UPMC
PARISUNIVERSITAS

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité: Sciences Physiques
ED 389: Physique de la particule à la matière condensée

Présentée par

Lionel TABOURIER

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Méthode de comparaison des topologies de graphes complexes. Applications aux réseaux sociaux.

soutenue le 5 Juillet 2010 devant le jury composé de:

M. Hugues CHATÉ, ingénieur, SPEC (CEA) ..... directeur de thèse
M. Paul BOURGINE, ingénieur général, CRÉA (École Polytechnique) ..... président
M. Éric FLEURY, professeur, LIP (ENS Lyon) ..... rapporteur
M. Pablo JENSEN, directeur de recherche, Laboratoire de Physique (ENS Lyon) ... rapporteur
M. Matthieu LATAPY, directeur de recherche, LIP6 (UPMC) ..... examinateur
M. Camille ROTH, chargé de recherche, CAMS (EHESS) ..... examinateur
Mme. Annick LESNE, chargée de recherche, LPTMC (UPMC) ..... invitée

Thèse préparée au SPEC (CEA Saclay), Bâtiment 772, Orme des Merisiers, F-91191 Gif-sur-Yvette

Résumé de la thèse

Les graphes des réseaux d'interactions sociales révèlent des propriétés topologiques dont nous cherchons à comprendre l'origine. Un des obstacles rencontrés tient au fait que nous manquons de modèles de référence qui permettraient de construire une échelle de comparaison de leurs caractéristiques géométriques. Cette thèse propose une méthode générique pour produire des graphes synthétiques dont les propriétés sont ajustables, dans l'ambition de réaliser un balisage de l'espace des graphes.

Nous procédons d'abord à une revue des définitions et outils classiques propres à l'étude des graphes, ainsi que ceux plus spécifiques aux données sociologiques que nous nous proposons d'étudier; cet exposé nous permet également de détailler le cadre d'hypothèses dans lequel ce travail s'inscrit. Puis, après une présentation des méthodes usuelles pour générer des graphes aléatoires, nous proposons une évolution de l'une d'entre elles vers une procédure algorithmique expérimentale polyvalente, dont nous expliquons le fondement et les limites. Enfin nous rendons compte de quelques applications pratiques dans divers domaines: réseaux technologiques, de collaborations, d'échanges commerciaux.

La procédure proposée dérive de méthodes à base de chaînes de Markov, dont l'étape élémentaire est un échange entre les extrêmités de deux liens, nous la généralisons en effectuant des échanges mettant en jeu k liens. Selon les contraintes que l'on souhaite imposer, une telle procédure doit être adaptée; nous discutons alors des difficultés inhérentes à sa réalisation pratique, l'étendue de son champ d'application, et les moyens à notre disposition pour estimer sa validité.

Le principe général mis en œuvre dans les illustrations que nous proposons consiste à construire une suite d'ensembles de graphes inclus les uns de les autres, obéissant à des contraintes de plus en plus exigeantes; puis à comparer les propriétés de chacun de ces ensembles aux données réelles, afin de déterminer quels éléments topologiques ont effectivement un rôle essentiel. Au fil de ces exemples, nous proposons des améliorations techniques et évolutions possibles de nos algorithmes qui permettraient d'élargir les possibilités de leurs utilisations.

Cette méthode serait suffisamment générale pour pouvoir d'une part décrire des réseaux d'interactions d'une autre nature, mais également pour intégrer des informations supplémentaires à la description graphique. En particulier, nous souhaiterions qualifier les nœuds à l'aide d'informations telles que des contenus sémantiques ou l'activité dans le temps - ce second point permettrait d'adapter la procédure à la génération de graphes dynamiques; nous proposons pour conclure quelques éléments de réflexion pour réaliser ces objectifs.

Mots-clefs: graphes complexes, matrices aléatoires, chaîne de Markov, réseaux sociaux, contraintes topologiques, méthode d'échange.

Abstract

Comparison method of complex graphs topologies applied to social networks analysis.

We are looking for the origin of the topological properties exhibited by social interaction graphs. To this end we are faced with a lack of benchmarks that could be compared with the geometrical features of real networks. This thesis aims at providing a method to produce artificial graphs whose properties are adaptable in order to draw a baseline of the graph space.

We first report the classical definitions and tools that can be used in graph analysis, and those specific to the sociological data we are interested in ; this review enables us to specify the hypothesis background of our study. After expounding usual methods to produce random graphs, we put forward an enhancement of one of them which leads to a versatile experimental algorithmic process, then we mark off its limitations and theoretical grounding. Lastly we report some practical applications to various fields: technological, collaborative or commercial networks.

The process derive from Markov chain methods whose elementary step would be the switch of the ends of two edges, we generalize them to swaps defined for k edges. Such a process should be adapted to the constraints we want to set, we thus discuss the difficulties inherent to its practical realization, the scope of its uses and the means we have to assess its validity.

The principle implemented in the instances we propose rests on the building of a sequence of sets complying with increasingly demanding constraints. Then we compare the properties of each ensemble to the real data in order to ascertain which topological features are essential to the graph global topology. As we proceed with the examples we propose technical improvements of our algorithms that could extend the scope of their applications.

This method should be general enough not only to describe other kinds of complex networks but also to aggregate extrinsic information to the graphical description. In particular, we would like to describe the nodes with information such as semantic contents or temporal activity - this second point should enable us to adapt this process to dynamical graph production ; as a conclusion we suggest a few ideas to carry out these purposes.

Keywords: complex graphs, random matrices, Markov chain, social networks, topological constraints, switching method.

Remerciements

Rétrospectivement, cette thèse n'a plus grand'chose de commun avec l'idée que je pouvais en avoir il y a quatre ans, que ce soit pour son contenu ou pour son déroulement. Du plan initial, j'espère avoir conservé les soucis de méthode et d'exhaustivité qui me tiennent à cœur dans la recherche; en revanche je n'avais pas imaginé à quel point d'autres y apporteraient leur contribution, car aujourd'hui elle ne m'apparaît plus seulement comme un projet individuel mais le résultat du travail, des idées ou simplement de la présence de beaucoup que je tiens ici à remercier.

À commencer par mes directeurs de thèse: Hugues Chaté et Annick Lesne qui m'ont accordé leur confiance et leur aide chaque fois que j'en ai eu besoin, tout comme Paul Bourgine qui m'a amené vers ces thématiques. Merci également à Éric Fleury et Pablo Jensen d'avoir accepté de rapporter ma thèse; ainsi qu'à Matthieu Latapy pour sa participation au jury et ses conseils.

Parmi ceux qui ont joué un rôle important durant ces quatre années, je tiens spécialement à remercier Camille Roth et Jean-Philippe Cointet pour leurs réflexions toujours intéressantes - scientifiques ou non - et leur patience à supporter mes questions et mes doutes; ainsi que Fernando Peruani, à qui je dois beaucoup de temps, d'énergie et d'encouragements.

Merci également à Christian Borghesi pour me pousser à aller chercher plus loin, dans ce projet et en-dehors et à ceux avec qui j'ai partagé quotidiennement les hauts et les bas de la condition de thésard: Diana Garcia, Kazumasa Takeuchi, Carla Taramasco, Benoît Lombardot et Damien Gredat.

La thèse m'a aussi permis de m'ouvrir un peu plus à d'autres horizons scientifiques au cours de divers projets et j'ai particulièrement aimé pouvoir travailler en ces occasions avec Fanny Henriet, Stéphane Hallegatte, Vincent Viguié, Alina Stoica et Christophe Prieur.

Que ce soit au SPEC, à l'ISC-PIF ou au CREA, les membres aussi bien scientifiques qu'administratifs des laboratoires entre lesquels j'ai gravité au cours de cette thèse ont aussi participé, pour un peu ou pour beaucoup, à son bon déroulement. J'en remercie Ivan, Robert, Roger, Marc, Adel, Francesco, Emmanuel, Masa, Raphaël, Thierry, Jean-Baptiste, Telmo, David, René, Pierre, Daniel, Romain, Marie-Jo, Nadiège, Geneviève, Noemi, Marcel et sans doute encore quelques autres qui m'excuseront de l'oubli.

Je remercie également mes amis et tous ceux qui ont contribué sur un plan plus personnel que scientifique à faire de cette thèse ce qu'elle est. Enfin, merci à mes parents et toute ma famille sur qui j'ai toujours pu compter.

Table des matières

Introduction

Nous assistons depuis la fin des années 90, à une explosion d'activité autour de la représentation et la modélisation de grandes bases de données d'interactions. Celle-ci n'est pas limitée à un domaine scientifique restreint mais s'étend à tous les systèmes qui peuvent être décrits comme une population d'agents associés entre eux par une certaine forme de relation; depuis les réactions métaboliques entre des molécules de la cellule jusqu'à l'organisation des sociétés humaines.

Le phénomène est d'autant plus remarquable que les acteurs de ce mouvement ne sont pas seulement les communautés traditionnellement intéressées à ces sujets, mais également des domaines du savoir faisant appel à un formalisme plus mathématisé comme l'informatique ou la physique statistique. On peut se demander pourquoi ces systèmes deviennent des objets d'études pour des disciplines dont ils semblent sortir du champ d'expertise et surtout, ce que ces dernières peuvent apporter aux représentations et aux catégories d'analyse existantes.

L'élément nouveau dans ce paysage tient à un faisceau de révolutions techniques: l'utilisation généralisée des technologies de l'information et de la communication a permis la création de grandes bases de données enregistrant une large variété de traces des activités humaines. Et grâce à l'Internet, l'archivage systématisé s'accompagne d'une bien meilleure accessibilité de ces inventaires. Enfin, les performances des ordinateurs actuels autorisent à manipuler des volumes d'information supérieurs de plusieurs ordres de grandeur à ce qui était faisable il y a seulement 10 ans.

