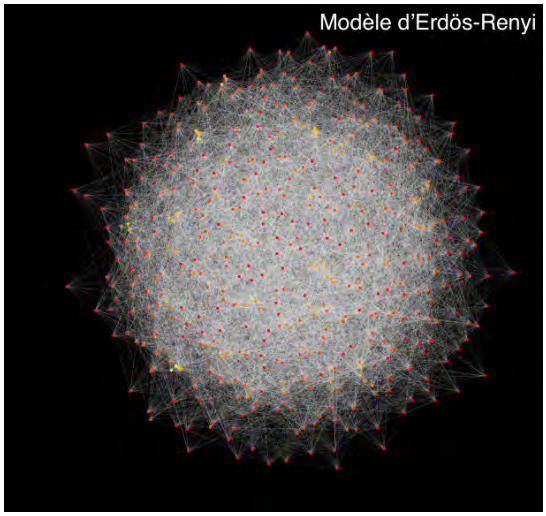


(a)

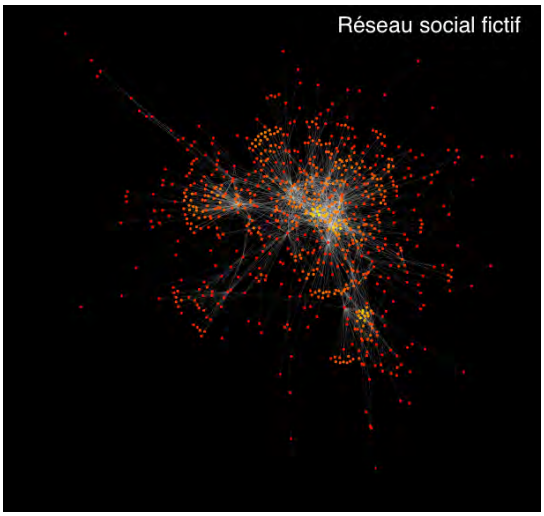
(b)



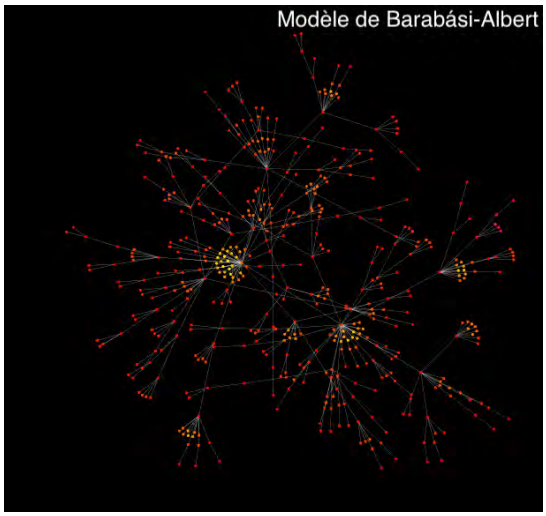
(c)



(d)

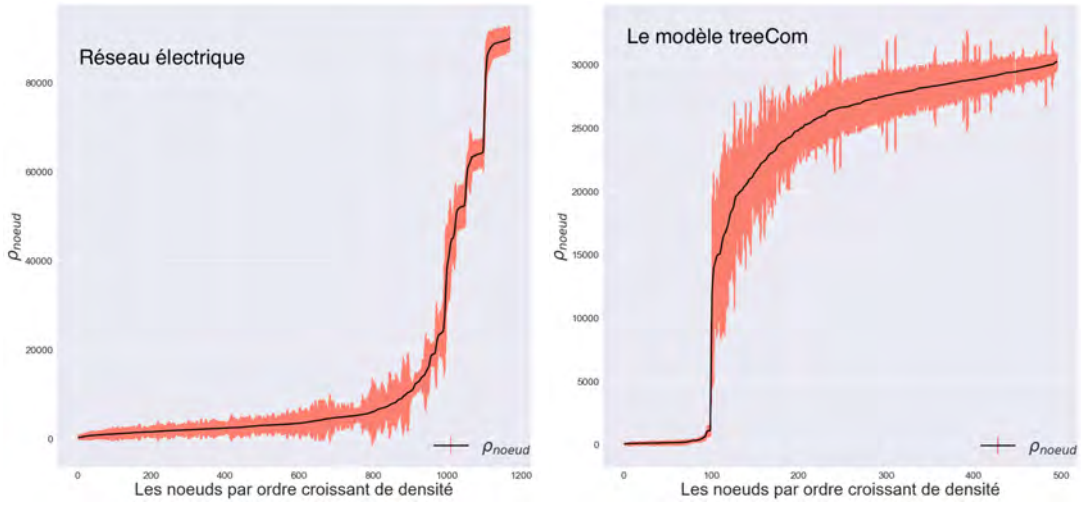


(e)

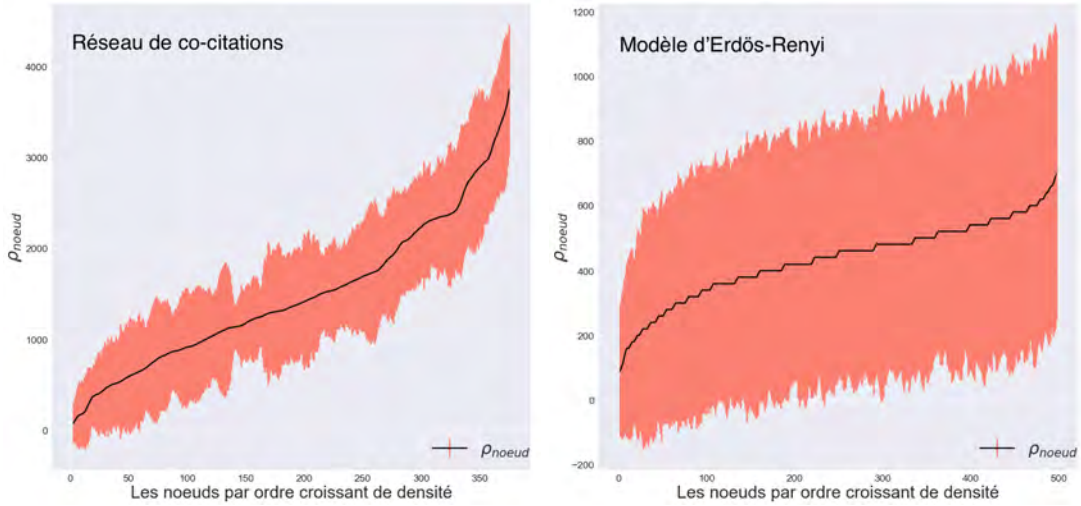


(f)

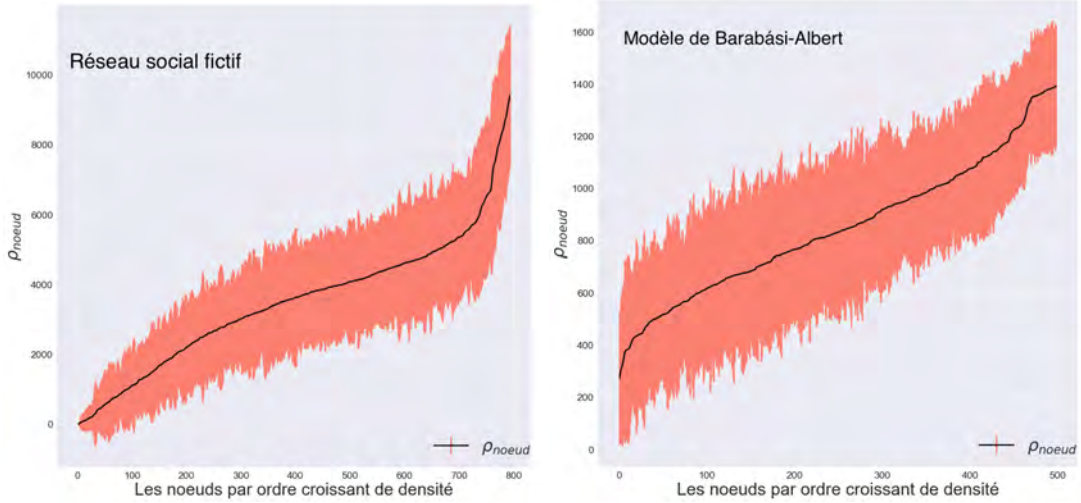
Figure 2.17: Positionnements et densités de trois réseaux réels (a),(c),(e) et trois modèles synthétiques (b),(d),(f)



(a) (b)



(c) (d)



(e) (f)

Figure 2.18: La courbe de la densité moyenne des nœuds, et son écart-type, calculée sur 1000 positionnements différents, de trois réseaux réels (a),(c),(e) et trois modèles synthétiques (b),(d),(f)

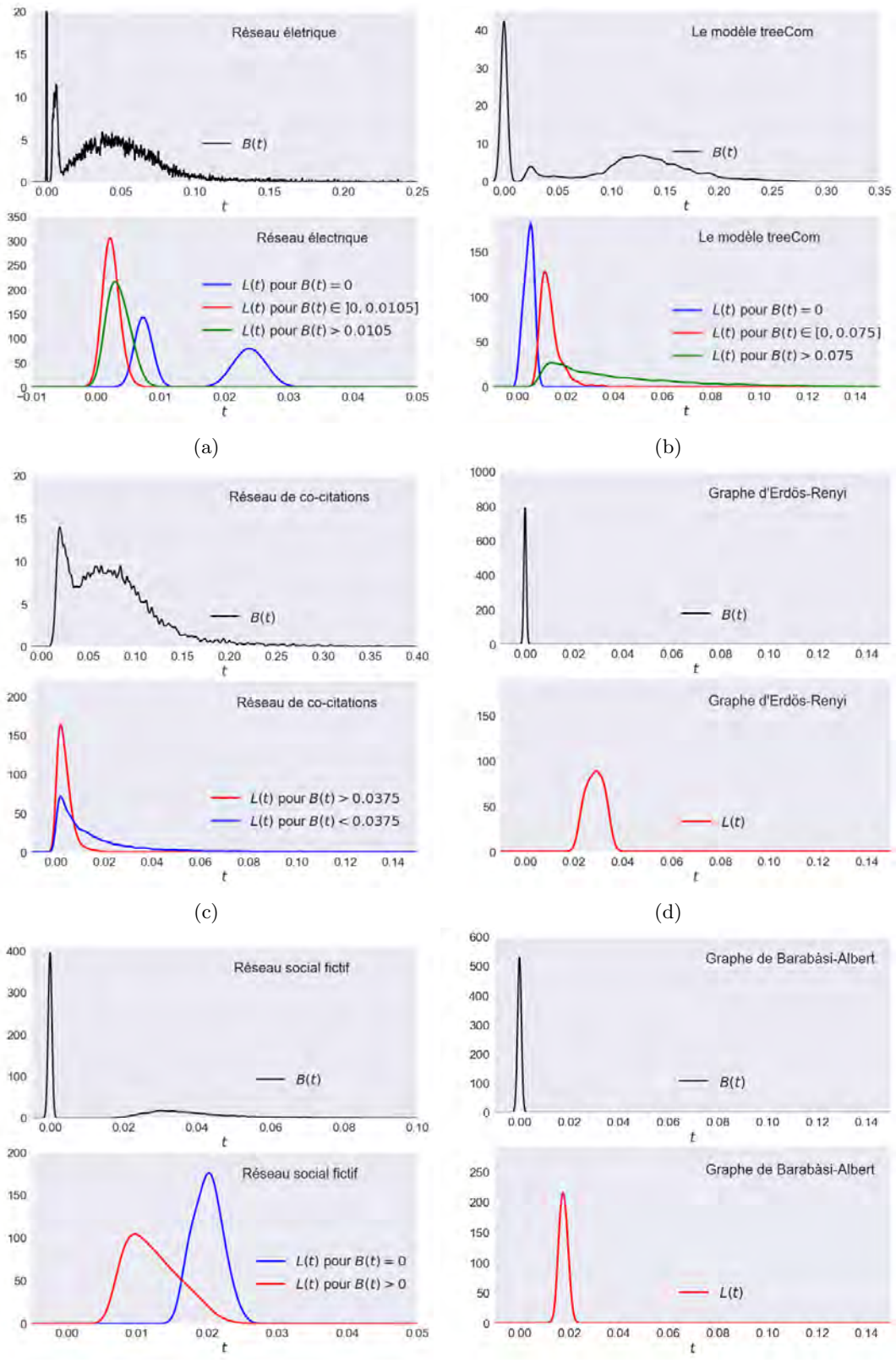


Figure 2.19: Distributions approchées des instants de naissance, et des durées de vies des cycles 1-dimensionnels obtenus sur 1000 positionnements différents de trois réseaux réels (a),(c),(e) et trois modèles synthétiques (b),(d),(f)

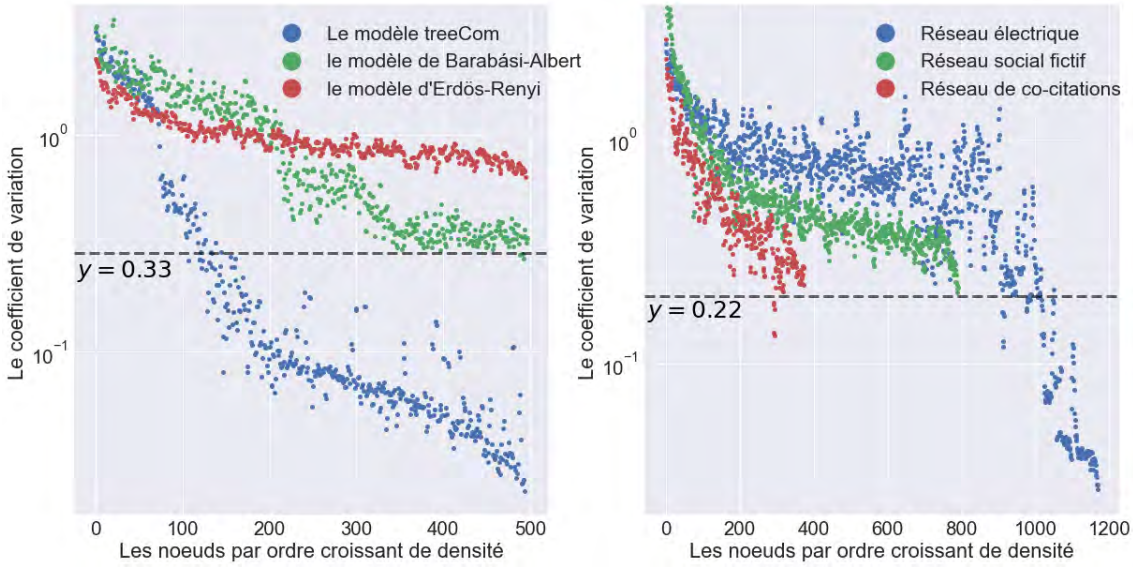


Figure 2.20: Les coefficients de variations, définis par le rapport entre l'écart-type et la densité moyenne obtenus sur les 1000 positionnements calculés pour trois réseaux synthétiques (figure de gauche) et trois réseaux réels (figure de droite). Le graphique est représenté sur une échelle semi-logarithmique.

2.7 Bilan

Nous présentons dans ce chapitre une mesure de densité appliquée aux nœuds dans les réseaux complexes, basée sur la densité spatiale de chacun d'entre eux, par l'intermédiaire d'une transformation qui plonge le réseau dans un espace métrique. Cette transformation est elle-même basée sur des analogies avec le domaine de la physique, et modélise le graphe comme un système de points en interactions, en rapport avec la structure du réseau.

La mesure de densité développée a été incorporée dans l'algorithme DBSCAN, et ceci nous a permis de passer de deux paramètres dans la version classique de ce dernier (M et ϵ) à un seul paramètre α, qui sert à délimiter l'enveloppe concave du nuage de points étudié. Les résultats montrent que notre version retravaillée de DBSCAN retrouve avec succès les clusters issus de nuages de points de différentes formes, et bien qu'ayant des critères plus sévères en termes d'identification du bruit, en comparaison avec d'autres algorithmes (*cf.* fig. 2.13), notre approche résout l'un de ses problèmes les plus importants : la capacité à détecter des clusters d'échelles différentes de densité.

Nous avons ensuite présenté puis développé une version plus générale de l'algorithme de Fruchterman et Reingold, concernant le positionnement des nœuds dans un espace métrique de dimension d (fixé à $d = 2$), qu'on utilise comme support pour définir la densité des nœuds dans un réseau. Cette approche a été analysée statistiquement, en calculant un grand nombre de positionnements différents sur le même réseau, afin d'avoir d'un côté une idée sur la variabilité de la mesure, compte tenu de l'aspect stochastique du processus de positionnement, et d'un autre côté d'établir un lien entre l'analyse topologique des données et le caractère multi-échelle de la densité spatiale d'un nuage de points, à l'aide de l'étude des distributions des instants d'apparition et des durées de vies des cycles 1-dimensionnels.

Cependant nos résultats sont à mettre en perspective avec plusieurs points importants, qui n'ont pas été mentionnés lors de cette analyse. D'abord le choix de la dimension $d = 2$ est arbitraire. Elle permet certes une visualisation facile mais n'est en rien un choix justifié mathématiquement. Il est possible qu'il existe une dimension différente de celle choisie, qui représente mieux les données étudiées. D'une manière générale, on sait juste qu'il ne faut pas que d soit trop grand, car on se retrouve confronté au fameux fléau de la dimension, qui rend incongrues les notions de distance et de densité. Par ailleurs, il ne faut pas toujours que d soit trop petit, car on pourrait finir par écraser le nuage de points dans un espace dont la dimension ne le représente pas au mieux. On ne connaît aucune technique permettant de décider du meilleur choix à prendre concernant la dimension du positionnement, et ce choix dépend des propriétés que l'on souhaite mettre en évidence et du type de réseau.

Il est aussi important de rappeler que l'aspect stochastique de l'algorithme de positionnement induit

une certaine variabilité, celle-ci peut poser problème dans le cas où il faudra se baser sur ces estimations pour sélectionner un ensemble de nœuds pour en construire un sous-réseau. En effet les propriétés structurelles des réseaux peuvent parfois être très sensibles à l'ajout et/ou la suppression des nœuds. On peut par exemple faire le choix d'étudier la structure du sous-réseau généré par les nœuds ayant une densité inférieure à 10% de la valeur maximale mesurée, mais ce réseau peut s'avérer significativement différent d'un calcul à l'autre. La raison est que la densité varie d'une exécution à l'autre de l'algorithme de positionnement. De plus, les résultats montrent que cette variabilité est plus importante quand la densité est faible (*cf.* fig. 2.20).

En somme on a montré dans ce chapitre qu'il est possible d'avoir une idée sur la structure d'un réseau complexe en nous concentrant sur le nuage de points issu de l'application d'un algorithme de positionnement. Nous en avons conclu que cette approche ne suffit pas pour autant à décrire toute la complexité que le réseau peut contenir.

Chapitre 3

Mesures déterministes : la densité topologique

Table des matières

3.1 Introduction

Après avoir vu une première approche qui consiste à estimer la densité des nœuds d'un réseau par leur densité spatiale dans un nuage de points, nous abordons ce chapitre dans le but de trouver une approche alternative. Notre but est donc de mettre au point une mesure de la densité qui corrige le défaut majeur d'un plongement géométrique. Ce défaut réside dans le fait qu'il puisse y avoir plusieurs configurations qui réalisent l'optimum de la fonction de coût. De fait, il est difficile de savoir si une configuration est meilleure qu'une autre. Il faudra donc se concentrer uniquement sur les propriétés structurelles des réseaux, tout en s'assurant que cette mesure possède un certain nombre de propriétés fixées au préalable, et caractéristiques de la notion de densité. Ceci nous fixe un cahier des charges qu'il nous faudra respecter, et dont nous allons établir les différents points.

Tout d'abord, il faut préciser que le terme densité renvoie à la densité en connectivité des nœuds dans le réseau. Nous voulons dire par là que plus la densité en liens d'un cluster est élevée, plus la valeur de la densité des nœuds qui s'y trouvent est élevée. Ensuite, une caractéristique très importante qu'une mesure de densité quelconque doit posséder est son caractère local. Il est largement admis que la densité d'un réseau s'obtient par le rapport entre le nombre de liens qu'il contient, et le nombre de paire de nœuds qu'il y a dans le réseau. Ceci admet implicitement (en considérant que la densité moyenne d'un réseau est la moyenne des densités de chaque nœud [92]), que la densité d'un nœud s'obtient par le rapport entre son degré et le nombre total de voisins auxquels il peut être rattaché. On peut cependant facilement se rendre compte des faiblesses d'une telle mesure, car même si le degré joue un rôle important dans la caractérisation de la densité, il ne révèle rien de plus que le nombre de voisins que possède un nœud. Il faudrait pour être plus précis, inclure une connaissance de l'état du voisinage de ce dernier. Quelles sont les informations d'intérêt que l'on peut en tirer ? À titre d'exemple, nous pouvons considérer un nœud qui a un grand nombre de voisins, sans que ces voisins ne soient reliés deux à deux. Selon la mesure précédemment évoquée, ce nœud va avoir une forte densité, alors qu'en principe il ne devrait pas. La raison est que si on isole le sous-réseau généré par ce nœud et son voisinage, on retrouve une étoile sans aucune fermeture transitive. C'est pour cela qu'il faut étendre la mesure de la densité à quelque chose de plus que le degré uniquement. À l'inverse, une mesure comme le coefficient de clustering nous informe sur la cohésion du voisinage d'un nœud, mais ne prend pas en compte la taille de celui-ci. Par exemple, un nœud dont l'ego-graphe génère un triangle à un coefficient de clustering égal à 1.
Enfin comme pour toute mesure, il serait préférable que celle-ci ne dépende d'aucun paramètre, de telle sorte que celui-ci n'ait pas à être fixé arbitrairement.

Ce chapitre est divisé en quatre parties : la première est dédiée à l'introduction de deux mesures de densité, basées toutes les deux sur les propriétés des nœuds. La seconde à l'introduction d'une autre mesure s'appuyant quant à elle sur les propriétés des liens. Cette dernière sera d'ailleurs celle que l'on retiendra par la suite, pour des raisons que l'on donne plus loin dans ce chapitre. Ensuite la troisième partie est consacrée à l'introduction de deux modèles de réseaux synthétiques : le premier a été construit pour répondre aux besoins des différents tests effectués au cours de cette thèse, et dont le positionnement a déjà servi dans le chapitre précédent, et le second est le modèle bien connu des bloques stochastiques. Enfin la quatrième partie comporte des tests sur les réseaux synthétiques qu'on vient d'évoquer, ainsi qu'un réseau réel dont les données sont disponibles en ligne.

3.2 Mesures basées sur les propriétés des nœuds

Nous allons dans cette partie présenter deux mesures qui s'appuient sur les propriétés des nœuds pour en estimer la densité. Nous nous concentrons ici sur la description des mesures, ainsi que sur les idées qui ont motivé leur mise au point. Les tests et leurs résultats seront quant à eux présentés plus loin dans ce chapitre.

3.2.1 Extension du coefficient de clustering : la mesure γ

Le coefficient de clustering est, rappelons-le, un indicateur sur le degré de cohésion du voisinage d'un nœud. C'est une mesure moins locale que le degré, car elle est le résultat d'un calcul qui ne se concentre pas uniquement sur le nombre de voisins du nœud en question, mais aussi sur l'état de son voisinage à travers la proportion de liens reliant ses voisins. Il est de ce fait possible d'exploiter ce coefficient

pour l'utiliser comme élément central d'une mesure de densité, à la place du degré comme ça a déjà été expliqué.

Nous choisissons cependant de ne pas nous restreindre au voisinage direct du nœud examiné. La nouvelle mesure que nous proposons étend le coefficient de clustering aux voisins successifs, en pondérant l'apport de chaque sous-ensemble de voisins proportionnellement à leur distance du nœud évalué. Cette pondération est effectuée par une fonction décroissante choisie empiriquement, qui a pour rôle de donner un poids plus grand aux voisins les plus proches.

La raison d'un tel choix est qu'en effet, pour qu'un nœud soit dans une zone dense, il ne lui suffit pas toujours d'avoir un grand coefficient de clustering, il faut que ce soit aussi le cas pour ses voisins, les voisins de ses voisins dans une moindre mesure, etc. Nous avons ainsi une vision centrée autour du nœud que nous évaluons, et qui s'étend petit à petit à tous les nœuds du graphe.

Soit $G = \{V, E\}$ un graphe non dirigé et non pondéré, et $u \in V$ un nœud de ce graphe. Nous appelons $\zeta_i(u)$ l'ensemble des nœuds dans V à distance i de u.

$$\zeta_i(u) = \{v \in V | d(u, v) = i\} \tag{3.1}$$

par convention nous posons $\zeta_0(u) = \{u\}$.

Nous définissons $\gamma(u)$, la mesure de densité du nœud u basée sur le coefficient de clustering comme :

$$\gamma(u) = \sum_{i=0}^{L} f(i) \cdot \frac{1}{|\zeta_i(u)|} \sum_{v \in \zeta_i(u)} C_v \tag{3.2}$$

où L désigne le diamètre du graphe G, C_v le coefficient de clustering d'un nœud v, et $f(i)$ le poids attribué à l'ensemble de nœuds $\zeta_i(u)$ se trouvant à distance i de u par une fonction décroissante f. Le choix de cette fonction est arbitraire, et on doit faire attention à ce que sa décroissance ne soit ni trop rapide, auquel cas $\gamma(u)$ ne sera pas très différent du coefficient de son clusteing, ni trop faible, auquel cas les différents ensembles de voisins $\zeta_i(u)$ se retrouvent avec des poids de valeurs proches. Cette dernière configuration résulterait en une distribution de γ sans variation significative d'un nœud à l'autre du réseau. Nous choisissons $f(i) = \frac{1}{i+1}$ lors des tests effectués plus loin dans ce chapitre.

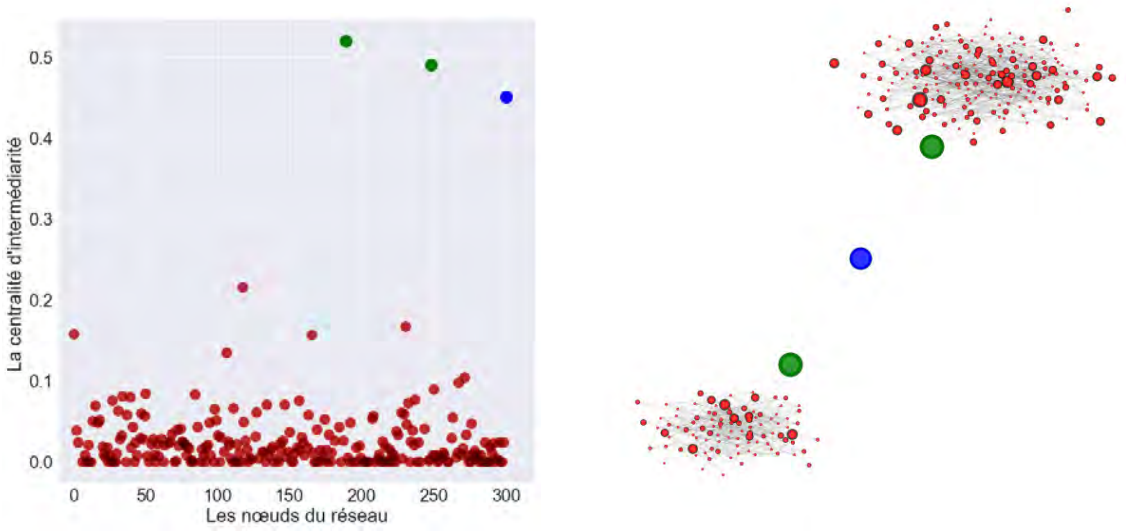
Cette mesure présente l'avantage de ne pas dépendre explicitement du degré. On peut imaginer le cas d'un réseau constitué de deux clusters de tailles différentes dont le plus petit présente une plus grande densité en connectivité en liens. Cette mesure fait abstraction de ce genre de détails et retourne une densité moyenne plus importante pour les nœuds du cluster les plus densément liés, car les coefficients de clustering des nœuds y sont plus élevés [1].

3.2.2 Betweenness pondérée : la mesure β

Pour cette seconde partie, nous établissons une mesure qui est basée sur la centralité d'intermédiarité (qu'on appellera aussi betweenness), en suivant le même raisonnement que pour la mesure γ, c'est à dire en l'adaptant pour qu'elle rende compte des propriétés de la densité évoquées dans l'introduction de ce chapitre. Nous pouvons commencer par montrer les résultats d'un test simple, en calculant la centralité de betweenness sur un réseau fabriqué à partir de deux grandes composantes, chacune étant un réseau d'Erdös-Rényi de tailles respectives 100 et 200 nœuds, avec un paramètre $P = 0.05$. Ces deux composantes sont reliées par un pont composé de deux liens et un nœud, comme il est montré sur la figure 3.1.

Nous pouvons voir sur cette figure que certains nœuds des clusters (représentés en vert sur la figure 3.1), ont une centralité de betweenness supérieure à celle du nœud intermédiaire, qui se situe pourtant entre les deux clusters (représenté en bleu sur la figure 3.1). Ceci est dû au fait que pour passer d'un cluster à l'autre les chemins doivent forcément passer par le nœud se situant à l'intermédiaire des deux clusters, mais aussi par les nœuds qui relient celui-ci aux clusters. Ces nœuds sont au nombre de deux et ont par conséquent au moins la même centralité que le nœud intermédiaire. Il se trouve que ces derniers sont aussi contenus dans des géodésiques reliant les paires de nœuds qui se trouvent au sein du même cluster, ce qui n'est pas le cas du nœud intermédiaire. Ceci est à l'origine de la différence de centralité entre le nœud intermédiaire et les deux nœuds le reliant aux clusters, comme on peut le voir sur la figure 3.1b.

[1] Sous réserve que la fonction de pondération décroisse suffisamment vite



(a) (b)

Figure 3.1: Les valeurs de la centralité de betweenness, (a) on remarque qu'il y a trois points dont les valeurs sont nettement supérieures à celles des autres, (b) la taille de chaque nœud est proportionnelle à sa centralité.

On vient de voir sur un exemple simple que la centralité de betweenness n'est pas toujours un outil adéquat pour mesurer de la densité. Il n'est cependant pas tout à fait juste de dire qu'elle est complètement inadaptée, car elle attribue quand même un score relativement élevé aux nœuds se trouvant entre les clusters, même si ce n'est pas le plus élevé parmi les nœuds du réseau. Pour corriger ce défaut, nous pouvons faire le choix de ne pas compter les chemins les plus courts qui ont une longueur inférieure ou égale à 2, comme dans [70]. Ce choix permettrait d'attribuer le même score aux deux nœuds représentés en vert sur la figure 3.1b, et celui (en bleu) qui sert de pont entre les deux clusters.

Nous choisissons finalement d'adopter une approche plus générale, que nous décrivons ci-dessous : soit

$$f_i(l) = \begin{cases} 1 & \text{si} \quad l \geq i \\ 0 & \text{sinon} \end{cases} \tag{3.3}$$

nous pouvons donc écrire la betwenness modifée dans [70] comme :

$$B^*(u) = \sum_{v \neq w \neq u} \frac{\sigma(v, w|u)}{\sigma(v, w)} \cdot f_2(l_{v,w}) \tag{3.4}$$

où $l_{v,w}$ décrit ici la valeur du chemin le plus court entre les nœuds v et w. D'une manière plus générale nous définissons notre mesure de la densité, basée sur la centralité de betweenness comme :

$$\beta(u) = \sum_{i=2}^{L} \sum_{v \neq w \neq u} \frac{\sigma(v, w|u)}{\sigma(v, w)} \cdot f_i(l_{v,w}) \tag{3.5}$$

où L désigne le diamètre du réseau. Nous pouvons ensuite facilement vérifier que la somme décrite ci-dessus se transforme en :

$$\beta(u) = \sum_{v \neq w \neq u} \frac{\sigma(v, w|u)}{\sigma(v, w)} \cdot (l_{v,w} - 1). \tag{3.6}$$

Ainsi cette mesure est équivalente à une centralité de betweenness à laquelle on rajoute un terme qui attribue un poids à chaque paire de nœuds comptée dans la somme, qui est proportionnel à la valeur du chemin le plus court les séparant. Ceci a pour but d'attribuer des scores faibles aux nœuds qui sont impliqués dans des géodésiques de petites longueurs, qui est une caractéristique que l'on retrouve chez les nœuds se situant dans des clusters. En revanche, cette mesure attribue des scores élevés aux nœuds

qui sont impliqués dans des géodésiques de valeurs élevées, comme les nœuds qui servent de pont aux clusters, qui sont de faible densité.

Nous pouvons noter qu'à l'inverse de la mesure γ qu'on a décrit dans la partie précédente, la mesure β attribue des scores faibles aux nœuds se situant dans les zones denses, mais il est possible d'inverser la mesure β pour qu'elle soit cohérente avec la mesure γ. Il est aussi important de rappeler que ces deux mesures demandent le choix empirique d'une fonction de pondération, dans le cas de γ, ce choix est explicitement décrit par la fonction décroissante $f(i)$, et dans le cas de β, nous pouvons choisir une autre fonction $f_i(l)$ dans eq. (3.3), ce qui donnerait une définition de β différente de celle exprimée par eq. (3.6).

3.3 Mesures basées sur les propriétés des arêtes

On présente maintenant une troisième mesure de densité, issue d'une approche différente de celles qui ont permis de mettre au point γ et β. Au lieu de se concentrer sur les propriétés du nœud concerné par la mesure, on évaluera d'abord l'une après l'autre les arêtes qui lui sont rattachées. Pour cela, on va d'abord définir une mesure sur les liens, qui nous servira pour calculer la densité des nœuds.

3.3.1 L'indice $\omega_{u,v}$

On considère qu'un lien relie deux nœuds de forte densité si les deux propriétés suivantes sont satisfaites en même temps :

1. les deux nœuds à son extrémité sont de degrés élevés
2. les deux nœuds à son extrémité partagent un nombre élevé de voisins en commun.

on définit le score de densité relatif à chaque arête, qu'on appelle $\omega_{u,v}$ comme :

$$\omega_{u,v} = A_{u,v} \cdot k_u \cdot k_v \cdot S_{DCS}(N_u, N_v) \tag{3.7}$$

où

- $A_{u,v} = \frac{a_{u,v}}{(N-1)^2(N-2)}$ est une constante de normalisation, et $a_{u,v}$ vaut 1 si la paire (u, v) est reliée dans le graphe, 0 sinon.
- k_u et k_v les degrés respectifs des nœuds u et v
- N_u et N_v, les ensembles de nœuds représentant les voisinages respectifs de u et de v.
- $S_{DCS}(N_u, N_v) = 2 \cdot \frac{|N_u \cap N_v|}{k_u + k_v}$ l'indice de similarité de Dice-Czekanowski-Sørensen [40] entre N_u et N_v

Il est alors facile de vérifier que l'on peut réécrire $\omega_{u,v}$ comme :

$$\omega_{u,v} = A_{u,v} \cdot |N_u \cap N_v| \cdot H\big(k_u, k_v\big) \tag{3.8}$$

avec $H\big(k_u, k_v\big) = \frac{2 \cdot k_u k_v}{k_u + k_v}$ la moyenne harmonique des degrés k_u et k_v.

Notons ces deux propriétés de la moyenne harmonique $H(a, b)$:

1. $H(a, a) = a$ la moyenne harmonique de deux quantités égales est cette même quantité.
2. si $a \gg b$ alors $H(a, b) \approx b$

Cette moyenne harmonique rend donc faible la valeur attribuée à une arête dont l'une des extrémités est un nœud de fort degré, et la deuxième un nœud de faible degré. D'un autre côté, le terme $|N_u \cap N_v|$ rend faible la contribution d'une arête reliant deux nœuds qui ne partagent pas un grand nombre de voisins en commun. On a finalement une mesure qui satisfait les conditions évoquées plus haut.

3.3.2 La mesure δ

À l'aide de ω nous définissons la mesure δ de la densité comme la somme des ω attribuées à chacun des liens qui sont rattachés au nœud mesuré :

$$\delta(u) = \sum_{v \in V} w_{u,v} \tag{3.9}$$

par convention, nous mettons $\delta(u) = 0$ si $k_u = 0$.

De plus, le poids $\omega_{u,v}$ d'un lien est contenu dans l'intervalle $[0; \frac{1}{N-1}]$ et est égal à 0 lorsque les nœuds u et v n'ont pas de voisins en commun. En revanche $w_{u,v} = \frac{1}{N-1}$ si, et seulement si, u et v sont tous les deux voisins avec tous les autres nœuds du graphe, c'est-à-dire $N_u \setminus \{v\} = N_v \setminus \{u\} = V \setminus \{u, v\}$. Ainsi, $\delta(u) = 1$ si, et seulement si, chaque lien incident à u a un poids égal à $\frac{1}{N-1}$, ce qui n'est le cas que lorsque le graphe est une clique. En général, nous avons $0 \leq \delta(u) \leq 1$.

Soit maintenant m le nombre de liens dans le réseau, alors nous obtenons la valeur moyenne de ω sur tous les liens du réseau comme :

$$\overline{\omega} = \frac{1}{2} \cdot \frac{1}{m} \sum_{u,v} \omega_{u,v}$$

le terme $\frac{1}{2}$ vient du fait que chaque lien est compté deux fois dans la somme.
D'un autre côté nous avons aussi le degré moyen qui s'écrit :

$$\overline{k} = \frac{1}{N} \sum_{u \in V} k_u = \frac{2m}{N}$$

Nous pouvons alors obtenir la valeur moyenne de δ comme une expression faisant intervenir les termes $\overline{k}$ et $\overline{\omega}$:

$$\overline{\delta} = \frac{1}{N} \sum_{u \in V} \delta(u)$$

que nous réécrivons à l'aide de eq. (3.9) et eq. (3.8) :

$$\overline{\delta} = \frac{1}{N} \sum_{u \in V} \sum_{v \in V} \omega_{u,v} = \frac{1}{N} \sum_{u,v} \omega_{u,v}$$

ce qui donne finalement :

$$\overline{\delta} = \overline{\omega} \cdot \overline{k} \tag{3.10}$$

Nous obtenons finalement que la moyenne de δ sur tous les nœuds du réseau est le produit entre la valeur moyenne de ω et celle du degré. Cette propriété est importante et sera exploité lors de l'utilisation de δ dans une application qu'on décrira dans le prochain chapitre.

3.4 Jeux de données

Nous allons maintenant nous occuper de tester les trois mesures γ, β et δ sur des réseaux synthétiques et réels pour en dégager les principales propriétés. Mais d'abord, il est important de bien décrire les réseaux sur lesquels ces tests sont effectués. Nous allons donc décrire deux modèles, le premier ayant été développé pour les besoins de cette étude, et le second qui est un modèle bien connu : le modèle de blocs stochastiques.

3.4.1 Le modèle treeCom

Nous introduisons ici un modèle jouet, qui a été créé dans le but de tester les mesures décrites précédemment, sur un graphe ayant une structure simple qui permet sans difficulté d'interpréter les résultats. Sa simplicité réside dans le fait qu'il soit composé d'une partie dense et d'une partie non dense aisément identifiables. La première induit comme sous-réseau plusieurs clusters séparés les un des autre, avec une forte connectivité à l'intérieur de chacun, et la seconde induit un arbre. On peut ainsi établir un premier niveau de tests, permettant de voir les propriétés de base des mesures qui ont été présentées dans les sections précédentes.

Nous décrivons ci-dessous le modèle treeCom composé de plusieurs blocs, chacun étant généré séparément à travers un modèle d'Erdös-Rényi, avec son propre paramètre P. Nous connectons ensuite les nœuds de l'arbre à chacun des blocs avec une certaine probabilité. Les paramètres du modèles sont :

- N_b : le nombre de blocs b_i du modèle
- N_i : le nombre de nœuds dans le bloc b_i, $i \in \{1, 2, .., N_b\}$
- P_i : la probabilité de créer une arête entre chaque paire de nœuds à l'intérieur du même bloc b_i
- N_t : le nombre de nœuds contenus dans l'arbre (la partie non dense du modèle)
- p_i^t la probabilité de créer une arête entre un nœud de l'arbre et un nœud du bloc b_i.

Cela génère un réseau avec un nombre de $\sum_{i=1}^{N_b} N_i + N_t$ nœuds. Nous générons d'abord la partie non dense, qui peut par exemple être un arbre aléatoire (dans les exemples que l'on verra plus tard, il est obtenu en calculant l'arbre couvrant d'un réseau aléatoire d'Erdös-Rényi) dont le nombre de nœuds est N_t. Ensuite, nous générons chacun des N_b blocs avec ses paramètres correspondants, et finalement nous connectons chaque nœud de l'arbre à chaque nœud du bloc b_i avec la probabilité p_i^t.

Il est plus intéressant pour nous d'effectuer nos tests sur des réseaux connexes, il faut dès lors choisir les paramètres du modèle pour que le graphe qui en résulte le soit. Pour s'en assurer il suffit de choisir p_i^t de sorte que le nombre moyen de liens, appelons le M, qui relient l'arbre aux blocs soit très grand par rapport à 1. Il est facile de vérifier que

$$M = \sum_{i=1}^{N_b} N_t \cdot N_i \cdot p_i^t$$

Il suffit alors de prendre $p_i^t \gg \frac{1}{N_t \cdot N_i}$ pour que le résultat du réseau jouet soit connexe.

3.4.2 Le modèle de blocs stochastiques (SBM)

Description du modèle

Soit $G = (V, E)$ un réseau, on appelle G un modèle de blocs stochastiques [65] s'il est composé de r blocs conformément à une certaine configuration. Il est caractérisé par les paramètres suivants :

- une partition de l'ensemble de nœuds $V = \{1, \ldots, N\}$ en sous-ensembles disjoints $b_1, \ldots, b_r$ appelés blocs, ou communautés
- une matrice P de taille $r \times r$ symétrique, contenant les probabilités P_{ij} qu'un nœud quelconque du bloc b_i soit relié à un nœud quelconque du bloc b_j
- Le nombre de nœuds $N = \sum_{i=1}^{r} n_i$, avec n_i le nombre de nœuds à l'intérieur du bloc b_i

Un cas particulier du modèle de blocs stochastiques est le modèle à partition plantée, noté PPM pour "planted partition model" en anglais. Dans ce modèle, la matrice de probabilité est remplie avec un paramètre p constant sur sa diagonale, et un autre paramètre q constant sur les termes non diagonaux. De surcroît, le nombre de nœuds à l'interieur de chacun des blocs est le même et est noté n. Les calculs ainsi que les tests concernant le modèle de blocs stochastiques rencontrés plus bas sont en particulier effectués sur un PPM.

Pour simplifier les calculs sur ce modèle, nous ajoutons une notation pour l'architecture du PPM, en utilisant la matrice B de taille $r \times r$, qui peut être vue comme une matrice d'adjacence des blocs, définie par $B_{i,j} = 0$ si les nœuds du bloc b_i ont une probabilité non nulle ($q > 0$) d'être liés à ceux du bloc b_j et 0 sinon. La matrice de probabilité peut être notée comme suit

$$P = p \cdot \mathbb{1}_{r \times r} + q \cdot B$$

où $\mathbb{1}$ est le symbole de la matrice identité.

Ainsi, deux nœuds qui sont dans le même bloc partagent une arête avec une probabilité égale à p, alors que deux nœuds qui sont dans dans des blocs différents partagent une arête avec une probabilité q, si leurs blocs respectifs sont adjacents. Nous distinguons deux cas particuliers, le premier où $p > q$ qui

est le cas assortatif, par opposition au second où $p < q$ représentant le cas disassortatif [2]. Comme les nœuds d'un même bloc partagent en moyenne les mêmes propriétés, ce qui est calculé dans les formules suivantes se réfère aux valeurs moyennes. ainsi, nous notons k_i le degré moyen d'un nœud dans le bloc b_i, et K_i le nombre de blocs adjacents au bloc b_i.

L'avantage avec l'utilisation d'un tel modèle, est qu'il est possible d'obtenir les expressions analytiques des valeurs moyennes de certaines mesures comme le degré, où encore le coefficient de clustering et le δ. Il n'est cependant pas évident d'obtenir les expression des mesures β et γ. Pour la première, c'est dû au fait qu'elle soit dépendante du nombre de chemins les plus courts (soit ceux traversant un nœud donné, soit la totalité des chemins les plus courts) entre toutes les paires de nœuds. Il n'est en général pas possible de lui trouver une expression analytique, sauf dans les cas où les réseaux sont triviaux.
Pour ce qui est de γ, la difficulté vient du fait qu'il faille identifier l'ensemble de voisins à distance l de chaque nœud. Cet ensemble dépend à la fois de l'architecture du PPM donné par la matrice B, mais aussi de la valeur des paramètres p et q (le voisinage quand p et q valent tous les deux 1 n'est pas le même que celui où ils sont de faibles valeurs), rendant impossible l'obtention d'une expression de γ dans un cadre général.

Le calcul formel

Pour rappel, nous nous restreignons au cas où les blocs sont tous composés du même nombre de nœuds, et les paramètres p et q remplissent respectivement les éléments diagonaux et non diagonaux de la matrice de probabilités P.

Le degré moyen

Pour le degré, il est facile de se convaincre que chaque nœud aura des voisins à l'intérieur du même bloc, ainsi que des voisins dans les blocs adjacents suivant l'expression :

$$k_i = n \cdot (p + K_i \cdot q). \tag{3.11}$$

Ainsi l'apport des nœuds internes au bloc est le même pour tous les nœuds, mais plus celui-ci possède de blocs adjacents, plus le degré de ses nœuds augmente.

Le coefficient de clustering

Nous savons que le coefficient de clustering est obtenu par le rapport entre le nombre de triangles fermés dans le voisinage du nœud considéré, et le nombre de paires issues de ce voisinage. Il est donc nécessaire de pouvoir quantifier ces deux grandeurs en fonction des paramètres du modèle. Pour le nombre de paires formées, il est facile de voir qu'il est égal à $\frac{k_i \cdot (k_i - 1)}{2}$, et on connaît l'expression de k_i (*cf.* eq. (3.11)). Nous ajoutons aussi quelques approximations, en nous limitant au régime où $p \gg \frac{1}{n}$ et $n \gg 1$. Ceci nous permet de considérer que $\frac{k_i \cdot (k_i - 1)}{2} \approx \frac{k_i^2}{2}$.

Il reste alors à donner l'expression du nombre moyen de triangles fermés dans lesquels un nœud est impliqué. Le problème est décomposé en identifiant les différents types de triplets : d'abord il y a les triangles composés d'un nœud et de deux voisins du même bloc b_i, leur nombre moyen est $\frac{n^2}{2} \cdot p^3$. À ceux-ci s'ajoutent les triangles dont les trois nœuds sont dans deux blocs différents b_i et b_j. Leur nombre dépend du degré K_i du bloc b_i et est égal à $3 \cdot \frac{n^2}{2} \cdot K_i p \cdot q^2$ (ici le facteur trois désigne le nombre de permutations selon lesquelles il est possible de distribuer trois nœuds sur deux blocs). Le dernier est le terme qui détermine le nombre de triangles dont les trois nœuds sont dans trois blocs différents. Leur nombre dépend du nombre de triades de blocs dans lesquels se trouve le bloc b_i. En effet, on obtient un nombre égal à ce nombre de triades multiplié par $n^2 \cdot q^3$. On note ce terme μ_i et il s'exprime ainsi :

$$\mu_i = \frac{1}{2} \cdot |\{(j,k) : B_{i,j} = 1, B_{i,k} = 1, B_{j,k} = 1\}|$$

[2] Nous précisons que c'est ici les noms donnés à chacun des deux régimes, mais que ça n'est pas directement en rapport avec les propriétés d'assortativité et de dissortativité introduites dans le premier chapitre

En combinant ces trois termes, on obtient pour un nœud i dans un bloc b_i :

$$C(i \in b_i) = \frac{p^3 + 3K_i \cdot pq^2 + 2q^3 \cdot \mu_i}{(p + K_i \cdot q)^2}. \tag{3.12}$$

On remarque ici que l'expression du coefficient de clustering dépend nettement (en comparaison avec l'expression du degré donnée précédemment) de l'architecture du PPM. Celle-ci est donnée par la matrice B d'adjacence des blocs, car non seulement le terme désignant le "degré" K_i de chaque bloc apparaît, mais aussi le terme μ_i qui est proportionnel au nombre de triangles dans le graphe représenté par la matrice B. Cette matrice dépend directement de l'architecture que l'on donne au PPM.