Il s'agit donc avant tout d'un changement d'échelle. Aussi variées que soient ces informations, leur abondance suscite une interrogation commune à toutes: comment en extraire de la connaissance? Cela impose d'adapter voire d'inventer nos outils: il n'est pas possible de soumettre chaque unité constitutive de l'information à une analyse humaine. Il devient alors nécessaire de ne conserver qu'une forme simplifiée d'un signal très complexe, et donc d'accepter de perdre une partie du contenu pour permettre un traitement automatisé.

Ce point de vue implique l'accès à un autre ordre de connaissance: on renonce à une description microscopique qui nous donnerait du comportement des agents une description détaillée; on lui substitue une image globale, moyennée sur une population.

C'est pourquoi cette nouvelle génération d'études met en avant les structures formelles qui peuvent décrire ces données plutôt que la nature de ce qu'elles représentent.

Cette évolution semble se construire essentiellement par la pratique: en transposant ponctuellement des méthodes, en observant des régularités dans les données, en extrapolant le vocabulaire existant. On définit des concepts, on teste des outils sans savoir *a priori* à quel point ils s'avéreront informatifs; et au travers de cette approche quasi-expérimentale, il se constitue un domaine du savoir à part entière. L'unité qui se dégage peu à peu de cet ensemble encore mal défini est regroupée sous la dénomination des réseaux complexes.

Nous considérerons un mode de description où l'on fait le choix de ne modéliser que l'existence des relations mises en jeu. Il est dans ce cas possible de représenter un réseau par un graphe, c'est-à-dire un ensemble de noeuds (les agents du réseau) associés entre eux par des liens (les interactions). L'extrême simplicité du modèle autorise une grande polyvalence. Il pourrait alors être tentant d'examiner toute la zoologie des réseaux de manière transversale, car les problèmes spécifiques posés par chacune des disciplines présentent des attraits particuliers.

Mais nous essayons de conserver - jusqu'à un certain point - une unité dans les thèmes abordés. En effet, si les outils considérés se veulent polyvalents, lorsque l'analyse est poussée au-delà de la ressemblance superficielle entre les diverses familles de graphes, elle demande d'ajuster les méthodes et l'interprétation à l'objet de l'étude. Les réseaux que nous avons choisi de considérer dans ce travail sont de nature sociologique. Et, sans prétendre à une analyse experte dans ce domaine, les questions d'interprétation que nous souleverons auront donc trait à cette discipline.

Le principal enjeu dans ce contexte est la description des mécanismes qui dirigent la constitution et le fonctionnement du réseau. Il s'agit d'une part de déterminer s'il existe des règles d'interactions entre les agents dont on trouverait la trace dans la topologie du graphe, et d'autre part de comprendre comment la structure du réseau conditionne les processus qui s'y produisent, tels que la transmission de l'information. En modélisant les deux phénomènes, il serait possible d'aboutir à des prévisions sur l'évolution dynamique de ces grands réseaux d'interaction.

Cette thèse cherche à contribuer à la réalisation d'outils méthodologiques normalisés, dont l'objectif est de produire des graphes artificiels. En comparant leur structure à celle des graphes réels, on pourrait identifier les caractéristiques essentielles pour comprendre l'assemblage et le fonctionnement des réseaux sociaux. À plus long terme, le but serait d'évaluer à quel point la représentation en graphe est informative sur ces mécanismes; c'est-à-dire non pas vraiment de répondre à la question "qu'est-ce que ces données nous apprennent?", mais plutôt d'évaluer ce que les données *peuvent* nous apprendre.

Dans la première partie nous présentons le vocabulaire à notre disposition pour définir le système et ses caractéristiques. Nous mêlons à cette exposition nécessaire quelques utilisations qui en sont faites dans la littérature pour l'analyse des réseaux réels. Notre intention est ainsi de montrer vers quel type de description converge l'usage pratique de ces outils. Ainsi, nous dégagerons peu à peu quelles informations nous pouvons y rechercher, et cela nous amènera à énoncer explicitement les hypothèses sur notre cadre de travail. Nous constatons alors qu'une standardisation méthodologique du domaine passerait par la localisation des caractéristiques du réseau réel vis-à-vis de références qui sont encore manquantes; c'est pourquoi nous cherchons à définir des points de repère dans l'espace des graphes, qui seront des graphes aléatoires dont on peut ajuster les caractéristiques.

L'aspect technique de ce projet fait l'objet de la seconde partie: après un bref tour d'horizon des algorithmes usuels de génération de graphes, mettant en évidence leurs intérêts mais aussi leurs limites, nous décrivons la contribution principale de cette thèse consistant en une méthode de synthèse qui autorise une grande liberté sur le choix des contraintes imposées.

Nous la mettons ensuite en pratique dans le contexte de réseaux d'interactions sociales. Nous verrons alors que la méthode peut remplir l'objectif que nous lui destinons: elle permet de produire avec une certaine confiance des graphes satisfaisant des contraintes flexibles, et donc d'observer comment celles-ci conditionnent les propriétés géométriques du graphe. On peut ainsi rechercher de manière systématique un ensemble d'éléments topologiques suffisants pour justifier les autres caractéristiques du réseau réel. Puis, à l'aide d'une évolution simple des algorithmes, nous pourrons tester comment les processus qui se déroulent sur le réseau sont affectés par des modifications de la structure.

Enfin la méthode s'avère aussi utilisable comme instrument pour sonder l'environnement dans les espaces de graphes contenant le réseau réel, et nous essaierons en conclusion de mettre en lumière comment il serait possible d'en tirer partie.

Chapitre 1

Description de données d'interaction

Sommaire

Ce chapitre est consacré à l'état de l'art sur la description des réseaux d'interaction: nous y présentons les définitions générales de théorie des graphes correspondant au lexique courant du domaine et les utilisations qui en ont été faites au cours des dernières années pour en tirer une information statistique sur les bases de données d'interaction.

Nous exposons le vocabulaire élémentaire nécessaire à la description de l'objet graphe et ses diverses déclinaisons (orientés, pondérés etc.). Puis nous définissons les observables dont nous ferons usage par la suite, qui sont extraites du corpus de mesures traditionnellement employées pour la description des réseaux complexes. Nous présenterons l'information qui est tirée de ces mesures dans la littérature et discuterons leur intérêt, compte-tenu d'une part de leur pertinence et d'autre part du coût algorithmique qui leur est associé. Enfin nous examinerons plus spécifiquement des descriptions et des mesures adaptées aux données sociologiques sur lesquelles porte ce travail.

1.1 Familles de graphes

Nous définissons tout d'abord les différents types de graphes dans le but de lever d'éventuelles ambiguïtés sur l'utilisation et les notations des termes qui ne font parfois pas consensus. Nous utilisons autant que possible le vocabulaire conventionnel de la théorie des graphes tel qu'on le trouve dans l'ouvrage de référence de Berge [Ber70], ou en donnant les équivalents plus usités dans la littérature des graphes de réseaux complexes, ainsi que leur traduction anglaise usuelle.

1.1.1 Graphes

En théorie, le terme **graphe** G se réfère à ce que nous définirons par la suite comme "graphe orienté"; cependant par commodité, nous qualifierons de graphe (ou graphe monoparti) l'objet mathématique constitué à partir d'un ensemble $V = \{x_1, ..., x_n\}$ de sommets, pouvant être associés deux-à-deux par des arêtes: $E = \{u_1, ..., u_m\}$ et $u_i \in V \times V$, et nous autorisons dans cette définition l'existence d'une arête unique entre deux sommets. On notera alors: $G = \{V, E\}$.

Par la suite, nous utiliserons davantage les termes de **nœud** *(node)* et de **lien** *(link)* plutôt que sommet et arête. La **densité** du graphe désigne le rapport du nombre de liens $L(= |E|)$ au nombre de nœuds $N(= |V|)$.

Deux nœuds liés sont dits **voisins** *(neighbours)* l'un de l'autre. On qualifie de **chemin** *(path)* entre deux nœuds, une séquence de liens consécutifs dont ils sont les extrêmités, la **longueur** de ce chemin sera le nombre de liens qu'il comporte; la **distance** entre deux nœuds sera le minimum des longueurs de chemins allant de l'un à l'autre.

Si un graphe ne contient pas de liens associant un nœud à lui-même (**boucle** - *self-loop*), on parle de **graphe simple** *(simple graph)*, mais nous nous dispenserons de cette précision quand il n'y a pas d'ambiguïté sur le caractère simple ou non du graphe.

Nous utiliserons abondamment la notion de **degré** *(degree)* d'un nœud (δ), il s'agit du nombre d'extrêmités de liens partant de celui-ci. Remarquons que pour un graphe simple, le degré s'assimile au nombre de voisins et le degré moyen ($\overline{\delta}$) à la densité du graphe.

La **matrice d'adjacence** *(adjacency matrix)* **A** est une représentation matricielle exactement équivalente au graphe. Cette matrice ($N \times N$) est binaire, $m_{ab} = 1$ si il existe un lien entre les nœuds x_a et x_b, $m_{ab} = 0$ dans le cas contraire. On notera qu'elle est nécessairement symétrique selon notre définition d'un graphe, et pour un graphe simple ses éléments diagonaux sont nuls.

De plus, nous pouvons dès maintenant faire l'observation que lorsqu'il s'agira de considérer des graphes de plus grande taille, les matrices d'adjacences associées seront souvent **éparses**: le degré moyen est très petit devant le nombre de nœuds du graphe ($\overline{\delta} \ll N$).