La mesure δ

Nous procédons de la même façon pour la mesure du δ, en décomposant le problème en plusieurs sous-problèmes. La différence entre δ et le coefficient de clustering est que ce dernier est calculé directement sur les nœuds, alors que δ est obtenu comme la somme des poids des liens attachés à chaque nœud. Il faut dès lors identifier pour un PPM les différents types de liens, puis faire la somme de leurs poids pour chaque nœud en fonction de l'architecture du PPM. On rappelle l'expression

$$\omega_{i,j} = |N(i) \cap N(j)| \cdot H(k_i, k_j)$$

qui donne le poids d'un lien (i, j) comme le produit du nombre de voisins communs aux nœuds i et j, et de la moyenne harmonique de leurs degrés respectifs. Ces deux quantités doivent donc être exprimées selon les différents cas que l'on peut identifier.

Liens (i, j) ayant les extrémités dans le même bloc

Dans cette configuration, nous avons des voisins communs qui peuvent être dans le même bloc que les nœuds i et j dont le nombre moyen est (quand n est assez grand) égal à $n \cdot p^2$, nous avons aussi des voisins communs qui appartiennent à un autre bloc, le nombre moyen de ceux-ci est égal à $n \cdot q^2 \cdot K_i$. Dans ce cas, nous pouvons aussi donner une expression pour le terme :

$$H(k_i, k_j) = n \cdot (p + K_i \cdot q)$$

où K_i est le degré du bloc dans lequel les nœuds i et j sont situés. En agrégeant tous les termes précédents, on obtient une expression pour $\omega_{i,j}$ soit :

$$\omega_{i,j} \approx \frac{1}{n \cdot r^3} \cdot (p^2 + K_i \cdot q^2) \cdot (p + K_i \cdot q). \tag{3.13}$$

L'approximation est effectuée sur le facteur de normalisation $A_{u,v}$ de eq. (3.7)
$A_{u,v} = \frac{a_{u,v}}{(N-1)^2(N-2)} \approx \frac{a_{u,v}}{N^3} = \frac{a_{u,v}}{n^3 \cdot r^3}$

Liens (i, j) ayant les extrémités dans deux blocs différents

Maintenant nous traitons le cas où les deux nœuds ne sont pas dans le même bloc. Ici nous avons des voisins communs qui peuvent être dans l'un ou l'autre des blocs contenant les deux extrémités, car il peut s'agir d'un troisième bloc auquel n'appartiennent ni i ni j. Dans le premier cas (voisins dans l'un ou l'autre des blocs contenant les extrémités) nous avons :

$$| N(i) \cap N(j) | = 2 \cdot n \cdot p \cdot q$$

dans le cas où les voisins communs se trouvent dans un troisième bloc :

$$| N(i) \cap N(j) | = n \cdot q^2 \cdot \Gamma_{i,j}$$

avec $\Gamma_{i,j}$ le nombre de "blocs voisins communs" entre le bloc b_i et b_j qui contiennent respectivement i et j. Les degrés moyens sont :

$$k_i = n(p + q \cdot K_i) \textbf{ et } k_j = n(p + q \cdot K_j)$$

il s'ensuit que

$$\omega_{i,j} \approx \frac{1}{n \cdot r^3} \cdot (2 \cdot p \cdot q + \Gamma_{i,j} \cdot q^2) \cdot \frac{2(p + K_i \cdot q)(p + K_j \cdot q)}{(p + K_i \cdot q) + (p + K_j \cdot q)} \tag{3.14}$$

Le calcul de δ

Finalement, il suffit de combiner ces termes pour obtenir le δ décrit dans l'équation eq. (3.9), en séparant les voisins comme dans les deux parties précédentes, et en estimant le nombre de paires qui répondent à ces critères. Par exemple, la somme sur les voisins appartenant au même groupe devient alors l'expression obtenue dans eq. (3.13) multipliée par le nombre de voisins qu'il a dans son propre bloc. Il est facile de se convaincre que le nombre moyen de voisins que possède un nœud dans son propre bloc est égal à $n \cdot p$, et que le nombre de voisins qui proviennent d'un autre bloc b_j est $n \cdot q \cdot B_{ij}$ (on suppose que i est le bloc dans lequel le nœud évalué est situé).

En collant toutes les pièces on trouve l'expression suivante pour le nœud i du bloc b_i :

$$\boldsymbol{\delta}(i)_{i \in b_i} = \{p(p^2 + K_i q^2)(p + K_i q) + 2q \sum_j \frac{B_{ij}(2pq + \Gamma_{i,j} \cdot q^2)(p + K_i q)(p + K_j q)}{2p + (K_i + K_j)q}\} \frac{1}{r^3}. \tag{3.15}$$

Le cas particulier des architectures en forme d'arbre

On verra par la suite un exemple de calcul effectué sur un PPM dont l'architecture est celle d'un arbre. Ce choix n'est pas anodin car les arbres ne contiennent pas de cycles, ce qui permet de simplifier les expressions du coefficient de clustering (*cf.* eq. (3.12) avec $\mu_i = 0$) et de δ (*cf.* eq. (3.15) avec $\Gamma_{ij} = 0$) dont nous donnons les résultats :

$$C_{i \in b_i} = \frac{p^3 + 3K_i \cdot pq^2}{(p + K_i \cdot q)^2} \tag{3.16}$$

$$\boldsymbol{\delta}(i)_{i \in b_i} = \{p(p^2 + K_i q^2)(p + K_i q) + 2q \sum_j \frac{B_{ij} \cdot 2pq(p + K_i q)(p + K_j q)}{2p + (K_i + K_j)q}\} \frac{1}{r^3} \tag{3.17}$$

On remarque que dans ce cas particulier, on obtient une expression du coefficient de clustering qui ne dépend que du degré K_i du bloc dans lequel se trouve notre nœud examiné, alors que dans l'expression de δ la valeur obtenue dépend à la fois du degré K_i du bloc b_i dans lequel le nœud évalué se trouve mais aussi des degrés K_j de ses blocs voisins.

3.4.3 Le réseau mondial des transports aériens

Les données sont récoltées sur le site `https://openflights.org/data.html`, dont la dernière mise à jour date de Janvier 2017. Elles génèrent un réseau composé de 3330 nœuds et 19080 arêtes, représentant l'ensemble des aéroports et des aérogares répertoriées au moment de cette mise à jour. Nous travaillons cependant sur la plus grande composante connexe de ce réseau, qui contient 3304 nœuds et 19054 arêtes. Il est important de noter que nous ne prenons pas en compte l'aspect pondéré qui pourrait être rajouté au réseau. En effet, un aéroport peut aisément être distingué d'un autre en comparant par exemple leurs nombres respectifs de voyageurs annuels. On peut aussi se dire qu'un vol civil partant d'un aéroport A vers un aéroport B, avec une fréquence de 5 vols par semaine et une moyenne de 200 passagers par vol, n'est pas équivalent à un vol de ravitaillement reliant deux bases militaires, ou deux villes isolées.

Nous les considérons tout de même ici comme étant identiques et les modélisons par une arête simple. Nous choisissons de ce fait de nous concentrer uniquement sur la topologie du réseau, en négligeant son aspect dirigé et pondéré.
Cette représentation aura des répercutions sur les résultats que nous montrons ici, car en mettant sur le même pied d'égalité des vols de fréquences différentes, il se peut que l'on privilégie parfois la diversité des vols à leurs fréquences et quantités.

3.5 Tests sur des réseaux synthétiques

Nous commençons d'abord par montrer et commenter les résultats des mesures β,γ et δ sur les modèles synthétiques. Pour le modèle treeCom, nous nous concentrons sur la capacité de la mesure donnée, à bien séparer les valeurs attribuées aux nœuds des clusters et ceux de l'arbre. Le but est alors d'observer un saut de valeurs clairement identifiable lors du passage de l'arbre aux clusters. De plus, on cherche aussi à observer une différence de valeurs entre les nœuds qui appartiennent à des clusters présentant des caractéristiques différentes (en particulier la taille et le paramètre p).
Pour le modèle de blocs stochastiques, nous profitons du fait qu'il soit possible d'obtenir les expressions analytiques des valeurs moyennes de δ, du coefficient de clustering et du degré pour dégager certaines des propriétés des mesures évaluées. Nous présentons les résultats des différentes mesures sur un cas particulier de PPM, dont l'architecture et les paramètres ont été choisis pour mettre en évidence les propriétés des différentes mesures.

3.5.1 Le modèle treeCom

Nous montrons ici les résultats de nos mesures sur trois jeux de paramètres différents. Dans les trois cas, le réseau jouet issu du modèle treeCom est composé d'un arbre de $N_t = 100$ nœuds, et de $N_b = 4$ clusters. En premier lieu, nous avons un modèle dont les 4 clusters sont de tailles identiques : $N_i = 50$ nœuds pour $i \in \{1, 2, 3, 4\}$. Nous attribuons cependant à chacun d'entre eux un paramètre P_i différent : $P_1 = 0.3, P_2 = 0.6, P_3 = 0.75, P_4 = 0.9$. Ainsi nous pouvons à l'aide de ce choix de paramètres isoler l'effet du paramètre P_i sur les différentes mesures introduites, dans le cadre du modèle treeCom. Le second jeu de paramètres consiste à attribuer les mêmes valeurs $P_i = 0.5$ pour chaque cluster pour $i \in \{1, 2, 3, 4\}$. Nous faisons cependant varier la taille des clusters avec $N_1 = 200, N_2 = 150, N_3 = 100, n_4 = 50$. Ainsi nous pouvons isoler l'effet de celle-ci sur les mesures à tester. Enfin le dernier jeu de paramètres consiste à attribuer une probabilité P_i inversement proportionelle à la taille N_i de chaque nœud, tout en gardant un degré moyen constant à l'intérieur de chaque cluster. Ceci nous permet d'avoir une superposition des deux premiers tests, dans laquelle les deux paramètres P_i et N_i varient en même temps, tout en supprimant l'influence potentielle du degré moyen (qu'on a choisi de garder constant) : $N_1 = 200, N_2 = 150, N_3 = 100, n_4 = 50$ et $P_1 = 0.225, P_2 = 0.3, P_3 = 0.45, P_4 = 0.9$. Ces valeurs sont marquées sur les figures 3.2a, 3.2b et 3.2c.

Nous précisons que pour les trois cas évoqués, nous choisissons $p_i^t = \frac{N_i}{N_t \cdot \sum_j N_j}$, ce qui vérifie la condition de connexité du réseau résultant, qui rappelons-le est $p_i^t \gg \frac{1}{N_t \cdot N_i}$. Ainsi nous avons bel et bien $N_i \gg 1$ pour chacun des blocs. Les résultats sur les trois jeux de paramètres précédemment décrits du modèle treeCom sont montrés sur la figure 3.2.

Résultats

On remarque sur la figure 3.2 qu'il existe des caractéristiques propres à chacune des mesures. Pour le degré (première ligne sur les figures 3.2a, 3.2b et 3.2c), il est facile de voir qu'il est distribué autour de la valeur correspondant au produit de la taille et du paramètre de probabilité pour chaque cluster. Il prend des valeurs faibles pour les nœuds de l'arbre.
Il n'est donc pas nécessaire de longuement s'attarder sur l'analyse de ces résultats, ce que nous faisons par la suite pour les mesures β, γ et δ.

Analyse des résultats de β

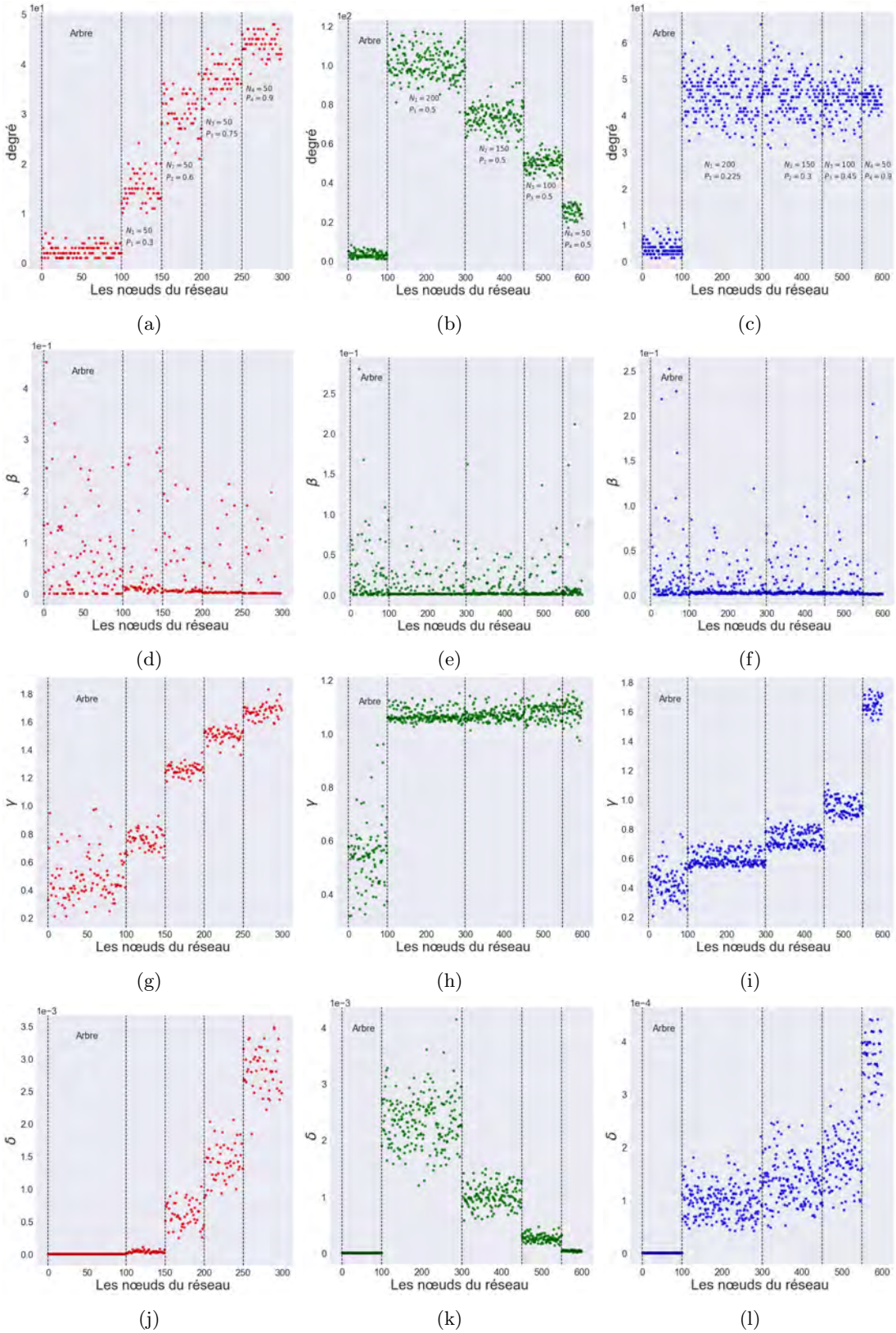


Figure 3.2: Les résultas des mesures β, γ et δ, ainsi que du degré de chaque nœud, dans des réseaux treeCom issus de 3 jeux de paramètres différents. Chaque configuration se différencie par sa couleur : rouge pour le cas où les clusters sont de même taille, verte pour celui où les clusters ont les mêmes paramètres de probabilité, et bleue dans le cas où les tailles sont inversement proportionnelles aux paramètres P_i.

Les résultats de la mesure β montrent qu'en modifiant la centralité de betweenness, on obtient bien une valeur maximale pour un nœud se situant dans l'arbre (*cf.* fig. 3.2d, fig. 3.2e et fig. 3.2f). On n'observe cependant pas de domination de la part des nœuds de l'arbre en terme de valeurs. En effet, on aurait pu s'attendre à ce que les nœuds de l'arbre soient ceux dont les valeurs sont les plus élevées, mais ceci n'est pas tout à fait le cas. La raison est qu'une partie importante parmi ces nœuds n'est que peu ou même pas du tout impliquée dans des géodésiques reliant des nœuds de différents clusters. À l'inverse, il y a quelques rares nœuds parmi ceux de l'arbre qui sont au milieu d'un grand nombre de ces géodésiques. Il est cependant à noter que la proportion de nœuds dont le score β est supérieur à 10% de la valeur maximale est en moyenne 5 fois plus grand pour les nœuds de l'arbre que pour ceux des clusters. Ceci nous laisse croire que β favorise dans une certaine mesure les nœuds qui se situent entre les clusters.

Analyse des résultats de γ

Pour ce qui est de γ on distingue plusieurs propriétés : sa caractéristique la plus remarquable est qu'elle n'est pas sensible au degré, d'après la comparaison qui peut être faite entre la figure 3.2h et la figure 3.2b. Ceci est appuyé par le fait que même si les clusters sont de tailles différentes (les mesures dont les résultats sont sur la colonne centrale en vert correspondent aux cas ou les clusters sont de tailles différentes mais de même paramètre P_i), tous les nœuds des clusters ont en moyenne une valeur constante de γ, indépendamment du cluster auquel ils appartiennent.
En revanche en observant les résultats des calculs effectués sur des clusters ayant des paramètres de probabilité P_i qui varient (*cf.* fig. 3.2g où tous les clusters ont la même taille, et fig. 3.2i où les clusters ont des tailles N_i inversement proportionnelles au paramètre P_i), on remarque que les résultats de la mesure γ sont corrélés avec les paramètres de probabilité P_i. Ce résultat est compréhensible quand on sait que chacun des clusters, quand il est considéré à part en tant que sous graphe induit, est un graphe aléatoire de paramètres (N_i, P_i). On sait par ailleurs que que pour pour ce dernier type de graphes, on retrouve un coefficient de clustering moyen égale à P_i.
La mesure γ étant une moyenne pondérée des coefficients de clustering des voisinages successifs, il est normal de retrouver une corrélation positive entre γ et P_i et non entre γ et N_i. On peut aussi rajouter que la mesure γ attribue aux nœuds de l'arbre des valeurs faibles (en comparaison avec les valeurs obtenues sur les nœuds des clusters) mais non nulles. Certains des nœuds de l'arbre ont même parfois des valeurs supérieures à celles d'un nœud se trouvant dans un cluster (*cf.* fig. 3.2g, fig. 3.2i).

Analyse des résultats de δ

Nous pouvons voir en comparant la figure 3.2a et la figure 3.2j que pour des clusters ayant la même taille, le choix du paramètre P_i influe sur les résultats de δ, de sorte que plus le P_i est grand, plus le δ augmente. Ceci est également le cas pour la taille N_i des clusters, en comparant les figures 3.2b et 3.2k on voit que pour des clusters ayant la même valeur de P_i mais des tailles différentes, plus un cluster est grand plus son δ est élevé. Étant donné la définition de δ (*cf.* eq. (3.9) et eq. (3.7)) ces résultats sont compréhensibles. Nous rappelons que cette définition dépend à la fois du nombre de voisins communs entre chaque nœud et ses voisins, ainsi que de la valeur du degré. Ceci explique à la fois le rôle de N_i et de P_i.

En ce qui concerne le troisième jeu de paramètres (celui dont les nuages de points sont sur la colonne de droite et de couleur bleue), on peut observer que lorsqu'on fixe le degré moyen (*cf.* fig. 3.2c), on obtient un δ croissant comme on peut le voir sur la figure 3.2l. Ceci signifie que la sensibilité présumée de δ à la taille des clusters (*cf.* fig. 3.2k) était en réalité due à la différence des degrés moyens des nœuds qui les composent. Car pour une taille de clusters décroissante et un P_i croissant, le résultat est que le δ moyen est lui aussi croissant.

On peut finalement remarquer que contrairement à la mesure γ, la mesure δ attribue des valeurs nulles à tous les nœuds faisant partie de l'arbre. Ceci peut être expliqué par le fait qu'il n'existe aucun cycle dans un arbre, donc aucune paire ayant des voisins communs, et par conséquent toutes les arêtes de l'arbre sont de poids nuls. Le reste des liens qui relient les nœuds de l'arbre à ceux des clusters ne sont (compte tenu du choix des paramètres p_i^t) pas assez nombreux. Ceci laisse une faible probabilité pour qu'un nœud de l'arbre et celui d'un cluster aient des voisins en commun, laissant le δ de l'ensemble des nœuds de l'arbre de valeur nulle.

3.5.2 Le modèle de blocs stochastiques

Nous nous tournons à présent vers l'analyse d'un exemple de modèle de blocs stochastiques. En premier lieu, nous commençons par fixer les paramètres p et q, ainsi que l'architecture du réseau à générer. Les blocs sont de tailles assez grandes pour que les approximations qui permettent d'aboutir aux résultats des eq. (3.15) et eq. (3.12) soient justifiées. Nous fixons donc à 250 nœuds la taille de chacun des blocs dans l'exemple que l'on va décrire. Nous choisissons ensuite d'attribuer une architecture arborescente aux blocs, pour simplifier la compréhension des résultats. L'arbre en question est un arbre binaire d'une profondeur égale à 3, comme représenté sur la figure 3.3.

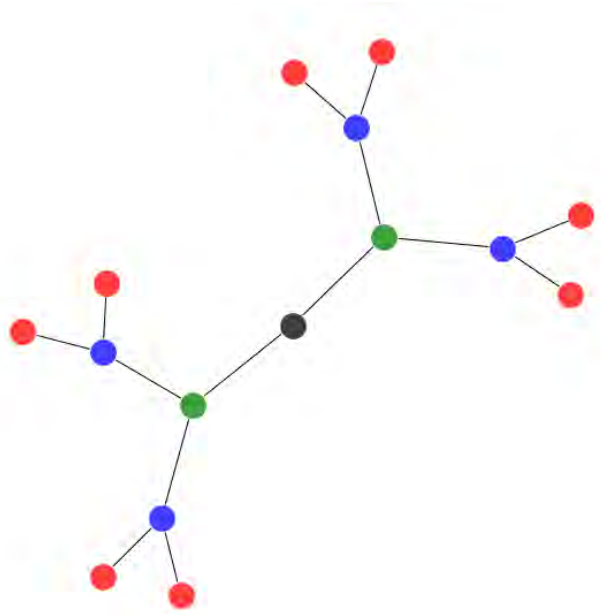


Figure 3.3: Architecture du modèle de blocs stochastiques étudié, chaque nœud est colorié soit en vert, en bleu ou en rouge suivant sa distance à la racine, ici représentée en noir.

Nous fixons ensuite les paramètres p et q du modèle, ce choix peut être fait arbitrairement, mais nous choisissons de guider ce dernier à l'aide de la formule donnée par eq. (3.15). Nous calculons ainsi pour chaque paire de valeurs (p, q) l'ensemble $\{\overline{\delta_1}, \overline{\delta_2}, ..., \overline{\delta_m}\}$ des valeurs moyennes de chaque bloc (ici la notation en dessous de la barre indique que l'on calcule la valeur moyenne de δ dans chaque bloc). Ensuite, nous calculons pour cette même paire la valeur de g_δ, représentant le plus petit saut de valeurs séparant deux valeurs successives de δ *i.e* : $g_\delta = \min\{\overline{\delta_2} - \overline{\delta_1}, \overline{\delta_3} - \overline{\delta_2}, ..., \overline{\delta_m} - \overline{\delta_{m-1}}\}$. Nous calculons ensuite le pourcentage que représente g_δ par rapport à la valeur $\overline{\delta_{max}}$ qui représente pour la paire (p, q) la valeur maximale du δ moyen. Nous montrons sur la figure 3.4 les valeurs prises par g_δ en fonction de p et q.

Pour l'exemple qui va suivre, nous choisissons la paire (p, q) de sorte que g_δ soit égal à la moitié de sa valeur maximale, soit à 10.7% de $\overline{\delta_{max}}$. Ceci s'obtient par les valeurs $p = 0.82$ et $q = 0.61$. Nous notons donc que l'on dispose d'un modèle assortatif de blocs stochastiques, car nous avons $p > q$.

Nous montrons ensuite la matrice d'adjacence du modèle résultant d'un tel choix de paramètres sur la figure 3.5

Résultats

Nous montrons ici, comme dans le cas du modèle treeCom, les résultats des mesures β, γ et δ ainsi que ceux du coefficient de clustering, que l'on a calculé sur le modèle dont les paramètres ont été fixés plus haut. Pour chacune de ces mesures, nous montrons le nuage de points résultant du calcul de ces grandeurs sur le modèle, ainsi que sa distribution estimée par un noyau gaussien d'une largeur égale au millième de la plage totale de valeurs obtenues sur chacune des mesures. Nous indiquons aussi, pour δ et le coefficient de clustering les valeurs obtenues analytiquement à l'aide des eq. (3.15) et eq. (3.12).

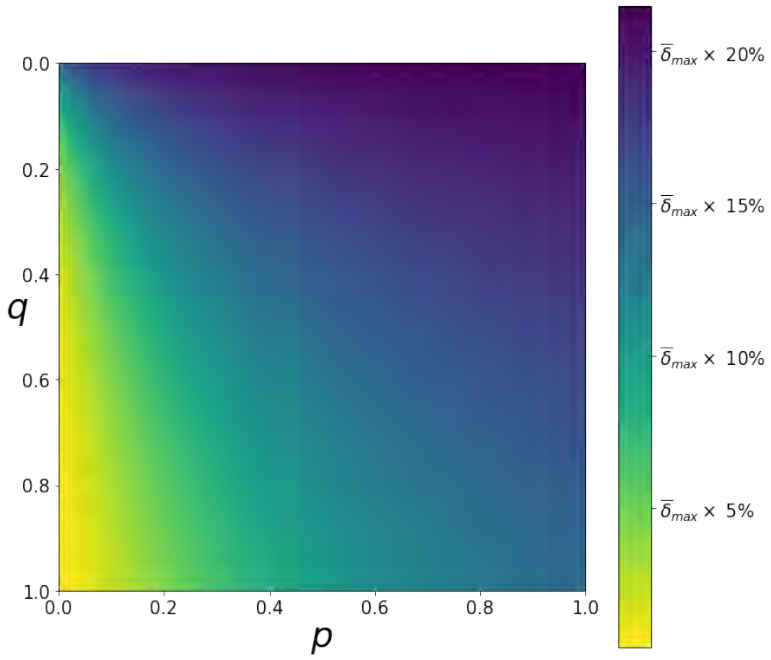


Figure 3.4: Les valeurs prises par $_{\delta}$ en fonction des paramètres p et q du modèle de blocs stochastiques.

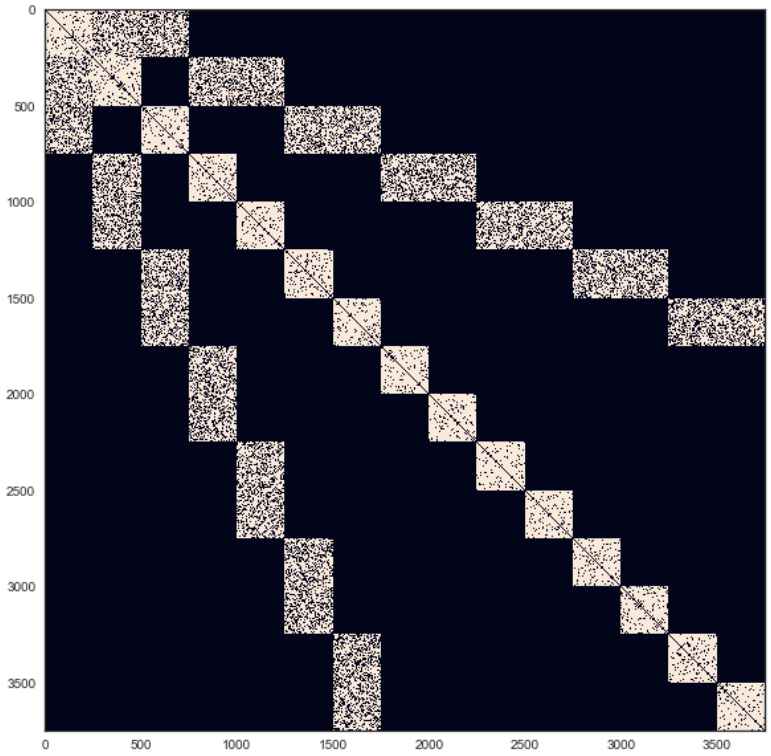


Figure 3.5: La matrice d'adjacence du modèle résultant des paramètres décrits précédemment.

Les couleurs des nuages de points sur la figure 3.6 sont choisies de sorte que chaque nœud ait la couleur du bloc correspondant, tel qu'il est décrit sur la figure 3.3.

Nous observons que les trois mesures β, γ et δ contiennent quatre modes dans leurs distributions respectives (même s'il est difficile de le voir avec précision sur la distribution de β, *cf.* fig. 3.6d), contrairement au coefficient de clustering qui n'en affiche que trois. Nous pouvons expliquer les trois

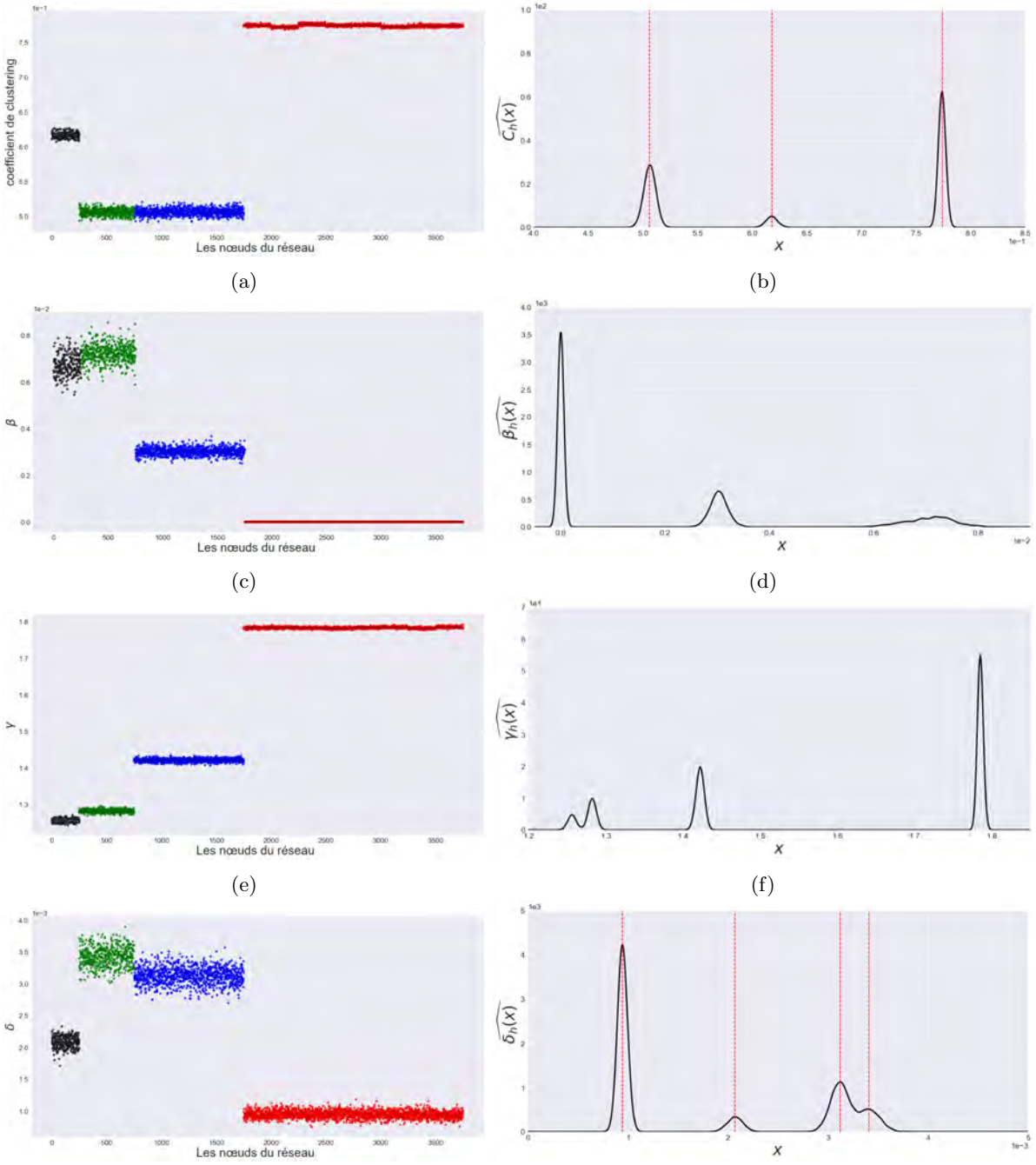


Figure 3.6: Les nuages de points représentant les valeurs du coefficient de clustering (a), les mesures β (c), γ (d) et δ (e). Chaque point est de la même couleur que le bloc auquel il appartient, suivant la représentation de la figure 3.3. À droite nous avons les distributions approchées par la méthode de l'estimation par noyau Gaussien, notés $\widehat{C_h(x)}$ pour le coefficient de clustering (b), $\widehat{\beta_h(x)}$ (d) ,$\widehat{\gamma_h(x)}$ (f) et $\widehat{\delta_h(x)}$ (h). Les valeurs calculées à l'aide des eq. (3.15) et eq. (3.12) sont indiquées par les lignes verticales en pointillés rouges sur (b) et (h).

modes de la distribution du coefficient de clustering par le fait que sa valeur moyenne à l'intérieur d'un bloc donné ne dépende que du degré de celui-ci eq. (3.16). Comme les degrés des nœuds de l'arbre représentés sur la figure 3.3 appartiennent à l'ensemble $\{1, 2, 3\}$, il en résulte une distribution du coefficient de clustering ayant trois modes distincts uniquement. L'amplitude de chaque mode croit

avec le nombre de blocs ayant le degré dont il est représentatif. Ainsi, les deux modes qui représentent les blocs de degrés 1 et 3 (on les voit rouge pour ceux de degré 1, verts et bleus pour ceux de degré 3 sur la figure 3.3) ont une amplitude plus grande que le mode représentant le bloc de degré 2 (représenté par le nœud noir sur la figure 3.3).

Pour les mesures que l'on a développé lors de ce chapitre, on va tenter d'expliquer la présence de 4 modes. Les blocs ayant un degré 3 peuvent être dissociés en deux types distincts, suivant les caractéristiques respectives de chaque mesure. Pour le cas de δ nous pouvons expliquer cette séparation à l'aide de la formule donnée par eq. (3.17). En effet, nous y avons une expression du δ moyen qui pour chaque bloc prend en considération le degré de celui-ci, mais aussi le degré de ses voisins. Ainsi nous pouvons séparer les blocs coloriés en bleu (dont deux blocs voisins sont de degré 1 et un de degré 3) de ceux coloriés en vert (dont deux voisins sont de degré 3 et un de degré 2).

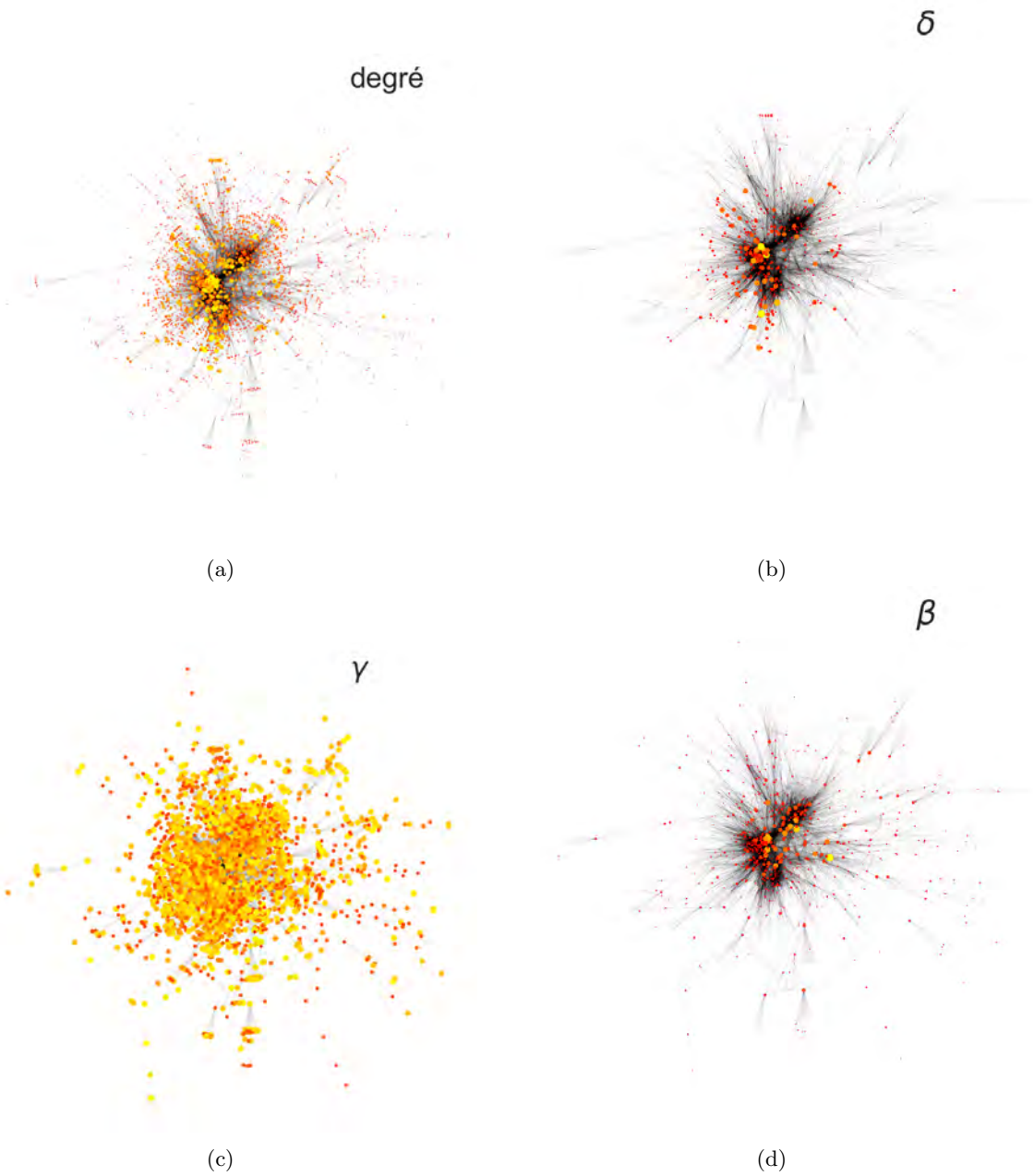
Il est aussi possible d'expliquer cette séparation dans les résultats de γ et β, même si l'on ne dispose pas de formule explicite pour les calculer sur un PPM. Dans le cas de γ, ceci s'explique par le fait que suivant la position d'un nœud dans un bloc donné, la distance maximale le séparant de n'importe quel autre nœud du PPM va varier. Par exemple, être sur un bloc représenté en rouge sur la figure 3.3 signifie que le nœud le plus éloigné est à une distance relativement élevée (comparable au diamètre du PPM). Alors que pour un nœud positionné à l'intérieur du bloc représenté en gris, la distance maximale au reste des nœuds se trouve fortement réduite. Cependant dans la définition de γ donnée par eq. (3.2), la somme est effectuée sur tous les voisinages successifs. On découvre alors une autre caractéristique de γ, qui est que plus la centralité de proximité d'un nœud est élevée, plus sa valeur de γ sera faible. La raison est que ce nœud a moins de termes dans la somme donnée par eq. (3.2) à calculer. Le cas de β peut aussi être expliqué par des arguments similaires, concernant la centralité de betweenness de chaque bloc (équivalente à la centralité de betweenness des nœuds de l'arbre dans la figure 3.3). Plus un bloc est central, plus les nœuds qui sont à l'intérieur de celui-ci le sont aussi, et donc plus leur β augmente.

Il est très important de souligner que l'ordre attribué aux nœuds des différents blocs varie d'une mesure à l'autre. Celui de δ attribue les scores les plus élevés aux nœuds se situant dans des blocs qui ont en même temps des degrés élevés, et des blocs voisins de degrés élevés. La mesure γ privilégie quant à elle les nœuds se situant dans des blocs dont la centralité de proximité est faible. Ceci aboutit à une contradiction car nous avons construit le modèle SBM de la figure 3.3 pour que les nœuds dont la densité est la plus faible soient ceux représentés en rouge, or nous observons l'inverse sur la figure 3.6e. La même remarque peut s'appliquer sur les résultats de la mesure β affichés sur la figure 3.6c. Nous rappelons que la construction de la mesure β est faite de telle sorte que plus la densité d'un nœud est élevée, plus sa valeur de β est censée être faible.
Enfin nous pouvons aussi observer que la largeur des modes est proportionnellement plus élevée dans le cas des mesures β et δ qu'elle ne l'est dans celui du coefficient de clustering. Celle-ci connaît significativement moins de variabilité car son expression dépend uniquement du degré de chaque bloc, et par extension de la mesure γ qui est basée dessus.

3.6 Tests sur un réseau réel : le réseau des aéroports

Nous présentons dans cette section les résultats de nos mesures appliquées à un réseau réel de transports aériens, de 3304 nœuds (aéroports) et 19055 arête (liaisons aériennes) . Il est difficile d'associer une réalité de terrain aux résultats que l'on va montrer, pour les raisons évoquées dans la section 3.4.3 (les liens ne sont ni pondérés ni dirigés). On se concentre dès lors uniquement sur l'aspect topologique du réseau.
En premier lieu nous donnons une représentation graphique des résultats, afin d'en avoir une image globale, ensuite nous donnerons dans l'ordre décroissant les nœuds ayant obtenu les scores les plus élevés. Nous tentons finalement de donner une interprétation à ces résultats. D'abord on peut facilement voir sur la figure 3.7 que δ et β sont de meilleurs filtres que le degré et γ. Nous pouvons le confirmer à l'aide des résultats de la figure 3.8 qui montrent l'évolution relative de chacune de ces mesures, en partant du nœud ayant le score le plus élevé jusqu'à celui ayant le plus faible. En effet nous voyons, en nous restreignant aux 20 premiers points de chacune des mesures, que les décroissances de β et de δ sont effectivement plus rapides que celles du degré et de γ. Nous donnons sur la table 3.1 les classements associés à chacune des mesures.

On peut observer une similitude entre les résultats des tests effectués sur des réseaux réels et ceux sur les réseaux synthétiques. Par exemple nous voyons que les aéroports les mieux classés par γ sont



(a) (b)

(c) (d)

Figure 3.7: Le réseau mondial des aéroports. Les nœuds sont d'une taille proportionnelle aux mesures du degré (a), δ (b), γ (c) et β (d). Les couleurs sont aussi représentatives du score de chaque nœud, plus celui-ci est élevé, plus la couleur va du rouge vers le jaune.

pour la plupart de petits degrés (les trois premiers dont deux sont situés au Groenland et un en Norvège induisent tous les trois un triangle comme égo-graphe). Ceci va dans le sens d'une propriété mise en évidence dans la section 3.2, qui est que cette mesure attribue des scores élevés aux nœuds les plus isolés dans le réseau, pour peu que leurs voisins directs (même s'ils sont en petit nombre) soient bien connectés entre eux.
Quant aux résultats de β, nous voyons que certains des aéroports qui se situent dans les grandes villes (Paris CDG, Los Angeles, Amsterdam, etc.) se trouvent dans le top 20. Ceci va dans le sens opposé des propriétés que l'on souhaitait avoir pour β qui, rappelons-le, est censé attribuer des scores faibles aux nœuds de forte densité. Ceci est supposément le cas pour les principaux aéroports de capitales européennes.
Finalement nous observons aussi une certaine ressemblance entre les résultats du classement par degré et

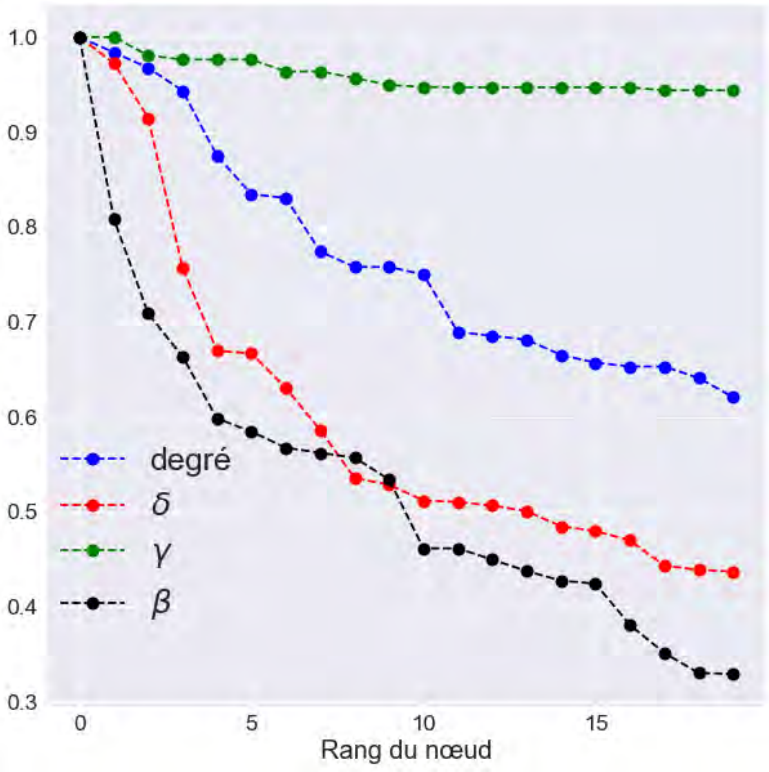


Figure 3.8: Les valeurs normalisées des mesures δ, γ, β ainsi que celles du degré, dans l'ordre décroissant. On restreint chaque courbe aux 20 aéroports les mieux classés par la mesure correspondante.

celui obtenu par δ, avec un indice de Jaccard d'une valeur de 0.54 entre leurs deux top 20. Nous notons aussi la présence de quelques différences intéressantes dans les rangs attribués. En effet les résultats montrent que le classement fourni par le degré ne suit pas strictement le classement fourni par δ. À titre d'exemple, le top 20 fourni par δ est celui qui contient le plus grand nombre d'aéroports européens (13 dont 8 dans le top 10, contre 10 dont 5 dans le top 10 pour le classement donné par le degré). On explique partiellement cela par la faible distance géographique séparant les métropoles européennes, et donc une plus grande interconnectivité, ce qui est favorable à une valeur élevée de δ.

On finit par montrer sur la figure 3.9 l'évolution de la taille de la plus grande composante connexe du réseau étudié (qui contient initialement 3304 nœuds), une fois les i nœuds les mieux classés supprimés (notée $|G_i^{cc}|$), dans l'ordre résultant des différentes mesures, ainsi que dans un ordre aléatoire. Cette figure montre que l'ordre attribué par les mesures δ et β produit un fractionnement du réseau en petits morceaux, et ce après avoir supprimé un nombre de nœuds inférieur au tiers du nombre total. L'efficacité suivant ces deux dernières mesures est ici comparable à celle que l'on obtient en supprimant les nœuds dans l'ordre décroissant des degrés, bien que l'on observe une meilleure efficacité suivant l'ordre fourni par la mesure β au début du processus (jusqu'à 240 nœuds supprimés), ce qui n'est pas surprenant compte tenu de la construction de cette dernière. Par ailleurs, l'efficacité suivant l'ordre des mesures β et δ est nettement supérieure à celle obtenue quand on cible les nœuds aléatoirement. Ceci n'est pas le cas pour la mesure γ, qui fournit un ordre dont l'efficacité est inférieure à l'aléatoire, et ce sur la majeure partie du processus de suppression (nombre de nœuds supprimés inférieur ou égal à 2284).