Pour illustrer ces définitions, imaginons un groupe de sept individus dont on cherche à décrire les interactions et les possibles échanges d'information. Le graphe simple ci-dessous - et la matrice d'adjacence qui lui est associée - pourrait représenter l'existence de discussions entre les membres du groupe sur une période donnée, ce qui se prête bien à un modèle de graphe non-orienté puisque l'échange d'information peut être réciproque.

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

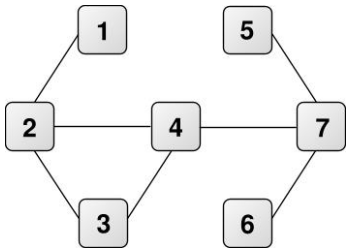


FIG. 1.1: Exemple de graphe simple, non-orienté et matrice d'adjacence associée.

1.1.2 Graphes orientés

Dans un **graphe orienté** *(directed graph* ou *digraph)* un lien est dirigé d'un nœud vers un autre, on parle alors d'**arc** *(arc)* d'une **source** vers une **destination** (ou extrêmités initale et terminale). La notion de degré sera alors généralisée en **degré**

sortant *(outdegree)* et **degré entrant** *(indegree)*. Pour définir un chemin orienté (ou **chaîne**) entre les nœuds x_a et x_b, il faut pouvoir construire une suite d'arcs de la forme: $\{(x_a, x_1); (x_1, x_2); ...; (x_{k-1}, x_k); (x_k, x_b)\}$.

Dans le contexte des graphes orientés, on définira également une matrice d'adjacence, les lignes correspondront aux nœuds sources et les colonnes aux destinations. Contrairement au cas précédent cette matrice n'est pas symétrique.

Reprenant le groupe d'individus de l'exemple précédent, le graphe simple et orienté suivant pourrait représenter l'existence d'échanges par e-mails entre ses membres, pour lesquels on identifie clairement un émetteur et un récepteur de l'information.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

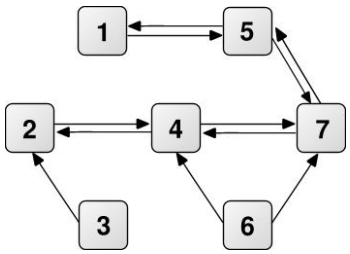


FIG. 1.2: Exemple de graphe simple, orienté et matrice d'adjacence associée.

1.1.3 Graphes pondérés

Le multigraphe est une extension de la notion de graphe autorisant l'existence de liens multiples entre les nœuds. Nous utiliserons plutôt le terme de **graphes pondérés** *(weighted graphs)*. Il est bien sûr aussi possible de construire des graphes pondérés et orientés. On trouve parfois une distinction entre multigraphe et pseudographe, le second pouvant comprendre des boucles, mais le terme multigraphe étant souvent utilisé en ce sens, nous ne ferons donc pas la nuance.

La matrice d'adjacence est toujours définie mais elle n'est plus binaire: la valeur de m_{ab} étant le nombre de liens (la **multiplicité** ou **poids**) existant entre les nœuds x_a et x_b. Selon cette définition $m_{ab} \in \mathbb{N}$, on peut généraliser la notion de graphe pondéré à des cas où $m_{ab} \in \mathbb{R}$ si l'on souhaite quantifier une relation autrement que par le nombre d'interactions entre agents.

Dans notre exemple, on pourrait vouloir quantifier les échanges d'information; il existe alors de nombreuses mesures imaginables: le nombre d'e-mails échangés, leur tailles (en caractères, en octets), ou encore une estimation de leur contenu informatif par les membres du groupe. On pourrait obtenir par exemple:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 3 & 1 & 0 & 0 \end{pmatrix}$$

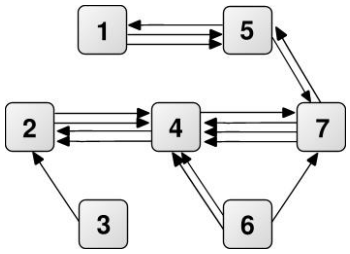


FIG. 1.3: Exemple de graphe pondéré, orienté et matrice d'adjacence associée.

1.1.4 Hypergraphes et graphes multipartis

L'**hypergraphe** *(hypergraph)* est une généralisation de la notion de graphe où l'on substitue au lien - qui peut être considéré comme un sous-ensemble à deux éléments de l'ensemble des nœuds - un hyperlien: un sous-ensemble comportant un nombre quelconque de nœuds.

L'hypergraphe pourra être représenté par un nouveau type de matrice binaire qu'on appelle **matrice d'affiliation** *(affiliation matrix)*. On fait figurer dans les colonnes les nœuds de l'hypergraphe et en ligne les hyperliens (ou l'inverse), $m_{ij} = 1$ signifiant que l'hyperlien y_i contient le nœud x_j.

Dans le cadre de notre exemple, nous représentons les réunions auxquelles participent certains membres du groupe, ce seront alors les hyperliens qui associeront les acteurs.

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

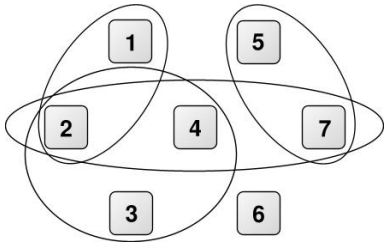


FIG. 1.4: Exemple d'hypergraphe et matrice d'affiliation associée.

Un **graphe biparti** *(bipartite graph)* est un graphe non-orienté pour lequel on peut identifier une partition des nœuds en deux sous-ensembles tels qu'il n'existe pas de lien entre des nœuds du même sous-ensemble. On l'utilise pour modéliser des réseaux dans lesquels on identifie deux types distincts de nœuds (typiquement des agents et des événements).

Il existe une bijection entre la classe des graphes bipartis et celle des hypergraphes.

En effet, soient $X = \{x_1, ..., x_n\}$ et $Y = \{y_1, ..., y_p\}$ les deux familles de nœuds du graphe biparti, l'application qui à chaque y_i associe le sous-ensemble X_i des nœuds de X qui sont connectés à y_i est bijective et définit l'ensemble des hyperliens de l'hypergraphe $\{\{x_i\}; \{X_i\}\}$.

C'est pourquoi le contenu en terme d'information d'un hypergraphe est strictement identique à celui d'un graphe biparti. En revanche, selon le contexte, on préférera l'une ou l'autre représentation qui ne favorisent pas les mêmes intuitions.

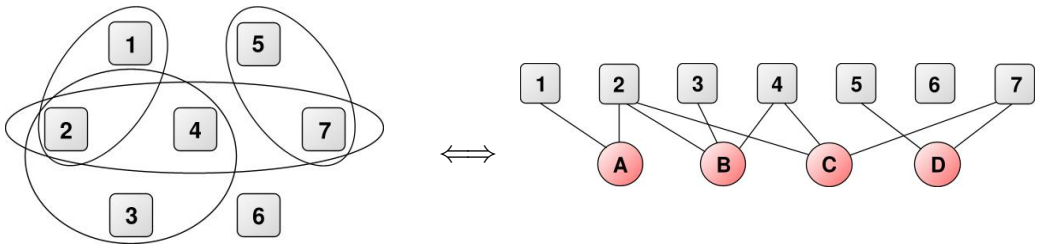


FIG. 1.5: Exemple d'hypergraphe et graphe biparti équivalent.

remarque : Il est également possible de construire une injection de l'ensemble des graphes orientés vers celui des graphes bipartis précédemment défini. En effet, on peut faire correspondre aux nœuds $\{x_i\}$ du graphe orienté deux ensembles de nœuds $\{y_i\}$ et $\{z_i\}$ d'un graphe biparti tel que y_a serait connecté à z_b si et seulement si il existe un arc de x_a vers x_b.

1.1.5 Notion de projection

Nous qualifions de **graphe projeté** *(projected graph)* le graphe monoparti obtenu à partir d'un hypergraphe - ou d'un graphe biparti - en liant entre eux les nœuds appartenant au moins une fois au même hyperlien. Et nous parlerons de **graphe projeté pondéré** *(weighted projected graph)* pour désigner le graphe monoparti pondéré obtenu à partir d'un hypergraphe en créant un lien pour chaque hyperlien commun aux deux nœuds.

Ainsi, depuis l'hypergraphe résumant la participation des acteurs du groupe aux réunions, on peut établir les projections biparties pondérée et non-pondéree, qui conservent partiellement l'information relative aux interactions entre individus:

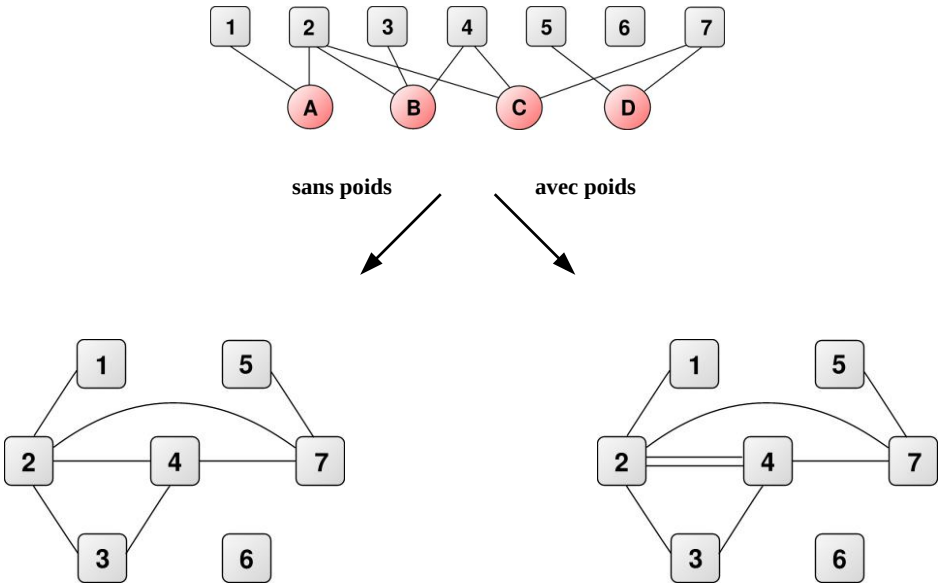


FIG. 1.6: Projections non-pondérée et pondérée d'un graphe biparti.

1.2 Caractérisation de la topologie d'un graphe complexe

Nous souhaitons tirer des informations de la structure des graphes, et dans ce but, nous disposons d'une grande étendue de mesures possibles, mais pas d'un ensemble d'observables simples d'usage qui permette de saisir systématiquement les particularités de la topologie. Alors comment définir les mesures qui sont pertinentes pour un graphe de réseau complexe?

Nous décrivons dans cette section les critères - essentiellement pratiques - qui guident le choix des mesures employées. Puis nous en présentons certaines, dont nous ferons usage par la suite, et l'information qui en est extraite sur quelques exemples issus de la littérature classique du domaine. Cela permettra de donner une idée de la variété d'applications possibles aux outils que nous allons manipuler.

Dans la majorité des exemples qui étayent cette partie, on comprend intuitivement ce que représente le graphe et nous n'en détaillerons pas l'interprétation. Néanmoins, pour avoir des précisions sur les réseaux auxquels nous faisons allusion ici, on pourra se reporter à l'Annexe A. Les bases de données qui y sont décrites sont spécifiées par un astérisque * dans le texte.

1.2.1 Quels choix d'observables?

Le premier critère pratique de sélection d'une mesure est bien sûr son **caractère informatif**, c'est-à-dire à quel point celle-ci nous renseigne sur la structure et le fonction-

nement du graphe. Cela suppose aussi que son interprétation soit relativement simple, ce qui explique que la plupart de ces grandeurs soient des scalaires ou des distributions.

Par ailleurs, les succès récents de l'analyse des réseaux d'interactions tiennent en partie au fait qu'elle ne repose pas sur des mesures très spécialisée de phénomènes précis, mais au contraire sur une approche générique à base d'outils polyvalents, qui permettent de traiter transversalement tous types de données: sociales, biologiques, linguistiques... On recherche donc des mesures **flexibles**, qu'il ne soit pas nécessaire d'adapter au contexte.

Pour saisir le plus d'aspects possibles de la géométrie du graphe, nous allons chercher à le décrire par un faisceau de mesures **complémentaires**. En particulier, pour avoir une image de la structure à diverses échelles, nous utiliserons des mesures locales et d'autres globales. La distinction entre ces deux échelles est arbitraire car, comme nous le verrons, deux nœuds du graphe sont rarement très distants l'un de l'autre; mais on comprend bien qu'une mesure qui qualifie le comportement d'un nœud vis-à-vis de ses voisins puisse être considérée locale par rapport à une autre qui nécessiterait l'examen de tous les nœuds du graphe.

En pratique, la capacité à pouvoir réaliser de manière systématique une mesure sur de grands graphes est aussi un critère essentiel pour le choix des observables. Autant que possible on essaye d'employer des mesures **peu coûteuses**. Or, la réalisation d'une mesure qui caractérise le graphe à grande échelle nécessite - sauf exception - de se déplacer dans une grande partie de la base de données. Ces mesures globales ont donc souvent un coût algorithmique élevé, c'est pourquoi celles-ci peuvent être inadaptées à la réalisation de statistiques sur des bases de données de grandes tailles. Mais même des mesures plus locales peuvent devenir problématiques lorsque nous travaillons sur des graphes comptant des milliers de nœuds et que nous cherchons à effectuer ces mesures en série sur des échantillons de graphes. L'emploi d'une mesure sera donc conditionné par l'objet et la fréquence à laquelle elle serait utilisée.

1.2.2 La structure des données

Comme le critère du coût algorithmique joue un rôle important dans le choix des obsrvables employées, il est nécessaire de détailler le type de structure des données que nous pouvons utiliser pour décrire le graphe, car il conditionne la complexité des différentes mesures.

a. Type matrice

Il s'agit de stocker les données au format de la matrice d'adjacence (éventuellement d'affiliation), donc sous la forme d'un tableau à deux entrées dont l'élément d'indice (i, j) correspond à l'élément m_{ij} de cette matrice. Dans ce format, les liens du graphe sont accessibles en temps constant; en contrepartie, la taille des tableaux étant fixe, on occupe un espace mémoire proportionnel à N^2.

b. Type tableau de liste

Selon ce second point de vue, le graphe est stocké sous la forme d'un tableau dont la ligne indexée i contient la liste des voisins du nœud i. Ce format est adapté aux graphes épars comme ceux que nous étudions, le degré moyen $\overline{\delta}$ des nœuds étant petit devant N. En effet il nécessite seulement un espace mémoire en $N\overline{\delta}$; en revanche, la recherche d'un voisin particulier du nœud i requiert le parcours de la liste et se fera donc en moyenne en $\overline{\delta}$.

Selon les cas, il sera préférable d'employer l'une ou l'autre solution pour améliorer les performances des algorithmes. D'autres types plus élaborés permettent un compromis entre les deux précédents - par exemple un tableau de liste dont la taille de chaque liste est connue - mais nous évaluerons la complexité des algorithmes en référence à ces deux types simples et standards.

1.2.3 Description de la structure à l'échelle locale

Nous qualifierons une mesure de locale si celle-ci décrit le proche environnement d'un nœud, c'est-à-dire un sous-graphe constitué par des voisins ou des voisins de voisins de celui-ci.

a. Distribution de degrés

Le nombre de voisins d'un nœud est une caractéristique fondamentale du graphe, car elle conditionne la structure entière de celui-ci. C'est pourquoi la description d'un réseau d'interactions commence presque systématiquement par la mesure des degrés des nœuds. Quel que soit le format de stockage du graphe, on peut réaliser cette tâche par un simple parcours des L liens.

Sur une large variété de réseaux complexes, on observe que la forme de la distribution des degrés évoque l'allure d'une loi de puissance, c'est-à-dire que la fonction de probabilité $\mathcal{P}$ de tirer aléatoirement un nœud de degré δ serait de la forme $\mathcal{P}(\delta) = a.\delta^{-\gamma}$, avec a et γ des constantes. Cette observation est par exemple vérifiée avec un bon ac-

cord sur les graphes de liens hypertextes du *world wide web* [KKR+99] ou sur le réseau Internet physique [FFF99] - c.f. figure 1.7.

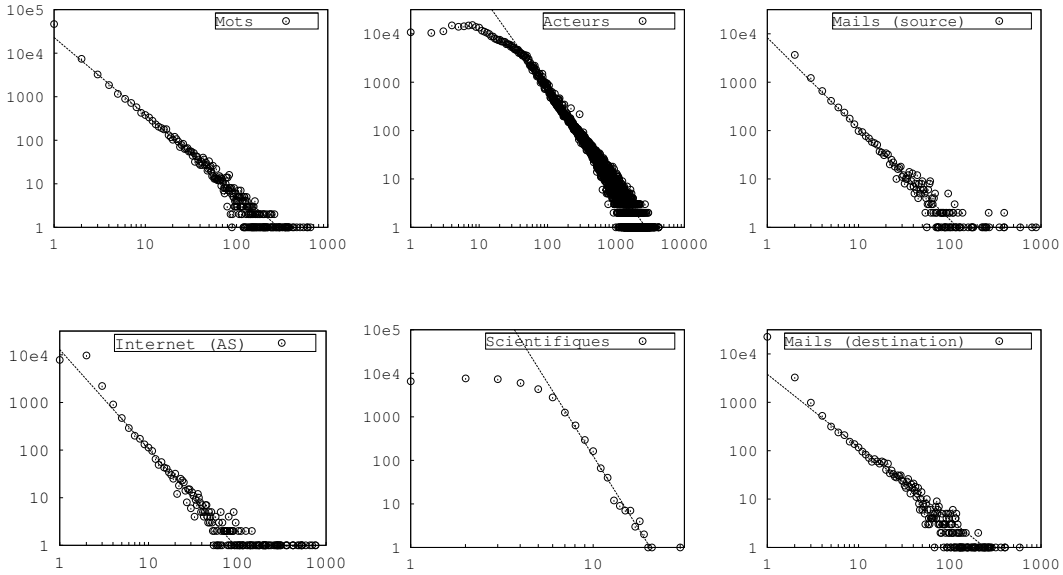


FIG. 1.7: Distributions de degré de réseaux réels*. Coocurrence de mots dans des définitions (*Wordnet*). Internet, niveau AS (*Route Views*). Collaborations d'acteurs (*Notre-Dame*). Collaborations de scientifiques (*Medline*). Échanges de mails, source et destination (*Kiel*).

Cette observation peut être rapprochée des distributions des suites numériques décrites par la loi de Benford, pour lesquelles la distribution de probabilité est telle que $\mathcal{P}(\delta) = log\left(\frac{\delta+1}{\delta}\right)$. Celle-ci serait en fait fréquente dans la nature, Benford l'avait identifiée sur la distribution d'occurrences des premiers chiffres dans les prix de produits boursiers [Ben38], puis sur de nombreuses grandes bases de données statistiques comme les superficies de lacs, ou les poids de composés moléculaires.