Pour finir on rappelle que ces tests ont été effectués sur un réseau réel, et fournissent des résultats qui peuvent difficilement être mis en perspective avec une réalité de terrain. En effet, tout ce que nous avons ici se résume à une liste de vols d'un aéroport A à un aéroport B. Ces résultats permettent cependant de relever certaines propriétés qu'on ne pouvait pas avoir à partir des tests effectués sur les réseaux synthétiques. Nous pouvons citer à titre d'exemple la résilience du réseau à une attaque ciblée *cf.* fig. 3.9, les réseaux réels étant en général significativement plus hétérogènes que ceux générés synthétiquement.

Le rang	**Classement par degrés**	**Classement par δ**	**Classement par β**	**Classement par γ**
1	Amsterdam	Amsterdam	Ted Stevens	Nanortalik
2	Frankfurt	Frankfurt	Los Angeles	Alluitsup Paa
3	Paris CDG	Paris CDG	Charles de Gaulle	Valan
4	Atatürk	Munich	Dubai	Ambler
5	HJ Atlanta	London Heathrow	Tacoma	Shungnak
6	Beijing	Atatürk	Frankfurt am Main	Kobuk
7	Chicago	Leonardo da Vinci	Lester B. Pearson	Poplar Hill
8	Munich	Barcelona	Chicago O'Hare	North Spirit Lake
9	Moscou Domodedovo	Adolfo Suárez	Amsterdam Schiphol	Fort Severn
10	Dallas	Chicago	Beijing	Kasigluk
11	Dubai	HJ Atlanta	Guarulhos	Hooper Bay
12	London Heathrow	Zürich	Atatürk	Chevak
13	Denver	J. F. Kennedy	Sydney	Scammon Bay
14	Houston	Brussels	Brisbane	Buckland
15	London Gatwick	Beijing	Narita	Deering
16	Barcelona	Dubai	London Heathrow	Bob Baker
17	J. F. Kennedy	Dublin	Singapore	Robert Curtis
18	Leonardo da Vinci	Vienna	Montreal	Rennell/Tingoa
19	Adolfo Suárez	Newark Liberty	Denver	Deer Lake
20	Shanghai Pudong	Düsseldorf	Dallas	Gambell

Table 3.1: Les aéroports correspondants aux nœuds dont les scores sont les plus élevés.

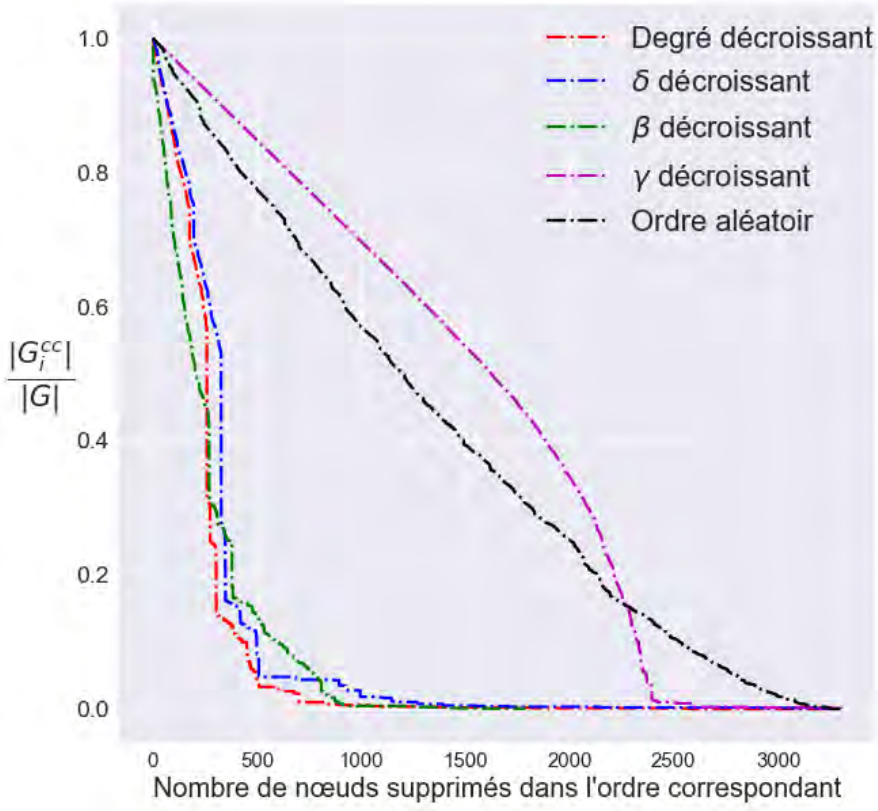


Figure 3.9: Taille de la plus grande composante connexe, après la suppression des i nœuds les mieux classés par le degré, δ, β, γ ainsi que dans un ordre aléatoire. Le résultat est divisé par la taille du réseau, et fourni en fonction de i.

3.7 Bilan

Ce chapitre a été consacré à la mise au point et à l'analyse de trois mesures de densité, qui contrairement à la densité spatiale développée lors du chapitre précédent, sont des mesures déterministes qui attribuent des valeurs fixes aux nœuds d'un réseau. Ces mesures sont construites sur la base d'un ensemble de propriétés évoquées plus haut (déterminisme, localité, favorisation de la connectivité). Nous avons d'un côté présenté β et γ, qui sont des mesures directement applicables sur les nœuds, mais qui sont aussi paramétriques car comportant une fonction de pondération dans leur définition respective. Le choix de cette dernière est justifié dans le cas de β, mais peut tout à fait être arbitraire comme c'est le cas pour la mesure γ. D'un autre côté nous avons la mesure δ, qui est basée sur les propriétés des arêtes qui sont attachées à chaque nœud, et qui a en plus l'avantage de ne pas être paramétrique.
Ces mesures ont toutes les trois été testées sur deux modèles jouets, le premier est constitué de quatre blocs reliés entre eux par un arbre, et a été construit pour les besoins de notre thématique. Le second est un modèle de blocs stochastiques, sur lequel on a d'abord calculé une expression analytique de δ (mais pas de γ et β), et ensuite donné les distributions de β, γ et δ qu'on a restreint à un cas particulier de modèle de partition plantée, dont nous avons fixé l'architecture. Ensuite nous avons effectué quelques tests sur le réseau mondial des aéroports, et nous avons montré les différents classements attribués par chacune des mesures présentées, ainsi que celui obtenu en utilisant le degré. Les résultats de cette comparaison peuvent être expliqués par les propriétés mathématiques de chacune des mesures, mais il nous a été difficile de les appuyer par une réalité de terrain, même si nous avons donné une hypothèse concernant les différences de classement entre le degré et δ. Celle-ci s'appuie sur le fait qu'il y ait en Europe la plus grande concentration mondiale de grandes villes, ce qui engendre une forte connectivité entre les aéroports européens, car ces derniers sont proches géographiquement les uns des autres.

Les résultats des tests ont permis de mettre en évidence quelques propriétés : pour β, on a pu voir que même sur des cas simples comme le réseau treeCom, il est difficile de différencier les nœuds de l'arbre de ceux des clusters. La raison est qu'un petit nombre de chacun des deux groupes se trouve au milieu de la majorité des chemins les plus courts. Pour γ on a vu que celle-ci était fortement corrélée au coefficient de clustering, mais indépendante de la taille du cluster dans lequel le nœud évalué se situe. On a aussi constaté qu'il y a des cas dans lesquels trois nœuds qui forment un triangle isolé, se retrouvent avec une valeur de γ parmi les plus élevées de tout le réseau, ce qui est contradictoire avec l'une des propriétés souhaitées pour γ.

En ce qui concerne δ, les résultats des tests sont satisfaisants sur l'ensemble des jeux de données, on a pu voir sur le modèle treeCom qu'elle était à la fois corrélée avec la taille des clusters, au paramètre de probabilité P sans exclusivement dépendre du degré. On a aussi vu qu'elle permettait de séparer les différents types de clusters dans le modèle de blocs stochastiques, en plus d'attribuer un ordre globalement plus intuitif que celui fourni par β et γ. De plus on obtient des résultats auxquels il est possible de donner une interprétation (bien que partielle) sur le réseau réel. Ainsi dans les prochains chapitres nous utiliserons uniquement δ comme mesure pour la densité, et nous nous baserons sur ses propriétés pour mettre au point un algorithme qui permet de séparer le graphe en deux
sous-ensembles : d'un côté la partie dense et de l'autre la partie non dense.

Chapitre 4

Classification par la densité : l'algorithme ItRich

Table des matières

4.1 Introduction

Nous sommes maintenant en mesure de calculer la densité des nœuds, à l'aide d'une mesure vérifiant des critères prédéfinis et pertinents. Nous allons poursuivre par la présentation d'un algorithme qui s'occupe de classer chaque nœud du graphe soit dans la partie dense, soit dans la partie non dense. Pour cela, nous avons développé deux approches : la première consiste à optimiser une mesure de qualité d'une partie non dense, en rajoutant progressivement et dans un ordre pertinent des nœuds au sous-réseau évalué. Cette démarche est similaire à celle adoptée par la plupart des algorithmes de détection de communautés, en ce sens où il s'agit d'abord de donner une mesure de qualité du résultat visé, et ensuite de choisir la partition qui rend celle-ci optimale. Cette approche s'est avérée moins efficace pour cette tâche que pour celle de la détection de communautés, pour des raisons qu'on détaillera dans ce chapitre.
La seconde approche est quant à elle inspirée de l'algorithme du rich club, à la différence que le réseau dont les rich clubs sont à extraire est pondéré par les poids $\omega_{i,j}$. Ceci assure que la somme des poids attachés à l'extrémité de chaque nœud soit égale à sa valeur de δ. Cette approche bien que inédite, donne des résultats plus satisfaisants que la première, et sera la plus détaillée des deux dans la suite du chapitre. Celui-ci est développé dans l'ordre suivant : d'abord nous présentons une première approche d'extraction des parties non denses, celle-ci est inspirée des méthodes classiques de détection de communautés, nous testons l'algorithme correspondant sur un modèle jouet simple, ainsi que sur un réseau réel. Ensuite nous détaillons l'approche qui se base sur l'extraction des rich clubs pondérés, en incluant quelques nouveautés par rapport aux algorithmes qu'on retrouve dans la littérature, avant de décrire en détails l'algorithme correspondant. Nous finissons le chapitre par la description du protocole expérimental qui nous permet de valider notre algorithme, en comparant ses performances à celles d'un algorithme présentant des similitudes avec le nôtre. Ces expériences sont effectuées sur des réseaux synthétiques, en l'absence de vérité de terrain dans les réseaux réels.

4.2 Première approche : optimisation d'un seuil et mesure de qualité de la partie non dense

Nous abordons cette section en résumant le problème à la question suivante : comment comparer deux ensembles de nœuds différents, et décider lequel d'entre eux est le meilleur ensemble en termes de non-densité ? Lors des deux précédents chapitres, nous avons tenté de répondre à une question équivalente, mais celle-ci était restreinte à l'échelle du nœud. Il nous faut maintenant la généraliser à l'échelle plus large du réseau, afin de séparer les parties denses des parties non denses. L'approche développée ici consiste à définir une mesure de la qualité d'une partition, et ensuite de construire un algorithme qui permet de bien séparer un réseau en deux parties distinctes, relativement à cette mesure de qualité.

L'idée est ici de combiner deux propriétés simples : d'un côté la partie non dense doit avoir une faible transitivité moyenne[1] par rapport à la transitivité du réseau entier, et, d'un autre côté, il est préférable que le sous-réseau induit par la partie non dense ait une composante géante dont la taille est significative. On va formaliser ces idées pour donner une mesure qui permet d'estimer la qualité d'un ensemble de nœuds, en tant que partie non dense du réseau.

Soit $G = \{V, E\}$ un graphe. On appelle triplet connecté de V l'ensemble constitué de trois nœuds $i, j, k \in V^3$ tels que $i \sim j$ et $j \sim k$.
Soit T_V l'ensemble des triplets connectés de V et t_V l'ensemble des triangles de V, constitué des triplés $\{i, j, k\} \in V$ tel que $i{\sim}j{\sim}k{\sim}i$. La transitivité notée $T_r(G)$ d'un graphe G, qui est aussi égale à la moyenne du coefficient de clustering calculée sur tous les nœuds du graphe, est obtenue par :

$$T_r(G) = \frac{3 \cdot t_v}{T_V}.$$

La transitivité d'une clique est égale à 1 alors que la transitivité d'un arbre est nulle.

Soit maintenant $V_F \subset V$ un sous-ensemble de nœuds, et F le sous-réseau de G induit par V_F. Appelons F^{cc} la plus grande composante connexe de F.

[1] On rappelle que la transitivité moyenne est donnée par le rapport entre le nombre de triangles et le nombre de triplets connectés contenus dans le réseau

Nous définissons ensuite la mesure de qualité d'un sous-réseau $F = (V_F, E_F)$ induit par V_F, en tant que partie non dense par la quantité :

$$S_G(F) = \frac{|F^{cc}|}{|F|} - \frac{T_r(F)}{T_r(G)} \tag{4.1}$$

ici $|F|$ désigne le nombre de nœuds dans le graphe F et $|F^{cc}|$ le nombre de nœuds contenus dans sa plus grande composante connexe. Cette mesure a une valeur qui appartient à l'intervalle $]-1, 1]$, et augmente quand les critères de qualités évoqués plus haut sont vérifiés. En extrapolant, elle vaut 1 quand F est un arbre connexe et 0 quand ce dernier est une clique.

Il reste à déterminer un algorithme permettant de calculer le meilleur ensemble de nœuds F. Pour cela nous allons nous servir de la mesure δ de la densité, ou plus précisément de l'ordre dans lequel elle classe l'ensemble des nœuds du graphe. Soit un graphe $G = (V, E)$, on souhaite calculer sa partie non dense. Pour cela, on ajoute l'un après l'autre les nœuds de V à un ensemble V_U initialement vide, dans l'ordre croissant de δ. [2] On extrait les parties non denses comme étant l'ensemble qui maximise l'indice de qualité $S_G(F)$ donnée par l'eq. (4.1). Voici un algorithme qui retranscrit les étapes qu'on vient de décrire :

Algorithm 1: Calcul de la PND en optimisant une mesure de qualité

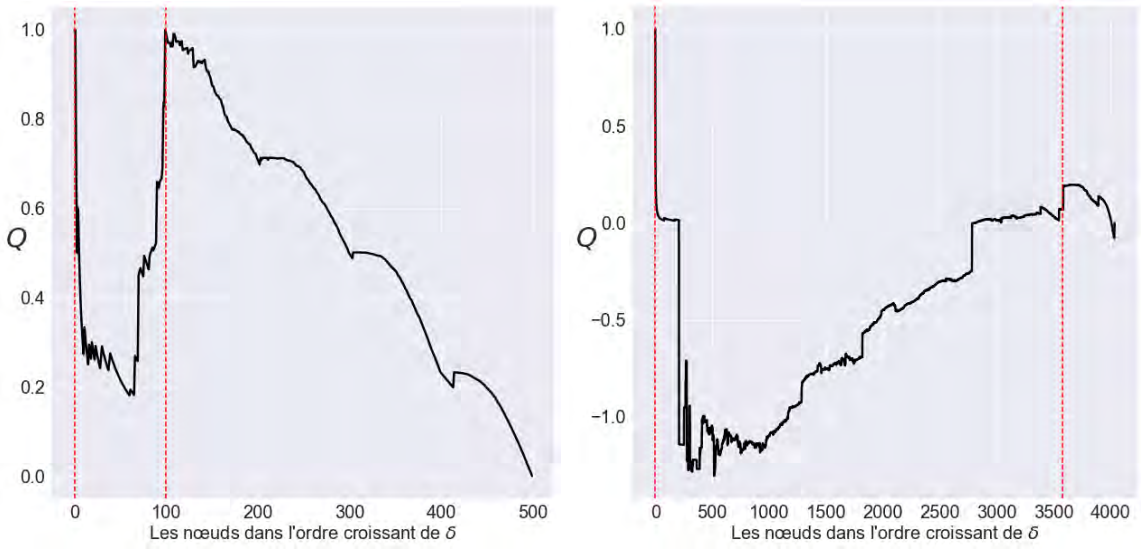
$G = (V, E)$;
$V_F = \emptyset$;
$Vals = \emptyset$;
Calculer $V^{(N)} = \{v_1, v_2, ..., v_N \mid \forall i \in [1, N-1]$ **;** $\delta(v_{i+1}) > \delta(v_i)\}$;
for $i \in [1, N]$ **do**
  $V_F = V_F \cup \{v_i\}$;
  $F = (V_F, E_F)$ Le sous-réseau induit par V_F dans G;
  $Vals = Vals \cup S_G(F)$;
end
Calculer $m = \text{argmax}(Vals)$;
return $V^{(m)} = \{v_1, v_2, ..., v_m \mid \forall i \in [1, m-1]$ **;** $\delta(v_{i+1}) > \delta(v_i)\}$;

Cet algorithme a été testé sur le modèle jouet treeCom composé de quatre clusters et d'un arbre de 100 nœuds, et sur un réseau réel de 4039 nœuds représentant un échantillon du réseau social Faceboook, dont les données sont à retrouver sur [79]. Les résultats obtenus sont représentés sur la figure 4.1.

Dans le cas du modèle jouet, on a un maximum global (hormis celui calculé sur le sous-réseau composé d'un seul nœud) obtenu sur un sous-réseau de taille 100. Celui-ci est précisément composé des nœuds de l'arbre qui sont au même nombre et d'une valeur de $Q = 1$. Ceci s'explique par le fait que l'arbre soit connecté, en plus de ne pas contenir de fermeture transitive. On retrouve donc avec succès le résultat souhaité par l'algorithme décrit dans cette section, puisque le modèle est construit de sorte que l'on puisse affirmer si un nœud appartient ou non à une partie non dense. Cette dernière est, dans notre modèle, exclusivement composée des nœuds de l'arbre.

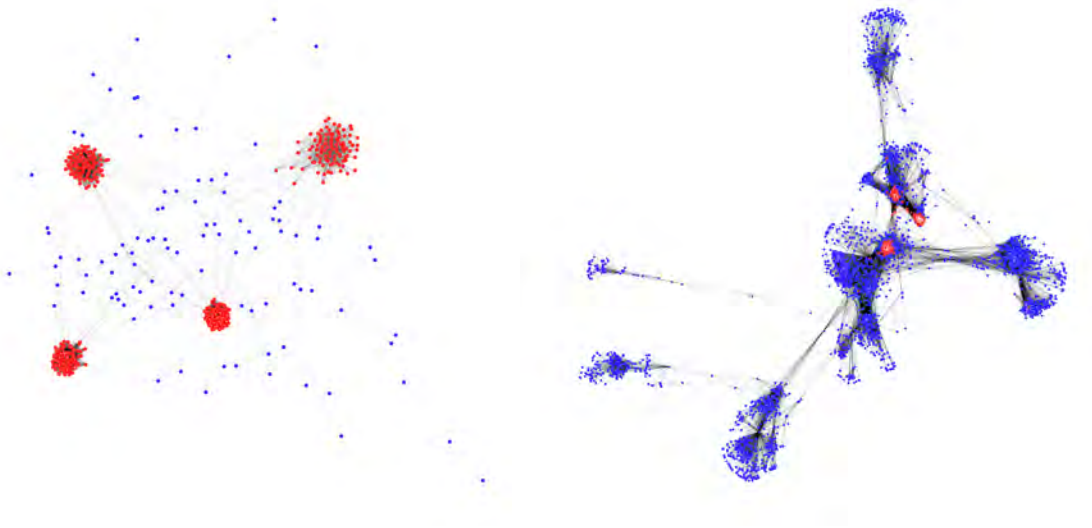
Dans le cas du réseau réel, nous voyons apparaître plusieurs maxima locaux. Certains sont positionnés autour de sous-réseaux composés d'un petit nombre de nœuds, et ont une valeur négative Q, ce qui les rend peu intéressants. On voit cependant apparaître un maximum constitué par un sous-réseau d'une taille de 3582 nœuds, et avec une valeur de Q positive. En excluant le maximum obtenu sur un sous-réseau composé d'un seul nœud, on obtient comme résultat une partie non dense composée de 3582 nœuds. Il reste cependant difficile de confirmer ou bien de réfuter les résultats obtenus, car le réseau sur lequel l'algorithme a été testé n'inclut aucune réalité de terrain. On ne peut que se baser sur des caractéristiques structurelles, comme par exemple le fait que la transitivité moyenne de la partie dense est 1.9 fois supérieure à celle de la partie non dense, ou que ce même rapport pour la mesure δ est quant à lui supérieur à 45.

[2] On choisit ici δ parce que c'est notre meilleure mesure, mais l'algorithme peut aussi fonctionner avec d'autres mesures de densité.



(a) Un modèle jouet

(b) Un réseau réel



(c) Un modèle jouet

(d) Un réseau réel

Figure 4.1: L'évolution de Q sur un modèle jouet (a) et un réseau réel (b), ainsi qu'une représentation sur leurs graphes respectifs des résultats, avec en bleu les nœuds de la partie non dense et en rouge le reste, qu'on considère comme la partie dense des réseaux (c) et (d).

On peut aussi souligner le fait que cet algorithme a été testé en utilisant l'ordre qui résulte de la mesure δ, mais il aurait tout à fait été possible d'utiliser une autre mesure. Cependant l'algorithme reste sensible à l'ordre attribué aux nœuds, et est par conséquent incompatible avec une mesure non déterministe de la densité, comme c'est le cas de ρ qu'on a introduit dans le chapitre 2. Il est aussi à noter que notre algorithme ne garantit pas l'optimalité du résultat en sortie, car il est tout à fait possible qu'il existe un autre sous-ensemble $\tilde{F} \neq F$, tel que $S_G(\tilde{F}) > S_G(F)$.
En revanche, l'algorithme qu'on a décrit dans cette section est fait de telle sorte que les nœuds classés dans la partie dense aient tous un δ supérieur à ceux classés dans la partie non dense. Il peut donc être assimilé à un algorithme qui détermine un seuil de la mesure δ (ou de toute autre mesure de densité dont l'ordre est celui utilisé en entrée) qui optimise certaines caractéristiques de S_G. Nous ne développerons pas cet algorithme au-delà de ce qui a été fait dans cette section. D'abord parce qu'il présente les points faibles qu'on vient de mentionner, mais aussi parce que nous allons présenter dans la section suivante une nouvelle approche, dont les résultats sont plus riches. Chacun de ces niveaux est constitué d'un ensemble de nœuds dont on évalue finalement la qualité.

4.3 Approche par rich club

Nous avons décrit dans le premier chapitre de cette thèse le phénomène du rich club dans un réseau. On a détaillé ses caractéristiques à l'aide d'un algorithme dont nous avons donné les points importants, et cela dans un cadre restreint aux réseaux non pondérés. Ici on va s'intéresser au même phénomène, mais en l'étendant au cas des réseaux pondérés. La définition qu'on donne par la suite aux parties non denses y est liée.

On part donc de la proposition suivante : au lieu de se concentrer sur l'analyse d'un réseau non pondéré pour en extraire la partie non dense, on choisit de passer par une représentation intermédiaire du réseau, dont les liens sont pondérés de sorte qu'ils reflètent les caractéristiques structurelles de la densité. Ces poids ont été introduits dans le précédent chapitre, à travers la mesure ω_{ij} (qui sert à calculer δ) et qui rappelons-le, octroie à chaque lien un poids en fonction des degrés et du nombre de voisins communs entre les nœuds de ses extrémités. Pour résumer, soit un réseau $G = (V, E)$, le réseau intermédiaire correspondant, qu'on notera W est défini par :

$$W = (V, E, \Omega) \text{ tel que } \Omega = \{\omega_{ij} \ \forall i, j \mid i \sim j\}. \tag{4.2}$$

On appelle *force d'un nœud* dans un réseau pondéré, la somme des poids des arêtes qui lui sont rattachés. On utilise aussi le terme poids d'un nœud pour désigner la même quantité, ce qui n'est pas à confondre avec les poids des arêtes $\omega_{i,j}$. Suivant cette définition, le réseau W a pour particularité que chaque nœud possède une force égale à sa valeur de δ.
On définit finalement les parties denses d'un réseau G par les rich clubs successivement[3] extraits de son réseau intermédiaire W. Il nous faut alors faire en sorte d'adapter l'algorithme présenté dans le chapitre 1 [124] au cas des réseaux pondérés. Pour cela, on choisit de prendre pour paramètre de richesse la valeur de δ, et ainsi faire en sorte qu'un rich club soit constitué d'un ensemble de nœuds fortement connectés, et dont les δ sont élevés.

Avant de donner une description de notre l'algorithme, on doit d'abord définir le modèle nul qui servira à la comparaison, en s'assurant que celui-ci ne contienne pas de corrélations $\delta - \delta$ entre paires de nœuds voisins. Car on rappelle que le modèle nul introduit dans [124] n'a pas d'équivalent pondéré, pour plusieurs raisons qu'on expliquera plus bas dans la partie 4.3.1. On retrouve cependant plusieurs modèles nuls pour les réseaux pondérés dans la littérature [125, 94, 109, 81], mais aucun d'entre eux ne fait consensus. On peut expliquer ceci par la diversité des objectifs dans l'ensemble des approches proposées. Car dès lors qu'on étudie des réseaux pondérés, on peut choisir de se concentrer sur la topologie en ignorant les poids, ou à l'inverse de se concentrer sur les poids en ignorant la topologie. Comme nous allons le voir, il est difficile de concilier les deux approches. On va présenter dans ce qui va suivre les différents modèles nuls proposés dans la littérature, et on justifiera la pertinence du choix pour lequel nous avons opté dans notre problématique.

4.3.1 Le cas du modèle nul

Nous cherchons à transformer un réseau pondéré W en un réseau aléatoire W^{null}, en éliminant les corrélations $\delta - \delta$ entre les nœuds voisins. Il faut de surcroît supprimer les corrélations degrés-degrés qui peuvent exister entre les nœuds voisins de W, de sorte à ce que les plus riches des nœuds (ceux ayant les δ les plus élevés) ne soient liés entre eux que par l'effet du hasard. Il faut tout de même que ce modèle préserve certaines caractéristiques cruciales, qui permettent de faire le lien entre W et W^{null}, et ainsi rendre pertinente leur comparaison.
Si on se réfère au cas des réseaux non pondérés, seules les corrélations degré-degré sont à prendre en compte, ce qui permet de bien exploiter un modèle nul comme le modèle de configuration. Rappelons que ce dernier consiste à réarranger aléatoirement les liens du réseau, tout en gardant la même séquence des degrés pour les nœuds. En revanche dans le cas des réseaux pondérés, la distribution des degrés n'est pas la seule quantité à conserver. On doit ajouter à celle-ci la distribution des poids (les $\omega_{i,j}$ dans notre cas) et la distribution des forces (la séquence des δ) de chaque nœud, pour caractériser au mieux le modèle nul. C'est précisément cet aspect qui comporte toute la difficulté. Il s'avère que trouver un algorithme qui permet d'obtenir un modèle nul répondant à toute ces caractéristiques n'est pas un problème simple, et on ne peut au mieux que s'en approcher.

[3] On expliquera ce qu'on veut dire par successivement dans la partie 4.4.1 de ce chapitre.

Nous donnons quelques exemples de modèles nuls dont nous soulignons les principales caractéristiques. Dans [125], les auteurs construisent un modèle nul avec une distribution de forces donnée, mais changent à la fois la distribution des poids et celles des degrés. Ici le modèle est approprié pour un cas où la topologie du réseau n'est pas prise en compte, car on y modifie le nombre de liens (en général le nombre de liens dans le modèle nul est supérieur à celui dans le réseau étudié), et par conséquent la séquence des degrés. De plus, les poids des liens sont eux aussi changés par la procédure qui est décrite.

Les auteurs de [94] proposent une procédure de randomisation qui consiste à considérer chaque arête comme deux arêtes dirigées de sens opposés, puis d'intervertir aléatoirement les arêtes sortantes de chaque nœud. Ce procédé conserve la valeur de la somme des poids des arêtes sortantes, mais pas celle des arêtes rentrantes.
Plusieurs remarques peuvent être émises sur cette procédure, d'abord le fait d'effectuer l'interversion uniquement entre les arêtes sortantes du même nœud fait que le recâblage est local, et ceci limite clairement l'aspect aléatoire du modèle nul. En effet, si on étudie un réseau dans lequel on prend en compte à la fois la topologie et les poids des liens, et qu'en plus les deux propriétés sont positivement corrélées, alors en intervertissant localement les liens on finit par obtenir un modèle nul dans lequel les rich clubs n'ont pas significativement changé. La raison est que les liens de fort poids étaient initialement partagés entre les nœuds de forces élevées, et continuent à l'être.

Un autre modèle est proposé dans [109]. Il génère un réseau avec des distributions de degrés et de forces qui convergent vers celles du réseau dont il est tiré, lorsque celui-ci est suffisamment grand. Les observations empiriques que nous avons faites sur des réseaux du monde réel suggèrent que cette limitation de taille a un impact plus important lorsque les réseaux étudiés sont caractérisés par des distributions de degrés et/ou de forces qui obéissent à une loi de puissance.

Notre choix de modèle nul

Rappelons qu'il est prioritaire que le modèle nul à adopter puisse à la fois éliminer les corrélation δ-δ et degré-degré entre les nœuds voisins, tout en préservant les séquences des degrés k_i, des poids $\omega_{i,j}$ et des forces $\delta(i)$.

Dans le modèle décrit par [81] et [94], la démarche consiste en partant d'un réseau pondéré W, à randomiser à la fois la topologie et la répartition des poids. Ceci peut s'effectuer en deux étapes simples : d'abord modifier la topologie en recâblant aléatoirement les liens de W, afin de rompre les corrélations de degrés, et ensuite redistribuer les poids de manière aléatoire sur les arêtes du réseau résultant. Par construction, ce modèle nul préserve la séquence des degrés et celle des poids, et donc la valeur moyenne de δ, mais ne préserve pas la séquence de celle-ci.
On se retrouve donc face un problème de taille, car en modifiant la séquence des δ, on n'est plus en mesure d'assurer que le filtre de l'algorithme des rich club produise un modèle nul ayant le même nombre de nœuds que le réseau analysé. En effet rappelons que nous avons :

$$\phi(s) = \frac{E_{>s}}{F(s)}$$

qui représente le rapport entre la somme des poids des liens du sous-réseau $W_{>s} = (V_{>s}, E_{>s})$ induit par l'ensemble $V_{>s}$ des nœuds ayant une force supérieure à s, et un facteur de normalisation $F(s)$[4]. Par ailleurs, si la séquence des forces n'est pas conservée, alors l'équivalent de l'ensemble $V_{>s}$ pour le modèle nul n'est pas nécessairement constitué du même nombre de nœuds, ce qui rend la comparaison moins pertinente. Pour parer à cela, nous allons introduire des modifications dans le calcul du coefficient du rich club $\phi(s)$, qui sont détaillées dans la prochaine section.

4.4 L'algorithme ItRich

Dans cette section, nous présentons un algorithme pour l'extraction des parties denses, sous forme d'un sous-ensemble de nœuds de fort δ, en utilisant une approche par rich clubs pondérés. Comme le modèle nul ne préserve pas la séquence des δ calculée sur le réseau initial, nous devons aussi introduire une stratégie de filtrage des nœuds, différente de celle qui consiste à sélectionner les nœuds dont la force est supérieure à une certaine valeur δ.

[4]On retrouve parfois des articles désignant par $E_{>s}$ le *nombre de liens* et non la somme de leurs poids respectifs

Ainsi, étant donné un réseau G, notons par V_n l'ensemble constitué des n nœuds [5] de plus grandes valeurs de δ

$$V_n = \{v_1, v_2, ..., v_n \mid \delta(v_i) > \delta(v_{i+1})\}.$$

Nous définissons le coefficient pondéré du rich club $\phi(n)$ comme la somme des poids de tous les liens contenus dans le sous-réseau induit par V_n divisée par la valeur de cette même somme calculée sur l'ensemble des nœuds du réseau initial :

$$\phi_G(n) = \frac{\sum\limits_{(i,j)\in E\cap(V_n\times V_n)} \omega_{i,j}}{\sum\limits_{(i,j)\in E} \omega_{i,j}}. \tag{4.3}$$

Nous pouvons voir $\phi_G(n)$ comme le rapport entre la quantité de ressources ω partagées entre les n nœuds ayant les δ les plus élevés, et la valeur totale des ressources disponibles dans le réseau. En comparant ce coefficient du rich club pondéré à celui obtenu à partir du modèle nul (par analogie avec le cas des réseaux non pondérés), cette définition de $\phi_G(n)$ garantit que les deux quantités sont calculées à partir de réseaux ayant le même nombre de nœuds n, comme dans le cas non pondéré.

Nous définissons maintenant notre paramètre de rich club pondéré $R_G(n)$ comme :

$$R_G(n) = \phi_G(n) - \phi_G^{null}(n) \tag{4.4}$$

où $\phi_G^{null}(n)$ est le coefficient du rich club pondéré, calculé à partir du modèle nul.

Nous définissons ensuite un rich club pondéré comme étant l'ensemble des nœuds qui maximise $R_G(n)$:

$$V_M \ : \ M = \arg\max_{n\in[1,N]}\{R_G(n)\}. \tag{4.5}$$

En comparant notre méthode à la méthode classique des réseaux non pondérés, il convient de noter que nous avons la différence $\phi_G - \phi_G^{null}$ au lieu du rapport ϕ_G/ϕ_G^{null}.
Ce choix est justifié dès lors que nous avons $\phi_G(n) \ll 1$ et $\phi_G^{null}(n) \ll 1$, et en même temps une valeur élevée pour le ratio $\phi_G(n)/\phi_G^{null}(n)$. Ce cas se produit pour de petites valeurs de n, et est souvent observé pour les réseaux dont le modèle auxiliaire W_G contient une forte corrélation δ-δ entre les nœuds voisins, tout en ayant une distribution δ à queue lourde. Cela s'explique par la présence d'un petit nombre de nœuds qui ont des valeurs de δ beaucoup plus élevées que la moyenne, et qui sont également liés les uns aux autres, comme le suggère la corrélation. Cette configuration engendre à la fois une croissance rapide de $\phi_G(n)$ lorsque n est petit, et une croissance lente de $\phi_G^{null}(n)$ en raison de la randomisation. Par conséquent, le ratio $\phi_G(n)/\phi_G^{null}(n)$ atteint son maximum pour des petites valeurs de n, ce qui donne un petit rich club pondéré (qu'on appellera $\delta-$*rich club* par la suite), ne contenant parfois que deux nœuds et une seule arête, ce qui n'est pas très pertinent comme résultat [6].

Pour récapituler, notre méthode introduit deux nouveautés méthodologiques par rapport à celles que l'on trouve dans la littérature : la première concerne le remplacement du paramètre de richesse continue par un entier représentant le nombre de nœuds n ayant le δ le plus élevé, et la seconde est dans la mesure de la distance au modèle nul à travers la différence entre les valeurs des coefficients du rich club, au lieu de leur rapport.

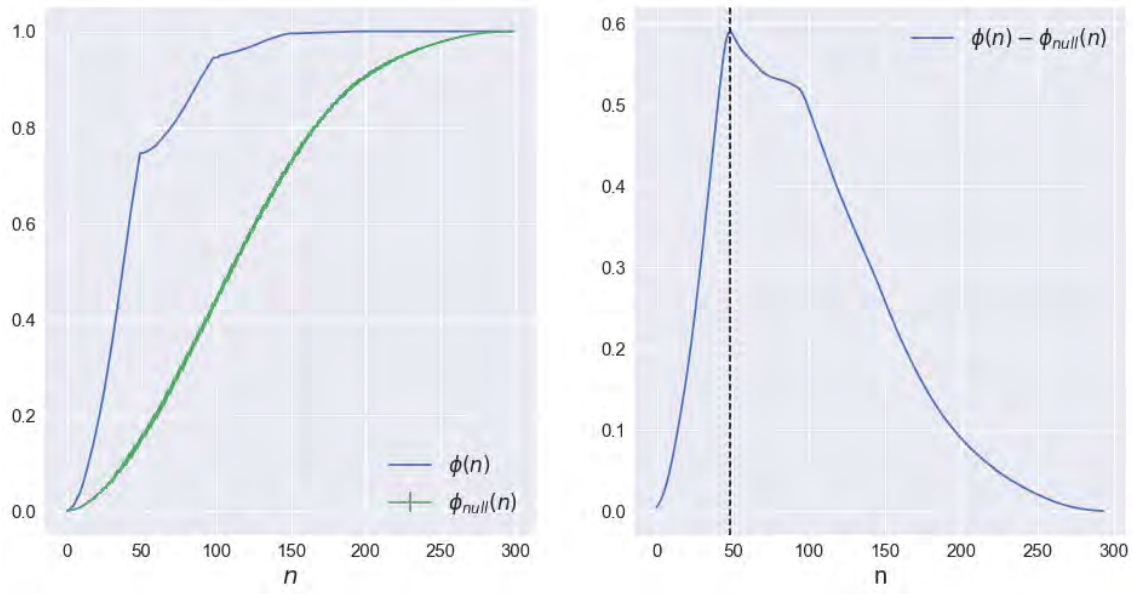
4.4.1 Insuffisance d'une seule itération

Nous allons maintenant montrer qu'une seule exécution de l'algorithme décrit ci-dessus ne suffit pas à produire une information complète sur la structure globale d'un réseau.

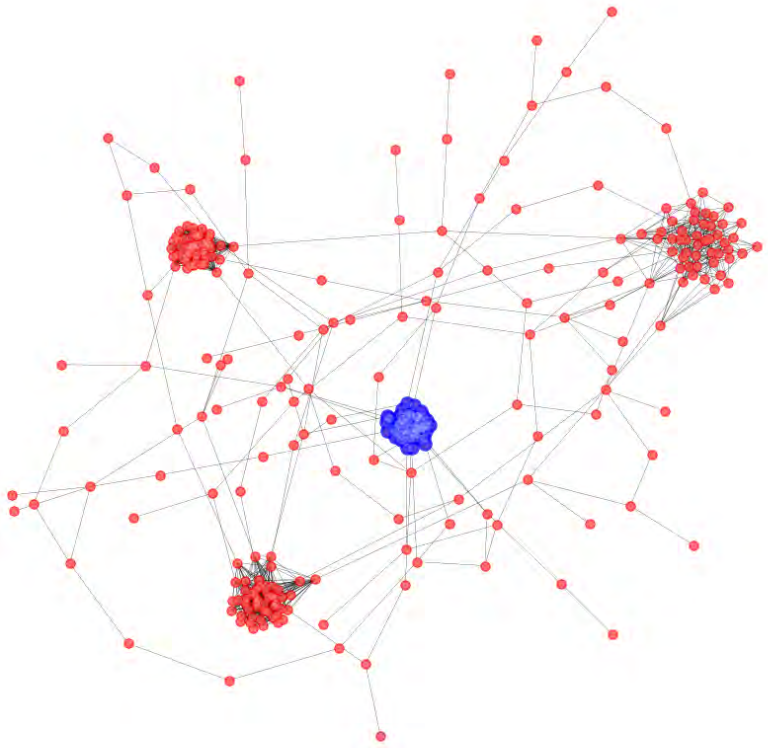
Considérons un modèle treeCom généré à partir des paramètres suivants : un arbre aléatoire de $N_t = 100$ nœuds, $N_c = 4$ blocs de $N_b = 50$ nœuds chacun, et de probabilités $P_1 = 0.8$, $P_2 = 0.6$, $P_3 = 0.4$, $P_4 = 0.2$. On prend $p_i^t = \frac{1}{4\cdot N_t}$. Une fois que ce réseau jouet créé, la procédure décrite plus haut est appliquée, et les résultats du calcul sont représentés sur la figure 4.2. Nous voyons sur la figure 4.2 qu'il ne suffit pas de calculer un seul $\delta-$*rich club* pour mettre en évidence toute la partie dense d'un réseau. En effet, des quatre clusters que contient le graphe jouet, seul le premier qu'on génère à

[5] Si plusieurs nœuds ont le même δ, nous choisissons aléatoirement l'ordre dans lequel ils sont ajoutés. En pratique, cette situation est très peu probable et l'ordre de choix a un impact négligeable dans les grands réseaux réels.

[6] Une autre justification de ce choix est d'ordre pratique, et concerne l'algorithme que nous allons introduire. Celui-ci devient plus rapide lorsque nous évaluons la différence que lorsqu'on calcule le rapport.



(a)



(b)

Figure 4.2: (a) L'évolution du paramètre et du coefficient du rich club pondérés, pour le réseau jouet décrit ci-dessus, et son modèle nul correspondant. (b) Le résultat de l'exécution de l'algorithme . Les nœuds détectés comme constituant le $\delta-$*rich club* sont ceux en bleu, ne constituant qu'un seul des quatre clusters de la partie dense.

partir du paramètre P le plus élevé, est ressorti comme sous-ensemble dense. C'est pour cela que nous introduisons dans la prochaine partie un processus itératif permettant de récupérer les uns après les autres tous les sous-ensembles denses, et d'en estimer la qualité.

4.4.2 Calcul itératif et qualité d'un $\delta-$*rich club*

Nous répétons le processus décrit ci-dessus de manière itérative, tout en supprimant, à chaque itération i, le $\delta-$*rich club* calculé en utilisant eq. (4.5) à l'itération $i-1$, et en conservant les mêmes poids sur les liens restants après la suppression. Cela permet d'extraire un par un les sous-ensembles de nœuds à δ élevé, et ce processus est répété jusqu'à ce que le critère d'arrêt de l'algorithme soit atteint. Une fois ces itérations terminées, on obtient une série de sous-ensembles denses.
Afin d'évaluer la qualité de chacun d'entre eux, nous calculons la valeur moyenne de la fonction donnée par eq. (4.4), qui donne la distance entre le coefficient ϕ sur le réseau et son modèle nul :

$$Q_i = \frac{1}{N_i} \cdot \sum_{n=1}^{N_i} R_{G_i}(n) \tag{4.6}$$

où G_i est le réseau obtenu après avoir supprimé les $i-1$ premiers $\delta-$*rich clubs* de G et N_i est le nombre de nœuds de G_i. Rappelons que les poids de G_i ne sont pas recalculés mais hérités de G. La mesure Q_i suit généralement le schéma suivant : en partant de sa valeur maximale, qui est la qualité mesurée sur le premier $\delta-$*rich club* extrait, on passe à des valeurs inférieures se référant aux qualités mesurées sur les $\delta-$*rich club* de rangs inférieurs. Cette diminution n'est pas garantie, bien qu'elle reste valable sur des réseaux ayant la propriété d'avoir plusieurs sous-ensembles de δ moyen élevé. Elle peut toutefois présenter un comportement imprévisible [7]. Dans l'exemple du modèle jouet, la mesure de qualité tombe à (presque) zéro une fois que tous ses clusters ont été extraits, comme nous pouvons le voir sur la figure 4.3. Ceci a lieu d'abord car le poids des liens qui restent dans le réseau (après l'extraction des clusters) est très faible, et ensuite car il n'y aura plus de clusters dans le réseau. Cela entraîne la diminution des valeurs $R(n)$ et de la mesure de qualité Q.

Jusqu'à présent, nous avons décrit un processus qui calcule une séquence de rich clubs pondérés, chacun étant ensuite associé à une mesure de qualité. Celle-ci est obtenue par la différence de valeur moyenne qu'il y a entre le coefficient du rich club pondéré, calculé sur le réseau et son modèle nul. Elle est ensuite normalisée de sorte qu'elle soit toujours inférieure à 1.
Pour la suite, nous utilisons cette mesure de qualité afin d'accepter ou au contraire de refuser les différents $\delta-$*rich clubs* dans la partie dense. Il s'agit alors de fournir un seuil en dessous duquel la mesure de qualité d'un $\delta-$*rich club* donné est considérée comme trop faible pour que le $\delta-$*rich club* correspondant puisse être compté dans la partie dense. Il n'y a pas de manière précise de définir un tel seuil, ce sera par conséquent un choix *ad hoc* qui varie en fonction du type de données étudiées et dépend souvent de la valeur Q_1 associée au premier $\delta-$*rich club* qu'on extrait. Il faut tenir compte du fait que les parties denses et les parties non denses obtenues en sortie dépendent de ce choix. Par exemple un seuil trop élevé signifie une forte sélectivité au sein de la partie dense, et par extension inclut dans la partie non dense des nœuds dont le δ peut être relativement élevé. Cette valeur (notée Q_{seuil}) caractérise le niveau de dichotomie du réseau. Par défaut, nous fixons Q_{seuil} au dixième de la valeur maximale atteinte par Q, (généralement égale à Q_1), ceci rend des résultats satisfaisants parmi les nombreux tests que nous avons effectués. Il est cependant préférable d'en choisir la valeur une fois terminée l'extraction des $\delta-$*rich clubs* du réseau (qui continue tant que Q est supérieur à zéro), et de choisir le seuil une fois toutes les valeurs de Q connues.

Nous appelons notre algorithme ItRich (pour Iterative rich club), et nous donnons ses différentes

[7] Par exemple, si le réseau n'a pas une structure particulière comme dans le cas d'un réseau Erdös-Rényi.

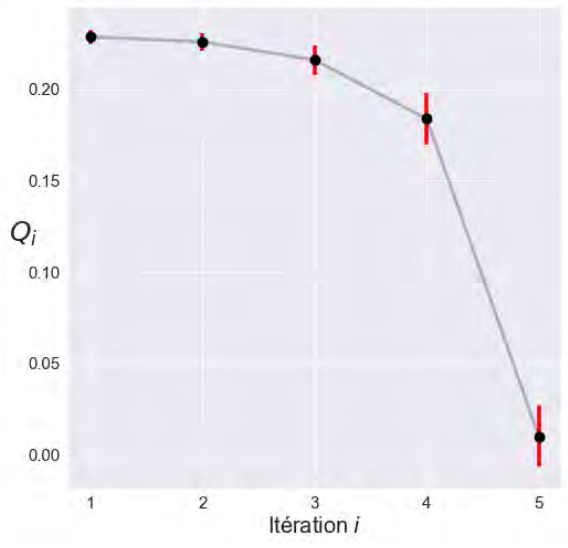


Figure 4.3: La valeur de la différence moyenne entre le coefficient $\phi(n)$ dans le réseau à l'itération i et son modèle nul, pour un modèle jouet treeCom à 4 clusters. Les moyennes sont recalculées 100 fois sur autant de modèles nuls différents, et les barres d'erreurs résultantes sont représentées en rouge sur la figure.

étapes comme suit :

Algorithm 2: ItRich

Data: Un réseau $G = (V, E)$

Result: Deux ensembles, le premier contenant les $\delta-$*rich clubs* qui constituent la partie dense, et le deuxième les nœuds de la partie non dense

Un ensemble de paires (U_i, Q_i) où $U_i \subset V$ est le $\delta-$*rich club* extrait à la $i-$ème itération, et Q_i son indice de qualité défini par eq. (4.6)

initialisation;

$V = \{v \mid v \in G\}$;

$E = \{(u, v) | u \in V, v \in V, u \sim v\}$;

$\Omega = \{\omega_{u,v} | (u, v) \in E\}$;

$W = (V, E, \Omega)$;

Collection $= \emptyset$;

$S = 1$;

while $S > 0$ **do**

  Calculer le modèle nul W^{nul} de W;

  Calculer $R_G(n),\ n \in \{1, 2, ..., |V|\}$ (*cf.* eq. (4.4));

  En déduire V_M (*cf.* eq. (4.5));

  Calculer Q (*cf.* eq. (4.6));

  Collection = Collection $\cup (V_M, Q)$;

  $V = V \setminus V_M$;

  $E = \{(u, v) | u \in V, v \in V, u \sim v\}$;

  $\Omega = \{\omega_{u,v} | (u, v) \in E\}$;

  $W = (V, E, \Omega)$;

  $S = \sum_{(u,v)\in E} \omega_{u,v}$;

Fixer la valeur de Q_{seuil}(Par défaut $\frac{Q_1}{10}$);

D=$\{V_M | (V_M, Q) \in$ Collection si $Q > Q_{seuil}\}$;

ND $= G \setminus D$;

return D,ND

4.4.3 Complexité de l'algorithme

Penchons-nous à présent sur l'étude de la complexité de l'algorithme, nous pouvons facilement remarquer que ce sont les deux premières lignes de la boucle de l'algorithme 2 qui constituent les étapes les plus coûteuses. En effet calculer le modèle nul peut s'avérer assez lent. L'une des approches possibles consiste à tirer aléatoirement deux liens $e_1 = (i_1, j_1)$ et $e_2 = (i_2, j_2)$ du réseau, puis d'intervertir l'une des extrémités du premier lien avec celle du second, pour obtenir finalement une configuration du type $e_1^{'} = (i_1, i_2)$ et $e_2^{'} = (j_1, j_2)$. En répétant cette étape un assez grand nombre de fois, on est assuré d'obtenir le modèle nul décrit plus haut à la sortie. Cependant le nombre de fois où l'on doit répéter ce processus est de l'ordre de $O(m^2)$ (le nombre de paires de liens disponibles), ce qui rend la procédure assez lente, en particulier pour les réseaux ayant une grande densité de liens. Une autre possibilité consiste à passer directement par le modèle de configuration, dont le résultat est faisable en $O(m \cdot log(m))$, sur lequel on redistribue dans un deuxième temps les poids sur les liens de manière aléatoire. On trouve donc que le calcul du modèle nul est au moins aussi lent que le calcul d'un modèle de configuration. Il est cependant parfois possible, comme nous allons le montrer dans cette section, de réduire la complexité de cette étape, sous certaines contraintes que nous détaillons plus bas.