Une telle loi est dite invariante d'échelle *(scale-free)*, car la multiplication de la variable distribuée par un facteur α quelconque laisse la distribution invariante à un facteur près, uniquement fonction de α. Cette propriété est remarquable, elle traduit que la distribution considérée est amodale: elle n'admet pas d'échelle caractéristique. Cela induit que sa forme ne dépend pas de l'unité dans laquelle est exprimée la grandeur mesurée, par exemple les premiers chiffres des produits boursiers seraient répartis de manière analogue quelle que soit la monnaie dans laquelle ils sont évalués.

La forme générale d'une fonction vérifiant la propriété précédente est une loi de puissance: $\mathcal{P}(\delta) = a.\delta^{-\gamma}$. Lorsque $\gamma = 1$ on parle généralement de loi de Zipf, d'après le linguiste qui l'avait observée sur la fréquence d'occurrences des mots dans *Ulysses* de Joyce [Zip49]. Nous considérerons l'extension de la loi à toute distribution de la forme $\mathcal{P}(\delta) = a.\delta^{-\gamma}$ avec $\gamma \geq 1$, très largement observée dans une variété de domaines,

et qui porte d'ailleurs d'autres noms encore selon les contextes.

La présence d'une telle loi dans des champs si différents peut être perçue comme un argument en faveur de l'existence d'un processus générique sous-jacent. Depuis l'introduction d'un modèle simple par Simon [Sim56], parfois considéré comme une reformulation de modèles antérieurs [SR06], on sait qu'il est possible de produire des distributions de ce type à l'aide d'un mécanisme du type attachement préférentiel, sur lequel nous reviendrons plus en détails en 2.1.2.b.

L'allure en loi de puissance est bien sûr une version idéalisée de ce qui est pratiquement observé dans la nature; selon les réseaux considérés, la forme pourrait être décrite plus précisément par une loi log-normale, une loi de puissance avec une chute exponentielle ou autre [ASBS00, CSN10]. Par ailleurs les graphes sur lesquels il est vraiment possible d'observer une allure assimilable à une loi de puissance sur plus de deux décades sont en fait relativement rares, et pour les grandes valeurs - correspondant à des événements exceptionnels - les fluctuations sont considérables.

Dans notre optique, il n'est pas fondamental de chercher à savoir quelle loi offre la meilleure description de telles ou telles données réelles, nous retiendrons plutôt que la distribution de degré est généralement inhomogène et ***"heavy-tailed"***, c'est-à-dire que celle-ci décroît plus lentement qu'une exponentielle et présente donc des nœuds de très fort degré relativement à la moyenne (ou ***hubs***).

b. Corrélations de degré, assortativité

L'étude de la structure locale du graphe amène naturellement à étudier un analogue aux fonctions de corrélations spatiales: les distributions des corrélations de degré.

Lorsqu'on les étudie à deux nœuds, il s'agira soit de mesurer la distribution de probabilité conjointe $\mathcal{P}(\delta, \delta')$ pour qu'un lien associe des nœuds de degrés δ et δ', soit la probabilité conditionnelle $\mathcal{P}(\delta|\delta') = \mathcal{P}(\delta, \delta')/\mathcal{P}(\delta')$ de tirer un lien associé à un nœud de degré δ sachant que l'autre extrêmité est de degré δ'. Il est aussi possible de mesurer le degré moyen des plus proches voisins: $\sum_{\delta'} \mathcal{P}(\delta|\delta')\delta$ qui découle de la définition précédente, mais dont les fluctuations sont moindres. Une telle notion peut aussi être généralisée à des corrélations plus complexes (e.g. [MKFV06]), comme la distribution des motifs triangulaires ayant des degrés fixés aux sommets.

Les mesures de corrélations à deux nœuds nécessitent la connaissance des degrés des extrêmités de chaque lien du graphe. Avec un format de données en tableau de listes, il est possible d'accéder au degré en temps constant et donc de donner la distribution $\mathcal{P}(\delta|\delta')$ en $\mathcal{O}(L)$.

Pour interpréter des distributions de corrélations à deux nœuds, on peut recourir à des mesures scalaires. Parmi celles-ci la vraisemblance *(likelihood)*: $\mathcal{L} = \sum_{(i,j)\in E} \delta_i \delta_j$

ou des formes normalisées de cette mesure [LAWD04]. Mais la plus populaire reste l'**assortativité** *(assortativity)* **r**: dans son sens strict celle-ci rend compte de la tendance des nœuds à être connectés préférentiellement à des nœuds de degré comparable.

Il s'agit en fait du coefficient de corrélation linéaire de Bravais-Pearson, dont une formulation possible dans ce contexte est la suivante [New02]: on pose q_δ est la distribution de probabilité du degré "en excès" *(remaining degree)*, i.e. la probabilité pour que le degré du nœud connecté à un demi-lien tiré au hasard soit $\delta + 1$. $\sigma_q^2 = \sum_\delta \delta^2 q_\delta - [\sum_\delta \delta.q_\delta]^2$ est la variance associée à cette distribution. Alors si $e_{\delta_1\delta_2}$ est la probabilité conjointe d'observer un lien associant des noeuds de degré en excès δ_1 et δ_2, on définira l'assortativité par[1]:

$$\mathbf{r} = \frac{\sum_{\delta_1,\delta_2} \delta_1.\delta_2.(e_{\delta_1\delta_2} - q_{\delta_1} q_{\delta_2})}{\sigma_q^2}$$

Cette mesure appartient à l'intervalle $[-1; 1]$, 1 correspond à des degrés totalement corrélés, -1: totalement anticorrélés et dans ces deux cas, la connaissance du degré d'un nœud permet de donner avec certitude le degré de ses voisins. À l'inverse, $\mathbf{r} = 0$ signifie que connaissant le degré δ du nœud i, on n'a aucune information supplémentaire sur la valeur du degré d'un voisin j de i.

Assortativité ou anti-assortativité de réseaux réels. On mesure dans une grande variété de réseaux sociaux une assortativité positive. C'est notamment le cas de graphes de collaborations scientifiques ou artistiques [New02]. Cette propriété est corrélée à la structure communautaire de ces graphes [NP03], point sur lequel nous reviendrons en 1.3.3.

D'autres types de réseaux*: technologiques (Internet niveau AS: [PSVV01]), biologiques (résaux métaboliques, réseaux trophiques [New02]), révèlent plutôt de l'anti-assortativité *(dissassortativity)*. Nous donnons quelques exemples de mesures dans la table 1.1.

	r
Mots *(Wordnet)*	-0.004
Internet AS *(Routeviews)*	$-0,198$
Acteurs *(Notre-Dame)*	$+0,204$
Scientifiques *(Medline)*	$+0,135$

TAB. 1.1: Mesure d'assortativité sur certaines des bases de données* citées en 1.2.3.a.

Dans [MS02], les auteurs observent que des réseaux d'interactions entre protéines sont anti-assortatifs et suggèrent que cela puisse expliquer la réalisation de plusieurs fonctions simultanément et sans interférence dans le milieu cellulaire. Très récemment

[1] On trouve une forme généralisée aux réseaux pondérés de cette définition dans [LC07].

[JTMM10], il a été démontré qu'une explication possible à cette observation est un effet statistique: à distribution de degré fixée les graphes anti-assortatifs seraient simplement plus nombreux que les assortatifs.

Rich-club. Par ailleurs, on observe que les nœuds de fort degré forment parfois un groupe fermé, effet que l'on trouve sous le qualificatif de *rich-club phenomenon* dans la littérature [ZM03]. On peut le mesurer à l'aide d'une fonction du degré, le cœfficient de *rich-club*: en notant $N_{>\delta}$, le nombre de nœuds de degré supérieur à δ et $L_{>\delta}$ le nombre de liens entre de tels nœuds, il est défini par:

$$\boldsymbol{\phi(\delta)} = \frac{2L_{>\delta}}{N_{>\delta}(N_{>\delta}-1)}$$

c'est-à-dire le rapport du nombre de liens existants sur le nombre de liens possibles entre ces nœuds. Mais il ne suffit pas que la fonction ϕ soit croissante avec δ pour qu'il puisse exister une "oligarchie" dans le réseau, en effet même dans un graphe non-corrélé, la probabilité que deux nœuds soient liés croît avec leur degré [CFSV06], elle doit donc être d'abord normalisée à sa valeur dans le cas non-corrélé.

Bien que les deux mesures ne soient manifestement pas indépendantes, le lien entre ce coefficient et l'assortativité n'est pas trivial, et des réseaux présentant un *rich-club* peuvent être anti-assortatifs.

Généralisations. L'assortativité par degré peut être vue comme un cas particulier d'une tendance plus générale à s'associer entre nœuds ayant des caractéristiques semblables. Et à un degré de généralité encore supérieur, on peut mesurer de l'assortativité entre des grandeurs différentes.

Ainsi, pour deux caractéristiques scalaires et discrètes X et Y des nœuds, on pourra définir e_{xy} la probabilité conjointe pour un lien d'associer des nœuds dont $\mathrm{X} = x$ à des nœuds de $\mathrm{Y} = y$; $a_w = \sum_x e_{xw}$ sera donc la probabilité pour un lien d'avoir pour extrêmité au moins un nœud dont la caractéristique X prendra la valeur x, et $b_z = \sum_y e_{zy}$, l'équivalent pour Y. Alors, on définira un coefficient d'assortativité pour le couples de propriétés XY:

$$\mathbf{r}^{\mathrm{XY}} = \frac{\sum_{x,y} x.y.(e_{xy} - a_x b_y)}{\sigma_a \sigma_b}$$

Cette généralisation - d'ailleurs utilisable pour les distributions de degrés dans les cas orientés et bipartis [New03b] - permet de traiter d'autres caractéristiques topologiques des nœuds mais également des propriétés externes à la représentation graphique du réseau.

La tendance des agents du réseau à se lier préférentiellement lorsqu'ils partagent des caractéristiques communes est intuitive dans le domaine des réseaux sociaux: il est

plus fréquent d'observer des interactions entre des individus ayant des caractéristiques communes (âge, sexe, origine ethnique, niveau d'études), on se réfère à cette propriété par le terme générique d'homophilie (e.g. [MSLC01]).

c. Clustering

Le ***clustering*** cherche à rendre compte d'une propriété observée là encore de manière transversale dans les réseaux complexes: si le nœud a est connecté à b et c, on remarque que la probabilité pour que b et c soient connectés entre eux est élevée par rapport à un réseau connecté aléatoirement [WS98]. Ce trait est particulièrement typique des réseaux sociaux, où il témoigne de l'émergence de groupes ayant les mêmes goûts, on le comprend aisément dans le cadre de réseaux d'affinités où elle traduit grossièrement l'idée que "les amis de mes amis sont mes amis".

Nous définissons la mesure de théorie des graphes associée, qui consiste à évaluer la quantité de structures triangulaires observées dans le graphe, normalisée au nombre maximum de triangles qui seraient envisageables étant donné le nombre de voisins de chaque agent.

Définitions globale et locale. Deux définitions sont employées (pour les graphes non-orientés, non-pondérés), conventionnellement qualifiées de globale et locale. Celles-ci sont corrélées mais pas équivalentes, et d'ailleurs on peut observer communément des écarts considérables entre l'une et l'autre [New03d]:

- **clustering global**: il s'agit de trois fois le nombre de triangles observés sur le nombre de triplets de nœuds connectés mesurés sur tout le graphe. On peut donc écrire symboliquement cette quantité:

$$\mathbf{c_{3-g}} = 3.\frac{\triangle}{\wedge}$$

  Le facteur 3 assure la normalisation de la quantité $\mathbf{c_{3-g}}$, puisque chaque triangle comporte trois chemins distincts de longueur 2.

- **clustering local**: il se rapporte à un nœud spécifique a du graphe, c'est alors le rapport entre le nombre de triangles ayant a pour sommet sur le nombre de triplets connectés dont a est le nœud central. En général, l'intérêt se porte sur la moyenne de cette mesure sur l'ensemble des nœuds:

$$\overline{\mathbf{c}}_{\mathbf{3-l}} = \frac{1}{N}\sum_{a} \frac{a\triangle}{a\wedge}$$

  Plusieurs conventions sont possibles concernant les nœuds de degré 0 et 1: ils sont parfois exclus de la moyenne, parfois considérés comme nuls, nous choisirons cette seconde solution qui nous paraît plus usitée.

Dans les deux cas le clustering se situe dans l'intervalle $[0;1]$, 0 correspondant à une structure dépourvue de triangle; 1 signifie que si a est connecté à b et c, b et c sont nécessairement connectés entre eux.

On donne parfois un sens plus général au terme clustering en l'étendant aux cycles de taille quelconque, permettant ainsi d'obtenir une mesure normalisée de la quantité de cycles d'une taille k dans le graphe. Nous pouvons ainsi définir de manière analogue:

$$\mathbf{c_{4-g}} = 4.\frac{\diamond}{\diamond} \quad ; \quad \mathbf{c_{5-g}} = 5.\frac{\pentagon}{\pentagon} \quad ; \quad ...$$

Par ailleurs, lorsque nous ne précisons pas de quel cœfficient nous faisons usage, il s'agit du clustering global des cycles à trois nœuds. En effet le dénombrement de motifs semble souvent plus significatif que celui de ces grandeurs normalisées, en particulier dans le cas de réseaux sociaux, position que nous justifions en 1.3.3.a.

d. Motifs locaux

Les définitions précédentes s'attachent à l'importance des chemins et des motifs cycliques, mais nous nous intéresserons plus généralement à tout type de motif local, c'est-à-dire de sous-graphes mettant en jeu un "petit" nombre de nœuds (typiquement ≤ 5) - la **taille** du motif.

Informations associées aux motifs locaux. Le nombre et la forme des motifs font l'objet d'une grande attention, car ces caractéristiques sont associées aux mécanismes d'assemblage et de fonctionnement du réseau. Puisque ceux-ci varient beaucoup selon le contexte, il n'est pas surprenant que l'abondance relative des motifs permette de différencier des "superfamilles" de réseaux [MIK+04].

Dans le domaine des réseaux biologiques, des mesures de la topologie locale permettent de proposer des prédictions d'interactions possibles entre protéines en fonction des voies métaboliques identifiées [AA04]; dans les réseaux d'expression génétique, des études associent les fonctions régulatoires aux motifs locaux, au point de les décrire comme les briques élémentaires du réseau [SOMMA02, MSOI+02].

Les motifs locaux sont également un objet d'analyse prépondérant des réseaux sociaux, puisque le domaine s'est construit autour de l'étude des petits réseaux. On étudie par exemple depuis longtemps l'influence des schémas de communication sur la résolution de tâches [Bav50], ou encore les motifs correspondant à une organisation hiérarchique du réseau: selon l'allure des motifs on identifie le rôle des nœuds dans la structure [Miz94].

La dynamique de constitution des motifs locaux tient un rôle central pour la proposition de mécanismes régissant l'évolution de la structure du réseau social. Par exemple,

partant de l'hypothèse que les groupes sociaux tendent vers un état d'équilibre permettant de diminuer les relations conflictuelles entre les agents, la théorie de **l'équilibre structurel** *(structural balance)* [CH56, Dav63, KH07] propose un formalisme permettant de décider de la stabilité ou non d'un réseau, puis explique alors l'évolution de celui-ci par la recherche d'un état stable.

> **remarque :** On peut en comprendre le principe au travers d'une situation simple où les liens sont non-dirigées. Imaginons que les relations entre individus puissent être résumées de manière binaire: amitié ou hostilité, on affecte alors aux liens un signe respectivement $+$ ou $-$, puis on propose qu'un cycle stable soit un cycle dans lequel le produit des signes est $+$. Nous constatons alors qu'une triade (a, b, c) est stable si a et b sont amis et ont pour c le même type de sentiment, positif ou négatif. De telles hypothèses permettent ainsi de rendre compte de la tendance des réseaux sociaux au clustering et plus généralement au regroupement communautaire. Une analyse du même ordre dans le cas orienté explique également leur forte transitivité (si $a \to b$ et $b \to c$, alors $a \to c$).

L'accès aux réseaux sociaux en ligne a permis de renouveler ces questions en les examinant d'un point de vue statistique sur de vastes populations. Cela rend possible la comparaison à des modèles de mécanismes dynamiques expliquant par exemple la fermeture des cycles [LBKT08] et en particulier la transition des dyades aux triades.

Dénombrement pratique. Le problème algorithmique du dénombrement des motifs nécessite une réflexion sur la vitesse de calcul (voire la mémoire disponible) en fonction de la taille et de la géométrie du graphe à étudier.

Ainsi, il existe des algorithmes performants et spécialisés pour certains motifs et/ou certains types de graphes (sur les triangles dans les graphes épars: [Lat08], sur les cycles en général: [AYZ97, YZ04]). Et à l'inverse, des algorithmes d'une grande polyvalence, mais moins efficaces [MIK$^+$04]. Étant donné l'usage fréquent que nous ferons du dénombrement de motifs, nous développons cette question en Annexe B où nous proposons un bref comparatif des performances de quelques algorithmes.

Nous travaillerons généralement sur des graphes dont la taille n'excède pas quelques dizaines de milliers de nœuds et de liens; c'est pourquoi nous ferons usage d'algorithmes d'énumération simples et facilement adaptables pour des graphes épars. Celui proposé dans [Wer06] - également discuté en Ann. B - convient à cette situation.

e. Complétude, sous-groupes cohésifs

Un graphe est **complet** *(complete)* s'il existe un lien associant toute paire de nœuds du graphe. On se limite souvent à l'examen de cette propriété sur des sous-graphes de petite taille.

Cliques. Les sous-graphes maximum présentant la propriété de complétude sont nommés **cliques**. L'intérêt pour les cliques provient de l'idée que les sous-groupes cohésifs du graphe sont supposées être le siège d'interactions fortes entre les agents du réseau, définissant une entité collective autonome à laquelle on peut attribuer des caractéristiques, des fonctions etc. Ainsi, les cliques d'un réseau social pourraient représenter un groupe d'amis proches, dans lequel on observe des tendances à l'imitation ou à la diffusion rapide d'information.

Cette notion peut être informative sur d'autres réseaux complexes, ainsi dans les réseaux de régulation génétique où l'on observe que des structures construites comme des chaînes de cliques se superposent partiellement aux fonctions biologiques [PDFV05] - c.f. Fig. 1.8.

FIG. 1.8: Structure de cliques en chaîne au sens de [PDFV05]: chacun des groupes coloré est constitué de triangles adjacents partageant deux nœuds.

Cependant, leur usage reste limité en raison des difficultés pratiques à rechercher de tels objets sur de grands graphes. Ainsi, le problème consistant à trouver la clique de taille maximum contenue dans un graphe est de complexité exponentielle. Dans le cadre de notre étude, nous chercherons plutôt à énumérer des cliques de petite taille, question traitée en 1.2.3.d. et en Annexe B.

Autres mesures de cohésion. Par définition, les conditions pour obtenir une clique sont strictes, or l'existence d'un groupe ne nécessite pas systématiquement d'avoir des interactions entre tous les éléments du groupe. D'ailleurs le terme clique est souvent utilisé en sociologie dans ce sens élargi. Selon le contexte, on peut alors avoir recours à des outils d'analyse moins coercitifs que la clique (c.f. [WF94]).