La deuxième ligne de la boucle consiste à trier une liste de taille $|V| = N$, ce qui peut être fait en $O(N \cdot log(N))$ (rappelons que $N = |V|$ est la taille du réseau). On exploite ensuite l'ordre attribué par la mesure δ pour calculer $\phi(n)$, que l'on compare à $\phi^{nul}(n)$(la valeur calculée sur le modèle nul). On répète cela $|V|$ fois pour obtenir $R_G(n)$. Cette étape peut être raccourcie, en construisant en parallèle les N[8] sous-réseaux induits par $V_n \times V_n$ pour $n \in \{1, \cdots, N\}$. Ceci fixe la borne inférieure de la complexité pour cette étape à $O(m)$. Nous avons donc une première approche qui, dans le meilleur des cas, est obtenue en $O(m \cdot log(m))$.

Un métamodèle nul alternatif

Nous allons maintenant montrer qu'il n'est pas toujours nécessaire de passer par un modèle de configuration pour calculer ϕ^{null}. Nous nous basons sur la forte corrélation entre les valeurs de δ et celles du degré, que l'on observe chez les nœuds de W^{null} et qui sont affichées sur la figure 4.4. Cette figure montre qu'il existe une forte corrélation linéaire entre δ et le degré dans le modèle nul que nous avons choisi d'adopter. Le coefficient directeur de la droite de régression est égal à la moyenne $\frac{1}{m}\sum_{(i,j)\in E}\omega_{i,j}$ des poids topologiques calculés sur les liens du réseau.

On peut donc mettre au point un métamodèle nul permettant d'affecter une valeur de δ à chaque nœud, sachant que le degré de celui-ci reste inchangé. Il demeure cependant nécessaire de connaître l'ordre de grandeur de l'erreur qui sépare le nuage de points de la droite de régression.
On peut commencer en se représentant $\delta^{null}(k)$ par les valeurs de la variable aléatoire suivante :

$$X_\delta(k) = \sum_{i=1}^{k} X_{\omega,i} \tag{4.7}$$

avec $X_{\omega,i}$ une variable aléatoire, dont la loi de probabilité suit la loi des poids topologiques $P(\omega)$. Comme notre modèle nul est construit de sorte que les poids soient aléatoirement redistribués sur les liens, on peut facilement admettre que la variable aléatoire $X_\delta(k)$ est une somme de variables aléatoires $X_{\omega,i}$ indépendantes et identiquement distribuées.

On présente ici un modèle permettant d'estimer la valeur moyenne de $X_\delta(k)$ en utilisant la moyenne et la variance empirique de l'ensemble Ω des ω_{ij}, sous réserve que ces quantités puissent être prises pour estimateurs. En pratique, si la distribution des ω_{ij} dans le réseau suit une loi de puissance [9] , alors on doit faire attention à la valeur de l'exposant γ. Si celle-ci est inférieure à trois, la variance de la loi prend théoriquement une valeur infinie, et est mal estimée par sa valeur empirique. Il en est de même pour la moyenne $\overline{\omega}$, son estimateur empirique n'est efficace que si l'exposant est d'une valeur supérieure à 2. Il faut par conséquent d'abord analyser la distribution qui caractérise l'ensemble Ω, et dans le cas ou celle-ci suit une loi de puissance (ce qui est souvent le cas pour les données de réseaux réels), il faudrait s'assurer que l'exposant soit supérieur à trois. Les figures 4.6a et 4.6b montrent ces distributions pour les deux réseaux qu'on utilise comme exemple dans cette partie. Aucune des deux ne suit une loi de

[8] N étant le nombre de nœuds mais aussi le nombre maximal d'étapes pouvant être effectuées en parallèle
[9] du type $p(\omega) = a \cdot \omega^\gamma$, avec $\gamma > 0$

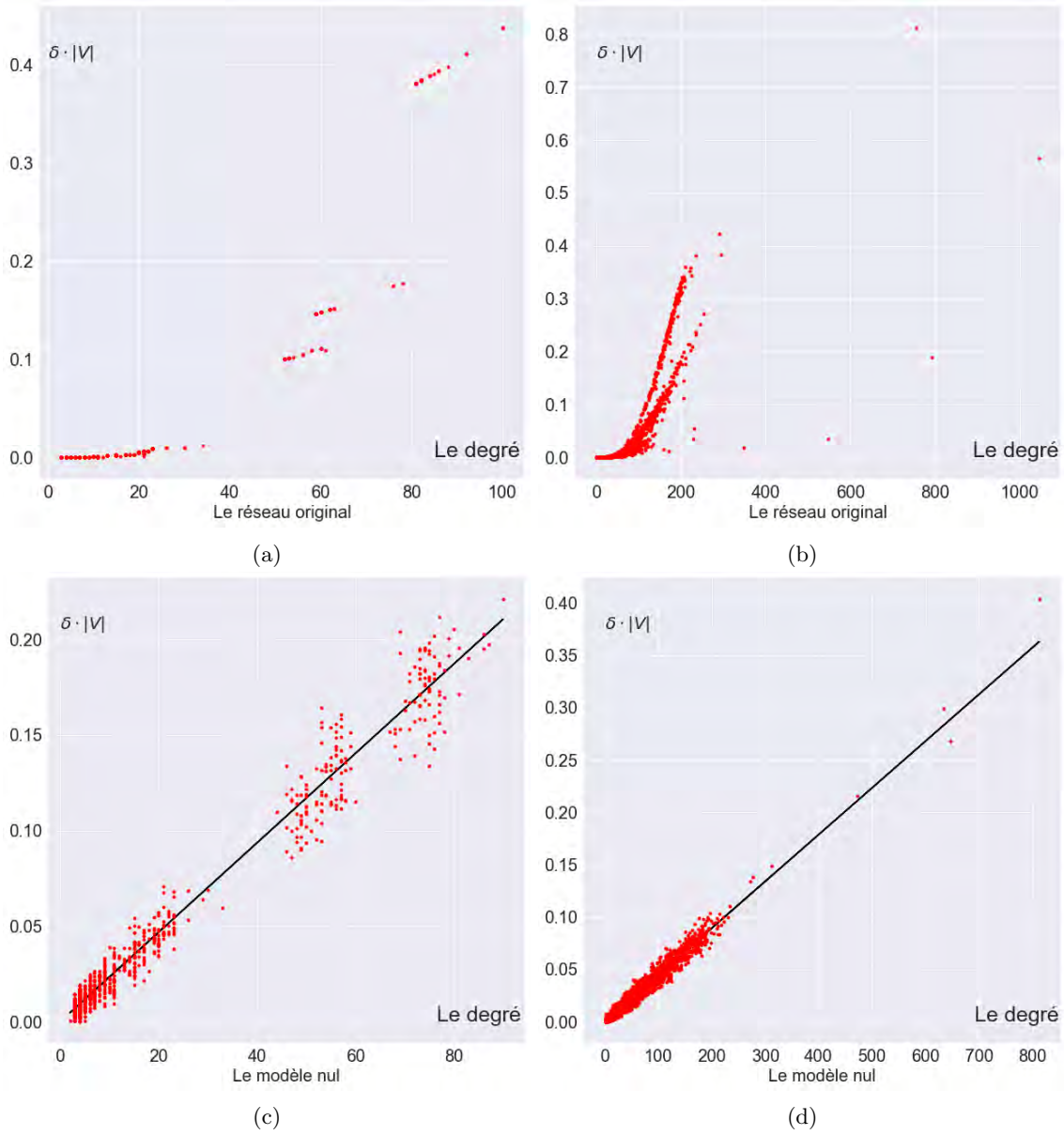


Figure 4.4: δ en fonction du degré dans un réseau électrique réel (a), un échantillon du graphe de Facebook (b) et de leurs modèles nuls respectifs (c) et (d). Dans les deux cas, la droite de régression a pour coefficient directeur la moyenne des poids topologiques $\omega_{i,j}$ du réseau.

puissance, même si la distribution de la figure 4.6b est assimilable à une loi de puissance tronquée par une exponentielle.

Si les hypothèses évoquées dans le paragraphe précédent sont vérifiées, alors à chaque nœud de degré k on attribue la valeur suivante de δ^{null} :

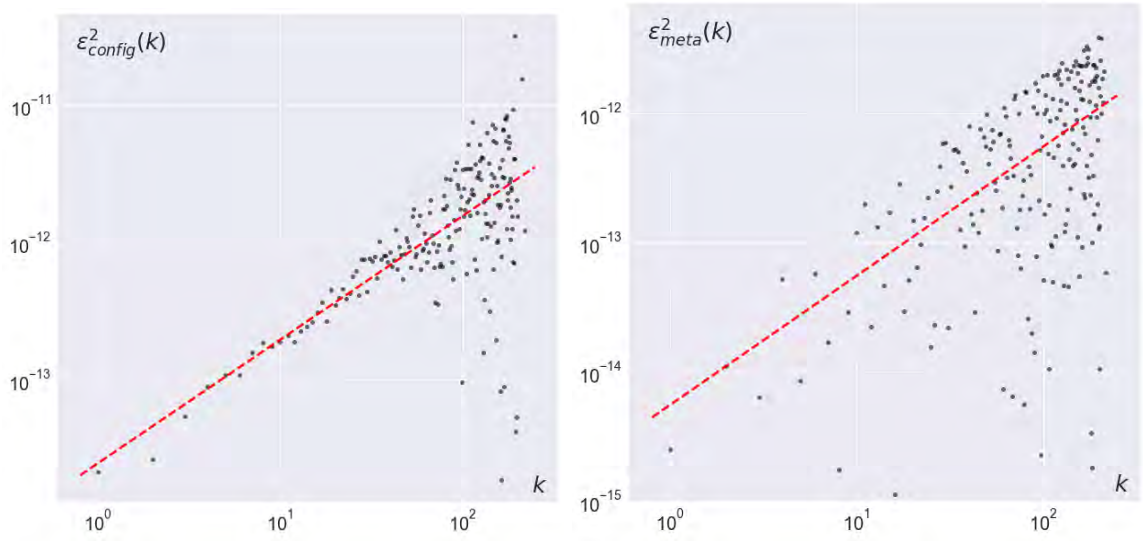
$$\delta^{null}(k) = \overline{\omega} \cdot k + \epsilon(k) \tag{4.8}$$

avec

$$\epsilon(k) \sim U([-\sqrt{k} \cdot \overline{\sigma}, \sqrt{k} \cdot \overline{\sigma}]) \tag{4.9}$$

avec $\overline{\omega}$ la moyenne empirique des ω_{ij} dans le réseau, $\overline{\sigma}^2$ la variance empirique, et $U[a, b]$ la loi uniforme continue sur l'intervalle $[a, b]$. La principale simplification du modèle est donc que la distribution du bruit suit une loi uniforme pour les nœuds ayant le même degré, d'une variance égale à k fois la variance

empirique de la distribution des ω_{ij} [10]. Cette simplification n'est évidemment pas toujours vérifiée, car la distribution de la variance dépend à la fois de la distribution des poids topologiques ω_{ij}, ainsi que de celle des degrés. Nous pouvons voir sur la figure 4.5b que notre modèle reproduit bien des termes d'erreurs ayant un ordre de grandeur proche de celui des termes calculés à travers le modèle de configuration (*cf.* fig. 4.5a), mais n'a pas la même distribution. Il est difficile de connaître la distribution de $\epsilon(k)$ dans le modèle de configuration sans avoir à calculer ce dernier, mais ceci ne constitue pas une contrainte majeure pour notre metamodèle. Dans le cas où $\overline{\omega}$ et $\overline{\sigma}^2$ sont eux aussi du même ordre de grandeur (à noter que ceci n'est plus vrai en dehors des hypothèses évoquées plus haut, où la variance peut devenir très grande devant la moyenne), et pour des graphes ayant une forte proportion de nœuds dont le degré est assez grand pour considérer k très grand devant $\sqrt{k}$, le terme d'erreur ne constitue qu'une petite fluctuation de δ^{null} (*cf.* eq. (4.8)) autour de sa valeur moyenne.



(a) (b)

Figure 4.5: (a) Le carré de l'erreur obtenu sur un modèle nul basé sur le modèle de configuration en fonction du degré k, (b) Le carré de l'erreur générée par une variable aléatoire distribuée comme décrit dans eq. (4.9), la droite en pointillés rouges est celle dont le coefficient directeur est la variance $\overline{\sigma}^2$. Le réseau dont on calcule le modèle nul est un échantillon du graphe de Facebook.

Une fois ces valeurs générées, on trie les nœuds dans l'ordre décroissant des δ^{null}, et il ne reste plus qu'à estimer la valeur de $\phi^{null}(n)$. On rappelle que celle-ci est égale à la somme des poids des ω_{ij} contenue dans le sous-réseau généré par V_n, qui contient les n nœuds les mieux classés dans le modèle nul. Or, en pratique nous n'avons besoin de connaître que le nombre de liens contenus dans ce sous-réseau (notons-le m_n). Sachant la probabilité que deux nœuds u et v soient reliés dans le modèle nul, qui est égale à $\frac{k_u \cdot k_v}{2m-1}$ [11], nous pouvons estimer le nombre de liens :

$$m_n = \sum_{u,v \in V_n \times V_n} \frac{k_u \cdot k_v}{2m-1}. \tag{4.10}$$

Nous pouvons finalement estimer la valeur de $\phi^{null}(n)$ en tirant aléatoirement m_n valeurs de $\omega_{i,j}$

[10] Ceci découle naturellement du fait que les variables aléatoires $X_{\omega,i}$ sont considérées comme indépendantes et identiquement distribuées

[11] On peut facilement s'en convaincre étant donné que la topologie de notre modèle nul est celle du modèle de configuration, qui est lui-même caractérisé par cette probabilité de former des liens entre les sommets de degrés connus.

dont nous calculons la somme. L'algorithme décrit dans cette partie est résumé ci-dessous.

Algorithm 3: Métamodèle nul

Data: Un réseau $G = (V, E)$
Result: $\phi^{null}(n)$ tel que $n \in \{1, 2, ..., |V|\}$
initialisation;
Calculer W et l'ensemble Ω des poids topologiques des liens de W;
Calculer $\overline{\omega}$ et $\overline{\sigma}$ de Ω
for $u \in V$ **do**
  $\underline{\delta}^{null}(u) = \overline{\omega} \cdot k_u + \epsilon(k_u)$
Calculer l'ensemble V_{sort} tel que $V_{sort} = \{v_0, v_1, ..., v_N\}$ et $\delta^{nul}(v_i) > \delta^{nul}(v_{i+1})$;
$m_0 = 0$;
$V_0 = \emptyset$;
$\phi^{null}(0) = 0$;
for $n \in \{1, 2, ..., |V|\}$ **do**
  $m_n = m_{n-1} + \sum_{u \in V_{n-1}} \frac{k_{v_n} \cdot k_u}{2m-1}$
  $V_n = V_{n-1} \cup \{v_n\}$;
  $\phi^{null}(n) = \phi^{null}(n-1) + \sum_{m_{n-1}}^{m_n} t(\Omega)$ tel que $t(\Omega)$ est la valeur d'un tirage aléatoire sans remise de l'ensemble Ω.
return $\phi^{null}(n)$ avec $n \in \{1, 2, ..., N\}$

Nous montrons sur les figures 4.6c et 4.6d une comparaison entre les valeurs de $\phi^{null}_{meta}(n)$ calculées par l'algorithme que nous présentons ici, et celles calculées à travers le modèle de configuration. Plusieurs modèles de configurations ont été calculés pour quantifier les éventuelles fluctuations de $\phi^{null}_{config}(n)$. Nous pouvons observer que, hormis une légère différence pour les faibles valeurs de n, l'algorithme proposé ici reproduit avec une grande précision la courbe de $\phi^{null}_{config}(n)$ qu'on obtient par la moyenne de 100 modèles de configuration.

Le gain potentiel de complexité

Ce métamodèle nul présente l'avantage d'être plus rapide que le calcul qui est basé sur le modèle de configuration, car il n'est pas nécessaire de passer par la randomisation des liens du réseau original pour le calcul de ϕ^{null}. Sa complexité est bornée par la seconde boucle de l'algorithme 3. Celle-ci est de l'ordre de $O(N^2)$, mais n'est pas calculable en parallèle contrairement à ce qui peut être fait pour le calcul de ϕ^{null} qui se base sur le modèle de configuration.
Ce métamodèle nul fait passer l'algorithme ItRich d'une complexité de $O(m \cdot log(m))$ à une complexité de $O(N^2)$. Ceci peut constituer un gain considérable dans le cas des réseaux à forte densité, dont le nombre de liens m est très grand devant le nombre de nœuds N.
Il ne faut cependant pas oublier les limites du métamodèle, que l'on peut résumer dans les points suivants : d'abord la distribution des poids topologiques doit être approchée par une densité de probabilité dont la variance n'est pas excessivement grande comparée à sa valeur moyenne, et ensuite, il faudrait avoir une proportion non négligeable de nœuds dans le réseau ayant un degré k assez grand pour que $\sqrt{k} \ll k$. Ces propriétés sont généralement bien vérifiées par les réseaux réels, dont la taille est supérieure à quelques centaines de nœuds. Dans le cas inverse, (réseaux de petite taille) le calcul de ϕ^{null} est à faire à travers le modèle de configuration, sans une grande différence de temps de calcul.

4.5 Tests sur des réseaux synthétiques

Nous allons tester l'algorithme qu'on vient de développer sur un ensemble de réseaux synthétiques. On modifie ces derniers en rajoutant des perturbations de la même manière que dans [78]. Nous mesurons la performance de notre algorithme, en calculant les valeurs des indices de rappel (proportion de vrais positifs) et de spécificité (proportion de vrais négatifs). Nous donnons plus bas les définitions de ces deux indices compte tenu des résultats de l'algorithme. Enfin, nous comparons ces performances avec

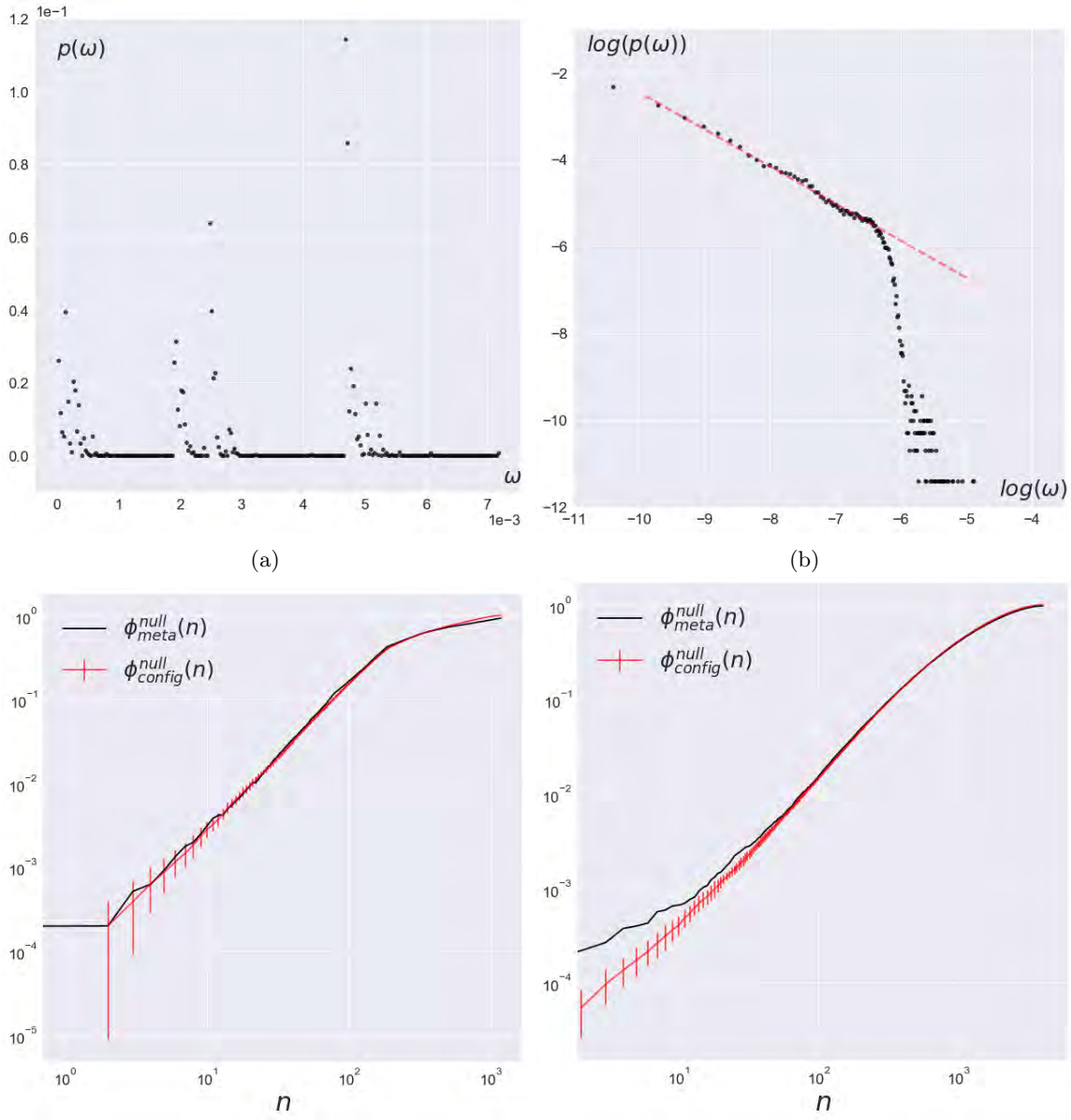


Figure 4.6: (a) la distribution $p(\omega)$ des poids topologiques dans le réseau électrique , (b) la distribution $p(\omega)$ de l'échantillon de Facebook sur une échelle logarithmique. Sur (c) et (d) La courbe noire indique les valeurs de $\phi^{null}(n)$ calculées par le métamodèle nul, en rouge la moyenne ainsi que l'écart type de $\phi^{null}(n)$ calculés à partir de 100 modèles de configurations, sur les mêmes réseaux (électrique (c), et échantillon du graphe de Facebook (d)).

celles obtenues par OSLOM [78], un algorithme que nous avons détaillé dans le chapitre 1, et qui est largement utilisé en analyse de réseaux.

Lors de ces tests, un seuil variable a été choisi. Une mesure de qualité Q_{ER_i} est calculée à chaque itération, obtenue en appliquant ItRich sur un réseau d'Erdös-Rényi, avec un paramètre de probabilité égal à $p_i = \overline{\delta}_i^{\frac{1}{4}}$, afin de s'assurer que le réseau Erdös-Rényi et le réseau évalué à l'itération i ont la même valeur de $\overline{\delta}_i$. Rappelons aussi que, pour un grand nombre de nœuds, un tel réseau d'Erdös-Rényi vérifie la valeur $\overline{\delta} = p^4$ (voir le calcul de δ sur un modèle de blocs stochastiques dans le chapitre 3). Si après un certain nombre d'itérations, la qualité Q_i d'un $\delta-$*rich club* extrait de G_i devient inférieure à Q_{ER_i}, nous décidons que nous avons atteint le point où le $\delta-$*rich club* n'est plus significatif. L'un des avantages qu'on tire d'un réseau Erdös-Rényi est le fait qu'il n'a pas de structure modulaire, ce qui nous assure

qu'il ne contient aucun sous-ensemble dont la densité est significativement plus élevée que celle du reste du réseau. Il y a cependant un autre avantage, qui est relatif aux type de données synthétiques étudiées ici. En effet, comme nous le verrons plus bas, la construction de ces données fait que le sous-réseau induit par la partie non dense possède une structure comparable à celle d'un réseau d'Erdös-Rényi, d'où l'intérêt de ce choix.

Données expérimentales
Les données synthétiques sont basées sur un modèle de Lancichinetti-Fortunato-Radicchi (LFR) [77] auquel des nœuds sont ajoutés. Les détails de la construction sont détaillés ci-dessous. Rappelons que dans un réseau de LFR, chaque nœud est assigné à une communauté, suivant les paramètres du modèle. Ces paramètres sont (entre autres) l'exposant γ de la distribution en loi de puissance des degrés, l'exposant β de la distribution en loi de puissance de la taille des communautés, et μ la proportion de liens qu'un nœud partage avec des nœuds en dehors de sa propre communauté.

Une fois le modèle LFR généré, nous ajoutons des nœuds de manière à créer deux classes : la première est la partie dense, composée des nœuds du modèle LFR initial, et la seconde, la partie non dense, contenant les nœuds qui ont été ajoutés. Ces derniers sont connectés suivant une procédure bien précise, de manière à ce que leur poids δ ne dépasse jamais celui du δ minimum mesuré sur le réseau LFR initial.

Cette contrainte permet d'assurer que nos deux classes ont bien été séparées suivant des critères pertinents, ce qui légitime la vérité de terrain qui sert à l'évaluation de l'algorithme.

Afin de connecter les nœuds ajoutés (on peut aussi qualifier ces nœuds de bruit) au réseau LFR d'origine, la méthode utilisée dans [78] consiste à attribuer un degré à chacun de ces nœuds, tiré d'une distribution identique à celle du réseau LFR. Cette distribution suit donc une loi de puissance avec un exposant d'une valeur égale à γ. Les nœuds ajoutés sont finalement reliés à ceux du LFR initial par attachement préférentiel. Les figures 4.7a et 4.7c montrent les valeurs ordonnées du logarithme de δ, obtenu à partir d'un modèle LFR de 1000 nœuds ($\gamma = 3$, $\beta = 2$, $\mu = 0,1$ et $\mu = 0,5$) auquel 1000 nœuds ont été ajoutés. Nous pouvons observer que la valeur moyenne de δ limitée aux nœuds de la partie non dense (sous la ligne pointillée noire) est beaucoup plus faible que celle limitée à la partie dense (au-dessus de la ligne pointillée rouge). Ceci rend la classification assez facile comme l'illustre la figure 4.7a qui montre qu'il existe un écart important entre la valeur minimale de δ dans le réseau LFR et la valeur maximale de δ dans l'ensemble des nœuds ajoutés. Le rapport entre ces deux valeurs est presque égal à 7.

Afin de rendre la classification plus difficile, nous modifions les liens incidents aux nœuds ajoutés comme suit :

- Nous commençons par calculer l'amplitude de l'écart $g_0 = \min(\delta_{LFR}) - \max(\delta_{noise})$ entre les valeurs de δ dans le réseau LFR initial et les nœuds ajoutés,
- nous comblons progressivement le gap g_0, en ajoutant des liens entre des paires sélectionnées aléatoirement à partir de l'ensemble des nœuds ajoutés. Ceci a pour effet de réduire la valeur de $\min(\delta_{LFR}) - \max(\delta_{noise})$, de sorte que nous pouvons tester notre algorithme à différents niveaux, entre le cas initial, avec un gap égal à g_0, et un cas final avec un gap nul.

En pratique, nous fixons un paramètre r entre 0 et 1 et ajoutons successivement des liens entre les paires de nœuds ajoutés (initialement mal connectés entre eux), jusqu'à ce que la valeur mesurée de $\min(\delta_{LFR}) - \max(\delta_{noise})$ atteigne $g(r) = g_0 \cdot (1 - r)$.
Lorsque $r = 0$ aucun lien n'est ajouté, et lorsque $r = 1$ l'écart initial g_0 est entièrement comblé. Les figures 4.7b et 4.7d donnent deux exemples de réduction totale de l'écart (c'est-à-dire $r = 1$) pour deux valeurs différentes du paramètre μ du modèle LFR.

Dans la figure 4.7, on peut constater que la partie de la courbe qui se trouve au-dessus de la ligne pointillée rouge est la même avant et après le remplissage du gap, alors que ce n'est pas le cas pour les nœuds du bruit (partie inférieure de la courbe). Ceci signifie que les liens ajoutés modifient la forme de la distribution δ des nœuds dans la partie non dense, à défaut de pouvoir seulement la décaler vers le haut. Cet effet est dû au fait que les liens sont ajoutés entre des paires de nœuds choisies au hasard, et non en ciblant particulièrement les nœuds qui ont des valeurs δ élevées, ce qui produit un ensemble de données pour lequel la séparation entre la partie dense et la partie non dense est moins facile à déterminer.

On peut également souligner que plus la valeur de g_0 est élevée, plus la distribution est sévèrement affectée lorsque le r augmente. On peut cependant constater des niveaux de perturbations plus faibles en fonction du paramètre μ utilisé pour générer le modèle LFR initial. Ceci peut se traduire par une distribution avec une faible valeur de g_0, et ce gap peut être réduit en ajoutant un nombre relativement faible de liens au réseau. Cela a un impact direct sur les performances de notre algorithme, comme nous le verrons sur la figure 4.9.

Protocole expérimental

Trois expériences sont rapportées ici, chacune avec une valeur différente du paramètre de mixage μ du modèle LFR. Pour chaque expérience, nous effectuons des séries de $N = 11$ calculs, chacune pour une valeur différente de r, allant de $r = 0$ jusqu'à $r = 1$ par pas de 0.1. Il est important de noter que pour une valeur de r supérieure à 1, la vérité de terrain devient contestable. En effet, cela voudrait dire qu'il existerait des nœuds dans la partie non dense avec une valeur de δ supérieure à celles de certains nœuds de la partie dense. Nous avons donc fixé une limite à $r = 1$.

Pour chacune des valeurs de r, les réseaux LFR initiaux ont un nombre de nœuds $N_{LFR} = 1000$, des paramètres $\gamma = 3$ et $\beta = 2$. Nous exécutons notre algorithme en rajoutant du bruit sous forme de nœuds de faibles valeurs de δ, au nombre de N_{noise} allant de $N_{noise} = 0$ à $N_{noise} = 1000$, par incréments de 10 nœuds. L'algorithme est donc exécuté 100 fois pour chacune des 11 valeurs de r. Chaque calcul est par ailleurs effectué avec une moyenne de 50 modèles nuls différents, tous calculés à l'aide d'un modèle de configuration.

Soit D l'ensemble des nœuds du modèle LFR initial, ND l'ensemble des nœuds rajoutés. Appelons D_r l'ensemble des nœuds détectés comme étant dans la partie dense par ItRich, et ND_r l'ensemble des nœuds classés dans la partie non dense. Nous définissons les ensembles de nœuds suivants :

- $VP = D \cap D_r$
- $VN = ND \cap ND_r$
- $FP = D \cap ND_r$
- $FN = ND \cap D_r$

Nous utilisons ensuite des mesures standard pour évaluer la performance des résultats :

- Le rappel[12] : $Sn = \frac{|VP|}{|VP|+|FN|}$
- La spécificité : $Sp = \frac{|VN|}{|VN|+|FP|}$

D'une part, la spécificité mesure l'indice de Jaccard entre l'ensemble de la partie non dense retournée par l'algorithme et l'ensemble établi par la vérité de terrain, et d'autre part, le rappel fait la même chose avec les nœuds de la partie dense.

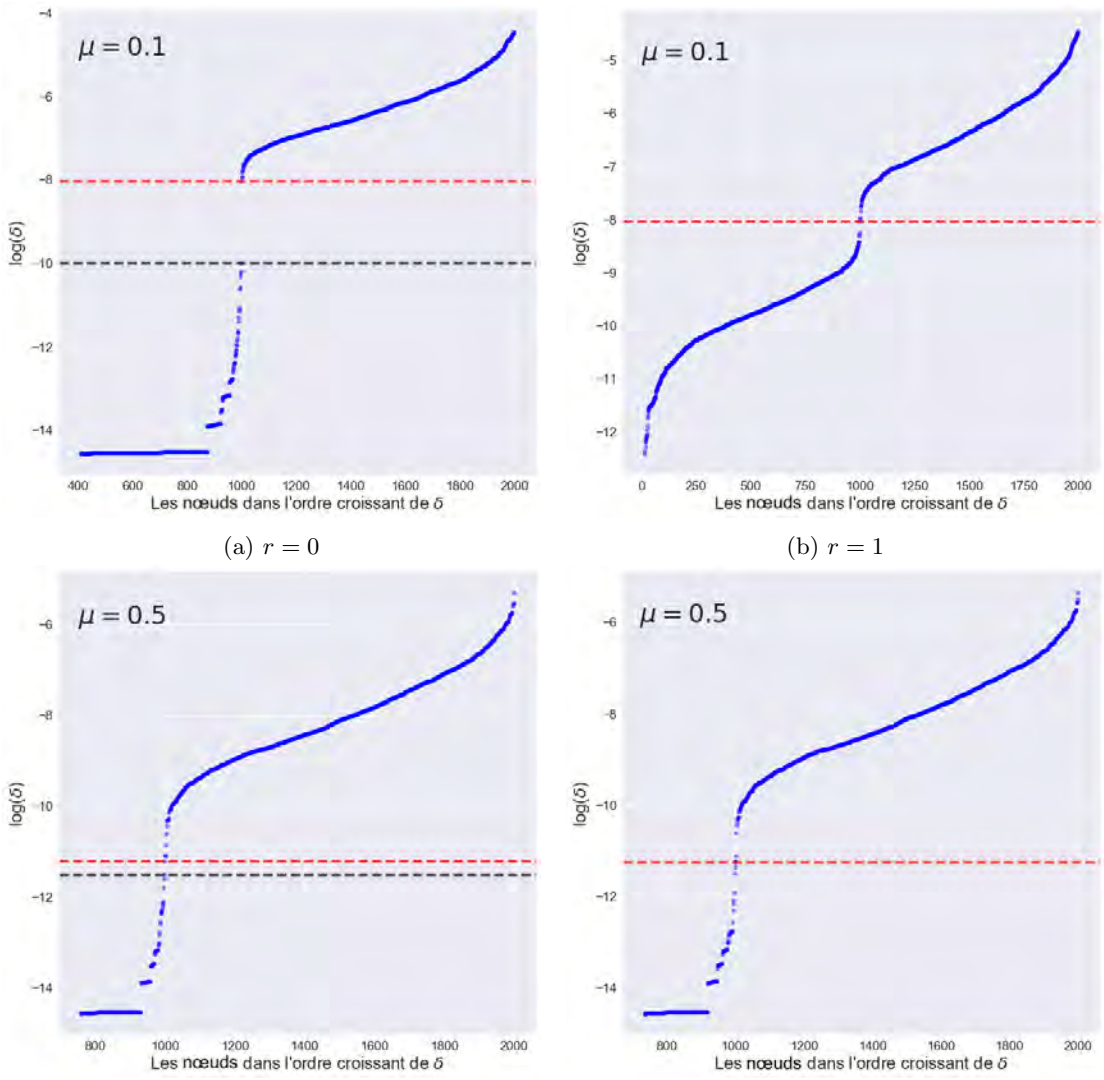
Les valeurs moyennes de Sp et Sn sur les 100 réseaux générés pour chacune des différentes valeurs de r sont calculées lors des trois expériences. Les résultats sont ensuite comparés à ceux de l'algorithme OSLOM [78]. Les nœuds "sans domicile" auxquels OSLOM n'attribue délibérément aucune communauté, sont considérés comme du bruit et comparés aux nœuds de la partie non dense obtenus par ItRich.

Résultats sur deux cas limite : $r = 0$ et $r = 1$

Nous testons d'abord notre algorithme sur des données dont les modèles LFR initiaux sont paramétrés de la façon suivante : $\mu = 0.1$, $\beta = 2$ et $\gamma = 3$. Nous traçons ensuite les résultats pour les deux valeurs limites de $r = 0$ et $r = 1$. Nous montrons ainsi en fonction du nombre de nœuds rajoutés, les évolutions du rappel et de la spécificité. La comparaison entre les résultats d'ItRich et ceux d'OSLOM est montrée sur la figure 4.8.

La figure 4.8a montre que lorsque $r = 0$, ItRich retourne une spécificité égale à 1 et qui reste constante à mesure que l'on augmente le nombre de nœuds ajoutés. Cela signifie que tous ces nœuds ont été correctement classés dans la partie non dense, quel que soit leur nombre qui va de 0 à 1000. À l'exception de quelques valeurs aberrantes, c'est également le cas lorsque $r = 1$ et que le nombre de

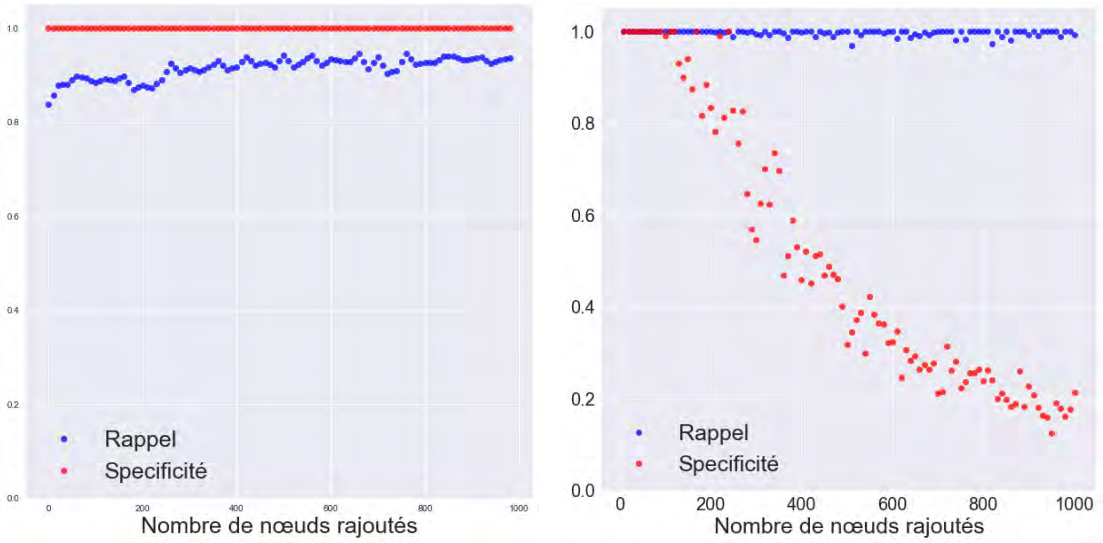
[12] Parfois aussi appelé sensibilité



(a) $r = 0$ (b) $r = 1$

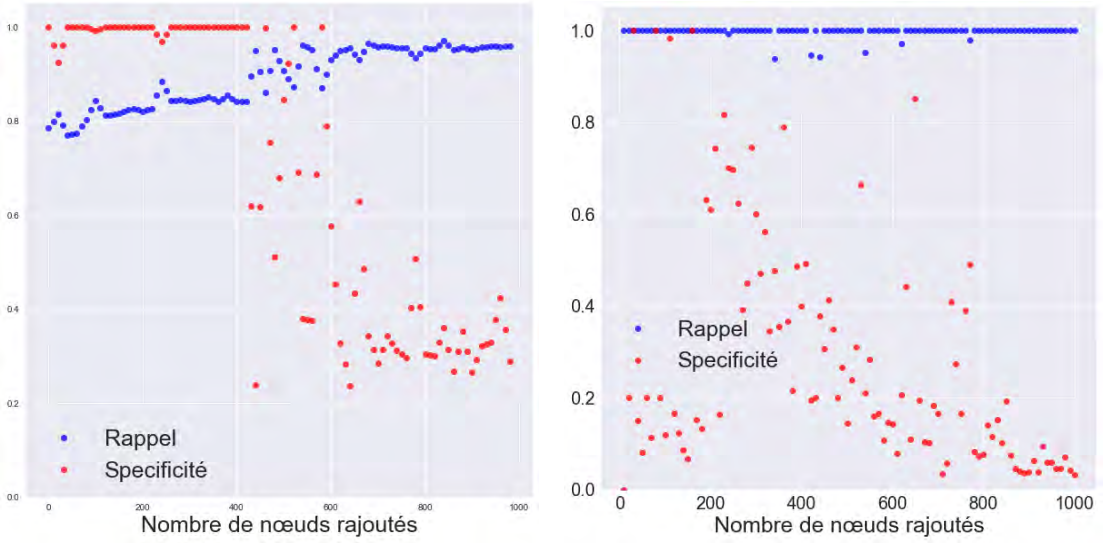
(c) $r = 0$ (d) $r = 1$

Figure 4.7: Le logarithme de δ de chaque nœud en fonction de son rang. Les lignes en pointillé noire représentent la valeur maximale de δ au sein de l'ensemble des nœuds ajoutés (le bruit), tandis que les lignes en pointillé rouges représentent la valeur minimale de δ dans le modèle LFR initial (qu'on construit à partir des paramètres $\gamma = 3, \beta = 2$). Sur (a) et (c), il existe un écart entre les valeurs de δ séparant les nœuds du modèle LFR initial l'ensemble des nœuds ajoutés, lorsque $r = 0$. Cet écart est plus important lorsque le paramètre de mixage μ est petit ($\mu = 0.1$ dans (a) et $\mu = 0.5$ dans (c)). Ces écarts sont comblés en (b) et (d), qui sont les graphes correspondant à une valeur de $r = 1$. On se retrouve dans les deux cas avec des réseaux dans lesquels la valeur minimale de δ dans le LFR est égale à la valeur maximale de δ dans la partie non dense.



(a) ItRich, $r = 0$

(b) OSLOM, $r = 0$



(c) ItRich, $r = 1$

(d) OSLOM, $r = 1$

Figure 4.8: Valeurs des mesures de rappel (points bleus) et de spécificité (points rouges) obtenues par ItRich et OSLOM, en fonction du nombre de nœuds rajoutés. Les paramètres du LFR initial sont fixés à $\mu = 0,1$, $\beta = 2, \gamma = 3$, et nous montrons les résultats pour les deux valeurs limites $r = 0$ (a),(b) et $r = 1$ (c),(d).

nœuds ajoutés est inférieur à la moitié du nombre de nœuds dans le LFR (*cf.* fig. 4.8c). À l'inverse, avec l'algorithme OSLOM, la valeur de la spécificité diminue de manière significative et régulière après l'ajout d'environ 200 de nœuds, suivant une valeur de $r = 0$ (*cf.* fig. 4.8b) ou bien $r = 1$ (*cf.* fig. 4.8d). Cela suggère que plus le nombre de nœuds ajoutés est important, plus certains d'entre eux sont affectés à une communauté par OSLOM, ce qui se traduit par un ratio plus faible de nœuds additionnels correctement affectés à la partie non dense.

Quant au rappel Sn, ses valeurs calculées par ItRich sont élevées mais inférieures à 100% du début à la fin du calcul, même lorsqu'il n'y a pas de nœuds ajoutés du tout et peu importe si $r = 0$ ou $r = 1$. Cela est dû au fait que certains nœuds du réseau LFR sont initialement classés comme étant dans la partie non dense, et le restent lorsque le bruit est ajouté. En comparaison, OSLOM maintient également une valeur constante de Sn avec une moyenne de 99%, et un écart-type de 1%. Cela montre que cet algorithme attribue une communauté à presque tous les nœuds du réseau LFR initial, peu importe la valeur de r.

Enfin, on peut noter deux comportements particuliers. Le premier est pour OSLOM lorsque $r = 1$, et après qu'environ 200 nœuds ont été ajoutés. Il y a un effet de seuil qui provoque un saut des valeurs de la spécificité, de valeurs inférieures à 0.2 à des valeurs autour de 0.8. Cela s'explique par le fait que tant que le nombre de nœuds ajoutés est suffisamment petit, OSLOM les regroupe incorrectement en une seule communauté. La seconde est pour ItRich lorsque $r = 1$ et que environ 500 nœuds ont été ajoutés. La spécificité semble prendre des valeurs fluctuantes entre 0 et 1 et le rappel entre 0.85 et 0.95. Ces fluctuations sont d'une amplitude assez importante, ce qui suggère qu'après l'ajout d'un nombre de nœuds équivalent à la moitié de la taille du réseau LFR initial, ItRich produit des résultats instables, mais qui le sont de moins en moins à mesure que l'on rajoute des nœuds. Ceci est accompagné d'une baisse des valeurs de la spécificité, qui signifie que plus de la moitié des nœuds rajoutés sont alors classés dans la partie dense du réseau. Nous expliquons cela par le fait que lorsque nous comblons un large gap, comme celui montré sur la figure 4.7a, nous nous retrouvons avec une partie non dense dont les nœuds ont une valeur de δ qui est faible lors de la première itération d'ItRich (nous avons construit les données pour que ceci soit toujours le cas). Lors des itérations suivantes, on retrouve dans le réseau certains nœuds qui sont propulsés plus haut dans la liste triée dans l'ordre décroissant de δ (après la suppression d'un certain nombre de rich clubs). Ceci conduit à l'ajout dans la partie non dense des nœuds ayant les valeurs les plus élevées de δ parmi les nœuds rajoutés. Nous verrons plus bas que plus le gap à combler est petit, moins ce dernier phénomène est observé.

Résultats des tests effectués avec un gap variable :

Nous montrons maintenant les performances moyennes calculées sur des modèles issus de trois différentes valeurs du paramètre de mixage μ, et pour des valeurs de r allant de 0 à 1 avec un pas de 0.1. Nous cherchons ainsi à observer l'évolution des performances des deux algorithmes que l'on compare, à mesure que l'on remplit l'écart initial g_0. L'expérience est réalisée pour des valeurs fixes de γ et β ($\beta = 2, \gamma = 3$) et trois valeurs du paramètre de mixage μ (0.1, 0.2 et 0.5). Les résultats de ItRich sont comparés à ceux d'OSLOM.

Nous remarquons sur les figures 4.9d,4.9e et 4.9f que pour les trois différentes valeurs de μ, le score de rappel de l'algorithme ItRich est inférieur mais proche de 1 (avec un minimum de 0.79 pour la valeur la plus basse lorsque $\mu = 0.5$). Ce résultat généralise les cas particuliers observés dans la figure 4.9 : l'algorithme ItRich classe dès le départ quelques nœuds du modèle LFR initial dans la partie non dense, et cela reste vrai pour différentes valeurs de μ à mesure que r augmente. D'autre part, OSLOM a un score de rappel très proche de 1 pour les valeurs $\mu = 0.1$ et $\mu = 0.2$, indépendamment de la valeur de r, ce qui signifie qu'il attribue une communauté à presque tous les nœuds du réseau LFR initial. Le rappel moyen diminue cependant pour atteindre des valeurs oscillant autour de 0.75 lorsque $\mu = 0.5$.
Quant aux valeurs de la spécificité, on observe sur les résultats des deux premières expériences, correspondant aux paramètres $\mu = 0.1$ et $\mu = 0.2$ (figures 4.9a et 4.9d), que les valeurs moyennes ont tendance à diminuer pour OSLOM et pour ItRich, avec de meilleures performances pour ItRich.

Sachant que la spécificité mesure le rapport entre le nombre de vrais négatifs (ou le nombre de nœuds ayant correctement été classés dans la partie non dense) et le nombre de nœuds ajoutés, on peut en déduire que plus r est élevé, plus OSLOM tend à attribuer une communauté aux nœuds ajoutés. Les résultats d'ItRich sur ces mêmes expériences montrent que l'indice est initialement décroissant, ce qui peut s'expliquer par les mêmes arguments que dans le cas d'OSLOM. On résume ces arguments par le fait que plus le paramètre r est élevé, plus le nombre de nœuds parmi ceux qui ont été ajoutés, qui se

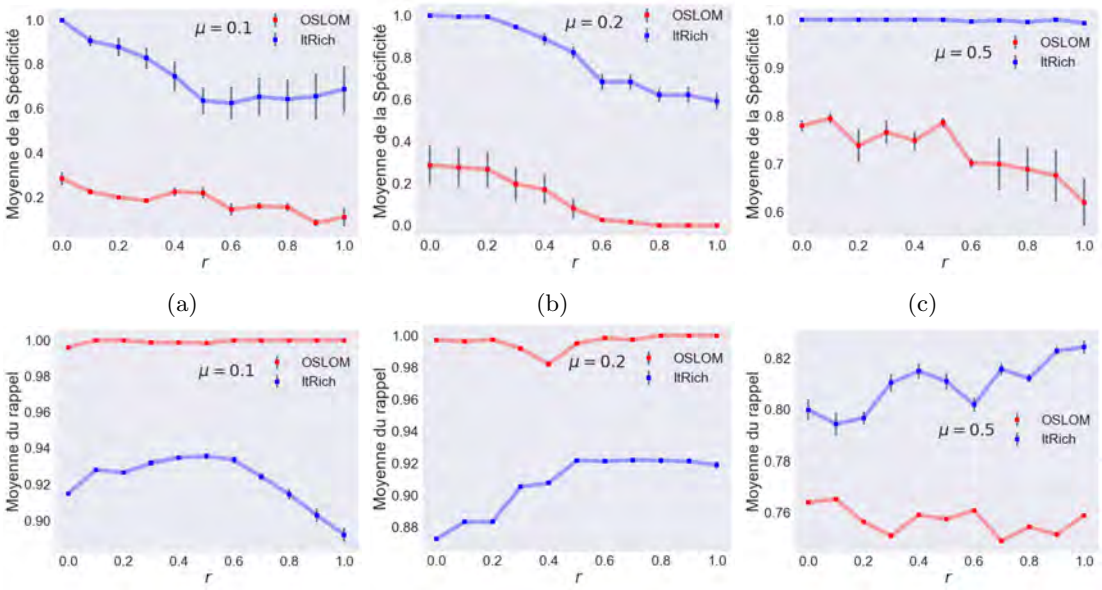


Figure 4.9: Évaluation des performances d'OSLOM et d'ItRich par les indices de rappel et de spécificité.

retrouvent classés dans la partie dense, croit lui aussi. On observe ensuite une quasi-stabilisation de la spécificité moyenne, pour les grandes valeurs de r (quand $r > 0.4$ pour $\mu = 0.1$ et pour $r > 0.5$ quand $\mu = 0.2$). Ce dernier effet est moins intuitif, mais peut s'expliquer par le fait que plus r est élevé, plus le critère d'arrêt devient efficace. Ce dernier a été choisi comme suit : si la qualité du $\delta-$*rich club* qui vient d'être calculé est inférieure à celle d'un modèle d'Erdös-Renyi ayant le même δ moyen, alors on arrête le processus itératif d'extraction.
Il se trouve qu'en augmentant la valeur de r, un nombre plus grand de liens aléatoires est créé parmi les paires de nœuds ajoutés. Ceci augmente progressivement la ressemblance entre le sous-réseau induit par l'ensemble des nœuds ajoutés, et un réseau d'Erdös-Rényi connecté (ce qui n'était pas le cas pour les faibles valeurs de r), ce qui rend plus efficace le critère d'arrêt. Cet effet est donc plus un artefact compte tenu du choix du critère d'arrêt Q_{seuil}, et des données sur lesquelles l'algorithme est testé, qu'une propriété intrinsèque de ItRich.

Pour la troisième expérience, correspondant au cas où $\mu = 0.5$ on observe une valeur moyenne de la spécificité très proche de 1 et un rappel oscillant autour de sa valeur moyenne qui vaut 0.81 pour ItRich. Les résultats d'OSLOM, en revanche, montrent un rappel moyen autour de 0.75 avec une spécificité qui varie entre 0.62 et 0.81. Les meilleurs résultats de ItRich sont obtenus lors de cette expérience, et cela s'explique par le fait qu'ils correspondent au cas où les perturbations sont les plus faibles (par rapport à $\mu = 0.1$ et $\mu = 0.2$). En effet, en augmentant le paramètre de mixage μ, on diminue la valeur moyenne de δ dans le réseau LFR initial. Ceci diminue la valeur de l'écart g_0 et permet de le combler sans ajouter trop de liens aléatoires (c'est-à-dire que la variance entre les réseaux générés de $r = 0$ à $r = 1$ diminue avec μ), comme nous pouvons le voir sur les figures 4.7c et 4.7d.

En général, nous observons de meilleurs résultats en moyenne pour ItRich que pour OSLOM. Ceci est particulièrement vrai en ce qui concerne la mesure de la spécificité. Celle-ci a une valeur moyenne minimale de 0.63 pour ItRich lorsque $\mu = 0.1$, ce qui signifie que dans le pire des cas, notre algorithme parvient à identifier en moyenne au moins 63% des nœuds "bruit". D'autre part, le rappel est meilleur pour OSLOM, mais cela s'explique facilement par le fait que dans ce cas, plus la spécificité est faible, plus le rappel est important.

L'algorithme ItRich est plus performant qu'OSLOM dans le cas des réseaux LFR synthétiques modifiés que nous avons introduits précédemment. Il renvoie un indice de rappel qui est parfois inférieur (*cf.* fig. 4.9b et fig. 4.9e) à celui d'OSLOM. Cependant, le modèle LFR de base peut être lui-même sujet à certaines variations, et contenir des nœuds de faible δ qui seront classés dans la partie non dense, en plus des nœuds "bruit" qui ont été ajoutés. Ceci est confirmé par le fait que les valeurs de l'indice de rappel renvoyées par ItRich restent clairement inférieures à 1 pour toutes les valeurs r même pour $r = 0$

(*cf.* fig. 4.8a). Ces résultats suggèrent qu'au tout début de chaque expérience, une partie des nœuds de faible densité dans le modèle LFR initial est classée dans la partie non dense.

Nous notons également que les écarts-types des résultats de l'algorithme ItRich sont de plus grande amplitude pour les petites valeurs de μ. Ceci est une conséquence de l'effet mentionné précédemment : plus le paramètre de mixage est grand, plus g_0 est petit, et moins nous avons à perturber le LFR initial pour combler cet écart, ce qui conduit à une plus grande stabilité des résultats de l'algorithme ItRich.

4.6 Bilan

Ce chapitre propose un nouveau point de vue sur l'analyse de la structure des réseaux et décrit une approche alternative et supplémentaire aux méthodes standard, telles que la décomposition en $k-$cores, ou les méthodes de détection de communautés.

Dans une première partie, nous avons développé une approche naïve qui s'inspire des algorithmes de détection de communautés. Celle-ci consiste à fixer des critères généraux sur la qualité d'une partie non dense, pour ensuite mettre au point un algorithme qui optimise le seuil de δ suivant lequel un nœud est considéré soit dans la partie non dense, soit en dehors.

Nous avons ensuite présenté une deuxième approche plus consistante, lors de laquelle nous avons utilisé la mesure de densité δ introduite dans le chapitre 3, dans le contexte particulier des "rich clubs" pondérés. Ceci nous a permit de mettre au point un algorithme qui fournit en sortie plusieurs sous-ensembles de nœuds, constituant l'ensemble de ce que nous appelons la partie dense. L'ensemble des sommets qui ne se trouvent dans aucun de ces sous-ensembles est appelé la partie non dense. Nous avons aussi proposé un métamodèle nul, qui dans les limites de certaines hypothèses, permet de réduire le temps de calcul d'ItRich. Nous avons par ailleurs comparé ce métamodèle avec les modèles nuls qui existent dans la littérature. Les performances de l'algorithme ItRich ont été testées sur un modèle de réseaux synthétiques de référence, autour duquel nous avons construit notre propre vérité de terrain. Pour ce faire, nous nous sommes inspirés d'une méthode présentée dans [78] pour identifier les nœuds de faible densité.
La comparaison entre la méthode OSLOM et l'algorithme ItRich montre que, dans ce cadre, ItRich produit de meilleurs résultats. Les tests effectués sur divers réseaux réels constituent le contenu du prochain et dernier chapitre de cette thèse.

Chapitre 5

Applications

Table des matières

5.1 Introduction

Tous les éléments de l'approche développée lors de cette thèse qui s'articule sur la mesure de la densité dans les graphes, et qui a abouti à un algorithme de détection des parties dense et non dense, ont été désormais présentés et étayés par quelques exemples d'application. On a restreint ces applications à des données synthétiques, ou bien à des données réelles, sans porter jusqu'ici d'attention particulière sur la vérité de terrain. Nous abordons maintenant ce dernier chapitre en nous tournant vers l'étude des réseaux réels ayant une vérité de terrain qui rend possible une interprétation des résultats. Des deux algorithmes présentés dans le dernier chapitre, nous ne considérons ici que l'algorithme ItRich, car il permet un découpage plus riche du réseau. Nous montrerons qu'en plus de scinder le réseau en deux parties (dense et non dense) suivant une mesure de densité, il fournit aussi un découpage particulier au sein de chacun de ces deux sous-ensembles.

Nous ne structurons pas ce chapitre autour d'un jeu de données particulier, car l'outil que nous proposons n'a pas été développé dans ce but. De ce fait, on choisit de montrer les résultats obtenus sur diverses données, qu'on répartit en plusieurs subdivisions afin que chacune illustre une propriété particulière de l'algorithme ItRich.
Nous commençons par deux réseaux de petite taille, la communauté de dauphins de Doubtful Sound étudiée dans [84], et le réseau constitué des équipes universitaires de football américain de Newmann [62]. Nous exploitons leur petites tailles afin d'illustrer certaines propriétés des nœuds de la partie non dense. Ces propriétés seront ensuite mises en avant sur des réseaux de taille plus importante, comme celui des blogs politiques utilisés par Adamic [2] ou bien le réseau mondial des transports aériens, tel qu'il était en 2017 (*cf.* chapitre 3).
Nous nous tournerons ensuite sur l'analyse des différents $\delta-$*rich clubs* de la partie dense. Les données exploitées à cette fin proviennent de la saga littéraire "Le trône de fer" qui a été adaptée en série télévisée ("Game of thrones") [1], où nous disposerons à la fois des données provenant des livres et de la série, qu'on comparera en utilisant une approche qui combine ItRich et un algorithme de détection de communautés.

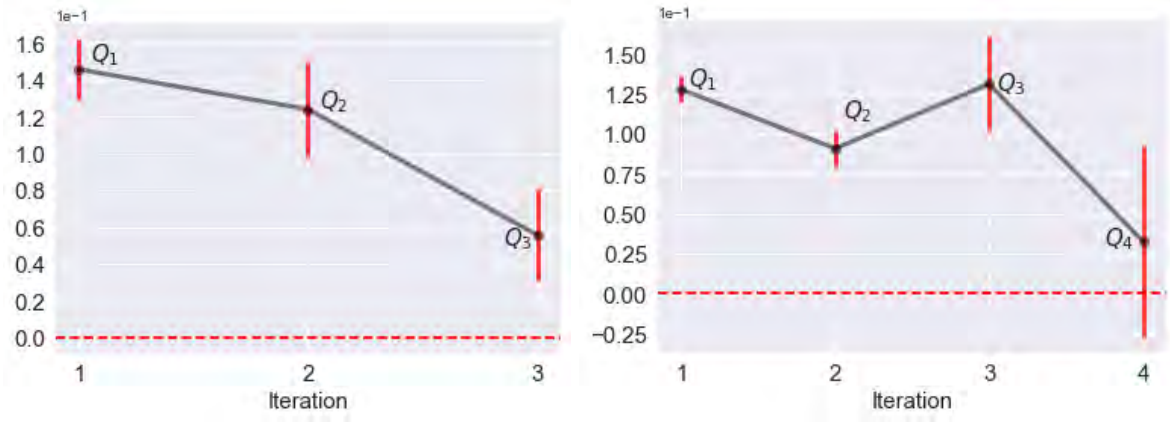
Nous finirons par l'analyse de trois réseaux dynamiques, dont les nœuds sont des élèves soit d'une école primaire soit d'une classe préparatoire. Un lien entre deux nœuds se crée si deux élèves ont été l'un en face de l'autre, et à faible distance de sorte que les émetteurs et les capteurs dont ils sont équipés puissent enregistrer une interaction. Cette interaction doit intervenir au moins une fois durant une fenêtre temporelle large de 20 secondes sans recouvrement. Là encore, il est facile sur ce type de données d'analyser les résultats d'ItRich, car il est toujours aisé de suivre dans le temps l'évolution de la partie dense, ou de la partie non dense, sachant que l'algorithme découpe le graphe en seulement deux ensembles à chaque instant. En comparaison, il est beaucoup plus difficile de suivre l'évolution d'une communauté, en raison de la possible versatilité des nœuds qui la composent. Une communauté identifiée à un instant t_1 peut ne plus être la même communauté à un instant ultérieur t_2, si la grande majorité des nœuds qui la composaient n'en font plus partie. Ce problème ne se pose pas pour ItRich, car même si les composants de la partie dense, ou de la partie non dense changent au cours du temps, ces sous-ensembles seront toujours identifiés, ce qui facilite l'interprétation.
Dans tous les cas étudiés dans ce chapitre, nous prenons la valeur par défaut de $Q_{seuil} = \frac{Q_1}{10}$. Il est parfois intéressant de faire varier Q_{seuil}, car sa valeur sert de filtre pour la partie dense, ainsi en l'augmentant on est de plus en plus sélectif, et *vice versa*. Ceci n'a pas été nécessaire, et nous montrerons que pour les données analysées ici, il est tout à fait raisonnable de garder cette valeur.

5.2 Analyse détaillée sur des réseaux de petites tailles

Nous commençons par montrer les résultats obtenus sur deux jeux de données qui sont largement étudiés dans le domaine : le premier est celui des grands dauphins de Lusseau [84, 83] constitué de 62 dauphins, et dont les données ont été récoltées sur une durée de plusieurs années par des spécialistes. Les liens du réseau sont entre les individus qui montrent davantage d'interaction que ce à quoi l'on s'attendrait si les interactions étaient aléatoires.
Le second réseau analysé ici est constitué de 115 équipes universitaires de football américain, un lien est créé entre chaque deux équipes si elles se sont affrontées au cours de la saison de l'année 2000. La

[1] `https://networkofthrones.wordpress.com/`

figure 5.1 montre l'évolution de la mesure de qualité, ainsi que le seuil représentant le dixième de sa valeur maximale, pour les deux réseaux étudiés dans cette section.



(a) (b)

Figure 5.1: Les valeurs de Q pour (a) les grands dauphins de Lusseau et (b) les équipes de football américain. La ligne pointillée rouge représente la valeur de $Q_{seuil} = Q_1/10$. Pour chaque réseau, le modèle nul est calculé 100 fois et les barres d'erreur représentent les écarts types de la mesure de qualité.

5.2.1 Les dauphins de Lusseau

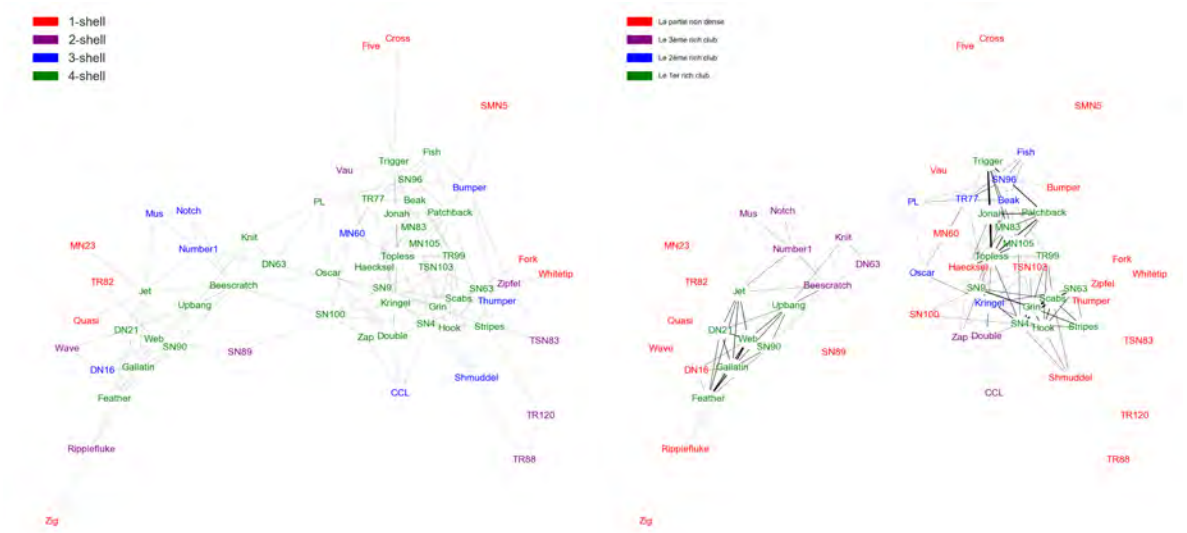
Dans les figures 5.2a et 5.2b le vert, le bleu et le violet correspondent respectivement aux premier, deuxième et troisième $\delta-$*rich clubs* identifiés par notre algorithme. Les nœuds qui ne sont affectés à aucun de ces $\delta-$*rich clubs* sont colorés en rouge et représentent la partie non dense du réseau. Ces résultats, comparés à la structure communautaire proposée dans [83] illustrent la différence entre ItRich et un algorithme de détection de communautés. Certains dauphins évoluent dans des groupes différents au sein du même $\delta-$*rich club* [2].

On observe que le premier $\delta-$*rich club* est composé de nœuds fortement connectés. Le deuxième et le troisième $\delta-$*rich clubs* sont composés de plusieurs sommets reliés, appartenant pour la plupart au voisinage du premier $\delta-$*rich club*. Ces observations impliquent de comparer la distribution des sommets au sein des $\delta-$*rich clubs* avec la distribution des sommets au sein des k-shells. Le réseau contient 4 k-shells, et une comparaison est faite entre la partie non dense produite par ItRich et la périphérie représentée ici par l'union de la 1-shell et de la 2-shell d'une part, puis entre la partie dense et l'union de la 3-shell et de la 4-shell.

On constate que la partie non dense produite par ItRich contient une partie importante des nœuds de la périphérie. Elle contient aussi toujours l'intégralité du 1-shell, car il s'agit d'un ensemble de nœuds dont le sous-réseau ne contient pas de fermeture transitive, ce qui implique que tous ses nœuds sont de poids nuls. De plus, dans cet exemple, la partie non dense contient également la totalité de la 2-shell. Nous observons un effet symétrique entre la partie dense et les k-shells, le premier $\delta-$*rich club* étant contenu dans la 4-shell. Notons que cette observation reste valable pour la comparaison entre la partie dense (l'ensemble contenant tous les $\delta-$*rich clubs* produits par ItRich) et le noyau, représenté ici par l'union du 3-shell et du 4-shell.

Le réseau des dauphins est un exemple éloquent en raison des différences entre les deux décompositions. Parmi les 62 sommets du réseau, 25 se trouvent dans la partie non dense. Ces 25 sommets sont de deux types différents : ceux qui ont une valeur nulle de δ et ceux dont la valeur de δ diminue fortement à chaque itération de ItRich, de sorte qu'elle reste inférieure à la valeur minimale requise pour être dans le $\delta-$*rich club* de l'itération en cours. Ces sommets de second type sont Ripplefluke, MN60, SN100, TSN103, DN16, Shmuddel, Haecksel, Thumper, Bumper. Les six premiers parmi ces derniers voient leurs valeurs de δ chuter à zéro dès lors que le premier $\delta-$*rich club* est retiré. Les sommets TSN103, SN100 et Haecksel sont étonnamment dans la 4-shell. Le sommet TSN103 a 4 voisins, chacun d'eux de

[2] voir [84] pour le dendrogramme complet de la décomposition en communautés du réseau, la décomposition admise comme vérité de terrain sur ce réseau est la meilleure coupe possible de ce dendrogramme



(a) (b)

Figure 5.2: Les grands dauphins de Lusseau (a) La décomposition du réseau en k-shell. (b) La décomposition obtenue par ItRich, la largeur de chaque arête est proportionnelle à son poids ω. Le positionnement est obtenu à l'aide d'un algorithme de type force de ressort.

haut degré mais faiblement liés entre eux. En fait, TSN103 est à l'intersection de deux ensembles de sommets de deux communautés différentes, et a donc une position intermédiaire très particulière. Le sommet SN100 est le sommet ayant la plus grande centralité de betweenness dans le réseau, et en même temps le plus petit coefficient de clustering non nul. Il se trouve également dans une position centrale entre des groupes distincts de sommets et a un degré relativement élevé. Cependant, si on regarde les valeurs prises par les poids topologiques des liens qui lui sont accrochés, on se rend compte sur la figure 5.3 qu'il n'est lié qu'à deux arêtes de poids non nul, dont les deux voisins à l'autre extrémité appartiennent au premier $\delta-$*rich club*. Les 5 autres arêtes (dont l'une le reliant à un nœud du deuxième $\delta-$*rich club*, deux autres à des nœuds du troisième $\delta-$*rich club*, et les deux dernières à des nœuds de la partie non dense, comme on peut le voir sur la figure 5.4) sont toutes de poids topologiques nuls.
Enfin, Haecksel a une valeur de δ qui diminue progressivement après avoir retiré le premier et le deuxième $\delta-$*rich clubs*. Il est largement lié aux sommets du premier $\delta-$*rich club*, qui, lorsqu'il est supprimé après la première itération de ItRich, conduit à une configuration dans laquelle Haecksel a un coefficient de clustering nul. Ainsi, à l'intérieur de la 4-shell, la position de Haecksel est en quelque sorte spéciale. La structure du réseau correspondant est donnée sur la figure 5.2, et les poids topologiques des différentes arêtes reliant les nœuds que l'on vient de citer à leurs voisins sont représentés sur la figure 5.3

L'attention portée précédemment à la partie non dense du réseau des dauphins suggère que les sommets de celle-ci, dans un réseau quelconque, peuvent être divisés en deux catégories : ceux qui ont une faible valeur δ, et qui sont dès le début à la périphérie du réseau (faible k-shells) et les autres. Ces derniers n'ont pas une valeur de δ des plus élevées, mais sont liés à des sommets de δ élevés. Ces sommets semblent occuper une position assez spécifique dans l'organisation des liens du noyau du réseau. Nous portons une attention particulière sur ce type de nœuds dans les exemples de la section 5.3 de ce chapitre.

5.2.2 Les équipes universitaires de football américain

Nous examinons ici un réseau représentant les confrontations entre différentes équipes universitaires de football américain au cours de la saison 2000 [62]. Les nœuds représentent les équipes participantes, et un lien relie deux équipes qui ont joué l'une contre l'autre pendant la saison. Chaque équipe est étiquetée par son appartenance à une conférence qui contient de 8 à 12 équipes. Pour la plupart des conférences, les matchs en interne sont plus fréquents que les matchs en externe, ce qui donne au réseau une structure modulaire. Nous avons toutefois identifié deux propriétés qui rendent cet échantillon de

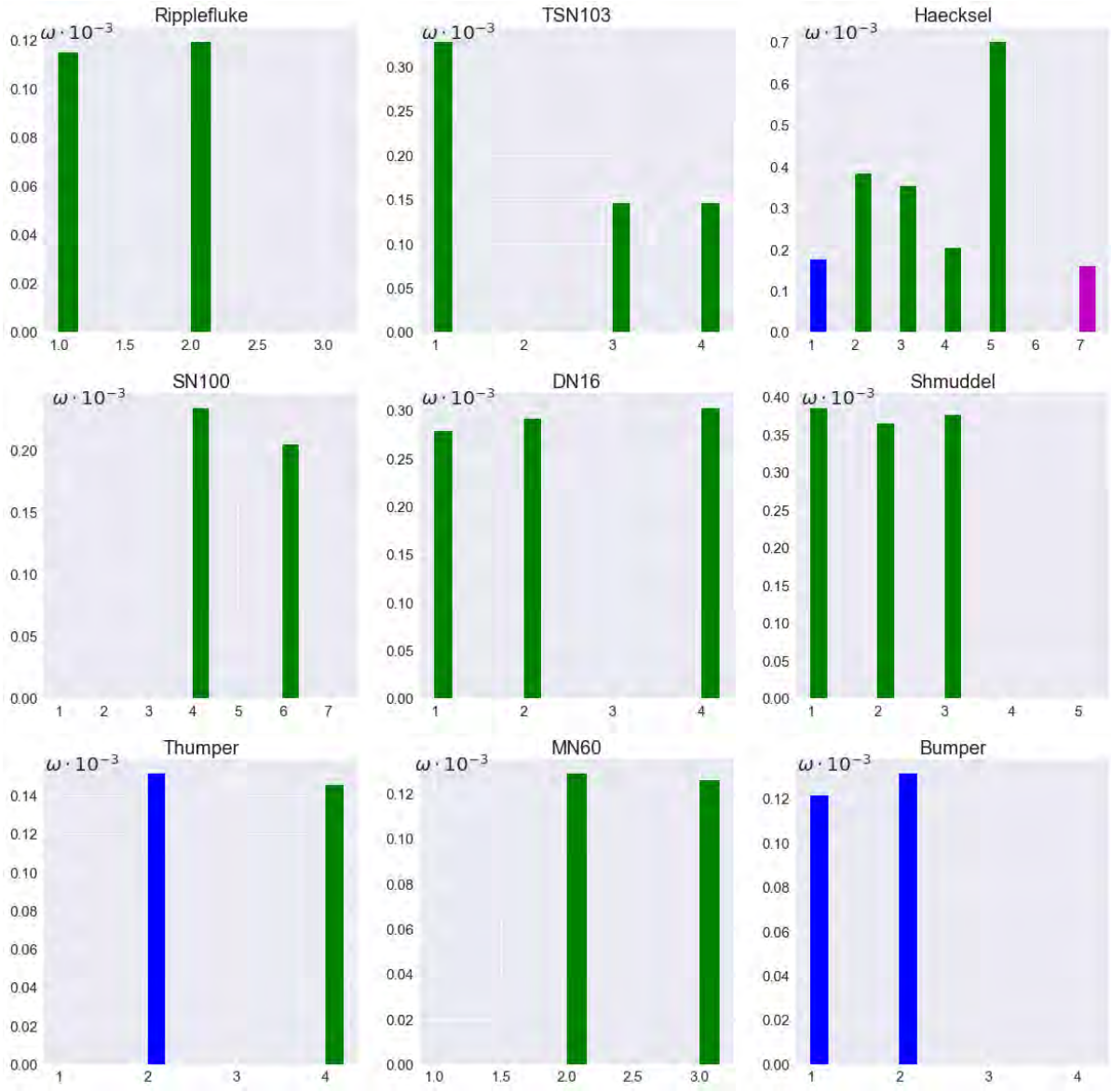


Figure 5.3: Les poids topologiques des arêtes reliées aux 9 nœuds de la partie non dense dont la valeur de δ est non nulle. La couleur de chaque barre correspond à la couleur du $\delta-$*rich club* auquel appartiennent les voisins de l'autre extrémité de l'arête, suivant le code couleur défini sur la figure 5.2

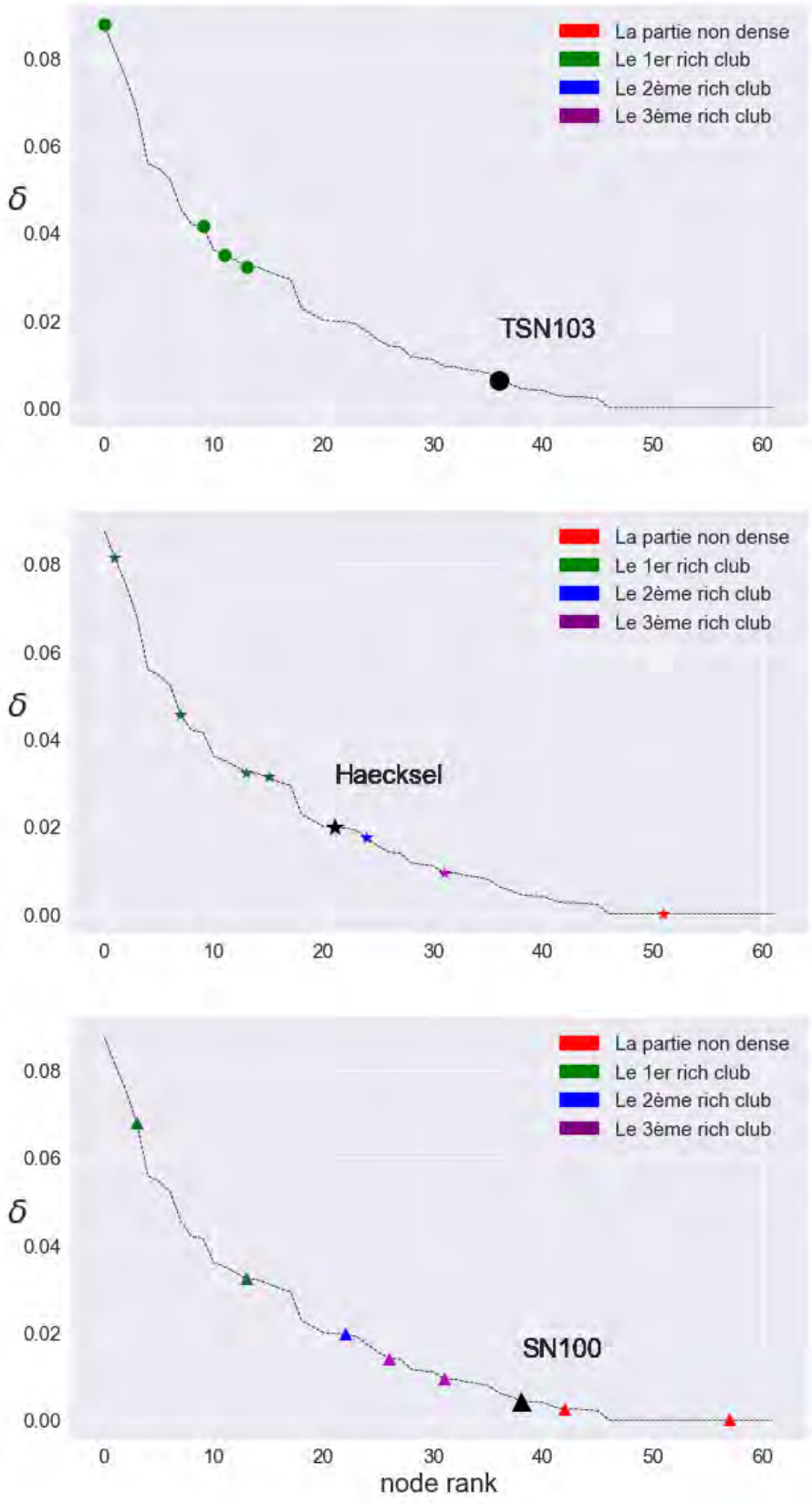


Figure 5.4: La position des 3 dauphins TSN103, Haecksel et SN100 dans la courbe décroissante des valeurs de δ. Les nœuds en question sont représentés en noir, et la couleur de leurs voisins respectifs correspond à celle du $\delta-$*rich club* auquel ils appartiennent.

données particulièrement adapté pour tester notre algorithme. La première est que tous les nœuds font partie du 8-shell, sauf un seul, qui est dans le 7-shell. Cela sous-entend que les résultats ne peuvent pas être trouvés par une décomposition en k-core. La deuxième propriété est qu'il y a 5 équipes qui ne font partie d'aucune conférence, qui ont le label "indépendant". Dans [62] il est également indiqué que les sept équipes de la conférence Sunbelt ont joué presque autant de matchs contre des équipes de la conférence Western Athletic que contre des équipes de leur propre conférence. Ils ont également joué une grande partie de leurs matchs inter-conférences contre des équipes de la conférence Mid-American. La figure 5.5 montre le $\delta-$*rich club* auquel appartient chaque équipe, ainsi que le nom de la conférence dans laquelle elle évolue.

ItRich révèle quatre $\delta-$*rich clubs*, les deux premiers étant composés respectivement de 58 et 42 nœuds, et les deux derniers étant plus petits avec respectivement 6 et 4 nœuds. Il reste cinq nœuds dans la partie non dense.

On peut noter que sur les 4 $\delta-$*rich clubs*, les deux premiers contiennent principalement des équipes qui jouent la majorité de leurs matchs contre des équipes de leur propre conférence, tandis que les deux derniers contiennent des équipes qui ont tendance à diversifier les conférences de leurs adversaires (par exemple la conférence SunBelt). Toutes les équipes du premier $\delta-$*rich club*, sauf une, sont celles qui ont été correctement classées dans [62], dans le sens où la composition de la communauté à laquelle elle appartient est exactement la composition de la conférence à laquelle elle appartient (l'equipe Texas Christian est dans le premier $\delta-$*rich club* mais elle est incorrectement classée dans [62]). Les deux derniers $\delta-$*rich clubs* contiennent des équipes qui jouent un grand nombre de matchs inter-conférences, y compris des équipes de la conférence Sun Belt, et quelques équipes de la conférence Western Athletic, qui sont, selon [62], des équipes dont la conférence ne forme pas vraiment une communauté, en ce sens où il y a peu de confrontations intra-conférence. On remarque également que les troisième et quatrième $\delta-$*rich clubs* sont les mêmes que certaines des communautés trouvées par l'algorithme de Girvan et Newman. Cela s'explique par le fait qu'ils induisent tous deux de petites cliques dans le réseau (avec respectivement 6 et 4 nœuds).

Pour quantifier les informations fournies par les relations entre les équipes et les conférences, nous utilisons une mesure empirique basée sur l'entropie de Shannon [110]. Soit k_i le degré du nœud $i \in V$, et $C = \{c_l\}_{l=1,\dots,12}$ l'ensemble des 11 conférences dans lesquelles les équipes jouent, plus l'ensemble des équipes indépendantes. Pour chaque nœud i nous appelons $p_l^{(i)} = \frac{|N(i) \cap c_l|}{k_i}$ le rapport entre le nombre de voisins du nœud i qui jouent dans la conférence c_l et le nombre total de voisins de i. On a alors

$$H(i) = -\sum_l p_l^{(i)} \cdot log(p_l^{(i)}) \tag{5.1}$$

Cette mesure est égale à zéro si tous les voisins de i jouent dans la même conférence, et a une valeur maximale de $log(12)$ qui est attribuée à un nœud uniquement lorsque tous ses voisins sont uniformément répartis sur toutes les conférences. La figure 5.6 montre pour chaque nœud la valeur de H en fonction de celle de δ.

Les deux premiers $\delta-$*rich clubs* sont caractérisés par des équipes qui ont de faibles valeurs de H et des valeurs élevées de δ. Cela reflète le fait que ces équipes jouent principalement des matchs intra-conférence. Au contraire, les équipes des deux derniers $\delta-$*rich clubs* et celles de la partie non dense jouent principalement des matchs contre des équipes de conférences variées.

La partie non dense est composée de 5 nœuds, dont 4 (Navy, Central Florida, Notre Dame et Connecticut) sont des équipes indépendantes, qui n'ont été classées dans aucune communauté dans [62]. La partie non dense contient ainsi toutes les équipes indépendantes, à l'exception de l'équipe de l'Utah State qui est classée dans le 3ème $\delta-$*rich club*, car elle appartient à une clique.

Le seul nœud de la partie non dense qui ne partage pas ces propriétés est l'équipe de Miami Florida (de la conférence Big East), dont le δ est élevé et le H est faible. Ce nœud se trouve dans la situation évoquée précédemment, à savoir que, malgré son poids élevé, il n'atteint pas celui de ses voisins, qui sont presque tous dans le premier $\delta-$*rich club*.

5.3 Le rôle topologique des nœuds de l'intervalle de chevauchement

Nous avons décrit dans la section 5.2 la répartition des nœuds d'un réseau parmi ses différents $\delta-$*rich clubs*, et la partie non dense. Nous avons pu voir que cette dernière est composée de deux types distincts

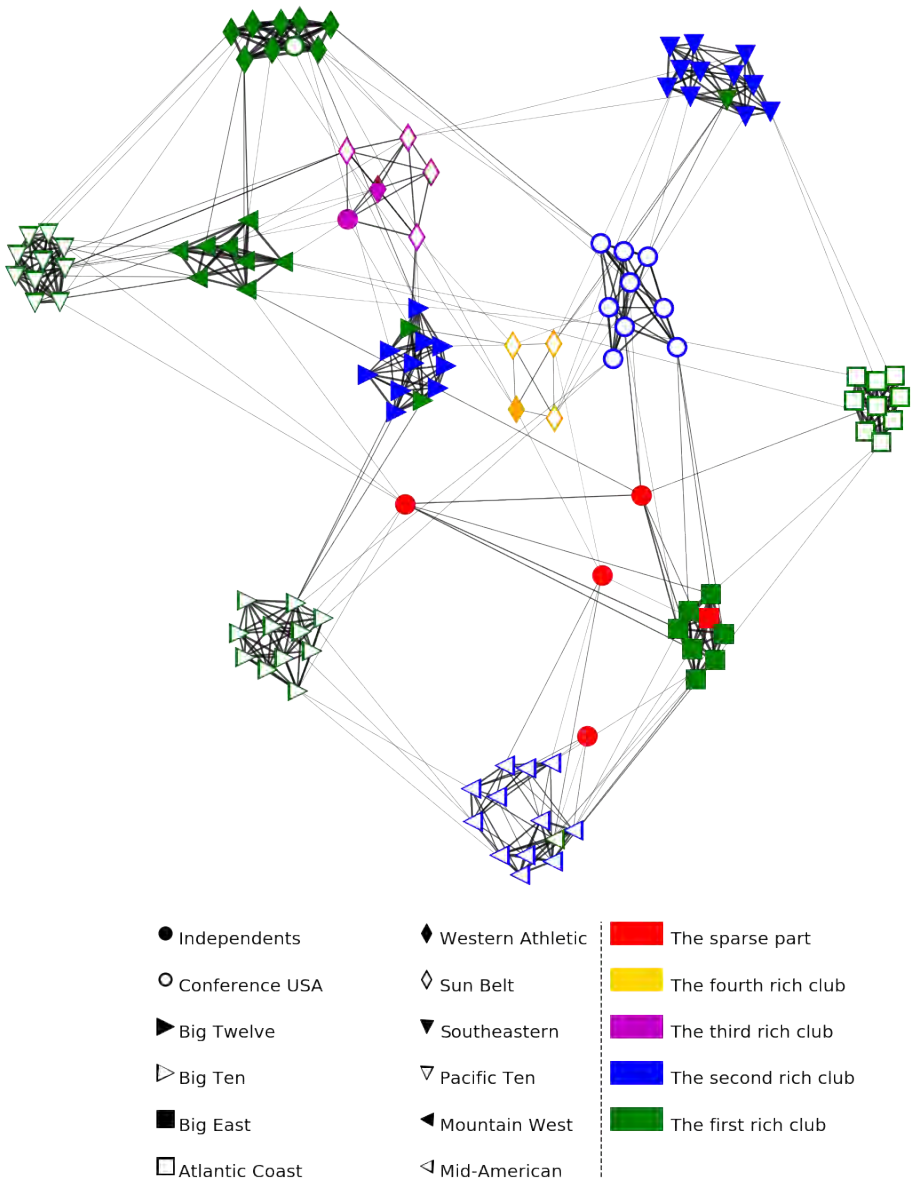


Figure 5.5: Le réseau des équipes universitaires de football américain, saison 2000. Chaque nœud est représenté par un marqueur différent représentant sa conférence, ainsi qu'une couleur différente selon le résultat de la classification d'ItRich. La largeur de chaque arête est proportionnelle à son poids ω. Le positionnement est obtenue à l'aide d'un algorithme de type force de ressort.

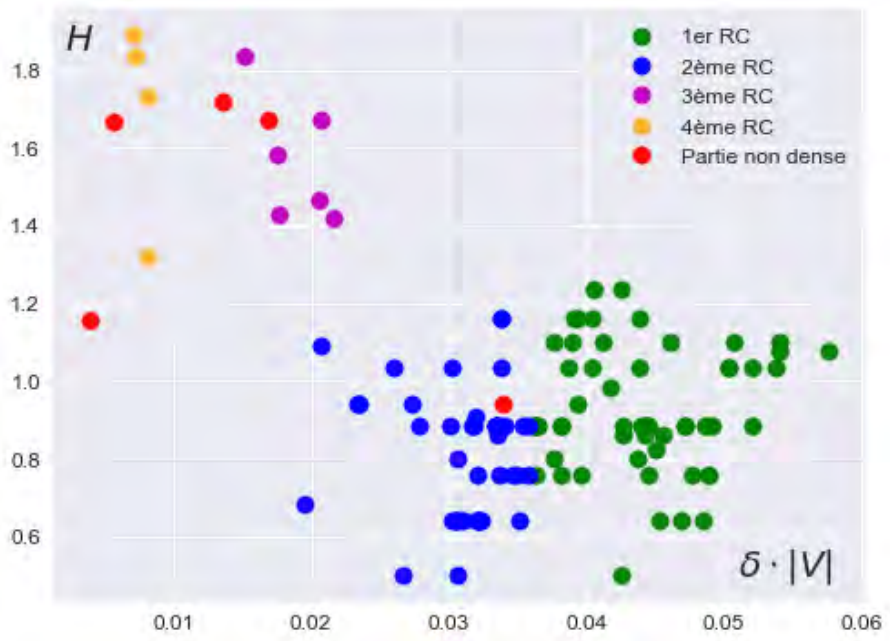


Figure 5.6: La quantité d'informations de chaque nœud par rapport à sa valeur en δ. Les nœuds sont distingués par leur couleur selon $\delta-$*rich club* auquel ils appartiennent.

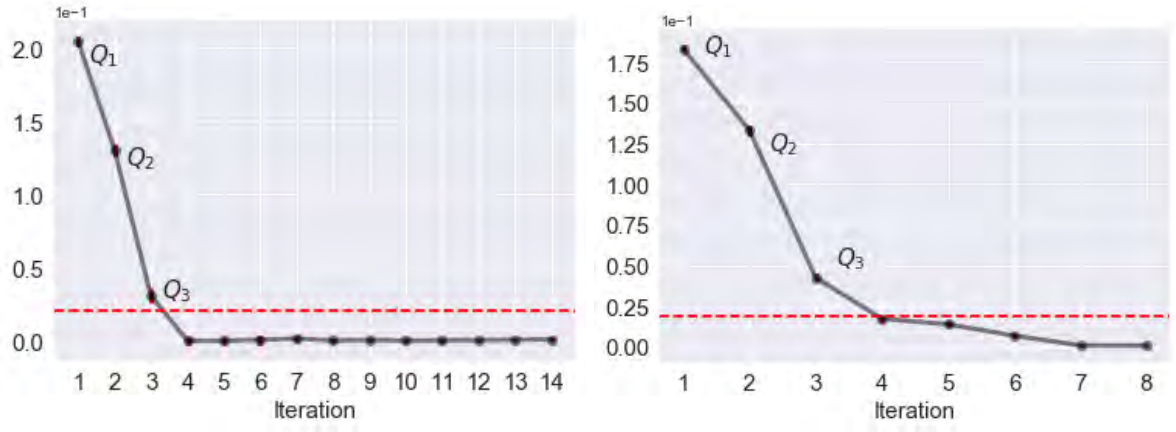
de nœuds. D'un côté nous avons les nœuds qui sont "légitimement" dans la partie non dense, à cause de leur faible valeur de δ, et de l'autre nous avons ceux qui ont un δ dont la valeur peut être relativement élevée, mais pas assez en comparaison avec celle de son voisinage. Une fois ce voisinage supprimé (si celui-ci finit dans l'un des $\delta-$*rich clubs* identifiés par ItRich), ces nœuds demeurent isolés dans le réseau résultant, et par conséquent dans la partie non dense. Nous allons montrer que cet effet se retrouve également dans des réseaux de tailles plus importantes, sur lesquels nous montrerons aussi quelques propriétés statistiques qu'on peut difficilement voir sur les réseaux de petites tailles. Nous analysons d'abord dans cette section un réseau composé de près de 1500 blogs politiques étiquetés et identifiés par les auteurs de [2]. Une arête relie deux blogs si l'un fait référence à l'autre. Le deuxième réseau analysé est celui qui est composé par tous les aéroports et aérogares du monde, référencés en 2017. Les arêtes du réseau relient les paires d'aéroports si un vol quelconque (civile, commercial, militaire, etc.) les relient. Ici aussi, nous montrons sur la figure 5.7 l'évolution de la mesure de qualité Q, ainsi que la valeur seuil choisie (qui dans les deux cas est la valeur par défaut).

5.3.1 Le réseau des blogs politiques américains

Composé d'un ensemble de 1490 nœuds représentant des blogs américains qui traitent de questions politiques, chaque blog dans ce réseau a été étiqueté par Lada Adamic [2] comme libéral (758 sommets) ou conservateur (732 sommets). Deux blogs sont liés si au moins l'un fait référence à l'autre. Les auteurs ont conclu que la plupart des liens se trouvaient au sein des deux communautés séparées, avec peu de liens allant d'une communauté à l'autre. Une propriété intéressante du réseau est que les blogueurs conservateurs créent plus de liens au sein leur communauté que ne le font les libéraux, mais aussi plus de liens vers la communauté opposée.

Comme pour les données analysées dans la section 5.2, nous observons une corrélation entre les k-cores et les valeurs de δ, avec un chevauchement de valeurs pour les parties dense et non dense (*cf.* fig. 5.8a). Cependant, le grand nombre de k-shells (avec k allant de 15 à 34) qui contiennent des nœuds qui sont à la fois dans la partie dense et la partie non dense, ne rendent pas la décomposition en k-cores efficace pour permettre la différenciation des $\delta-$*rich clubs* de la partie non dense du réseau.

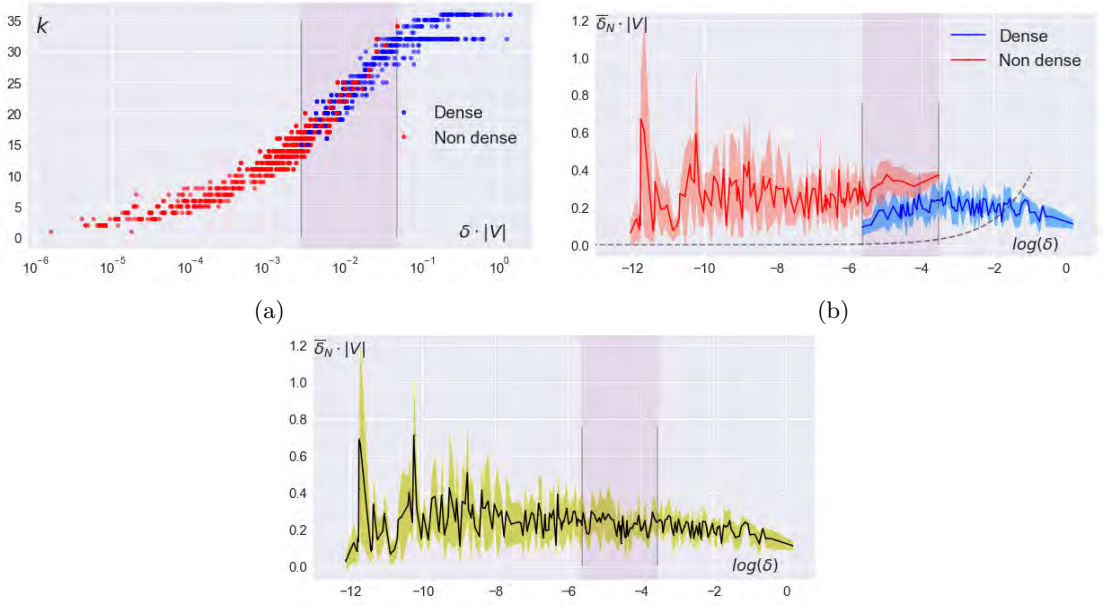
Nous portons maintenant une attention particulière aux nœuds dont le δ se trouve dans l'intervalle de chevauchement, couvert à la fois par les nœuds de la partie dense et ceux de la partie non dense. Les résultats affichés sur la figure 5.8b confirment que les valeurs moyennes de δ qu'on calcule sur le voisinage de chaque nœud et qu'on note $\overline{\delta}_N$, sont particulièrement importantes pour différencier les deux



(a) (b)

Figure 5.7: Les valeurs de Q pour (a) le réseau des blogs politiques (b) le réseau mondial des aéroports. La ligne en pointillées rouges représente la valeur de $Q_{seuil} = Q_1/10$. Pour chaque réseau, le modèle nul est calculé 100 fois et les barres d'erreur représentent les écarts types de la mesure de qualité.

types de nœuds de la zone de chevauchement. Pour la même valeur de δ, les nœuds de la partie non dense ont des voisinages qui ont des valeurs moyennes $\overline{\delta}_N$ supérieures à celles des nœuds de la partie dense. Comme beaucoup de ces voisins appartiennent à un $\delta-$*rich club*, les valeurs de δ initialement prises par ces sommets s'effondrent partiellement ou totalement après avoir supprimé certains $\delta-$*rich clubs* et cela explique pourquoi, *in fine*, ils sont classés dans la partie non dense du réseau.