Parmi celles-ci notons les n-cliques[2], sous-graphes maximum tels que la distance entre deux nœuds du sous-ensemble de nœuds soit toujours $\leq n$. Elles ont été définies dès les années 50 en sociologie [Luc50] pour donner une description hiérarchique affinée du réseau qui pourrait fournir des éléments d'explication sur son fonctionnement. D'autres

[2]Selon d'autres sources (e.g. [PDFV05]), une k-clique désigne une clique de taille k.

outils, souvent dérivés de ceux qui précèdent, peuvent être employés afin d'identifier des structures cohésives du graphe, selon la caractéristique du groupe sur laquelle on souhaite mettre l'accent (e.g. accessibilité des nœuds), nous pouvons évoquer par exemple le n-core: un sous-graphe dont chaque nœud est adjacent à au moins n autres [Sei83] ou le n-clan: une n-clique pour laquelle il existe toujours un chemin dans le sous-graphe de longueur inférieure ou égale à n entre tout couple de nœuds [Alb73, Mok79].

f. Communautés

Les mesures précédentes définissent de manière rigoureuse des sous-graphes cohésifs quelle que soit la topologie globale du graphe; il peut être plus approprié d'utiliser un élément structurel flexible pour répartir en sous-graphes les agents et relations du réseau, c'est à cette fin qu'on définit une **communauté**. Cette partie peut être lue comme une parenthèse car cette notion ne joue qu'un rôle secondaire dans la suite de ce travail; néanmoins nous l'évoquons car elle concentre une énorme activité, sans doute parce qu'elle permet d'organiser la structure du réseau.

Le terme "communauté" est chargé d'ambiguïtés dans le langage sociologique et il faut discuter son acception dans le contexte qui nous occupe. Pour l'analyse des graphes de réseaux, il n'y a pas de règle générale et c'est plutôt la procédure choisie qui amène à préciser la définition.

Régions denses du graphe. Parmi l'immense variétés de méthodes de détection communautaire (ou *graph clustering*, [For09]), un point de vue très répandu consiste à décrire une communauté comme un sous-ensemble du graphe pour lequel les liens internes sont plus fréquents que les liens vers l'extérieur [SMP90].

Les premières méthodes historiques consistent en des divisions successives du graphe en des ensembles (de tailles prédéterminées) selon le critère suivant: le partage est choisi de manière à conserver le plus de liens possibles dans chacun des groupes créés [KL70, Bar81].

Plus récemment, l'introduction de la **modularité Q** [NG04] fournit une grandeur scalaire évaluant la qualité d'une partition selon cette définition[3]. Le problème de la détection communautaire peut être alors traité comme une question d'optimisation: on se déplace dans un espace complexe de manière à trouver le maximum de la fonction $\mathbf{Q}$, en autorisant des tailles quelconques de groupes. Mais comme le parcours n'est pas exhaustif, il est fréquemment nécessaire de se satisfaire d'un certain maximum local de

[3]Si e_{IJ} est la fraction des liens associant un nœud de la communauté I à un nœud de la communauté J et $a_I = \sum_J e_{IJ}$ la fraction des liens ayant une extrêmité au moins dans I. Alors,

$$\mathbf{Q} = \sum_I (e_{II} - a_I^2)$$

la fonction.

On propose alors des algorithmes conservant le principe des séparations successives. Parmi ceux-ci, celui de Girvan et Newman [GN02] qui supprime itérativement des liens de fortes centralité (c.f. 1.2.4.c.) jusqu'à décomposer le graphe en ilôts fortement connectés; ou [RCC+04] dans lequel on supprime les liens par lesquels passent peu de cycles. À l'inverse, des algorithmes procèdent par agglomération successives de groupes [CNM04]. Ces procédés par itérations de divisions ou de fusions de sous-groupes permettent de définir une hiérarchie communautaire, correspondant à chaque niveau d'organisation, ce qui peut être utile pour décrire de manière plus détaillée un réseau réel.

Selon le contexte, les différentes procédures de parcours de l'espace ne sont pas toutes également efficaces (pour une comparaison des performances: [DDGDA05]). Les meilleurs en terme de qualité de l'optimisation utilisent des méthodes de complexité élevée comme le recuit simulé [GSPA04] ou dans une moindre mesure l'*extremal optimization* [DA05]. Les plus rapides, de complexité linéaire ou quasi-linéaire en N, sont basés sur des heuristiques gloutonnes (i.e. où l'on recherche un optimum local), qui conviennent à la détection de communautés sur de très grands graphes comme le *web* [BGLL08]. En contrepartie les algorithmes gloutons ont tendance à créer de très larges communautés [CNM04], et sont sensibles aux conditions intitales, c'est-à-dire qu'ils ne produiront pas le même partage communautaire en partant de deux points de départ différents.

Indépendamment des limites techniques à la recherche du maximum de modularité, des auteurs mettent en évidence une limite de résolution intrinsèque aux partitions reposant sur cette mesure qui tendrait naturellement à favoriser les grandes communautés [FB07].

Définitions alternatives. D'autres méthodes tirent parti d'analogies, en particulier avec la physique des matériaux magnétiques où l'on observe spontanément l'organisation de la matière en domaines [Jen06]. D'ailleurs, alors que les méthodes basées sur la modularité réalisent des partitions du graphe (chaque nœud appartient à une et une seule communauté), la comparaison aux milieux magnétiques permet de définir des communautés de manière plus floue [RB06a]. Plus généralement, certaines procédures autorisent la superposition partielle des communautés, comme dans [PDFV05] où elles sont construites comme des séquences de cliques.

Par conséquent, la détection de communautés repose sur un certain nombre de choix discutables: séparées ou se recouvrant partiellement, critère pour qualifier la qualité d'un découpage, procédure d'optimisation etc. Et à nouveau, la validité de ces choix dépend de l'interprétation que l'on souhaite privilégier du sous-groupe cohésif et donc de l'objectif dans lequel est réalisée la détection.

1.2.4 Description de la structure à l'échelle globale

Les mesures précédentes décrivent le proche environnement des nœuds. Pour saisir des aspects complémentaires de la topologie du graphe, on en mesure des caractéristiques à plus grande échelle, dont nous donnons dans cette partie quelques unes des plus classiques. Nous aurons alors les éléments suffisants pour obtenir une image schématique commune à de nombreux réseaux complexes, que l'on trouve souvent résumée par l'expression "petit monde". Nous évoquerons également les problèmes algorithmiques inhérents aux mesures globales qui limiteront ensuite leur utilisation.

a. Connexité, cohésion structurelle

On dit qu'un graphe est **connexe** *(connected)* si pour tout couple de nœuds, il existe un chemin de liens pour joindre l'un à l'autre. Dans le cas orienté, on parle de connexité lorsqu'il existe une chaîne joignant les nœuds. Un sous-graphe connexe maximum est nommé **composante connexe** d'un graphe.

Les graphes de réseaux réels, malgré leur faible densité, présentent souvent une composante connexe géante, c'est-à-dire qu'un nombre "important" de nœuds (de l'ordre de N) appartient à cette composante. C'est en fait une propriété que l'on retrouve chez les graphes aléatoires ayant une distribution de degré dont la moyenne est supérieure à 1 [CL02].

Cohésion d'un réseau et k-connexité. La notion de connexité se généralise à la **k-connexité** *(k-connectedness)*: un graphe est dit k-connexe si k est le nombre minimum de nœuds à retirer pour que le graphe obtenu soit non-connexe.

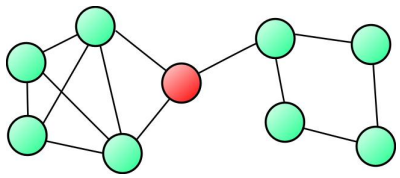


FIG. 1.9: Graphe 1-connexe, la suppression du nœud central (rouge) scinde le graphe en deux composantes respectivement 4- et 2-connexe.

Dans le contexte des graphes, on qualifie un groupe de cohésif s'il existe plus de liens entre ses membres que vers l'extérieur[4] [Col88], c'est le point de vue que nous avons adopté en 1.2.3.f. Des auteurs suggèrent que la k-connexité puisse être une estimation

[4]Ce terme a un sens souvent plus général dans le contexte sociologique: il peut par exemple faire référence au degré ressenti d'identification d'un individu à une communauté, ou à l'existence de relations fortes entre les agents; une revue sur cette notion est proposée dans [Fri04].

de cette cohésion structurelle du réseau social [MW03], car elle mesure à quel point un groupe dépend de ses constituants pour conserver son caractère de groupe.

La connexité simple peut être insuffisante dans ce but. En effet, considérons deux types de groupes: un où l'autorité est centrée sur un leader charismatique et l'autre où elle est dispersée entre les membres; des schémas de connectivité qui pourraient correspondre à cette situation sont représentés sur la figure ci-dessous. L'un et l'autre graphes sont connexes et de même densité, mais alors que la suppression du nœud central déconnecte la structure en étoile de gauche (1-connexe), il n'y a pas d'acteur jouant un rôle équivalent dans la structure de droite qui présente un cycle.

Ainsi, dans [MW03], les auteurs montrent que la structure hiérarchique induite par la k-connexité pourrait, mieux que le degré, la centralité (c.f. 1.2.4.c.) ou d'autres mesures topologiques, justifier des expériences concrètes comme des enquêtes portant sur l'attachement ressenti par des élèves à l'égard de leur école.

Notion de robustesse. Dans le même ordre d'idée, la capacité du graphe à rester connexe lorsqu'on supprime certains de ses agents aléatoirement ou de manière sélective, ou **robustesse** *(robustness)* du réseau, joue un rôle déterminant dans les problèmes de transport. Qu'il s'agisse de la circulation d'individus, de la diffusion d'un virus au sein d'une population ou d'une information dans un réseau social, il est primordial de savoir si les agents du réseau sont accessibles ou non. En effet, si le graphe est non-connexe, un phénomène de transport sera nécessairement cantonné à la composante dans laquelle il naît; cette situation peut s'interpréter selon les cas, comme une interruption du traffic, l'endiguement d'une épidémie etc.