(a) (b)

(c)

Figure 5.8: Réseau des blogs politiques américains : (a) tracé de k-core *vs.* δ avec les nœuds de la partie dense (resp. non dense) en bleu (resp. rouge). (b,c) La moyenne $\overline{\delta}_N$ de δ calculée sur le voisinage des nœuds par rapport à leur valeur δ sur une échelle logarithmique. Pour un δ donné, seule la moyenne (courbe en traits foncés) et son écart-type sont donnés. Sur (b), on distingue les nœuds de la partie dense de ceux de la partie non dense, ce qui n'est pas le cas sur (c). La zone de chevauchement entre les parties dense et non dense est mise en évidence, et délimitée par des lignes verticales.

Cependant, notons que la seule connaissance de δ et $\overline{\delta}_N$, sans utiliser ItRich, n'est pas suffisante pour fournir des informations permettant de caractériser les $\delta-$*rich clubs* ou la partie non dense, ou

même de trouver leur chevauchement (cf. fig. 5.8c).

Nous utilisons ce découpage pour calculer les proportions de nœuds et de liens dans chacun des $\delta-$*rich clubs*, et dans la partie non dense suivant l'orientation politique. Les résultats sont affichés dans la table 5.1.

	libéraux	conservateurs	total
$\delta-$*rich club* 1	113/3052 (48.2)	84/1682 (48.2)	197/5260 (27.2)
$\delta-$*rich club* 2	80/283 (9.0)	134/710 (8.0)	214/1083 (4.8)
$\delta-$*rich club* 3	50/42 (3.4)	96/116 (2.5)	146/165 (1.6)
Partie non dense	515/84 (0.06)	418/50 (0.05)	933/152 (0.03)
Total	758/7302 (2.5)	732/7841 (2.9)	1490/16718/ (1.5)

Table 5.1: Distribution des nœuds et des liens au sein du réseau des blogs politiques américains : A/B (C) correspond au nombre de nœuds pour A, à celui des liens pour B et au pourcentage de la densité de liens du sous-graphe induit pour C (c'est-à-dire $C = 2 \cdot \frac{B}{A \cdot (A-1)} \cdot 100$). Ici on calcule la densité sans distinguer la direction des liens, contrairement à ce qui est fait dans [2].

Pour $Q_{seuil} = Q_1/10$, la partie dense compte trois $\delta-$*rich clubs* qui représentent 37% des nœuds et 79% des liens du réseau total. Les trois $\delta-$*rich clubs* ont un nombre comparable de nœuds et, au sein de chacun d'entre eux, la densité des liens entre les nœuds libéraux, d'une part, et entre les nœuds conservateurs, d'autre part, est presque égale.

Nous observons aussi que la tendance qu'on mentionne plus haut (les blogueurs conservateurs créent plus de liens que les blogueurs libéraux), est le résultat d'un paradoxe de Simpson par rapport au découpage fourni par ItRich. Celui-ci est dû à une répartition non homogène des nœuds conservateurs et libéraux parmi les différents $\delta-$*rich clubs* identifiés [3]. En effet si on regarde la densité totale en liens, celle des blogs conservateurs est légèrement supérieure (que ce soit dans le sous réseau induit par les blogs conservateurs, ou bien en calculant la moyenne de la centralité des degrés, limitée à l'un ou l'autre des deux types de nœuds) à celle des libéraux. Cependant, à l'intérieur des $\delta-$*rich clubs* (et même dans la partie non dense), on retrouve une densité de liens supérieure ou égale chez les libéraux.

5.3.2 Le réseau mondial des transports aériens

Les résultats que nous allons montrer ici sont obtenus sur le réseau mondial des transports aériens, décrit dans la partie 3.4.3.
Nous pouvons voir sur la figure 5.7 que nous avons choisi de retenir trois $\delta-$*rich clubs* dans la partie dense, de tailles respectivement égales à 225, 257 et 346 nœuds (nous rappelons que la taille du réseau est de 3304 nœuds). Ceci donne une partie dense composée de 828 nœuds (25% de la taille du réseau) contre 2476 nœuds dans la partie non dense. Nous observons aussi les mêmes effets que ceux représentés sur la figure 5.8. sur cet exemple, au lieu de nous concentrer uniquement sur des propriétés topologiques, nous allons en plus de cela tenter de donner une interprétation pour les nœuds suivant leur appartenance, que ce soit dans les $\delta-$*rich clubs*, la zone de chevauchement ou bien le reste de la partie non dense.

Calcul d'un arbre de Steiner couvrant un ensemble aléatoire de terminaux

Le problème de l'arbre de Steiner est largement étudié dans le domaine de l'optimisation combinatoire [67]. Il peut être formulé de la manière suivante : étant donné un graphe G, dont les arêtes peuvent être pondérées, et un sous-ensemble S de sommets de G, nommé ensemble de nœuds terminaux, le problème consiste à trouver un ensemble d'arêtes de G de taille minimale (où de poids minimal si le graphe G est pondéré) tel que le sous-graphe induit soit connexe et contienne tous les sommets de S. Ce problème fait partie de la classe des problèmes NP-complets, mais il existe des algorithmes d'approximation qui fournissent un résultat en temps polynomial [33, 25].

Afin de mieux comprendre le rôle structurel que joue chacun des nœuds du réseau, une première solution serait de calculer les différentes centralités et d'en dégager des profils spécifiques de nœuds. Nous

[3] On ne prend pas en compte les arêtes reliant les nœuds conservateurs aux nœuds libéraux, car leur nombre est très faible et n'affecte pas les résultats ou leurs commentaires

allons utiliser une centralité que l'on a jugée pertinente pour l'analyse en cours, mais nous introduisons d'abord une mesure, que l'on note S_t.
Soit V l'ensemble de nœuds du réseau mondial des aéroports G, et soit N_{rep} un entier qui détermine le nombre d'itérations dans le procédé décrit par l'algorithme suivant :

Algorithm 4: Steiner score

Data: Un réseau $G = (V, E)$ et un entier N_{rep} qui détermine le nombre d'itérations dans cet algorithme.
Result: Le score $S_t(u)\ \forall u \in V$
$S_t(u) = 0\ \forall u \in V$;
while $i < N_{rep}$ **do**
  Tirer de V 300[a] nœuds au hasard pour constituer l'ensemble S de terminaux ;
  if $\frac{|G^{cc}[S]|}{|G[S]|} < \frac{1}{2}$ [b] **then**
    Calculer l'arbre de Steiner T dans G à partir de l'ensemble S de terminaux;
    $S_t(u) = S_t(u) + \frac{1}{N_{rep}}\ \forall u \in T \setminus S$;
    $i = i + 1$;
return $S_t(u)\ \forall u \in V$

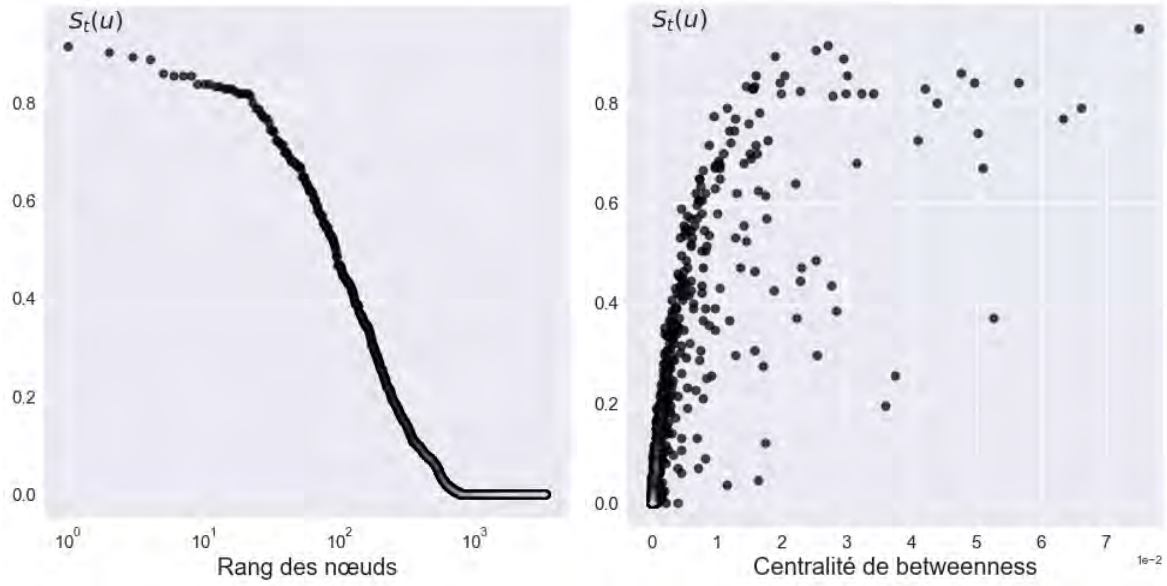
[a] Le nombre 300 est ici un choix arbitraire représentant approximativement le dixième de la taille du graphe sur lequel on fait nos tests, il est aussi préférable que ce chiffre soit petit devant le nombre de nœuds contenus dans G

[b] Avec $|G[S]|$ la taille du sous-réseau $G[S]$ généré par S dans G, et $|G^{cc}[S]|$ la taille de la plus grande composante connexe dans ce sous-réseau.

Il est important de préciser que le choix aléatoire de l'ensemble des terminaux n'est pas anodin. D'un côté cela permet de mettre sur le même pied d'égalité tous les nœuds du réseau (sous réserve que N_{rep} soit assez grand), et d'un autre côté, ce choix permet d'éviter de calculer l'arbre de Steiner d'un ensemble de terminaux qui induit dans G un sous-réseau connexe, ou contenant une composante géante (ce qui est de plus en plus probable à mesure que l'on augmente la taille de l'ensemble des terminaux, d'où le choix de 10% du nombre total de nœuds). Cela aurait pour effet de réduire la construction de l'arbre de Steiner à l'ajout d'un petit nombre de nœuds.
Un nœud avec un score élevé de S_t signifie que celui-ci joue un rôle médiateur important dans le réseau, qu'il participe à rendre connexes les différentes parties de celui-là. On serait tenté de croire que ceci n'est pas différent du résultat qu'on aurait obtenu d'une centralité de betweenness, mais comme vu lors du chapitre 3, la centralité de betweenness peut être altérée par la présence dans le réseau de nœuds ayant un grand degré, et qui se situent au milieu de clusters densément reliés. Ces nœuds sont certes importants car ils constituent les composants élémentaires des clusters denses, mais considérés individuellement, ils ne participent pas au maintient de la connexité du réseau, et n'ont pas de rôle médiateur parmi ses différents nœuds.
Considérons le problème de la façon suivante : s'il arrive qu'une panne géante rende impraticable tous les aéroports du monde, sauf quelques centaines d'entre eux, qui ont pu y échapper. Quels seraient alors les aéroports que l'on devrait en priorité rendre opérationnels, afin de rendre possible un voyage entre chaque deux aéroports parmi ceux qui sont encore opérationnels, sans pour autant créer de nouveaux liens que ceux qui existaient déjà avant la panne ?
On peut facilement se convaincre qu'il est nécessaire de réhabiliter certains des aéroports constituant les clusters du réseau de départ, mais il n'est pas nécessaire de le faire pour tous les nœuds ayant cette propriété, en raison de la redondance de leurs voisinages. Par exemple, l'aéroport de Frankfurt et l'aéroport d'Amsterdam sont tous les deux des éléments importants du cluster des aéroports européens, avec respectivement des degrés de 244 et 248. Il se trouve que ces deux aéroports ont également 165 destinations communes, ce qui réduit la nécessité de réhabiliter les deux en même temps.

Nous effectuons le calcul décrit plus haut, en prenant $N_{rep} = 1000$, et nous montrons sur la figure 5.9 les résultat obtenus pour chaque nœud, ainsi que les valeurs de $S_t(u)$ en fonction de la centralité de betweenness.

Nous pouvons voir sur la figure 5.9b qu'il existe effectivement une corrélation entre les valeurs de $S_t(u)$ et la betweenness, avec un coefficient de Pearson égal à 0.75.
Cela dit, ce résultat est très influencé par la corrélation qu'il y a entre les nœuds ayant à la fois de petites



(a) (b)

Figure 5.9: (a) Les valeurs normalisées de $S_t(u)$ dans l'ordre décroissant, (b) La valeur normalisée de $S_t(u)$ en fonction de la centralité de betweenness de chaque nœud.

valeurs de S_t et de betweenness. En restreignant par exemple la corrélation aux 1000 nœuds ayant les valeurs les plus élevées de $S_t(u)$, ce qui représente presque le tiers de l'ensemble des aéroports, nous obtenons un indice de corrélation égal à -0.11 et un autre de -0.08 si on retenait le même nombre mais suivant les valeurs les plus élevées de la betweenness. La corrélation est proche de zéro si on choisissait de considérer uniquement 100 nœuds (respectivement 0.04 et -0.09). Nous en déduisons que pour les nœuds dont les scores sont élevés (que ce soit ceux de la betweenness, ou de S_t), la corrélation est plus faible. De plus, il est important de noter que la tendance initiale se trouve inversée, car on passe d'une corrélation positive à une corrélation négative.

Ainsi nous nous concentrons désormais sur les nœuds ayant les plus grandes valeurs de S_t, et nous trouvons que sur les 100 nœuds les mieux classés, 76 sont dans la partie dense du réseau, avec 49 nœuds dans le premier $\delta-$*rich club*, 17 dans le second et 10 dans le troisième.
De l'autre côté, sur les 24 nœuds de la partie non dense qui sont identifiés, on en compte 9 qui se situent dans la zone de chevauchement, parmi lesquels se trouve le nœud ayant le score le plus élevé de S_t, représentant l'aéroport international Ted Stevens d'Anchorage, en Alaska. Ce rapport de 9/24 est près de trois fois supérieur au rapport distinguant les nœuds de la zone de chevauchement et ceux de la partie non dense $(387/2476)$[4]. Nous montrons sur la table 5.2 les 10 premiers aéroports identifiés dans la partie dense et la partie non dense.

À première vue, les aéroports listés sur la table 5.2 n'ont pas grand chose en commun, mais en regardant de plus près, on peut constater qu'ils ont quelques propriétés structurelles qui les rendent relativement similaires. Premièrement ils ont un coefficient de clustering significativement faible, sachant que la valeur moyenne de celui-ci dans le réseau vaut 0.49, nous avons un clustering moyen pour les nœuds de la partie dense affichés sur la table 5.2 égal à 0.24, et d'une valeur de 0.09 pour ceux de la partie non dense. Cela est naturellement dû au fait que les aéroports qu'ils desservent sont peu reliés entre eux, ce qui leur confère une position de points relais. Ceci n'empêche pas certains des nœuds de la table 5.2 d'être aussi reliés à un ou plusieurs clusters. Prenons par exemple l'aéroport Ted Stevens d'Anchorage, parmi ses voisins, on retrouve des aéroports américains ayant une forte valeur de δ : Chicago O'Hare, Las Vegas, Los Angeles, Phoenix Sky Harbor, Seattle, etc. tous font partie du premier $\delta-$*rich club*. Ces voisins forment eux-mêmes un cluster auquel appartient l'aéroport d'Anchorage, mais

[4]Sur les 100 nœuds de plus forte betweenness, 90 sont dans la partie dense et 10 dans la partie non dense

Partie non dense			**Partie dense**		
Aéroport	S_t	**Rang**	**Aéroport**	S_t	**Rang**
Ted Stevens Anchorage*	0.888	1	Domodedovo	0.880	2
Bethel*	0.868	4	Bogota-El Dorado	0.876	3
Faa'a*	0.828	15	Hartsfield Jackson Ata	0.844	5
Fairbanks	0.824	16	Montreal P.E. Trudeau	0.840	6
Yellowknife	0.792	23	Chicago O'Hare	0.840	6
Honiara	0.736	36	Juan Santamaria	0.840	6
Ndjili*	0.720	39	Stockholm-Arlanda	0.836	9
Bauerfield	0.708	42	Port Moresby Jacksons	0.836	9
Marcos A. Gelabert	0.668	47	Ninoy Aquino	0.836	9
Sioux Lookout	0.664	48	Dubai	0.836	9

Table 5.2: La liste des aéroports de plus forts S_t, la valeur de celle-ci et le rang des nœuds identifiés suivant cette valeur. Dans la moitié de droite nous avons les nœuds de la partie dense, et dans celle de gauche les nœuds de la partie non dense, avec une astérisque accompagnant les aéroports qui sont dans la zone de chevauchement des valeurs de δ.

son δ n'étant pas aussi élevé que celui obtenu par la valeur moyenne calculée sur son voisinage, il se voit relayé dans la zone de chevauchement une fois supprimé le $\delta-$*rich club* qui contient les aéroports qu'on vient de mentionner. Ces destinations ne sont pas les seules qu'on retrouve, l'aéroport d'Anchorage dessert aussi un grand nombre de destinations locales : des villes ou bien des îles qui se situent dans l'état de l'Alaska comme Kodiak, Kotzebue, Kenai, Unalakleet, etc. dont il est l'unique relais, sans doute à cause de leurs positions géographique isolées. On retrouve la même caractéristique chez le reste des nœuds de la partie non dense de la table, ainsi qu'une propriété très similaire chez les nœuds de la partie dense, à la différence que l'appartenance de ces derniers à un (ou plusieurs) clusters est largement plus prononcée. De plus, si les nœuds de la partie non dense desservent des destinations qui leurs sont (en moyenne) géographiquement proches, ceux de la partie dense sont en revanche des points de relais entre des aéroports de diverses régions du monde. Par exemple l'aéroport Domodedovo (Moscou) dessert l'extrême orient russe, l'Asie du sud-est, l'Asie centrale, l'Europe et quelques villes du continent américain. Chaque région est desservie via plusieurs aéroports qui forment localement un cluster.

En d'autres termes, les nœuds qui sont sur la table 5.2 et dans la partie non dense jouent le rôle de relais locaux, alors que ceux de la partie dense ont plutôt celui de relais globaux, comme l'appuient les valeurs moyennes des distances géographiques qui séparent les aéroports de leurs destinations respectives. On a une distance moyenne qui vaut $2258km$ pour les dix nœuds de la partie dense, contre $1039km$ pour les 10 de la partie non dense [5]. La distance moyenne calculée sur l'ensemble des nœuds du réseau étant de $1760km$. Les aéroports de la table 5.2 ainsi que leurs destinations sont représentées sur la figure 5.10

5.4 ItRich dans le contexte de la détection de communautés

5.4.1 Données

Nous avons vu dans les applications précédentes qu'ItRich propose une nouvelle manière de découper un réseau, dont nous avons exploré les différents composants. Après avoir vu en détails le résultat de ce découpage sur de petits réseaux, nous nous sommes ensuite intéressés à l'ensemble caractérisé par son appartenance à la partie non dense associée à des valeurs élevées de δ, qui auraient pu permettre aux nœuds de cet ensemble d'être classés dans la partie dense. Nous avons ainsi caractérisé topologiquement cet ensemble, en pointant le fait qu'il soit composé de nœuds dont les valeurs de δ sont significativement inférieures à celles de leur dense voisinage. Nous avons ensuite montré sur un exemple d'application que certains des nœuds qui sont dans cette position jouent un rôle médiateur important dans le réseau.

[5] La distance entre deux aéroports est ici représentée par une géodésique, ce qui n'est pas tout à fait vrai en réalité, où les vols sont parfois contraints à des trajectoires plus longues pour diverses raisons (politiques, météorologiques...)

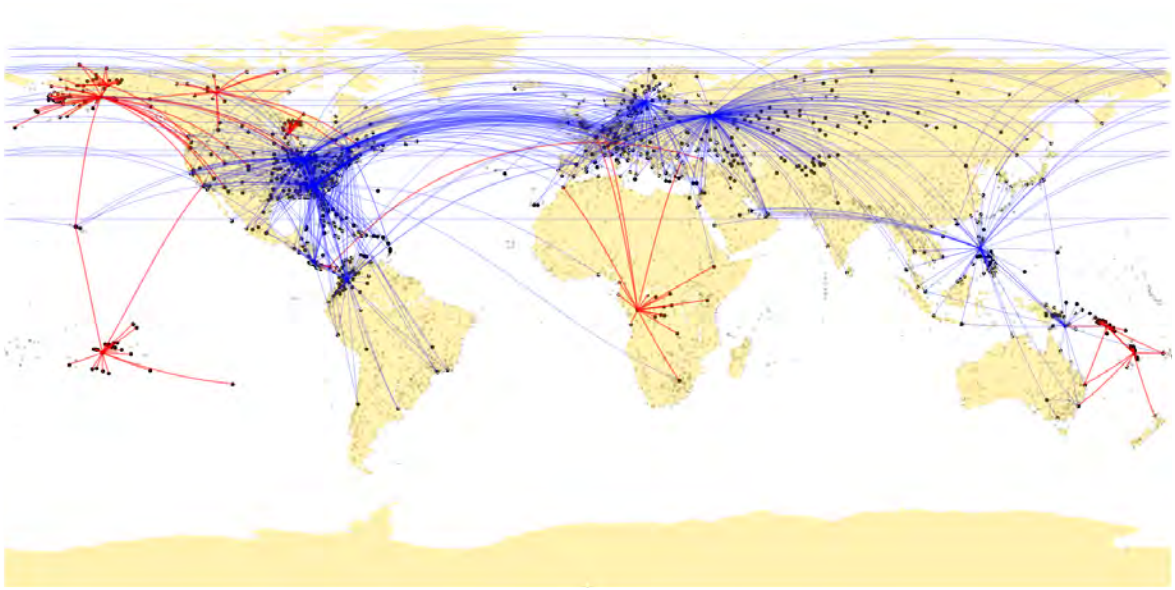


Figure 5.10: Les aéroports listés sur la table 5.2 ainsi que leurs destinations respectives, les arêtes bleues partent des nœuds de la partie dense et celles en rouge de ceux de la partie non dense.

À présent nous allons montrer qu'il existe un avantage à associer les résultats d'ItRich à ceux d'un algorithme standard de partitionnement en communautés. Nous faisons le choix délibéré de ne considérer que les algorithmes de détection non paramétrique, afin de renforcer l'utilité de cette association [6].

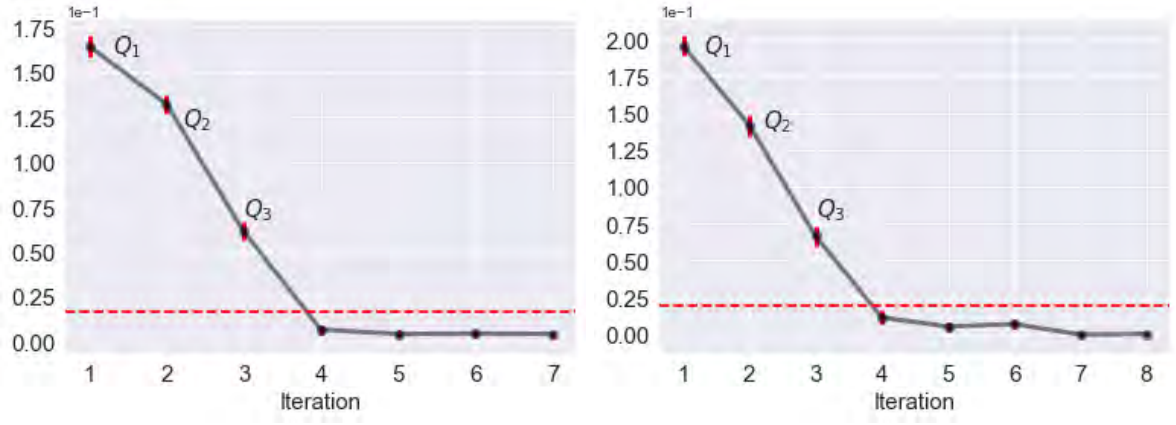
Pour cette partie, nous allons analyser des données provenant de la saga "Le trône de fer" ("A song of ice and fire" en anglais) qui contient à ce jour 5 volumes, et qui a été adaptée en série télévisée en 8 saisons ("Game of thrones"), dont les 6 premières sont directement inspirées des livres. Les données sont disponibles sur `https://networkofthrones.wordpress.com/` et ont été construites de la manière suivante : On crée un lien entre deux personnages si leurs noms (ou surnoms) sont mentionnés à moins de 15 mots l'un de l'autre au sein du même chapitre (ou séquence dans le cas de la série). Nous nous basons sur les données provenant du script des différents épisodes, rendu public après leurs sorties, pour construire le réseau provenant de la série télévisée. À cause de la multitude des lieux et le grand nombre de personnages intervenant dans la saga, les producteurs de la série ont dû faire des choix restrictifs pour rentrer dans leur budget. Ainsi, le nombre de personnages dans la série télévisée est nettement inférieur à celui de la saga, et certaines intrigues ont été délibérément délaissées. Nous montrons dans cette partie qu'il est possible de mettre en évidence une partie importante de ce qui différencie la série de la saga, et que nous y parvenons plus facilement en combinant un algorithme de détection de communautés [100] et ItRich.
Afin d'éviter qu'il y ait des divergences trop importantes, nous nous sommes restreints aux six premières saisons de la série, que l'on compare avec les 5 tomes qui leur servent de supports scénaristiques.
Les réseaux qui en résultent sont composés de 796 nœuds et 2823 arêtes pour le réseau de la saga qu'on notera ici G, et de 384 nœuds reliés par 2127 arêtes pour le réseau de la série que l'on notera H.

5.4.2 Résultats de la comparaison

Tout d'abord, nous pouvons voir sur la figure 5.11 que les deux jeux de données ont le même nombre de $\delta-$*rich clubs*, ce qui facilite leur comparaison. Le nombre de nœuds composant les $\delta-$*rich clubs* est cependant différent, nous comptons 53 nœuds dans le premier $\delta-$*rich club* pour les données de la saga contre 48 pour les données de la série, pour le second nous avons respectivement 113 et 56 nœuds et pour le dernier $\delta-$*rich club* nous avons 97 nœuds d'un côté et 47 de l'autre. Mis à part le premier $\delta-$*rich club*, dont la taille est presque la même dans un cas comme dans l'autre, il existe une différence de taille

[6] Les algorithmes paramétriques de détection de communautés nécessitent parfois une connaissance anticipée des données, voire de la vérité de terrain, comme par exemple la taille moyenne des communautés, la clique minimale que doit contenir chacune des communautés, etc.

remarquable dans les second et troisième $\delta-$*rich clubs*, avec plus de nœuds dans les ensembles provenant des données de la saga. Cela est normal car le réseau initial fait lui-même presque le double de la taille du réseau qui modélise la série. En revanche, on retrouve après la normalisation par le nombre respectif de personnages, que la partie dense contient 33% des nœuds de la saga et 39% dans le cas de la série télévisée.



(a) (b)

Figure 5.11: Les valeurs de Q pour (a) les données de la saga (b) les données de la série télévisée. La ligne en pointillés rouges représente la valeur de $Q_{seuil} = Q_1/10$. Pour chaque réseau, le modèle nul est calculé 100 fois et les barres d'erreur représentent les écarts types de la mesure de qualité.

Il est aussi intéressant de remarquer que la taille presque constante dans les deux cas du premier $\delta-$*rich club* découle du fait que l'ensemble des personnages principaux (ceux qui constituent le cluster le plus dense) est le même dans un cas comme dans l'autre avec un indice de recouvrement de 70%.

Afin de comparer les deux jeux de données nous lançons deux algorithmes de détection de communautés, le premier correspond à la méthode de la propagation de labels, et le second à celle de l'optimisation de la modularité. Nous lançons chacun de ces deux algorithmes d'abord sur l'intégralité des réseaux, puis en les restreignant à chaque fois aux différents ensembles retournés par ItRich.
Sur la figure 5.12 nous pouvons voir le nombre et la taille de chacune des communautés, issues de ces deux différentes partitions, quand le réseau en entrée est intégral.

Sur la figure 5.12 nous pouvons voir les tailles des différentes communautés retournées par les deux algorithmes. Il est difficile de comparer les résultats provenant de ces deux jeux de données, simplement parce que le faible nombre de communautés détectées, ainsi que les tailles significativement grandes de certaines communautés, nous empêchent d'effectuer une comparaison précise. À titre d'exemple, la communauté composée des 443 nœuds détectés par la méthode de propagation de labels, tout comme celle comptant 230 nœuds qui résulte d'une optimisation de la modularité, constituent dans les deux cas une fraction importante de la partie dense dans le réseau correspondant. Ceci nous laisse avec des petites communautés dont les nœuds ne constituent souvent pas des personnages d'une grande importance.
C'est pour cela que nous choisissons de lancer une détection de communautés sur les ensembles séparés, qui sont constitués par les $\delta-$*rich clubs* identifiés ainsi que par la partie non dense. Les résultats de cette analyse sont montrés sur la figure 5.13

En comparant chaque graphique de la figure 5.13 avec son équivalent de la figure 5.12, nous pouvons voir que la détection de communautés restreinte aux résultats d'ItRich (en traitant indépendemment chaque $\delta-$*rich club* ainsi que la partie non dense, puis en regroupant les résultats) résulte en un nombre plus grand de communautés détectées, ainsi qu'une une distribution moins hétérogène de leurs tailles. On retrouve une communauté "géante" sur la figure 5.13a et la figure 5.13c, qui résulte du fait que la méthode de propagation de labels considère comme une seule communauté tous les nœuds du premier $\delta-$*rich club*, que ce soit pour le jeu de données traitant de la saga comme celui qui traite la série télévisée. Ceci n'est le cas ni pour les $\delta-$*rich clubs* de rangs inférieurs, ni pour l'ensemble des résultats issus de l'optimisation de la modularité.

Sachant qu'il existe des différences plus ou moins importantes entre les deux jeux de données, nous n'allons pas chercher à comparer deux à deux toutes les communautés pour en tirer des informations.

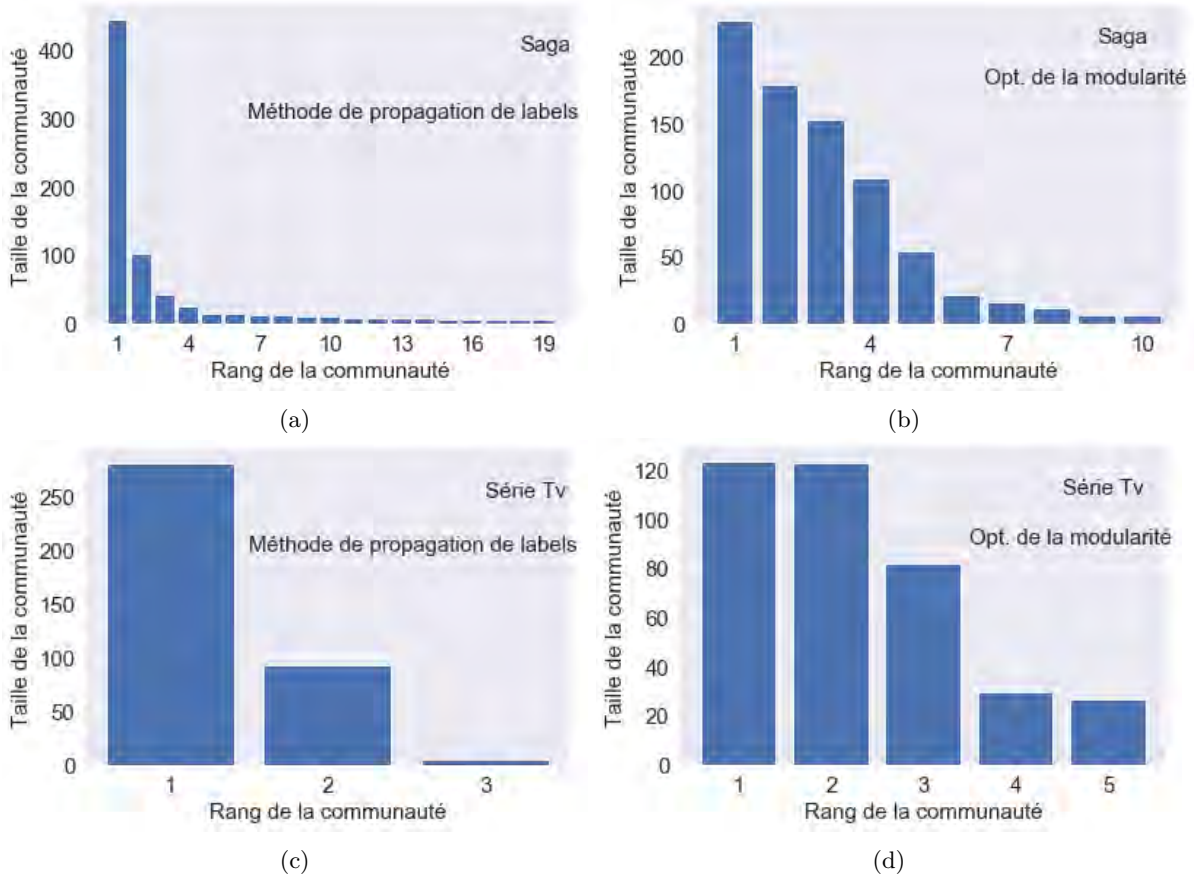


Figure 5.12: Tailles des communautés obtenues par l'algorithme de propagation de labels sur les données provenant de l'intégralité des réseaux de la saga (a) et de la série TV (c), ainsi que par un algorithme glouton d'optimisation de la modularité sur (respectivement) (b) et (d).

L'effort serait effectivement intéressant, mais il ne faut pas oublier que les algorithmes employés sont eux-mêmes sujets à critiques, et qu'il serait d'abord question de comparer les deux méthodes employées avant de comparer leurs résultats, ce qui n'est pas le but de cette application.
Nous allons donc nous concentrer sur les communautés dont tous les éléments sont uniquement dans l'un ou l'autre des deux jeux de données étudiés. Ceci a pour avantage de directement mettre en évidence les fragments d'histoires qui ont été rajoutés dans la série mais qui ne figurent pas dans la saga, ou à l'inverse les intrigues qui font partie de la saga mais qui ont été écartées lors de l'adaptation.

Pour cela, nous calculons l'indice de recouvrement entre chaque paire $(C_i^{saga}, C_j^{série})$ de communautés, tel que C_i^{saga} est une des communautés calculées sur le premier jeu de données, et $C_j^{série}$ du second jeu de données. Nous montrons sur la figure 5.14 les résultats d'un tel calcul.

Sur la figure 5.14, les colonnes dont la somme des éléments est nulle, représentent les communautés dont aucun des personnages n'est mentionné dans la saga, et ont donc été rajoutés lors de l'adaptation. De même que les lignes dont la somme est nulle représentent les communautés de la saga dont les personnages ont été écartés lors de l'adaptation. Nous identifions sur la 5.14b 16 communautés qui sont uniquement dans la saga contre 4 qui sont propres à la série télévisée. La figure 5.14a révèle une seule communauté propre à la saga et aucune qui n'existe que dans la série.

On pourrait objecter que ceci aurait pu être accompli sans passer ni par un découpage en $\delta-rich$ *clubs*, ni par une détection de communautés (en effet il suffirait d'identifier les éléments qui sont dans un ensemble sans être dans l'autre) mais on n'aurait alors eu aucune information sur les liens entre ces personnages. Effectivement, le fait que quelques-uns parmi ces nœuds soient dans la même communauté suggère que les personnages correspondants font aussi partie de la même intrigue.

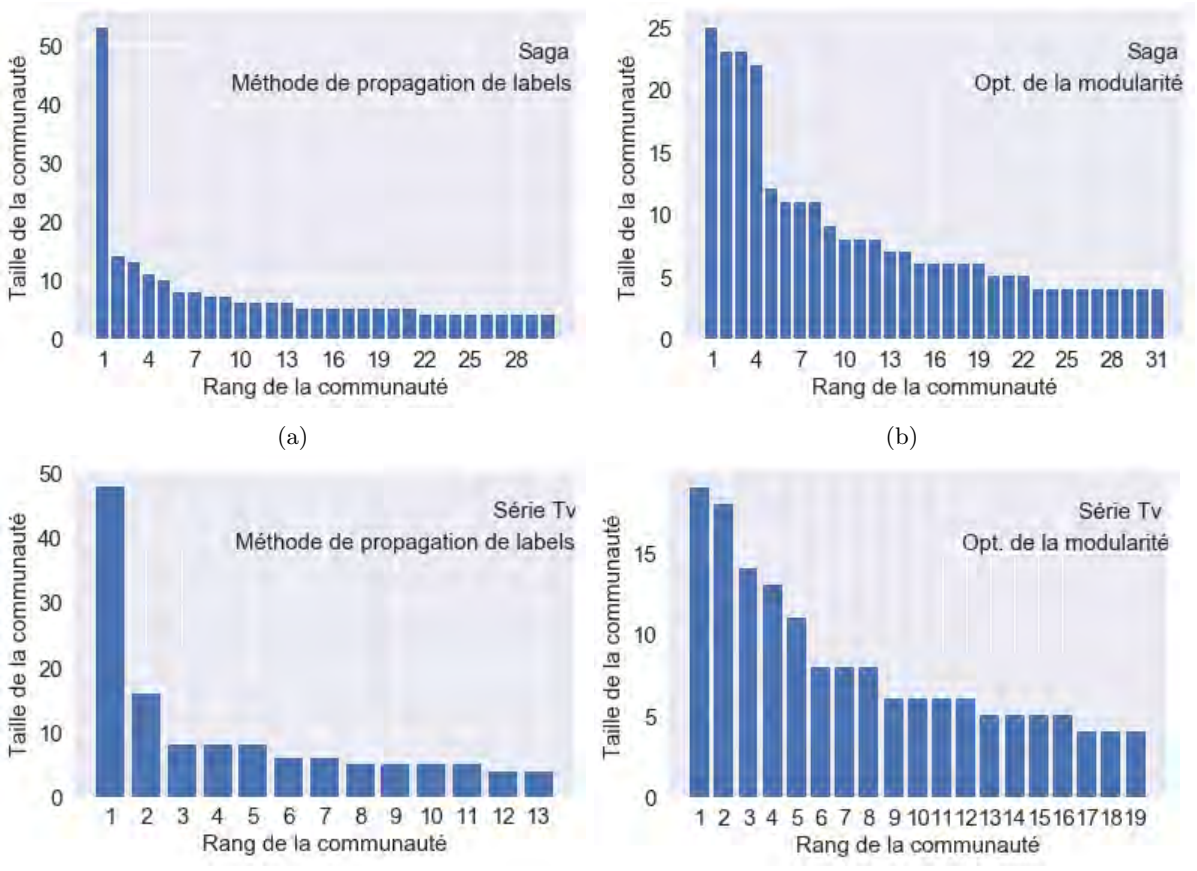


Figure 5.13: Tailles des communautés obtenues par l'algorithme de propagation de labels sur les résultats provenant des $\delta-$*rich clubs* calculés sur la saga (a) et la série TV (c), ainsi que par un algorithme glouton d'optimisation de la modularité sur (respectivement) (b) et (d). Seules les communautés dont la taille est supérieure à 4 nœuds sont affichées ici.

Nous allons maintenant donner quelques unes des communautés vérifiant les propriétés listées plus haut, en commençant d'abord par les communautés propres à la série[7] :

- *Bianca, Izembaro, Lady Crane, Bobono, Camello & Clarenzo* sont des personnages représentant une troupe de comédiens de la sixième saison, leur présence n'est pas surprenante sachant qu'une partie de l'histoire de la sixième saison dépasse déjà l'histoire originale. Ils sont classés dans le troisième $\delta-$*rich club*.
- *Loboda, White walker, Karsi, Night King & Leaf* Certains de ces personnages existent aussi dans la saga, mais leurs noms sont différents d'un jeu de données à l'autre (par exemple le roi de la nuit ne désigne pas le même personnage dans l'histoire originale, et Loboda joue un rôle similaire à celui de *Segorn* dans la saga). Ils sont classés dans la partie non dense.

Nous avons deux autres communautés qu'on retrouve pour des raisons sensiblement similaires à celles qu'on a mentionnées plus haut (une communauté de membres de la garde de nuit dont les noms ont été changés, et une autre composée de nains qui ont participé à un spectacle humoristique lors d'un mariage de la quatrième saison).

Nous faisons maintenant l'inverse en citant quelques unes des communautés qui sont propres à la saga, sachant qu'on en compte une vingtaine vérifiant la propriété (en comptant les communautés de taille

[7] Les noms des personnages sont en anglais

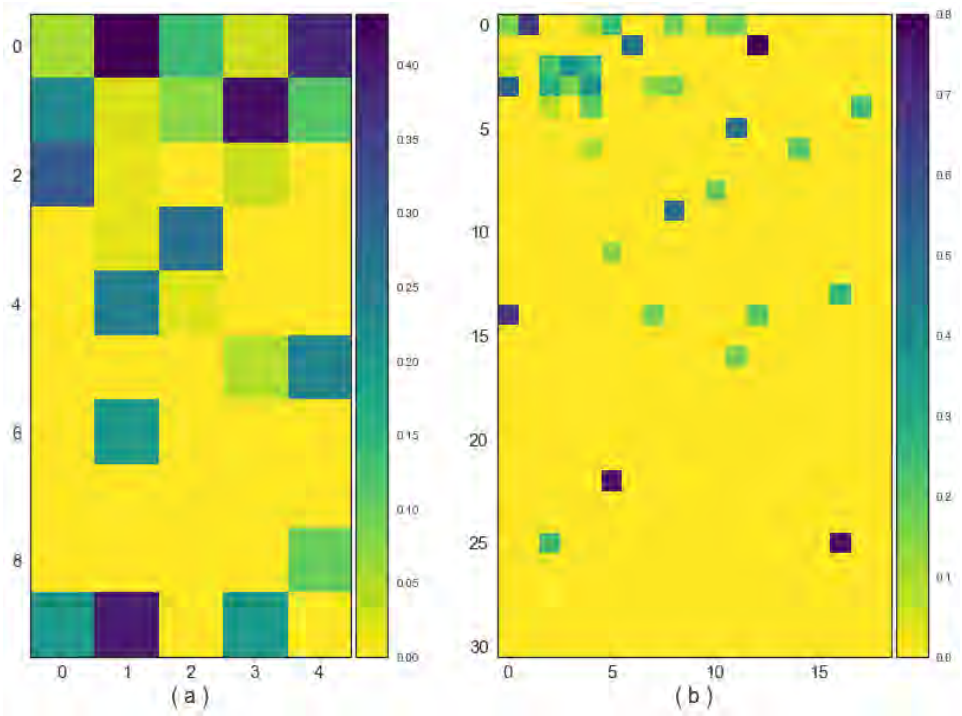


Figure 5.14: Matrices de recouvrement, chaque élément représente l'indice de recouvrement entre les communautés identifiées sur la saga (indices des lignes) et la série télévisée (indices des colonnes), par l'algorithme d'optimisation de la modularité. Sur (a) nous analysons les réseaux entiers, et sur (b) nous restreignons la détection de communautés à chaque portion du découpage fourni par ItRich, nous regroupons après coup l'ensemble des communautés ainsi obtenues.

inférieure ou égale à 4), nous allons citer ici les plus importantes et expliquer brièvement leur rôle dans l'histoire de la saga.

- *Armen, Leo Tyrell, Alleras, Pate (novice), Roone & Mollander* sont les personnages introduits dans le prologue du quatrième tome, et qui sont des étudiants de la citadelle de *Villevieille*, le décor dans lequel convergent plusieurs intrigues qui s'annoncent d'une grande importance pour la suite de l'histoire. Ils sont classés dans le troisième $\delta-$*rich club*.

- *Rodrik Harlaw, Aeron Greyjoy, Victarion Greyjoy& Asha Greyjoy* représentent les personnages les plus importants dans l'intrigue liée aux *îles de fer*, qui a été largement raccourcie (autant en termes d'histoire que de personnages impliqués) dans la série télévisée, et qui s'annonce importante pour la suite. Ils sont classés dans le second $\delta-$*rich club*.

- *Yandry, Lemore, Haldon, Aegon VI Targaryen (fils de Rhaegar), Rolly Duckfield, Ysilla & Jon Connington.* Cette communauté est sans doute la plus importante, car elle compte un personnage qui est l'héritier prétendu du trône, présumé mort en bas âge, ainsi que plusieurs autres personnages chargés de sa protection. Ils sont classés dans le second $\delta-$*rich club*.

Nous pourrions aller plus loin en rajoutant les communautés dont l'indice de recouvrement est faible (ce qui signifie que parmi tous ses membres, seuls quelques uns existent dans le jeu de données de comparaison) mais cela augmenterait significativement le nombre de communautés concernées.

Il est très important de préciser que la même comparaison effectuée sur les résultats issus d'une analyse sur l'intégralité des graphes n'en aurait pas autant révélé. En effet, les éléments des communautés que l'on retrouve ici (les trois communautés identifiées plus haut) sont noyés dans des communautés de grandes tailles, et qu'on ne peut expliciter de la même façon. Il est tout aussi important de rappeler que nous utilisons des algorithmes de détection de communautés qui ne sont pas paramétriques, le seul

paramètre qu'on fixe ici est celui de Q_{seuil}, qui est motivé par les résultats affichés sur la figure 5.11. Nous arrivons donc en combinant ItRich et un algorithme simple de détection de communautés, à identifier des différences majeures entre les deux jeux de données examinés. Cette comparaison est difficilement faisable en se restreignant à une détection de communautés lancée sur l'intégralité des réseaux [8], sans passer par le choix d'un paramètre d'échelle.

5.5 ItRich et graphes dynamiques : Étude des contacts entre élèves dans des établissements scolaires

Nous terminons ce chapitre en montrant les résultats d'ItRich appliqué à des données temporelles. Il s'agit de trois jeux de données que l'on peut retrouver sur le site `http://www.sociopatterns.org`, et qui enregistrent les interactions entre des élèves au sein de leurs établissements scolaires respectifs. Pour ce faire, chaque élève est équipé d'un badge contenant une puce de type RFID (identification par radio-fréquences), qui enregistre un contact à chaque fois qu'elle est à proximité d'une autre puce, portée en l'occurrence par un autre élève. Ceux-ci sont appelés à porter leurs badges sur leurs poitrines, afin que les interactions ne soient enregistrées que lorsque deux personnes se retrouvent l'une en face de l'autre, et à une distance maximale comprise entre $1m$ et $1.5m$ [115]. Cette dernière restriction a été rajoutée afin de pouvoir analyser les interactions à courte distance, susceptibles de jouer un rôle important en cas d'épidémies de maladies transmissibles par les voies respiratoires, à travers les postillons (éternuements, toux). Les paramètres de l'infrastructure sont réglés de sorte que la proximité entre deux personnes portant les badges RFID peut être évaluée et enregistrée avec une probabilité supérieure à 99% sur un intervalle de 20 secondes [115]. Cette échelle de temps permet une description adéquate des interactions entre les personnes, y compris si celles-ci sont brèves. Les interactions en dehors de l'établissement scolaire ne sont pas enregistrées.

Nous analysons les données enregistrées au sein d'une école primaire comptant 10 classes, étalées sur 5 années scolaires différentes, ainsi que deux jeux de données enregistrées pendant plusieurs jours sur des élèves de deuxième année en classes préparatoires. Les données étant récoltées à l'aide du même dispositif, nous pouvons ainsi comparer leurs résultats.

École primaire, 2009

Collectées durant deux journées successives d'octobre 2009, dans une école primaire de la ville de Lyon. Elles enregistrent des données relevées sur 242 personnes, dont 232 élèves en plus des 10 enseignants de chaque classe. Le nombre d'élèves par classe varie entre 21 et 26 élèves. Le nombre d'interactions enregistrées sur la durée totale de l'étude s'élève à 77602.

Classe préparatoire, 2011

Collectées durant cinq journées dans un lycée de la ville de Marseille. Le tôt de participation est de 100% [57], même si un faible nombre d'élèves refusent de porter en permanence leurs badges. Le nombre de classes est de 5, et le nombre d'élèves par classe varie entre 31 et 41 élèves. Le nombre total d'évènements enregistrés est quant à lui égal à 19774.

Classe préparatoire, 2013

Collectées durant cinq journées (une semaine de cours), dans le même lycée qui fournit les données de 2011. Celles-ci leur sont aussi très similaires, à la la seule différence qu'elles comptent les interactions des élèves issues de 9 classes différentes, contre 5 pour celles en haut. Le nombre d'élèves par classe varie entre 29 et 40 élèves [86].

5.5.1 Modélisation en graphes dynamiques

Comme nous l'avons mentionné, la résolution des données permet de générer des graphes dynamiques dont la fenêtre de largeur minimale est de 20 secondes (en combinant toutes les interactions qui ont pu

[8] La seule communauté qu'on retrouve ainsi, *cf.* la ligne 7 de la matrice représentée sur la figure 5.14a, est constituée de personnages secondaires (*Murch, Aggar, Gariss, Gelmarr, Gynir*), qui apparaissent uniquement dans le second tome et qui ne constituent pas une intrigue exclusive

avoir lieu pendant cette fenêtre dans un seul graphe). Nous choisissons cependant une largeur de 10 minutes, car les élèves sont dans leurs salles de classe pendant la majeure partie de la journée, et les évènements y sont plus rarement enregistrés (sauf pour les élèves de primaire qui comme nous le verrons montrent une activité importante durant les heures de cours).
Il reste important de rappeler qu'en élargissant la fenêtre temporelle, nous réduisons la précision du modèle. En représentant l'ensemble des interactions qui ont eu lieu au cours d'une fenêtre temporelle donnée par un seul lien non pondéré, il est possible de passer à côté de certaines interactions qui sont répétées plusieurs fois durant la fenêtre considérée. De plus, les contacts sont de durées variables, ce qui est de moins en moins facile à voir si les fenêtres sont de longues durées. Les analyses de [57] montrent cependant que la distribution des durées d'interactions est en loi de puissance, et qu'une grande majorité des contacts sont de courte durée (avec une durée moyenne de 33 secondes sur les données récoltées au sein de l'école primaire). Ainsi la largeur choisie doit être assez courte pour que la probabilité d'enregistrer plusieurs interactions entre la même paire d'individus durant la fenêtre considérée soit faible, mais aussi assez longue afin d'éviter que le graphe dynamique qui en résulte ne soit trop souvent composé uniquement de nœuds, avec peu ou pas de liens entre eux.
Le choix de la largeur de la fenêtre temporelle est le même pour toutes les données analysées (10 minutes), ainsi que la valeur de Q_{seuil} qui est là aussi prise par défaut, d'une valeur égale au dixième de sa valeur maximale.

5.5.2 Analyse temporelle des résultats d'ItRich

Tailles des parties en fonction du temps

Nous commençons par montrer l'évolution des tailles respectives des parties dense et non dense, sur une journée complète de présence dans l'établissement.

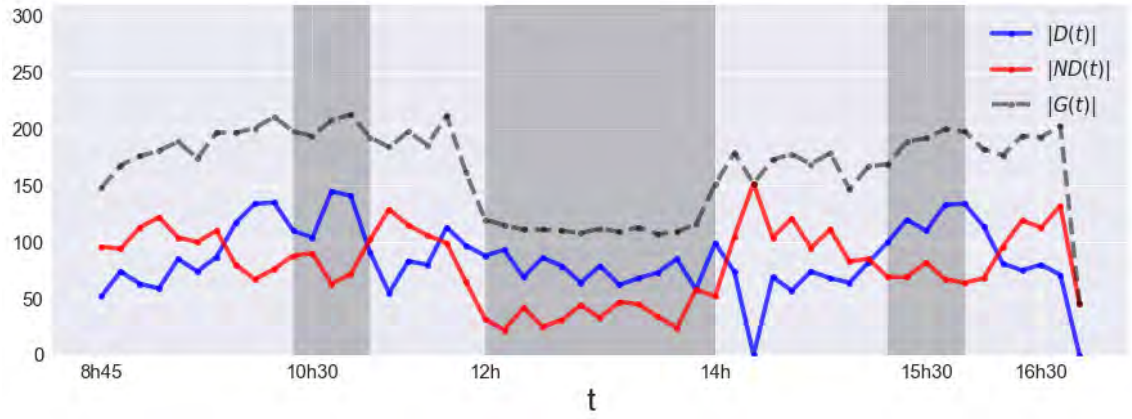
La figure 5.15 montre des similitudes entre les résultats obtenus sur les données provenant des élèves de classes préparatoires (*cf.* fig. 5.15a et fig. 5.15b). On remarque un pic principal qui se situe autour de 10h, donc durant la pause matinale, ainsi que des pics de moindre amplitude pendant la pause déjeuner et l'après-midi. Le reste du temps, la partie dense est soit vide soit contient un faible nombre de nœuds Parfois il suffit qu'il existe une clique de 4 individus ou plus pour qu'ItRich classe ses nœuds dans la partie dense. Ces résultats sont en revanche différents de ceux observés sur les données provenant de l'école primaire, où l'on remarque que la partie dense existe et a une taille non négligeable, et ce même durant les heures de cours. En revanche on remarque dans ce dernier cas que la taille de la partie dense est inférieure à celle de la partie non dense, durant la majeure partie des heures de cours. Cette tendance s'inverse systématiquement durant les récréations (une en matinée et une l'après midi) ainsi que durant la pause déjeuner. On remarque durant cette pause que la taille du graphe chute en raison du fait que certains élèves restent sur place, alors que d'autres rentrent chez eux pour déjeuner. Cette diminution de taille est plus significative sur la courbe rouge que sur la bleue de la figure 5.15a, suggérant une forte interaction entre les individus qui déjeunent à la cantine de l'établissement.

Nous précisons aussi que les heures auxquelles ont lieu les interactions ne sont pas enregistrées. Nous nous basons sur le fait que le premier contact enregistré de la journée doit être proche de l'heure à laquelle débutent les cours, et nous établissons notre référentiel temporel en fonction de cet évènement. Il est donc possibile d'avoir quelques imprécisions sur l'heure exacte à chaque pas de temps.

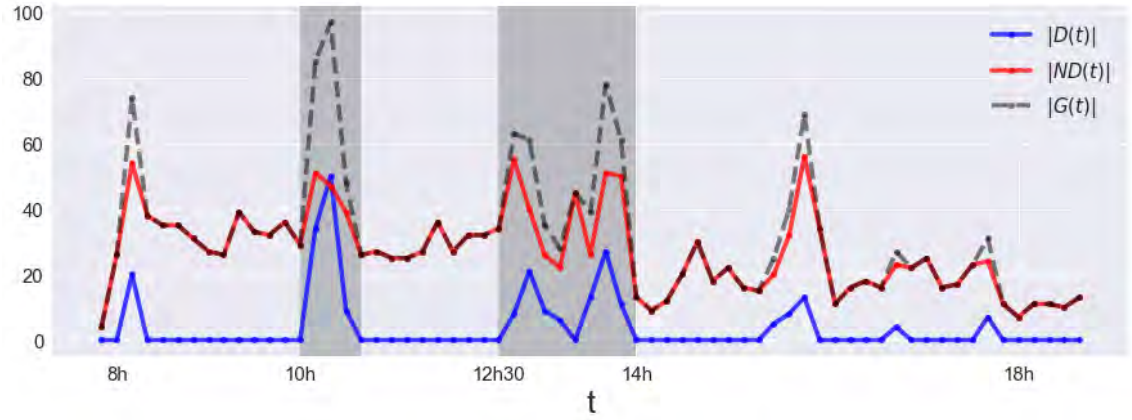
Dynamique et échanges entre parties

Si les tailles respectives de la partie dense et de la partie non dense nous informent sur la quantité d'interactions à différents moments de la journée, elles ne permettent pas en revanche de suivre avec précision l'évolution de l'une ou de l'autre. En effet les nœuds qui constituent par exemple la partie dense à l'instant t sont potentiellement différents de ceux qui la composaient à l'instant $t - \Delta t$ (avec Δt égal à $10mn$, la largeur de la fenêtre temporelle). Ainsi pour mieux comprendre la dynamique de la partie dense et de la partie non dense, nous devons examiner les échanges entre ces deux ensembles, ainsi qu'entre les nœuds qui qui ne manifestent aucune interaction pendant le pas de temps considéré.

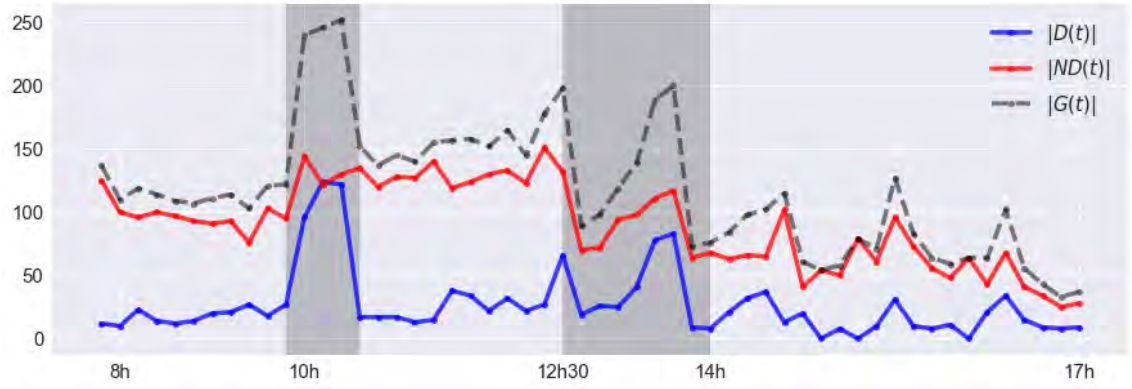
Notons $D(t)$ et $ND(t)$ les ensembles qui constituent respectivement la partie dense et non dense à l'instant t, et $D(t - \Delta t)$ et $ND(t - \Delta t)$ ceux du pas de temps antérieur. Sachant que certains individus ne forment aucun lien avec le reste du graphe pendant le pas de temps considéré, nous appelons ces nœuds les "absents" (en ce sens où ils sont absents au cours de la fenêtre de temps considérée, mais pas nécessairement absents toute la journée) et nous les notons $abs(t)$ et $abs(t - \Delta t)$.



(a)



(b)



(c)

Figure 5.15: Tailles respectives de la partie dense, de la partie non dense et de l'ensemble du graphe en fonction du temps. Les courbes correspondantes sont respectivement en bleu, rouge et gris. Sur (a) nous représentons les résultats obtenus sur les données récoltées dans une école primaire. Sur (b) et (c) les données récoltées dans un lycée sur des élèves de classes préparatoires. Les intervalles sur-lignés en gris représentent les heures durant lesquelles les élèves ne sont pas en cours. La durée de temps totale couvre une journée entière.

Nous calculons ensuite l'indice de Jaccard (nous rappelons que l'indice de Jaccard entre deux ensembles A et B est obtenu par le rapport entre la taille de l'intersection des deux ensembles et la taille de leur union) entre ces différents éléments, que nous représentons en fonction du temps sur la figure 5.16. Nous notons $J(A(t-\Delta t), B(t))$ l'indice de Jaccard entre l'ensemble A à l'instant $t-\Delta t$ et

l'ensemble B à l'instant t.

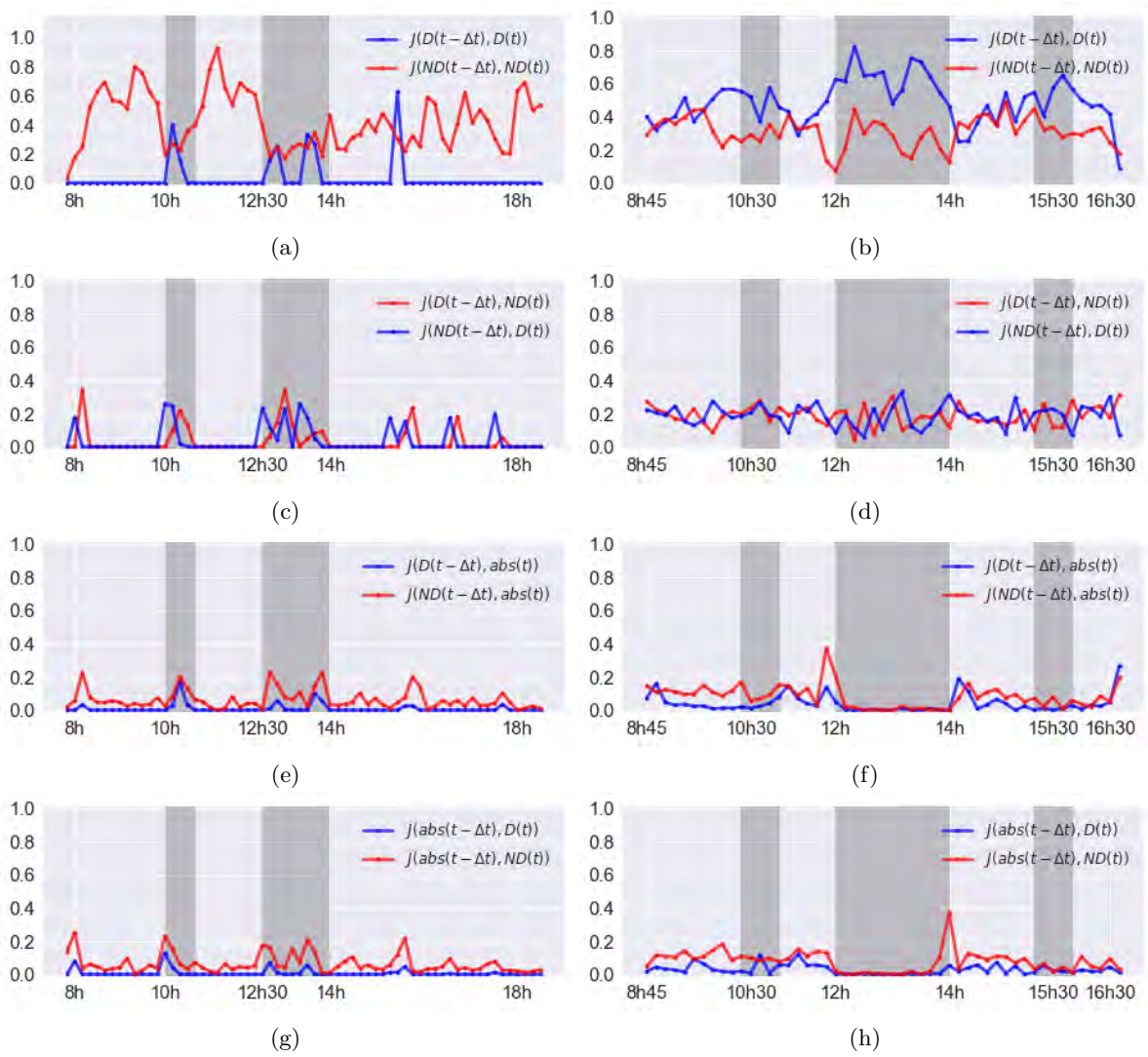


Figure 5.16: Échanges de nœuds entre la partie dense, la partie non dense et les absents, sur deux instants successifs, représentés par l'indice de Jaccard correspondant. (a),(c),(e) et (g) montrent les résultats obtenus depuis les données récoltées sur les élèves d'une classe préparatoire (données de 2011), et (b),(d),(f) et (h) les données récoltées sur les élèves de l'école primaire

Nous commençons par analyser les résultats obtenus sur les données de l'école primaire. Tout d'abord, nous voyons sur la figure 5.16b que l'amplitude de la courbe bleue devient plus importante durant les moments de pause, ce qui signifie que la partie dense y est d'autant plus persistante, car il y a une forte similitude entre $D(t - \Delta t)$ et $D(t)$. Ceci signifie que la partie dense calculée durant les heures de pause est plus stable que celle calculée durant les heures de cours. Les échanges entre la partie dense et la partie non dense représentés sur la figure 5.16d fluctuent autour d'une valeur stable de 0.2, ce qui suggère que le passage des heures de cours aux pauses n'affecte pas les échanges entre la partie dense et la partie non dense sur ces données-là.

En revanche, on peut observer deux pics remarquables de la courbe rouge (ainsi que deux pics de moindre amplitude de la courbe bleue, aux mêmes moments), le premier sur la figure 5.16f au début de la pause déjeuner indiquant qu'un certain nombre de nœuds de la partie non dense ne montrent plus de signes d'interaction (car ils passent de la partie non dense à l'ensemble des absents). Nous retrouvons ensuite un pic de la même amplitude sur la figure 5.16h qui suggère l'effet inverse : un certain nombre de nœuds qui n'interagissaient pas durant la pause déjeuner intègrent subitement la partie non dense. Ces nœuds représentent l'ensemble des élèves qui ne déjeunent pas à l'intérieur de l'établissement (il

est spécifié dans [115] que tous les élèves ne restent pas sur place pendant la pause déjeuner), et qui sont parmi les nœuds de la partie non dense (ou à moindre mesure, dans la partie dense) au moment de quitter l'école, puis à celui de leur retour.

En ce qui concerne les données générées par les élèves des classes préparatoires, nous pouvons voir que la courbe rouge sur la figure 5.16a affiche des valeurs plus importantes pendant les heures de cours que pendant les pauses (en particulier celle de 10h). Ceci veut dire que la partie non dense est moins stable durant les pauses, alors que les nœuds qui la composent sont d'une forte similitude, d'un pas de temps au suivant, le reste du temps.

L'amplitude et la position des pics de la courbe bleue (*cf.* fig. 5.16a et fig. 5.16c) montrent que la partie dense se forme en regroupant partiellement les nœuds qui étaient dans la partie non dense à l'instant précédent, ainsi que le montre le décalage d'un unique pas de temps entre les deux pics relevés à 10h sur les deux courbes bleues des figures 5.16a et 5.16c (le pic de la figure 5.16c apparaît exactement au moment où débute la pause, alors que celui de la figure 5.16a passe d'une valeur nulle à ce moment-là à sa valeur maximale l'instant d'après). Cela veut dire qu'à l'approche de la pause, une fraction de l'ensemble de la partie non dense va dans la partie dense. Une fois celle-ci formée, elle persiste en gardant à son tour une fraction de ses propres éléments d'un instant à l'instant suivant et ce tant que dure la pause. La partie dense finit par disparaître en déversant une fraction de ses nœuds dans la partie non dense, comme le montre le pic antérieur de la courbe rouge (antérieur par rapport à celui de la courbe bleue, durant la pause de 10h) sur la figure 5.16c. Bien que moins clairement identifiable, un comportement similaire à celui qu'on vient de décrire peut être observé durant la pause déjeuner.

Les figures 5.16e et 5.16g montrent quant à elles que les amplitudes des courbes rouges sont plus importantes que celles des courbes bleues au même instant. Ceci signifie que les nœuds qui n'interagissent pas à l'instant t font plus souvent la transition de l'ensemble des absents vers la partie non dense que vers la partie dense. De façon similaire, la partie non dense verse plus de nœuds dans l'ensemble des "absents" à l'instant suivant, que ne le fait la partie dense.
En combinant toutes ces observations, nous pouvons synthétiser un pattern dans lequel les nœuds passent d'abord d'un état où ils n'enregistrent aucune interaction, à la partie non dense (faible interaction) puis vers la partie dense lors des moments de fortes interactions, comme la pause de 10h.

Pour finir nous voyons des similitudes entre les résultats affichés par les deux jeux de données, comme le fait que les amplitudes de $J(abs(t-\Delta t), ND(t))$ et $J(ND(t-\Delta t), abs(t))$ soient respectivement supérieures à celles de $J(abs(t-\Delta t), D(t))$ et de $J(D(t-\Delta t), abs(t))$, pour les raisons expliquées en haut. Nous remarquons aussi une plus forte amplitude de $J(D(t-\Delta t), D(t))$ durant les périodes de forte activité (pauses), tout en enregistrant une baisse (forte sur la figure 5.16a, plus faible sur la figure 5.16b) de $J(ND(t-\Delta t), ND(t))$.
Malgré les différences de comportement qui peuvent exister entre des élèves de classes préparatoires et ceux d'une école primaire, leurs dynamiques affichent des similitudes. On peut d'un côté assimiler ces dernières à des phénomènes de densification, caractérisés par le passage des nœuds de l'ensemble des absents à la partie non dense, ou bien de la partie non dense à la partie dense [9]. D'un autre côté, nous observons aussi les phénomènes inverses, qu'on peut assimiler à de la dilution, caractérisée par les phases où un nœud passe de la partie dense à la partie non dense, ou bien de cette dernière à l'ensemble des nœuds absents.

5.5.3 Motifs d'interactions entre classes

Nous allons comme lors des analyses précédentes utiliser le découpage fourni par ItRich afin d'en illustrer la plus-value. Pour cela nous reprenons les analyses faites dans [57, 86] en calculant la densité moyenne des arêtes reliant les élèves de chacune des classes. Mais au lieu de se limiter à la moyenne temporelle calculée sur toute une journée, nous allons distinguer plusieurs régimes différents. En prenant en compte les moments de forte activité, et en prenant soin à chaque fois de séparer la partie dense et la partie non dense, et de comparer les motifs observés dans chacune.

Ainsi nous distinguons la moyenne obtenue lors des récréations (de 10h et de 15h30) de celle obtenue

[9] nous citons là des patterns fréquents, mais sans affirmer ou nier l'existence d'une quelconque corrélation entre eux, il faudrait pour cela une analyse plus poussée, ce qui n'a pas été fait ici.

pendant la pause déjeuner. Nous donnons finalement la moyenne sur l'ensemble de la journée.
Nous rappelons que la densité $\nu_t(A, B)$ de liens entre les élèves de la classe A et ceux de la classe B à l'instant t est donnée par :

$$\begin{aligned}\nu_t(A,B) &= \frac{|\{(i,j,t) \in E_t \ ; \ i \in A \text{ et } j \in A\}}{|A| \cdot (|A|-1)} \text{ si } A = B \\ &= \frac{|\{(i,j,t) \in E_t \ ; \ i \in A \text{ et } j \in B\}|}{|A| \cdot |B|} \text{ si } A \neq B\end{aligned}$$

avec $E_t = \{(i,j,t) | i \sim j \quad \text{à l'instant} \quad t\}$ l'ensemble des liens du graphe à l'instant t.
Nous calculons ensuite la moyenne temporelle de $\nu_t(A, B)$ sur une période $T = [t_1, t_2]$ par :

$$\overline{\nu_T(A,B)} = \frac{1}{T} \sum_{t=t_1}^{t_2} \nu_t(A,B)$$

Nous séparons l'ensemble des élèves suivant leurs classes respectives, et regroupons ensuite les résultats correspondant en fonction des différentes périodes sélectionnées. Nous précisons que les résultats affichés ci-dessous sont les moyennes obtenues sur une seule journée, choisie arbitrairement parmi celles sur lesquelles s'étend la durée de l'expérience. Nous commençons par montrer les résultats obtenus au sein de l'école primaire (*cf.* fig. 5.17)

Cette figure nous montre que les interactions enregistrées durant les récréations sont bien différentes de celles enregistrées pendant la pause déjeuner. En effet, sur la figure 5.17a nous pouvons voir que les interactions dominantes sont entre les élèves de mêmes classes, suggérant que ceux-ci sont peu enclins à se mélanger aux élèves des autres classes durant les pauses. Le motif observé durant la pause déjeuner (*cf.* fig. 5.17b) est en revanche différent. On distingue bien l'apparition de deux blocs, ce qui suggère que les élèves qui restent sur place pour le déjeuner sont séparés en deux grands groupes, qui interagissent fortement les uns avec les autres. Ces motifs sont observables à partir des densités de liens calculées sur l'ensemble du graphe, mais sont encore plus accentués dans le sous-graphe induit par la partie dense.
Rappelons que les auteurs de [115] avaient obtenus que les interactions enregistrées entre les élèves de la même classe étaient en moyenne trois fois plus importantes que celles enregistrées entre des élèves de classes différentes. Nous pouvons affirmer ici que ce résultat dépend du moment de la journée, ce rapport atteint en effet la valeur de 4.76 durant les récréations et 1.001 durant la pause déjeuner, et ce en prenant en compte les professeurs.

Nous faisons de même avec les données récoltées sur les élèves des classes préparatoires, dont les résultats sont affichés sur les figures 5.18 et 5.19 ci-dessous.

Nous pouvons voir sur ces deux figures des motifs différents de ceux observés sur les données générées par les élèves de primaire. En observant le motif obtenu par la moyenne de la partie dense que l'on calcule pendant la pause de 10h (*cf.* fig. 5.18a), on peut voir que la matrice affiche deux blocs distincts, le premier constitué par les élèves des classes MP^*1 et MP^*2, et le second par les classes PC, PC^* et PSI^*. Le premier bloc demeurant visible pendant la pause déjeuner mais pas le second (*cf.* fig. 5.18b). La moyenne calculée durant l'ensemble de la journée montre une forte similitude entre les résultats du graphe entier et ceux du sous-graphe induit par la partie non dense. Cela n'est pas surprenant compte tenu du fait que cette dernière est active durant toute la journée, alors que la partie dense connaît seulement quelques pics à des moments spécifiques (*cf.* fig. 5.15b).

Le motif affiché dans la partie dense de la figure 5.19a est similaire à celui observé sur la figure 5.18a, où l'on remarque l'existence de trois blocs composés par les trois classes de Biologie, les trois classes de MP et un dernier bloc composé des classes PC et PC^*. La classe PSI^* montre un faible taux d'interaction durant cette pause, ce qui nous laisse penser que leur emploi du temps est différent de celui des autres classes durant la journée choisie pour cet exemple [10].
Le motif que l'on observe dans la partie dense pendant la pause déjeuner (*cf.*fig. 5.19b) montre lui aussi une similitude avec celui observé durant la pause de 10h (*cf.*fig. 5.19a), avec un bloc dominant composé par les élèves des classes de Biologie, et un second par les trois classes MP ainsi que la classe PSI^*. Nous pouvons en conclure que les groupes qui se forment pendant la pause de 10h sont similaires à ceux

[10] On peut penser qu'un faible nombre d'élève seulement ont pu sortir prendre leur pause, ou bien que la classe PSI^* n'avait pas cours à ce moment-là, d'où le faible tôt d'interactions

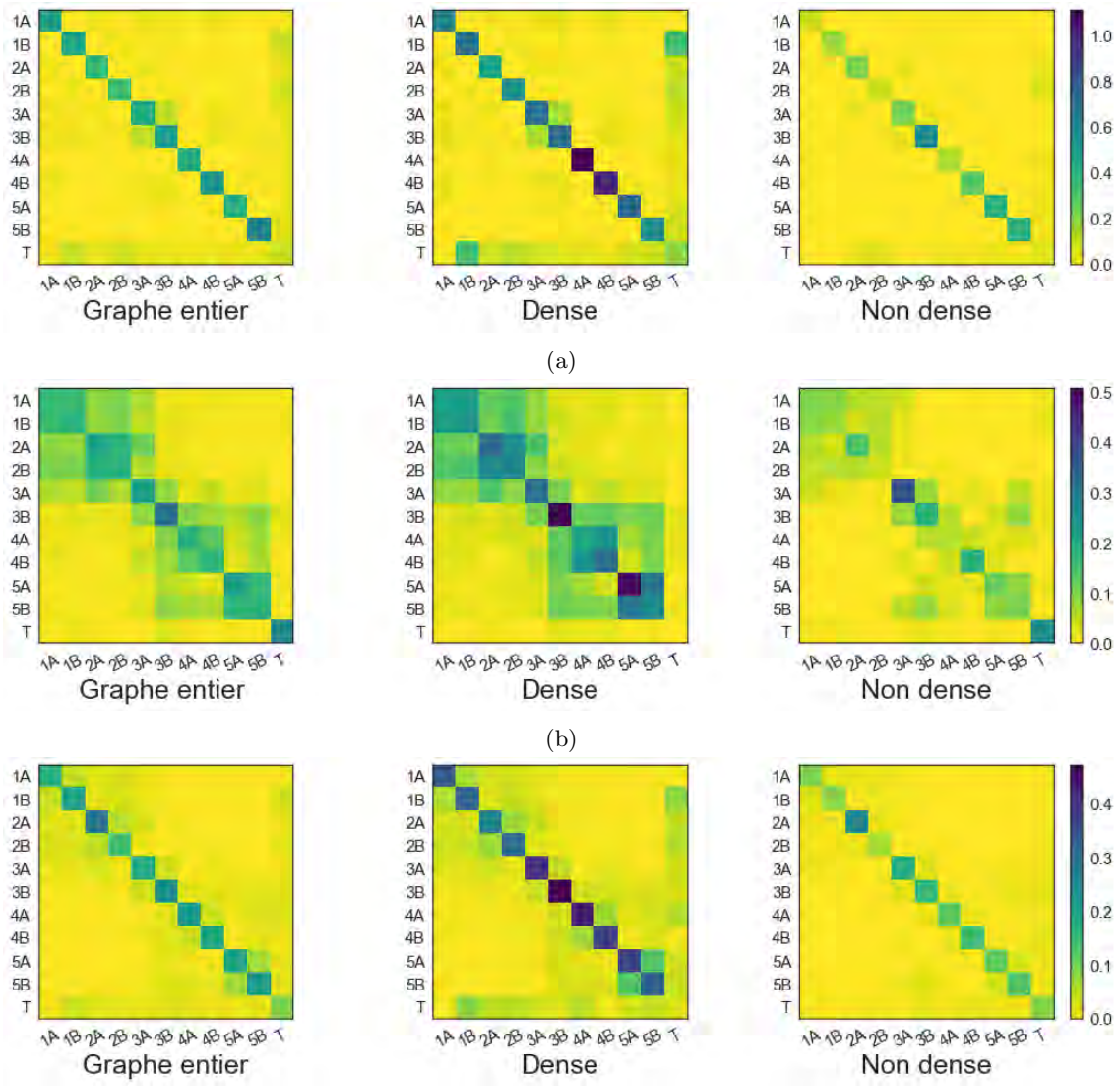
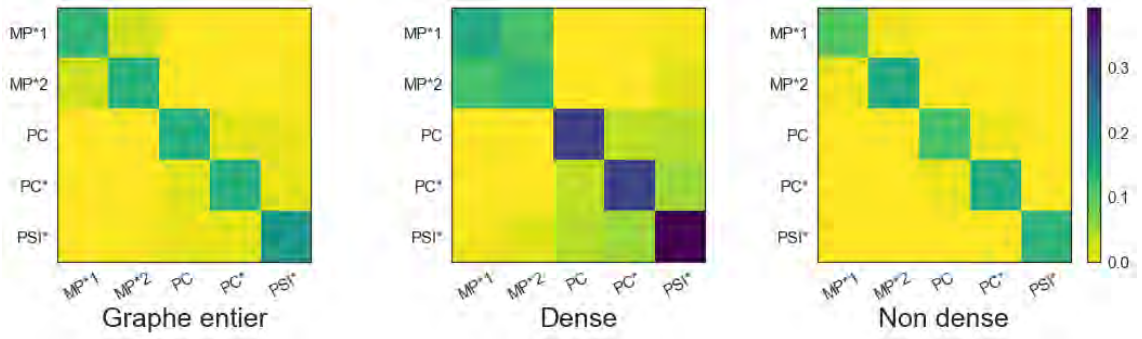
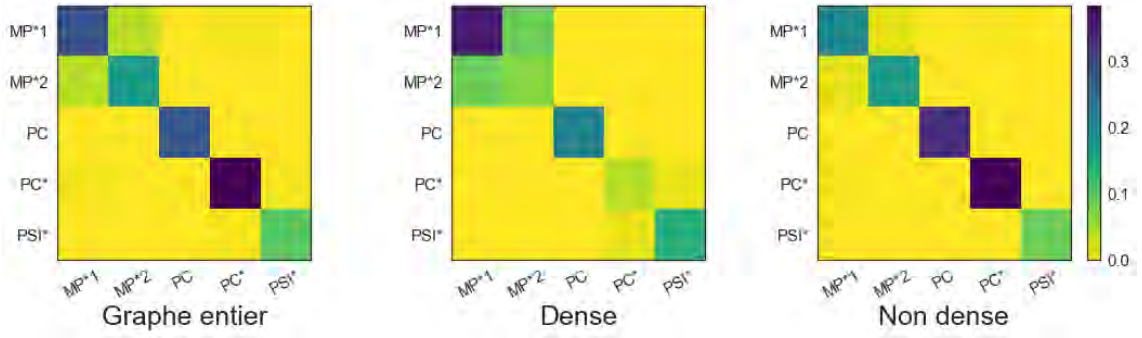


Figure 5.17: Densité moyenne de liens entre les élèves des différentes classes de l'école primaire, calculée (a) durant les récréations de 10h et de 15h30, (b) durant la pause déjeuner et (c) sur toute la durée de la journée. On donne la moyenne obtenue sur le graphe en entier, ainsi que celles obtenues sur les sous-graphes induits par la partie dense et non dense. Chaque classe est numérotée de 1 à 5 en indiquant l'année des élèves qu'elle contient, en plus du label A ou B qui indique la division de chaque année en deux classes. Le label T désigne les professeurs.

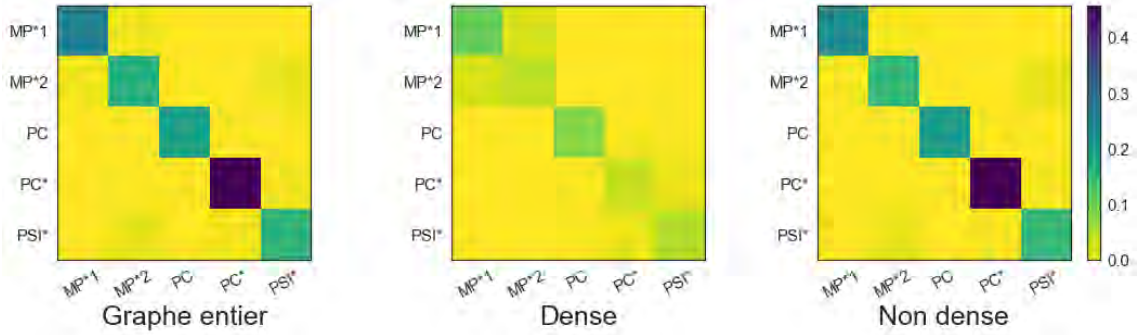
qui se forment pendant la pause déjeuner. Ce comportement est très différent de celui observé chez les élèves de l'école primaire (*cf.* fig. 5.17).
Ici aussi nous pouvons voir sur la figure 5.19c que le motif qu'on observe sur l'intégralité du graphe est similaire à celui qui est observé dans la partie non dense correspondante, mais différent de celui observé dans la partie dense. Les raisons sont les mêmes que celles citées plus haut pour la figure 5.19c, la partie dense étant moins active durant les heures de cours, la moyenne est dominée par la contribution de la partie non dense.



(a)



(b)



(c)

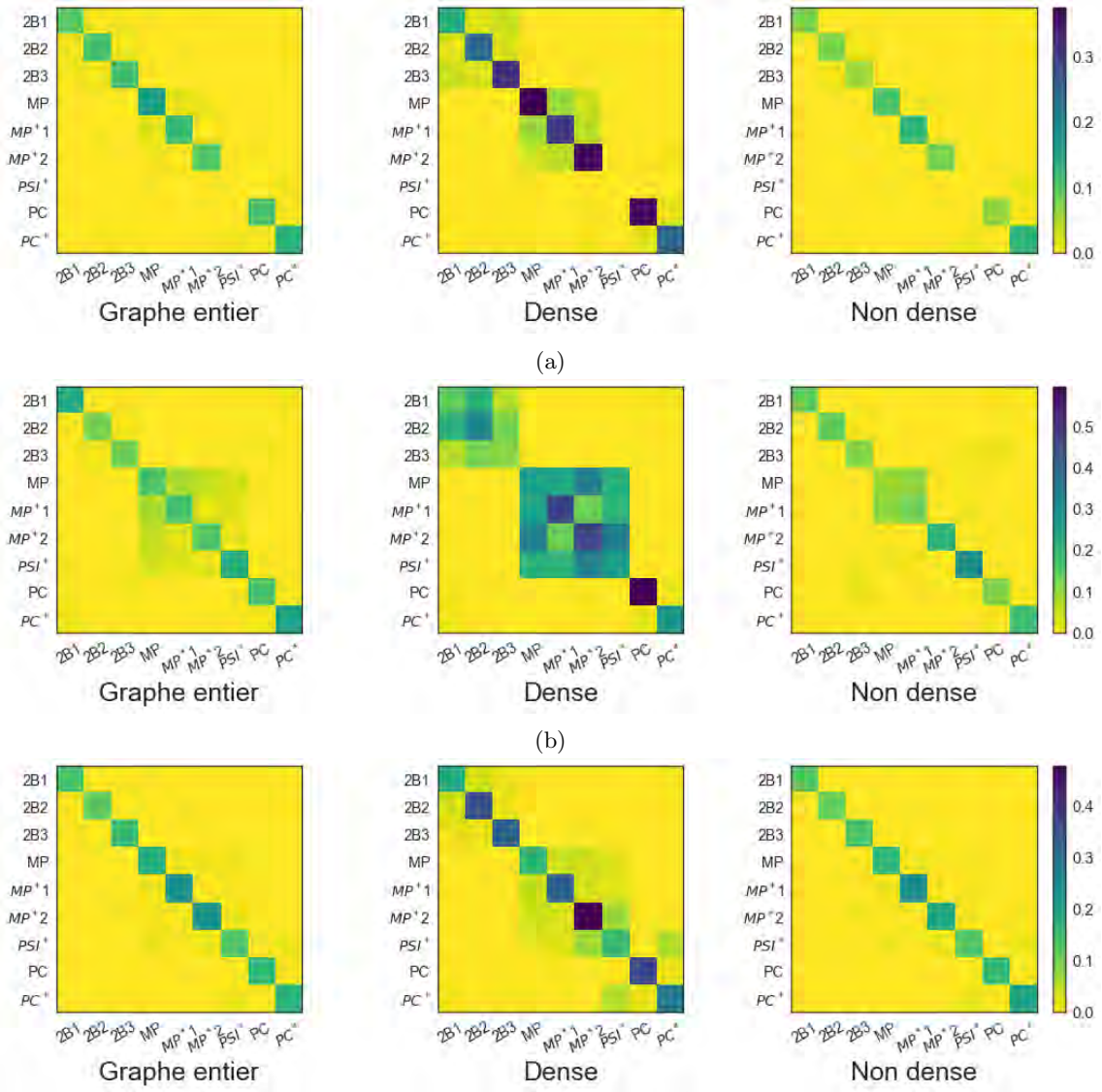
Figure 5.18: Densité moyenne de liens entre les élèves de 5 classes préparatoires, calculée (a) durant la pause de 10h , (b) durant la pause déjeuner et (c) sur toute la durée de la journée. On donne la moyenne obtenue sur le graphe en entier, ainsi que celles obtenues sur les sous-graphes induits par la partie dense et non dense. Les classes sont nommées en fonction des différentes spécialités.

5.5.4 Durée d'appartenance par individu

Pour aller plus loin, nous allons mesurer pour chaque individu sa durée d'appartenance à la partie dense et non dense [11] (cumulée sur toute la durée de l'expérience et normalisée par le nombre de jours qu'a duré l'expérience), celle-ci étant égale au nombre de pas de temps durant lesquels un nœud est classé dans l'une ou l'autre des deux parties. Nous montrons sur la figure 5.20 la durée d'appartenance à la partie dense de chaque nœud, en fonction de sa durée d'appartenance à la partie non dense.

Outre la corrélation positive plus ou moins forte que l'on remarque sur chacun des nuages de points (qui n'est pas tout à fait surprenante car elle signifie juste que plus un individu apparaît dans la partie dense, plus il est susceptible d'être apparu dans la partie non dense, ce qui se traduit par une forte tendance de cet individu à interagir avec les autres), nous pouvons voir sur la figure 5.20 que l'étendue de la durée d'appartenance à la partie non dense sur l'axe des abscisses est à peu de choses près la même

[11] On parle ici de durée approchée, car nous rappelons que les interactions qui ont lieu durant la même fenêtres temporelle sont toutes considérées comme étant équivalentes



(a)

(b)

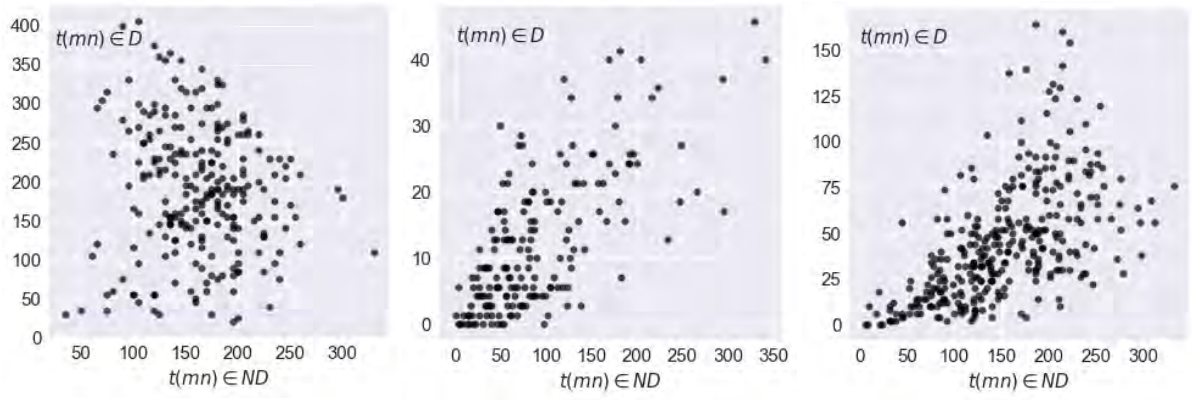
(c)

Figure 5.19: Densité moyenne de liens entre les élèves de 9 classes préparatoires, calculée (a) durant la pause de 10h , (b) durant la pause déjeuner et (c) sur toute la durée de la journée. On donne la moyenne obtenue sur le graphe en entier, ainsi que celles obtenues sur les sous-graphes induits par la partie dense et non dense. Les classes sont nommées en fonction des différentes spécialités.

sur les trois jeux de données. Elle diffère cependant d'un cas à l'autre suivant l'axe des ordonnées, ce qui démontre que les interactions formant des clusters denses sont moins importantes d'un cas à l'autre. Elles sont particulièrement élevées dans l'ensemble de données récolté au sein de l'école primaire, ce qui appuie les remarques faites plus haut : à savoir que les élèves de l'école primaire ont plus tendance à interagir que leurs homologues des classes préparatoires.

Nous pouvons aussi souligner le fait que la somme de la durée d'appartenance à la partie dense et non dense est parfois relativement faible devant les 8h d'enregistrement (560 minutes) d'une journée entière, ce qui montre bien que durant la majeure partie de la journée les élèves n'interagissent pas. Ceci est particulièrement vrai pour les élèves des classes préparatoires qui ont une charge de travail plus importante, et par conséquent moins de temps pour interagir les uns avec les autres.

Il est possible à partir de ces données de détecter les individus qui montrent des comportements singuliers. On peut par exemple extraire l'ensemble des élèves qui n'ont jamais été classés dans la partie dense durant les récréations, ou bien ceux qui montrent un taux d'interaction particulièrement faible,



(a) École primaire (b) Lycée 2011 (c) Lycée 2013

Figure 5.20: Durée moyenne d'appartenance en minutes à la partie dense (suivant les paramètres du modèle de graphe dynamique adoptés), en fonction de la durée d'appartenance à la partie non dense, le tout normalisé sur une journée, pour chacun des participants des trois data sets étudiés.

ou au contraire particulièrement élevé, et de comparer leurs profils avec une source complémentaire de données. C'est ce qui est d'ailleurs fait dans [86], où en plus de porter des badges, les élèves sont priés de communiquer (s'ils le souhaitent) la liste des individus avec lesquels ils sont amis dans la vraie vie, et/ou dans les réseaux sociaux. Mais comme cela peut prendre un certain temps, peu d'élèves répondent à ces demandes, pour ne pas trop perdre de temps, celui-ci étant moins disponible dans le contexte très compétitif des classes préparatoires. Cela a pour conséquence de rendre les données incomplètes, car une grande partie des élèves dont la spécialité est la Biologie ont répondu positivement à la demande, mais très peu parmi les classes $MP*$ ou PC. Ces données n'ont par conséquent pas été exploitées ici.

5.5.5 Durée d'appartenance par classe

Dans cette section, nous allons utiliser la durée d'appartenance de chaque individu aux parties dense et non dense, en regroupant les données montrées plus haut, afin de nous permettre de distinguer les différentes classes de chaque établissement par leur persistance au sein de chaque ensemble. Pour cela nous introduisons la fonction de répartition empirique de la durée d'appartenance, en regroupant les élèves de la même classe :

$$F^D_{>,\text{Classe}}(t) = \frac{|\{u \in \text{Classe} \quad \text{tq} \quad f^D(u) \geq t\}|}{|\text{Classe}|} \tag{5.2}$$

avec $f^D(u)$ la durée d'appartenance de l'individu u à la partie dense. La fonction de répartition pour la durée d'appartenance à la partie non dense est calculée de la même manière. Nous pouvons aussi réduire cette fonction à une certaine partie de la journée, en nous restreignant par exemple à la durée d'appartenance à la partie non dense pendant la pause de 10h, ou bien pendant la pause déjeuner. Nous rappelons que les calculs dont nous montrons ci-dessous les résultats sont les moyennes journalières qui résultent de toute la durée de l'expérience, étendue sur deux jours dans le cas de l'expérience menée au sein de l'école primaire, et une semaine pour celles des classes préparatoires [12]. Les résultats de ces mesures sont représentés sur les figures 5.21, 5.22 et 5.23.

Nous cherchons à identifier sur ces figures les classes qui se démarquent le plus parmi celles qu'inclut chaque ensemble de données. Nous précisions qu'une courbe peut se démarquer en étant significativement au dessus des autres. Ceci signifie que pendant la période considérée (récréation, pause déjeuner ou sur toute la journée), et à durée d'appartenance t_0 fixée, les individus qui sont présents dans la partie dense (ou de façon équivalente dans la partie non dense) pendant une durée au moins égale à t_0 représentent une plus forte proportion pour la classe correspondante en comparaison avec les autres classes. Dans le cas inverse, celui où la courbe se démarquerait vers le bas, les élèves de la classe correspondante représentent une plus faible proportion vérifiant les critères que l'on vient de mentionner.

[12] Les données enregistrées dans la classe préparatoire en 2011 l'ont été sur 7 jours mais nous n'en gardons que 5, évitant ainsi de compter deux lundis et deux mardis.

Ainsi, nous pouvons par exemple voir sur la gauche de la figure 5.22b qu'à durée d'appartenance égale, la proportion d'élèves de la classe MP^*1 est significativement plus élevée dans la partie dense pendant la pause déjeuner, comme indiqué par la courbe noire. La classe PC^* affiche le comportement inverse, en étant significativement en dessous des courbes qui représentent les autres classes.

Nous voyons par ailleurs sur les figures 5.23b et 5.23c que la courbe qui caractérise les élèves de la classe $2BIO3$ se démarque vers le haut dans la partie non dense (et à moindre mesure dans la partie dense) lors des pauses déjeuner et sur l'ensemble de la journée, mais ce même comportement n'est pas observé pendant les récréations où la partie non dense y est moins active. Cela signifie que les élèves de la $2BIO3$ sont impliqués de façon comparable au reste des classes dans la partie dense, mais significativement plus dans la partie non dense, ce qui confère aux étudiants de cette classe un statut particulier.

Les résultats obtenus sur les données qui proviennent de l'école primaire montrent qu'il n'y a aucune classe qui se démarque dans la partie non dense. En revanche, on apprend d'un côté que les professeurs ne restent pour la plupart pas sur place durant les pauses déjeuner, et qu'ils interagissent peu pendant les récréations, comme le montrent les courbes à gauche des figures 5.21a et 5.21b. D'un autre côté, on voit que les classes $1B$ et $4B$ se démarquent du reste concernant leurs moyennes journalières d'appartenance à la partie dense (*cf.* fig. 5.21c), la première vers le haut et la seconde vers le bas.

De cette façon, nous avons pu distinguer les classes par leurs durées d'appartenance moyennes à la partie dense ou non dense, sur une période réduite ou bien sur l'ensemble de la journée. Cependant, il reste difficile de fournir les raisons exactes de ces écarts. Celles-ci peuvent aussi bien être dues à des différences liées à l'emploi du temps de chaque classe pendant la durée de l'enregistrement des données (certaines classes peuvent par exemple avoir eu moins d'heures de cours que d'autres), qu'à des facteurs sociaux comme l'existence d'un conflit entre les élèves d'une classe donnée. Dans le cas des élèves de l'école primaire, on peut tout à fait considérer la possibilité d'avoir des méthodes d'enseignement qui diffèrent d'une classe à l'autre. On peut par exemple penser que l'enseignant d'une certaine classe favorise les échanges entre élèves, alors que c'est l'inverse pour un autre enseignant. Ces données nous étant inconnues, nous nous sommes contentés de fournir les résultats tels que nous les avons calculés.

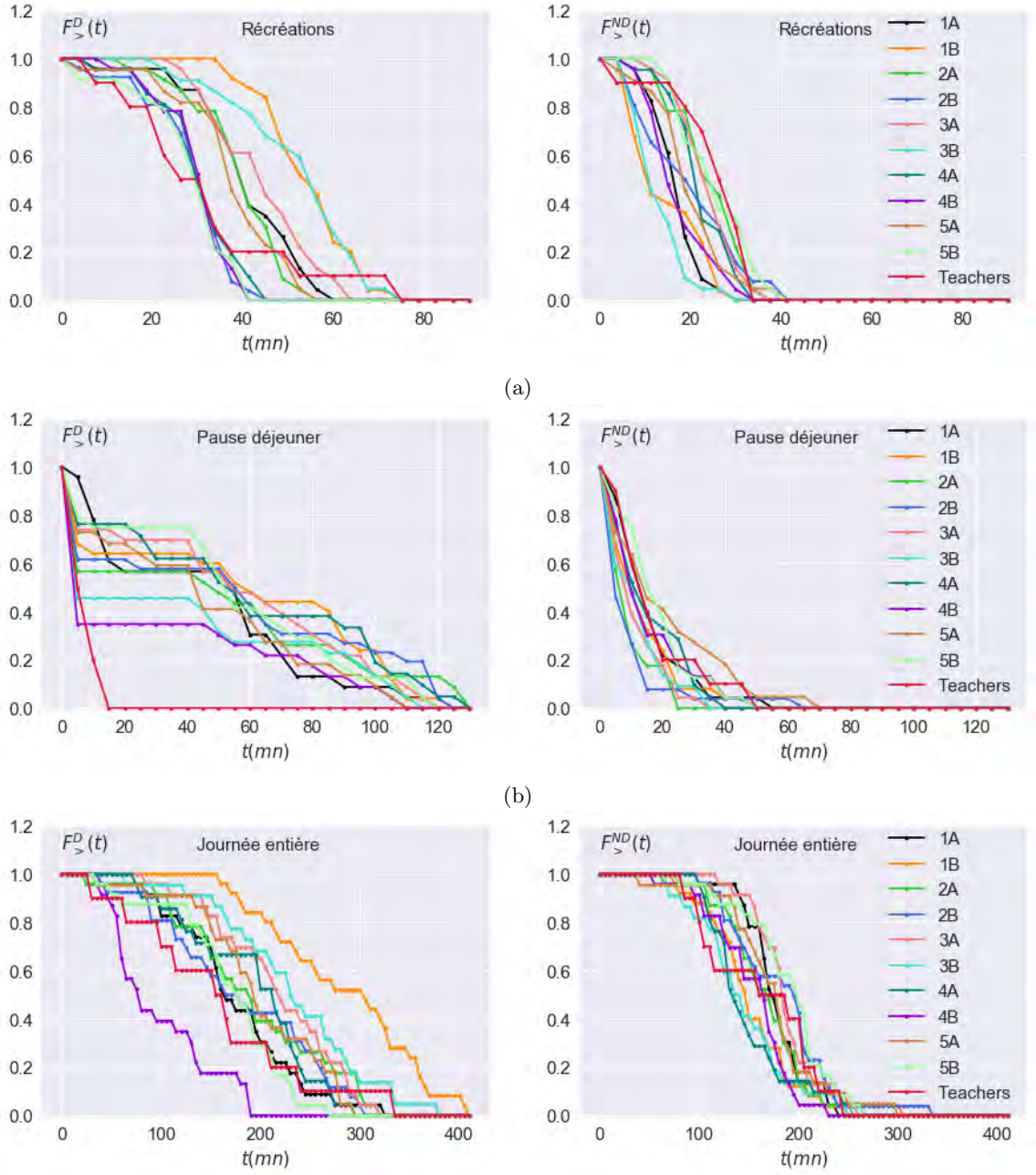
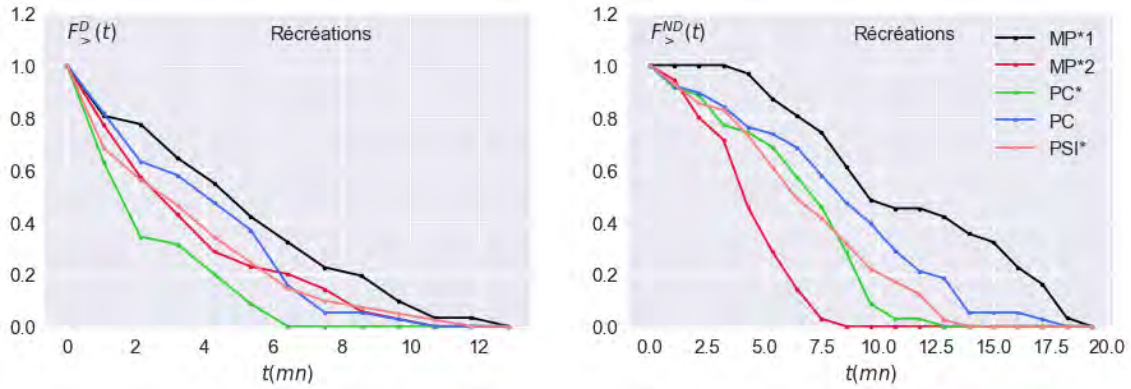
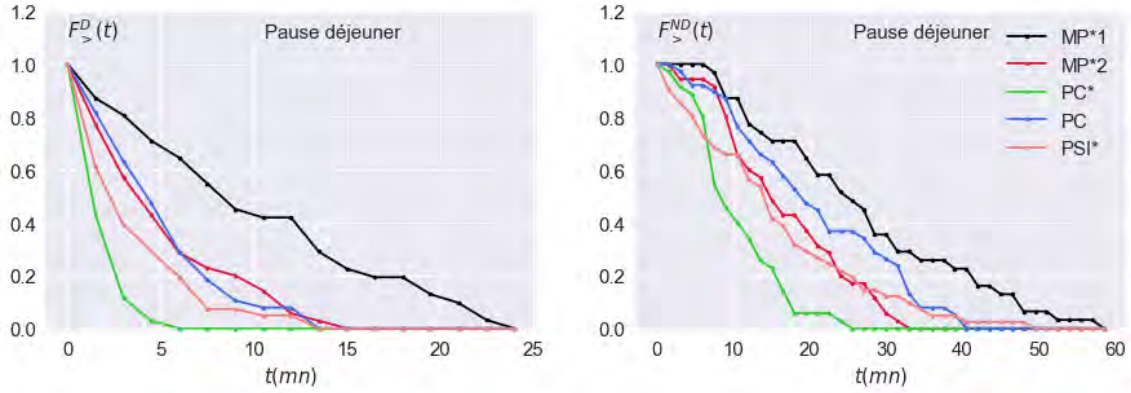


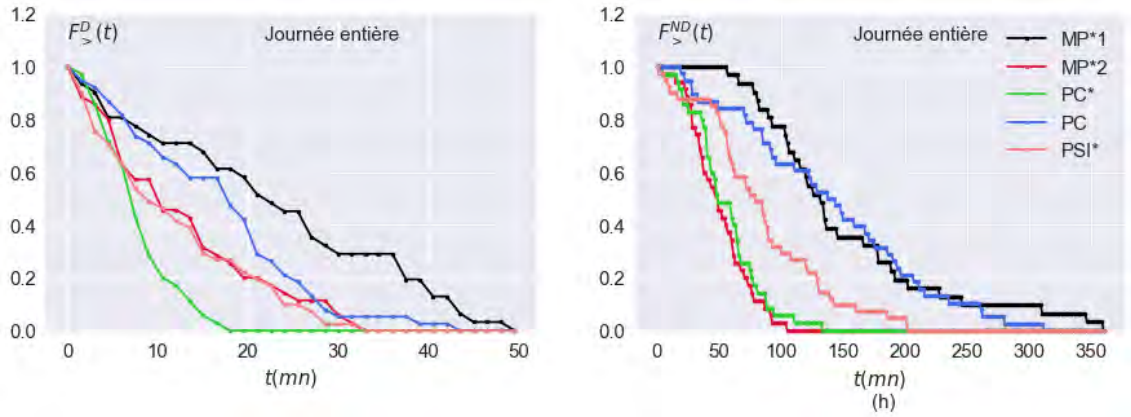
Figure 5.21: Fonction de répartition (*cf.* eq. (5.2)) de la durée d'appartenance à la partie dense (figures de gauche) et à la partie non dense (figures de droite) restreintes aux (a) récréations, (b) pauses déjeuner et (c) sur l'ensemble de la journée, pour les données récoltées sur les élèves de l'école primaire



(a)



(b)



(c)

Figure 5.22: Fonction de répartition (*cf.* eq. (5.2)) de la durée d'appartenance à la partie dense (figures de gauche) et à la partie non dense (figures de droite) restreintes aux (a) récréations, (b) pauses déjeuner et (c) sur l'ensemble de la journée, pour les données récoltées sur les élèves des classes préparatoires en 2011

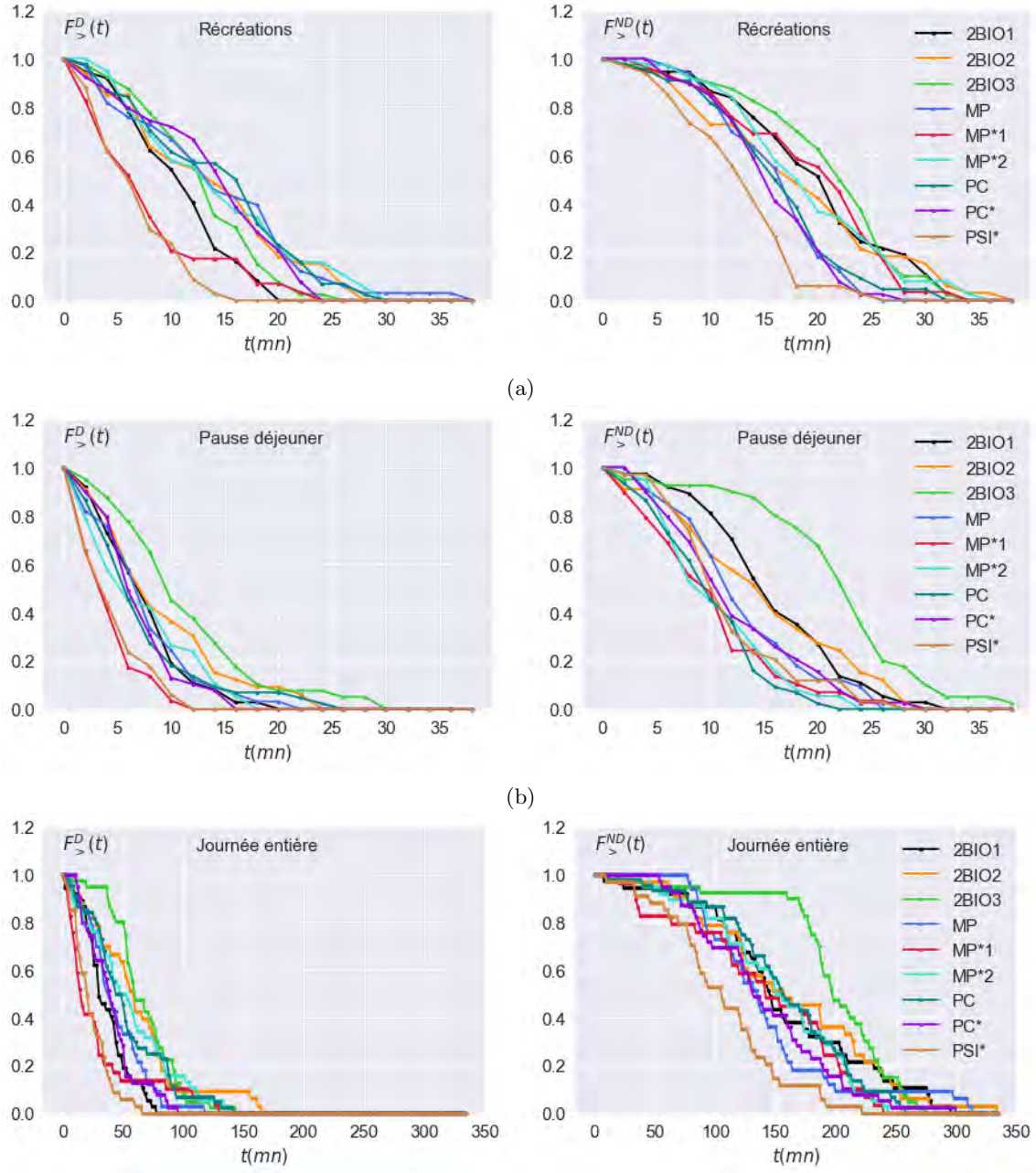


Figure 5.23: Fonction de répartition (*cf.* eq. (5.2)) de la durée d'appartenance à la partie dense (figures de gauche) et à la partie non dense (figures de droite) restreintes aux (a) récréations, (b) pauses déjeuner et (c) sur l'ensemble de la journée, pour les données récoltées sur les élèves des classes préparatoires en 2013

5.6 Bilan

Lors de ce chapitre, nous avons exploré un large éventail d'applications de l'algorithme ItRich, en analysant plusieurs ensembles de données et en insistant sur une ou plusieurs propriétés particulières suivant chaque cas. Nous avons commencé par montrer le fonctionnement détaillé de l'algorithme en portant notre attention sur les nœuds de la partie non dense, qu'on a pu séparer en deux catégories : d'un côté ceux dont la valeur de δ est faible et qui sont naturellement classés dans cette dernière, et de l'autre ceux qui y sont malgré une valeur élevée de δ. Nous avons par la suite expliqué les raisons d'une telle différentiation et avons comparé les résultats d'ItRich avec le découpage fourni par les k-cores. Ensuite, afin de montrer la différence entre ItRich et k-cores, nous avons appliqué notre algorithme à un réseau dans lequel presque tous les nœuds sont dans la même k-shell. Nous avons réussi à retrouver que la hiérarchie fournie par les différents $\delta-$*rich clubs*, ainsi que les nœuds de la partie non dense, était bien pertinente avec la réalité de terrain.

Nous avons ensuite insisté sur la deuxième catégorie de nœuds qui forment la partie non dense (ceux dont le δ est élevé), en montrant d'abord leurs propriétés topologiques, sur un réseau constitué de blogs politiques américains. Nous avons montré que ces nœuds partagent la propriété d'avoir un δ faible devant celui calculé en moyenne sur leurs voisinage respectifs. D'un autre côté nous montrons que certains de ces nœuds peuvent avoir un rôle important dans la reconnexion d'un réseau qui aurait été victime d'une large panne, et que l'on chercherait à rendre connexe à moindre coût (le coût étant le nombre de liens), à travers l'exemple du réseau mondial des transports aériens.

La troisième partie était portée sur l'association d'ItRich et d'un algorithme non paramétrique de détection de communautés (testé avec la méthode de la propagation de labels, et de l'optimisation de la modularité). Nous montrons dans cette application qu'il est possible à l'aide de cette combinaison de retrouver les différences d'intrigue majeures entre la saga littéraire et la série télévisée "Le trône de fer", que nous ne retrouvons pas en nous appuyant uniquement sur la détection de communautés.

Finalement nous avons appliqué ItRich à des données temporelles modélisées par des graphes dynamiques. Nous avons identifié les périodes de fortes activités dans la journée, que l'on a étudiées en détail. Nous avons ainsi mis en évidence un phénomène de densification qui permet le passage progressif de l'état de faible, à celui de forte densité. Nous avons ensuite regroupé les élèves par classes et avons souligné les motifs d'interaction dominants, à divers moments de la journée.
On a finalement pu distinguer chaque classe par le niveau de persistance moyen de ses élèves au sein de la partie dense et non dense, en montrant que les motifs d'interactions ainsi que la durée moyenne d'appartenance de chaque classe variaient selon les différents moments de la journée.

Conclusion

L'analyse de la structure des réseaux a connu un grand développement lors des deux dernières décennies, bien qu'un coup d'œil attentif nous montre qu'on est encore loin d'avoir atteint un niveau de compréhension nous permettant d'expliquer la diversité et la complexité des architectures des réseaux réels. Il est cependant largement convenu qu'il existe des éléments qui caractérisent un grand nombre de ces réseaux réels malgré leur diversité. Ces éléments sont caractérisés par des quantités mesurables à plusieurs échelles, dont l'organisation et la dynamique sont déterminantes pour la compréhension de la topologie des réseaux. Une approche possible pour capturer cette structure multi-échelles consiste à identifier les petits mondes qui correspondent souvent à des sous-parties denses en connexion (recouvrantes ou non) du réseau (appelées aussi communautés), et à étudier le réseau mésoscopique d'assemblage de ces communautés et sa dynamique. Il est ensuite possible d'étudier séparément chacun de ces petits mondes.

Les motivations initiales de cette thèse étaient de compléter ces outils de modélisation en prenant le contre-pied de cette approche par communautés, et en considérant que la partie non dense d'un réseau est aussi un élément structurant de son organisation, l'objectif étant de montrer que des études hybrides basées à la fois sur des parties denses et non denses permettent des avancées qualitatives et quantitatives dans la description des réseaux complexes et de leurs dynamiques. Nous avons ainsi développé une approche couvrant plusieurs échelles : l'échelle du nœud pour la mesure de la densité, l'échelle du graphe pour la dichotomie de ce dernier en deux parties, l'une dense et l'autre non dense, et finalement l'échelle intermédiaire via les sous-ensembles que comptent chacune de ces deux parties.

Contributions principales

La première contribution apportée lors de cette thèse a été de définir une nouvelle façon d'estimer la densité dans un espace métrique. Ceci fut fait à travers une mesure qui prend en compte l'état du voisinage de chaque nœud dans un nuage de points, en étendant le rayon d'exploration jusqu'à une valeur critique représentant un changement remarquable de densité. Cette mesure permet d'accentuer le contraste entre les points de faible densité et ceux de grande densité, et nous affranchit du choix d'un paramètre qui est souvent arbitraire (soit un nombre de points voisins qui fixent le rayon, ou à l'inverse un rayon fixé à l'intérieur duquel on calcule le nombre de points voisins).
En plus de cette mesure de densité, nous avons aussi revisité la fonction objectif employée pour le positionnement des nœuds dans un espace métrique, en incluant certaines des propriétés topologiques du graphe (la similitude entre les paires de nœuds). Cela nous a permis de généraliser l'expression des forces d'attraction et de répulsion qui régissent le déplacement des nœuds, dans le but d'améliorer la précision de l'algorithme de positionnement de Fruchterman et Reingold.
La deuxième contribution du début de cette thèse a été de mettre en relation la variabilité des résultats des algorithmes de positionnement et l'existence de plusieurs échelles de densité dans les nuages de points qui en résultent, à travers l'outil de l'analyse topologique des données.

Nous avons ensuite choisi de nous focaliser sur des mesures déterministes de la densité, en définissant un indice δ qui attribut un score à chaque lien du réseau, qui dépend des propriétés conjointes des deux nœuds qui sont à ses extrémités. On somme le score de chaque lien attaché à un nœud donné pour calculer son indice δ. Nous avons montré l'avantage d'utiliser cet indice au lieu d'autres indices plus classiques comme la mesure de degré.
Nous avons ensuite construit un algorithme innovant, ItRich, qui partitionne le graphe en deux parties :

la première dense et la seconde non dense. Chacune de ces deux parties est à son tour composée d'une ou plusieurs sous-parties, qui se distinguent elles aussi par des propriétés topologiques remarquables. Cet algorithme s'appuie sur une comparaison à un modèle nul, que nous avons soigneusement sélectionné parmi la multitude de modèles existants. Nous avons ensuite produit un modèle qui permet de le calculer dans un temps plus court, suivant certaines contraintes.
Cet algorithme est tout aussi applicable aux graphes statiques qu'aux graphes dynamiques. Il peut aussi être associé à d'autres algorithmes, notamment ceux de détection de communautés, pour une meilleure compréhension de la structure des réseaux représentés, comme le montrent les applications du dernier chapitre.

En résumé, ItRich est un outil supplémentaire à ajouter au large panel d'outils d'analyse de réseaux déjà disponibles. Les caractéristiques de ses résultats ont été identifiées et décrites tout au long de cette thèse, et peuvent avoir diverses interprétations qui varient suivant la diversité des données analysées. Ce n'est donc pas un algorithme qui a été développé dans le but d'analyser un ensemble particulier, et peut tout à fait s'appliquer à n'importe quel type de données représentables par un graphe.

On peut néanmoins identifier deux limites importantes à cette approche : la première réside dans sa complexité, qui est en $O(N^2)$. Ceci restreint l'application de l'algorithme à des graphes dont la taille ne doit pas être trop grande (jusqu'à quelques dizaines de milliers voire une centaine de milliers de nœuds), au risque d'avoir des temps de calcul rédhibitoires. Cette limite est d'autant plus valable quand il s'agit d'appliquer ItRich sur des graphes dynamiques, pour lesquels on applique l'algorithme sur un certain nombre de pas de temps successifs.
La deuxième limitation concerne le type de graphe que l'algorithme prend en entrée. Nous avons tout au long du manuscrit insisté sur le fait qu'ItRich ne s'intéressait qu'à la topologie des réseaux traités. Celle-ci est entièrement résumée par une représentation binaire des arêtes. Or, nous savons qu'en pratique, il est parfois difficile de dissocier le côté pondéré d'un graphe de sa topologie, et qu'il est de ce fait préférable d'associer ces deux aspects pour une meilleure représentation des données. ItRich n'a pas été testé sur des graphes pondérés, mais peut cependant être étendu pour s'adapter à cette contrainte.

Perspectives

Généralisation aux graphes pondérés

On peut penser à différentes manières de caractériser les parties denses et non denses dans un graphe pondéré. Il faut cependant définir au préalable l'aspect qui va être prioritaire dans une telle analyse. En effet, le poids topologique défini dans le troisième chapitre par la mesure ω est indépendant du poids "naturel " issu des données du graphe. Sachant cela, l'exercice qui consiste à intégrer à la fois l'information issue du jeu de données modélisé par le graphe, en plus de l'aspect topologique de ce dernier, se résume au choix d'une fonction de ces deux poids. Par exemple, dans le cas où ces deux aspects sont considérés comme étant d'importance égale, une fonction du produit de ces deux variables constitue un candidat potentiel. En pratique, il n'est pas facile de décider lequel des deux aspects (pondération issue des données *vs.* pondération topologique) est le plus important. De plus, ItRich s'appuie sur le fait que les nœuds qui constituent un sous-ensemble fortement connecté le sont à travers des arêtes qui par construction, possèdent des poids (topologiques) élevés. Cette propriété n'est plus garantie dès lors que l'on doit considérer une pondération des liens dont la valeur dépend à la fois du poids topologique et du poids "naturel", ce qui complique assurément une telle analyse.
C'est pourquoi la mise au point d'un algorithme qui généraliserait ItRich au cas des réseaux pondérés constitue une piste importante pour nos travaux futurs.

Les parties non denses comme outil de distinction des différentes topologies

La partie non dense d'un graphe peut prendre des formes très variées. Elle peut induire un sous-graphe connexe, ou à l'inverse être composée uniquement de nœuds de degrés nuls. Ces structures sont selon nous intéressantes à mettre en perspective avec la topologie du graphe dont elles sont extraites, ce qui n'a pas été fait au cours de cette thèse.

Il est cependant difficile de comparer tels qu'ils sont les sous-réseaux induits par les parties non denses de différents graphes. Nous prévoyons donc de mettre au point une transformation qui permet de rendre connexe la partie non dense, à travers, par exemple, un arbre couvrant l'ensemble des ces nœuds, tout en évitant au maximum les nœuds de la partie dense. Cet arbre serait une sorte de "squelette" du graphe, et devra être construit suivant des caractéristiques qui restent à établir. Le but est à terme de résumer l'information portée par la structure d'un graphe à celle contenue dans cet arbre, et dans les communautés denses. Il serait alors possible de proposer une classification des différentes structures à travers leurs quantités mesurables.

ItRich et détection de communautés dans un contexte dynamique

D'un autre côté, il est possible d'aller plus loin dans l'étude de la dynamique des parties denses et non denses, en combinant par exemple les résultats d'ItRich à ceux des algorithmes dynamiques de détection de communautés. Il serait alors intéressant de mettre en perspective les phénomènes de densification et de dilution identifiés par notre algorithme avec ceux plus complexes qui caractérisent l'évolution des communautés à travers le temps [103]. On bénéficierait alors d'une dynamique de comparaison qui pourrait permettre de mieux comprendre les évènements qui se produisent au niveau des communautés. Il serait par exemple intéressant de constater que les phénomènes de fusion/division, ainsi que ceux d'apparition/disparition, relevés au niveau des communautés sont les conséquences des phénomènes de densification/dilution observés à l'échelle des réseaux entiers. On pourrait ainsi plus facilement anticiper le comportement des communautés qui constituent le graphe, voire même d'ajuster les algorithmes de détection en fonction des résultats d'ItRich.

Bibliographie

[1] E. Achtert, H. P. Kriegel, and A. Zimek. "ELKI: A Software System for Evaluation of Subspace Clustering Algorithms". In: *Scientific and Statistical Database Management*. Ed. by Bertram Ludäscher and Nikos Mamoulis. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 580–585. ISBN: 978-3-540-69497-7.

[2] L. A. Adamic and N. Glance. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog". In: *Proceedings of the 3rd International Workshop on Link Discovery*. Chicago, Illinois: Association for Computing Machinery, 2005, pp. 36–43. ISBN: 1595932151.

[3] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann. "Link communities reveal multiscale complexity in networks." In: *Nature* 466 (2010), pp. 761–764.

[4] H. Akira, M. Kunioki, and D. V. Minh. "Plasma Distribution Function in a Superthermal Radiation Field". In: *Phys. Rev. Lett.* 54 (24 June 1985), pp. 2608–2610.

[5] N. Akkiraju et al. "Alpha shapes: definition and software". In: *Proceedings of the 1st International Computational Geometry Software Workshop*. 1995, pp. 63–66.

[6] R. Albert and A. L. Barabasi. "Statistical mechanics of complex networks". In: *Review of Modern Physics* 74.1 (2002), pp. 47–97.

[7] R. Albert, H. Jeong, and A. L. Barabási. "Diameter of the World-Wide Web". In: *Nature* 401 (1999), pp. 130–131.

[8] M. Ankerst et al. "OPTICS: Ordering Points to Identify the Clustering Structure". In: *SIGMOD Rec.* 28.2 (June 1999), pp. 49–60. ISSN: 0163-5808.

[9] T. Aynaud and J. L. Guillaume. "Static community detection algorithms for evolving networks." In: *WiOpt*. IEEE, 2010, pp. 513–519. ISBN: 978-1-4244-7523-0.

[10] A. L. Barabási. *Linked: The New Science of Networks*. Perseus Books Group, May 2002. ISBN: 0738206679.

[11] A. L. Barabási and R. Albert. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (1999), pp. 509–512. ISSN: 0036-8075.

[12] A. L. Barabási, R. Albert, and H. Jeong. "Scale-free characteristics of random networks: the topology of the world-wide web". In: *Physica A: Statistical Mechanics and its Applications* 281.1 (2000), pp. 69–77. ISSN: 0378-4371.

[13] A. L. Barabási, R. Erzsébet, and V. Tamás. "Deterministic scale-free networks". In: *Physica A: Statistical Mechanics and its Applications* 299.3 (2001), pp. 559–564. ISSN: 0378-4371.

[14] A. L. Barabási and M. Pósfai. *Network science*. Cambridge: Cambridge University Press, 2016. ISBN: 9781107076266.

[15] E. R. Barnes. "An Algorithm for Partitioning the Nodes of a Graph". In: *SIAM Journal on Algebraic Discrete Methods* 3.4 (1982), pp. 541–550.

[16] A. Bavelas. "Communication Patterns in Task Oriented Groups". In: *Journal Of The Acoustical Society Of America* 22.6 (Nov. 1950), pp. 725–730.

[17] J. M. Berthier and S. Semple. "Observing grooming promotes affiliation in Barbary macaques". eng. In: *Proc Biol Sci* 285.1893 (2018), pp. 20181964–20181964. ISSN: 1471-2954.

[18] V. D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[19] J. Y Boudec. *Rate adaptation, Congestion Control and Fairness: A Tutorial*. 2000.

[20] U. Brandes. "Drawing on Physical Analogies". In: *Drawing Graphs: Methods and Models - LNCS 2025*. Ed. by Michael Kaufmann and Dorothea Wagner. Springer, 2001, pp. 71–86. ISBN: 3540420622.

[21] T. F. Brittany, K. Jisu, and L. Fabrizio. *Introduction to the R package TDA*. Tech. rep. 2014.

[22] R. S. Burt. "Structural holes and good ideas". In: *American journal of sociology* 110.2 (2004), pp. 349–399.

[23] R. S. Burt. *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press, 1992.

[24] B. A. Burton. "Introducing Regina, The 3-Manifold Topology Software." In: *Experimental Mathematics* 13.3 (2004), pp. 267–272.

[25] J. Byrka et al. "An improved LP-based approximation for steiner tree." In: *STOC*. Ed. by Leonard J. Schulman. ACM, 2010, pp. 583–592. ISBN: 978-1-4503-0050-6.

[26] R. Campigotto, P. C. Céspedes, and J. L. Guillaume. "A Generalized and Adaptive Method for Community Detection." In: *CoRR* abs/1406.2518 (2014).

[27] A. Capocci et al. "Detecting communities in large networks". In: *Physica A: Statistical and Theoretical Physics* 352.2-4 (July 2005), pp. 669–676.

[28] G. Carlsson. "Topology and data". In: *Bulletin of the American Mathematical Society* 46.2 (2009), pp. 255–308.

[29] G. Carlsson et al. "Persistence Barcodes for Shapes". In: *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*. Nice, France: Association for Computing Machinery, 2004, pp. 124–135. ISBN: 3905673134.

[30] L. Cayton and S. Dasgupta. "Robust Euclidean Embedding". In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 169–176. ISBN: 1595933832.

[31] R. Cazabet. "Détection de communautés dynamiques dans des réseaux temporels. (Detection of dynamic communities in temporal networks)". PhD thesis. Paul Sabatier University, Toulouse, France, 2013. URL: `https://tel.archives-ouvertes.fr/tel-00874017`.

[32] W. Chao-Yang, W. Jian-Jun, and G. Zi-You. "Properties of Bottleneck on Complex Networks". In: *Communications in Theoretical Physics* 55.4 (Apr. 2011), pp. 725–728.

[33] C. Y. Chen and S. Y. Hsieh. "An efficient approximation algorithm for the Steiner tree problem". In: *Complexity and Approximation*. Springer, 2020, pp. 238–251.

[34] V. Colizza et al. "Detecting rich-club ordering in complex networks". In: *Nature Physics* 2.2 (Feb. 2006), pp. 110–115. ISSN: 1745-2481.

[35] L. D. F. Costa, F. A. Rodrigues, and A. S. Cristino. "Complex networks: the key to systems biology". In: *Genetics and Molecular Biology* 31 (2008), pp. 591–601. ISSN: 1415-4757.

[36] P. Courrieu. "Straight monotonic embedding of data sets in Euclidean spaces". In: *Neural Networks* 15.10 (2002), pp. 1185–1196. ISSN: 0893-6080.

[37] J. De Las Rivas and C. Fontanillo. "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks". In: *PLoS Comput Biol* 6.6 (June 2010), pp. 1–8. ISSN: 1553-7358.

[38] B. Delaunay. "Sur la sphere vide". In: *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* 7 (1934), pp. 793–800.

[39] L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (July 1945), pp. 297–302.

[40] L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (July 1945), pp. 297–302.

[41] R. Ding. "The Complex Network Theory-Based Urban Land-Use and Transport Interaction Studies." In: *Complexity* (2019).

[42] R. Ding et al. “Application of Complex Networks Theory in Urban Traffic Network Researches”. In: *Networks and Spatial Economics* 19.4 (2019), pp. 1281–1317. ISSN: 1572-9427.

[43] Y. Ding. “Applying weighted PageRank to author citation networks”. In: *Journal of the American Society for Information Science and Technology* 62.2 (2011), pp. 236–245.

[44] L. R. Duncan and A. D. Perry. “A method of matrix analysis of group structure”. In: *Prychometrika* (Oct. 1949).

[45] P. Eades. “A heuristic for graph drawing”. In: *Congressus Numerantium* 42 (1984), pp. 149–160.

[46] H. Ebel, L. I. Mielsch, and S. Bornholdt. “Scale-free topology of e-mail networks”. In: *Phys. Rev. E* 66 (3 Sept. 2002), p. 035103.

[47] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel. “On the shape of a set of points in the plane.” In: *IEEE Trans. Inf. Theory* 29.4 (1983), pp. 551–558.

[48] P. Erdös and A. Rényi. “On Random Graphs I”. In: *Publicationes Mathematicae (Debrecen)* 6 (1959), pp. 290–297.

[49] P. Erdős and A. Rényi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1 (1960), pp. 17–60.

[50] M. Ester et al. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. 1996, pp. 226–231.

[51] E. Estrada. “Network robustness to targeted attacks. The interplay of expansibility and degree distribution”. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 52.4 (2006), pp. 563–574. ISSN: 1434-6036.

[52] E. Estrada. “Topological structural classes of complex networks”. In: *Phys. Rev. E* 75 (1 Jan. 2007), p. 016103.

[53] B. T. Fasy et al. “Introduction to the R package TDA”. In: *ArXiv* abs/1411.1830 (2014).

[54] B. Fichet and G. Le Calve. “Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence”. In: *Statistique et analyse des données* 9.3 (1984), pp. 11–44.

[55] D. Firmani et al. “Computing Strong Articulation Points and Strong Bridges in Large Scale Graphs.” In: *SEA*. Ed. by Ralf Klasing. Vol. 7276. Lecture Notes in Computer Science. Springer, 2012, pp. 195–207. ISBN: 978-3-642-30849-9.

[56] S. Fortunato. *Community detection in graphs*. cite arxiv:0906.0612Comment: Review article. 103 pages, 42 figures, 2 tables. Two sections expanded + minor modifications. Three figures + one table + references added. Final version published in Physics Reports. 2009. DOI: 10.1016/j.physrep.2009.11.002. URL: http://arxiv.org/abs/0906.0612.

[57] J. Fournet and A. Barrat. “Contact patterns among high school students”. In: *PLOS ONE* 9.9 (2014), e107878.

[58] L. C. Freeman. “A Set of Measures of Centrality Based on Betweenness”. In: *Sociometry* 40.1 (Mar. 1977), pp. 35–41.

[59] T. M. J. Fruchterman and E. M. Reingold. “Graph Drawing by Force-directed Placement”. In: *Software - Practice and Experience* 21.11 (1991), pp. 1129–1164.

[60] Robert G. “Barcodes: The persistent topology of data”. In: *Bull. Amer. Math. Soc* (2007).

[61] N. Gaumont, M. Panahi, and D. Chavalarias. “Reconstruction of the socio-semantic dynamics of political activist Twitter networks—Method and application to the 2017 French presidential election”. In: *PLOS ONE* 13.9 (Sept. 2018), pp. 1–38.

[62] M. Girvan and M. E. J. Newman. “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences* 99.12 (June 2002), pp. 7821–7826. ISSN: 1091-6490.

[63] A. Globerson et al. “Euclidean Embedding of Co-occurrence Data”. In: *J. Mach. Learn. Res.* 8 (2007), pp. 2265–2295. ISSN: 1533-7928.

[64] S. Havlin et al. “Diffusion with a topological bias on random structures with a power-law distribution of dangling ends”. In: *Phys. Rev. A* 34 (4 Oct. 1986), pp. 3492–3495.

[65] P. W. Holland, K. B. Laskey, and S. Leinhardt. “Stochastic blockmodels: First steps”. In: *Social Networks* 5.2 (1983), pp. 109–137.

[66] P. Holme and B. J. Kim. “Growing scale-free networks with tunable clustering”. In: *Phys. Rev. E* 65 (2 Jan. 2002), p. 026107.

[67] F. K. Hwang and D. S. Richards. “Steiner tree problems.” In: *Networks* 22.1 (1992), pp. 55–89.

[68] P. Jaccard. “Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines”. In: *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (1901), pp. 241–272.

[69] P. Jensen. *Pourquoi la société ne se laisse pas mettre en équations*. Science ouverte. Editions du Seuil, 2018. ISBN: 9782021380118.

[70] P. Jensen et al. “Detecting global bridges in networks”. In: *Journal of Complex Networks* 4.3 (2016), pp. 319–329.

[71] S. Jung, S. Kim, and B. Kahng. “Geometric fractal growth model for scale-free networks”. In: *Phys. Rev. E* 65 (5 Apr. 2002), p. 056101.

[72] L. Katz. “A new status index derived from sociometric analysis”. In: *Psychometrika* 18.1 (Mar. 1953), pp. 39–43. ISSN: 1860-0980.

[73] B. P. Kent, A. Rinaldo, and T. Verstynen. *DeBaCl: A Python Package for Interactive DEnsity-BAsed CLustering*. 2013.

[74] B.W. Kernighan and S. Lin. “An Efficient Heuristic Procedure for Partitioning Graphs”. In: *The Bell Systems Technical Journal* 49.2 (1970).

[75] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by simulated annealing”. In: *science* 220.4598 (1983), pp. 671–680.

[76] R. Kraft. *Illustrations of Data Analysis Using the Mapper Algorithm and Persistent Homology*. 2016.

[77] A. Lancichinetti, S. Fortunato, and F. Radicchi. “Benchmark graphs for testing community detection algorithms”. In: *Physical Review E* 78 (2008), p. 046110.

[78] A. Lancichinetti et al. “Finding Statistically Significant Communities in Networks”. In: *PLOS ONE* 6.4 (Apr. 2011), pp. 1–18.

[79] J. Leskovec and A. Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. `http://snap.stanford.edu/data`. June 2014.

[80] L. Li et al. “Identification of type 2 diabetes subgroups through topological analysis of patient similarity”. In: *Science Translational Medicine* 7.311 (2015), 311ra174–311ra174. ISSN: 1946-6234.

[81] B. Liu et al. “Quantifying the Effects of Topology and Weight for Link Prediction in Weighted Complex Networks”. In: *Entropy* 20 (2018), p. 363.

[82] D. Luca and A. M. Miguel. “Detecting network communities: a new systematic and efficient algorithm”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2004.10 (Oct. 2004), P10012.

[83] D. Lusseau and M. E. J. Newman. “Identifying the role that animals play in their social networks”. In: *Proceedings of the Royal Society B: Biological Sciences* 271 (2004), S477–S481.

[84] D. Lusseau et al. “The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations”. In: *Behavioral Ecology and Sociobiology* 54 (Sept. 2003), pp. 396–405.

[85] D. M. MacKay. “Psychophysics of Perceived Intensity: A Theoretical Basis for Fechner’s and Stevens’ Laws”. In: *Science* 139.3560 (1963), pp. 1213–1216. ISSN: 0036-8075.

[86] R. Mastrandrea, J. Fournet, and A. Barrat. “Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys”. In: *PLOS ONE* 10.9 (Sept. 2015), pp. 1–26.

[87] P. Meo et al. “Generalized Louvain method for community detection in large networks.” In: *ISDA*. Ed. by Sebastián Ventura et al. IEEE, 2011, pp. 88–93. ISBN: 978-1-4577-1676-8.

[88] S. Milgram. “The Small World Problem”. In: *Psychology Today* 67.1 (1967), pp. 61–67.

[89] M. E. J Newman. "Assortative Mixing in Networks". In: *Physical Review Letters* 89.20 (Oct. 2002), p. 208701. ISSN: 0031-9007.

[90] M. E. J. Newman. "Detecting community structure in networks". In: *European Physical Journal* B 38 (2004), pp. 321–330.

[91] M. E. J. Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8577–8582.

[92] M. E. J. Newman. *Networks: an introduction*. Oxford; New York: Oxford University Press, 2010. ISBN: 9780199206650.

[93] M. H. Olyaee, A. Yaghoubi, and M. Yaghoobi. "Predicting protein structural classes based on complex networks and recurrence analysis". In: *Journal of Theoretical Biology* 404 (2016), pp. 375–382. ISSN: 0022-5193.

[94] T. Opsahl et al. "Prominence and Control: The Weighted Rich-Club Effect". In: *Phys. Rev. Lett.* 101 (2008), p. 168702.

[95] G. Palla et al. "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* 435.7043 (June 2005), pp. 814–818. ISSN: 0028-0836.

[96] V. Pareto. *Cours d'Economie Politique*. Genève: Droz, 1896.

[97] E. Parzen. "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 1065–1076. ISSN: 00034851.

[98] K. Pearson. "Note on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58 (1895), pp. 240–242.

[99] P. Pons and M. Latapy. "Computing communities in large networks using random walks". In: *International Symposium on Computer and Information Sciences*. Springer. 2005, pp. 284–293.

[100] U. N. Raghavan, R. Albert, and S. Kumara. *Near linear time algorithm to detect community structures in large-scale networks*. Sept. 2007. arXiv: 0709.2938. URL: http://arxiv.org/abs/0709.2938.

[101] A. Rodriguez and A. Laio. "Clustering by fast search and find of density peaks". In: *Science* 344.6191 (2014), pp. 1492–1496. ISSN: 0036-8075.

[102] C.A. Ronan. *Histoire mondiale des sciences*. Collection Points: Série Sciences. Ed. du Seuil, 1999. ISBN: 9782020362375.

[103] G. Rossetti and R. Cazabet. "Community Discovery in Dynamic Networks: A Survey". In: *ACM Comput. Surv.* 51.2 (2018). ISSN: 0360-0300.

[104] R. A. Rossi and K. A. Nesreen. "The Network Data Repository with Interactive Graph Analytics and Visualization". In: *AAAI*. 2015.

[105] M. Rosvall, D. Axelsson, and C. T. Bergstrom. "The map equation". In: *The European Physical Journal Special Topics* 178.1 (2009), pp. 13–23. ISSN: 1951-6355.

[106] G. L. Schuster, O. Dubovik, and B. N. Holben. "Angstrom exponent and bimodal aerosol size distributions". In: *Journal of Geophysical Research: Atmospheres* 111 (2006).

[107] S. B. Seidman. "Network structure and minimum degree". In: *Social Networks* 5.3 (1983), pp. 269–287. ISSN: 0378-8733.

[108] D. Serrano-Hernandez, J. Serrano-Hernandez, and D. Sanchez Gomez. "Simplicial degree in complex networks. Applications of topological data analysis to network science". In: *Chaos, Solitons Fractals* 137 (2020), p. 109839. ISSN: 0960-0779.

[109] M. Serrano, M. B. Ná, and R. P. Satorras. "Correlations in weighted networks". In: *Physical Review E* 74 (2006), p. 055101.

[110] C. E. Shannon. "A mathematical theory of communication". In: *SIGMOBILE Mob. Comput. Commun. Rev.* 5.1 (2001), pp. 3–55. ISSN: 1559-1662.

[111] P. C. de Simon et al. "Deciphering the global organization of clustering in real complex networks." In: *Scientific reports* 3 (2013), p. 2517.

[112] A. Singhal. "Modern Information Retrieval: A Brief Overview." In: *IEEE Data Eng. Bull.* 24.4 (2001), pp. 35–43.

[113] A. Solé-Ribalta, S. Gómez, and A. Arenas. "A model to identify urban traffic congestion hotspots in complex networks". In: *Royal Society Open Science* 3.10 (2016), p. 160098.

[114] T.J. Sørensen. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. Biologiske skrifter. I kommission hos E. Munksgaard, 1948.

[115] J. Stehlé et al. "High-resolution measurements of face-to-face contact patterns in a primary school". In: *PLOS ONE* 6.8 (2011), e23176.

[116] P. B. Stephen and G. E. Martin. "Models of core/periphery structures". In: *Social Networks* 21.4 (2000), pp. 375–395. ISSN: 0378-8733.

[117] P. R. Suaris and G. Kedem. "An algorithm for quadrisection and its application to standard cell placement". In: *IEEE Transactions on Circuits and Systems* 35.3 (1988), pp. 294–303.

[118] H. Takayasu, A. Provata, and M. Takayasu. "Stability and relaxation of power-law distribution". In: *Phys. Rev. A* 42 (12 Dec. 1990), pp. 7087–7090.

[119] J. Travers and S. Milgram. "An Experimental Study of the Small World Problem". In: *Sociometry* 32.4 (1969), pp. 425–443. ISSN: 00380431.

[120] L. Vietoris. "Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen". In: *Mathematische Annalen* 97 (1927), pp. 454–472.

[121] J. Wang et al. "Identification of Essential Proteins Based on Edge Clustering Coefficient." In: *IEEE/ACM Trans. Comput. Biology Bioinform.* 9.4 (2012), pp. 1070–1080.

[122] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684 (June 1998), pp. 440–442.

[123] H. Yu et al. "The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics". In: *PLOS Computational Biology* 3.4 (Apr. 2007), pp. 1–8.

[124] S. Zhou and R. J. Mondragón. "The rich-club phenomenon in the Internet topology." In: *IEEE Communications Letters* 8.3 (2004), pp. 180–182.

[125] V. Zlatic et al. "On the rich-club effect in dense and weighted networks". In: *The European Physical Journal B* 67 (2009), pp. 271–275.

[126] A. J. Zomorodian. *Topology for Computing*. USA: Cambridge University Press, 2009. ISBN: 0521136091.

[127] A. J. Zomorodian and G. Carlsson. "Computing Persistent Homology". In: *Proceedings of the Twentieth Annual Symposium on Computational Geometry*. Brooklyn, New York, USA: Association for Computing Machinery, 2004, pp. 347–356. ISBN: 1581138857.