On observe souvent une corrélation entre cette caractéristique et l'allure de la distribution de degré dont il était question précédemment. En effet, une propriété désormais bien connue des graphes aléatoires présentant une distribution *heavy-tailed* est leur grande résistance à la disruption face à une attaque non-ciblée [AJB00, CEBAH00, MGGP08]. Les conséquences sont visibles au travers de phénomènes réels, où les réseaux ont leurs degrés distribués de manière très hétérogène: l'Internet s'adapte à la défaillance de ses nœuds en trouvant de nouveaux chemins de routage [CR05], les virus informatiques sont capables de proliférer en dépit de faibles taux de contagion [PSV01], les organismes biologiques restent fonctionnels malgré les possibilités de mutations qui altèrent les mécanismes métaboliques [HHL$^+$99]. Ces indices révèlent qu'il est souvent possible de trouver des voies alternatives pour réaliser une fonction dans les réseaux complexes réels.

b. Distance et diamètre

Rappelons que la **distance** entre deux nœuds a et b désigne le nombre minimum de liens - ou éventuellement d'arcs orientés - à parcourir pour relier a à b; le **diamètre** désigne alors le maximum de distance, sur tous les couples de nœuds du graphe, des plus courts chemins (ou **géodésiques**) entre deux nœuds. De telles définitions n'ont de sens que pour une composante connexe.

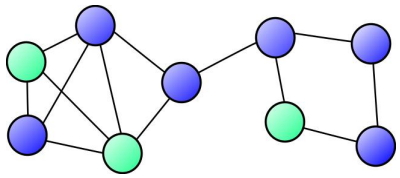


FIG. 1.10: Graphe de diamètre 6, le chemin reliant les nœuds bleus est une géodésique de taille maximum.

Contraintes d'utilisation. La distribution des distances entre les nœuds est riche en information, mais sa mesure pratique (et *a fortiori* celle du diamètre) sur de grands graphes est un problème techniquement difficile à résoudre. En effet, il n'est pas difficile d'imaginer des moyens de déterminer les distances dans un graphe (non-pondéré, non-orienté): ainsi elles pourraient être évaluées à l'aide de la matrice d'adjacence $\mathbf{A}$, en utilisant que le coefficient (i, j) de $\mathbf{A}^k$ est le nombre exact de chemins de longueur k liant i à j. Mais on devra aussi prendre en compte le coût algorithmique de la procédure si celle-ci doit être employée fréquemment, et le produit matriciel nécessitant N^3 multiplications et N^{2N-1} additions n'est pas du tout adapté à cette situation.

Nous emploierons régulièrement des procédures de **parcours en largeur** du graphe, qui consistent à examiner l'ensemble des nœuds voisins d'une source, puis les voisins de ces nœuds et ainsi de suite, "couche par couche" autour de la source. Quand on l'utilise pour évaluer les distances à partir d'un nœud, on parcourt tous les liens du graphe une seule fois donc la complexité temporelle d'un tel algorithme serait en $\mathcal{O}(L)$ pour chaque source, soit en $\mathcal{O}(LN)$ pour tout le graphe; pour une complexité spatiale en $\mathcal{O}(N^2)$ si l'on souhaite garder en mémoire toutes les distances entre couples de nœuds.

Il existe des méthodes plus rapides ou approchées (pour une revue: [Zwi01]), mais exigeant des algorithmes bien plus complexes, étant donnée la taille des graphes considérés dans les applications à venir, le parcours en largeur sera le plus souvent satisfaisant.

"Six degrés de séparation". Au cours des années 60, le psychologue Stanley Milgram et ses collaborateurs [Mil67, TM69] entreprenaient une série d'expériences, depuis restées célèbres. Il s'agissait de suivre le parcours d'un dossier entre deux individus quelconques, choisis dans une vaste population, étrangers l'un à l'autre, et ce

en faisant en sorte que l'expéditeur ne communique le dossier qu'à une de ses connaissances qu'il juge à-même de pouvoir s'approcher de la cible, moyennant un petit nombre d'informations sur celle-ci (métier, localisation).

Milgram observa alors que lorsque la chaîne entre la source et la cible était réalisée, celle-ci comportait entre 2 et 10 individus, avec une moyenne située entre 5 et 6. Résultat qu'on trouve souvent résumé par l'expression passée dans le langage courant des "six degrés de séparation" existant entre deux personnes.

Ce résultat est interprété dans le vocabulaire des graphes en affirmant que la distance moyenne entre deux nœuds d'un graphe connexe de réseau réel est "faible", tout comme le diamètre, c'est-à-dire qu'ils seraient en général de l'ordre de $log(N)$. Soulignons que cette interprétation graphique n'explique pas toutes les observations expérimentales, en particulier la capacité des agents à trouver - ou non - de "bons" chemins dans le réseau, la proportion de chaînes non-complétées restant très majoritaire.

Ces expériences ont été amplement reprises avec quelques variations (pour une revue critique: [Sch09]), et elles ont donné lieu à de nombreux commentaires dont certains mettent en question le protocole (e.g. [Kle02]). Mais l'essentiel des conclusions est maintenant appuyé par des mesures sur des réseaux de communication (*world wide web*: [AJB99], messageries instantanées: [LH08]) - même si la valeur 6 doit être comprise comme seulement indicative. De plus, la propriété de faible diamètre n'est pas exclusive aux graphes de réseaux sociaux, elle semble partagée par beaucoup de graphes de réseaux réels présentant une distribution de degré *heavy-tailed* [New03d], comme dans la figure 1.11.

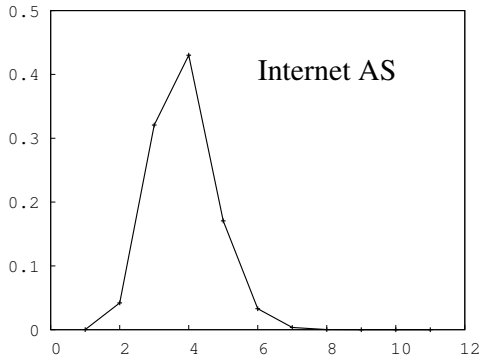


FIG. 1.11: Distribution de probabilité des distances entre deux nœuds de la plus grande composante connexe du graphe réel *Routeviews* de l'Internet au niveau AS*.

Les "petits-mondes". Le terme de "petit-monde" est employé par Milgram pour décrire l'expérience évoquée ci-dessus. L'interprétation associée varie en revanche selon le contexte et les auteurs. Dans [Mil67], on peut lire: *"We should think of the two*

points as being not five persons apart, but 'five circles of acquaintances' apart — five 'structures' apart"[5]. Autrement dit, l'idée de petit monde suppose déjà une structure en communautés du réseau de communication. Et l'information circulerait rapidement dans des cercles relativement limités, mais difficilement d'un cercle à un autre.

Nous utiliserons plutôt le sens restreint que l'on trouve dans les travaux de Watts et Strogatz [WS98], où la notion de petit monde est perçue avant tout au travers du graphe représentatif d'un réseau. Elle peut être alors résumée à la combinaison de deux propriétés: faible distance moyenne entre deux nœuds et fort clustering (c.f. Table 1.2).

	Réel		Aléatoire	
	$\overline{c}_{3-l}$	$\overline{d}$	$\overline{c}_{3-l}$	$\overline{d}$
Collaborations d'acteurs *(Notre-Dame)*	$0,79$	$3,65$	$0,00027$	$2,99$
Réseau neuronal *(C. Elegans)*	$0,28$	$2,65$	$0,05$	$2,25$
Réseau électrique	$0,08$	$18,7$	$0,005$	$12,4$

TAB. 1.2: Clustering local ($\overline{c}_{3-l}$) et distance moyenne ($\overline{d}$) dans des graphes réels et un échantillon aléatoire correspondant. Extrait de [WS98].

Les adjectifs "faibles" et "forts" sont ici ambigüs: la faible distance moyenne entre les nœuds se rapporte à ce que répondrait spontanément, selon Milgram, une majorité de gens non-informés sur cette question; en revanche, le clustering est élevé comparé à un graphe de même densité produit aléatoirement.

c. Centralité

La notion de **centralité** d'un nœud a été introduite par Bavelas [Bav47] pour caractériser l'importance de celui-ci relativement au réseau, initialement dans le cadre de réseaux de communication. Cette question prend encore une dimension supplémentaire avec l'émergence des réseaux en ligne; ainsi un moteur de recherche a pour principale fonction d'établir une classification des pages en fonction d'un certain estimateur de centralité [BP98, Kle99].

Nous en proposons ici une mesure classique, proposée par Freeman dans [Fre79]: la **centralité d'intermédiarité** *(betweenness centrality)* C_I est définie par la fraction des géodésiques du graphe contenant x_0. Soit, avec g_{ij} le nombre de plus courts chemins liant x_i à x_j (différents de x_0), et $g_{ij}(x_0)$, ceux d'entre eux qui passent par x_0:

$$\mathbf{C_I(x_0)} = \sum_{i,j \neq 0} \frac{g_{ij}(x_0)}{g_{ij}}$$

[5]Il faudrait voir deux points comme séparés non pas par cinq personnes mais par cinq cercles de relations - cinq structures.

Cette mesure est donc inspirée par l'idée que l'importance d'un nœud tient à sa capacité à jouer le rôle d'intermédiaire entre les sous-parties du graphe.

En pratique, les mesures de centralité sont usitées pour évaluer l'autorité notamment dans les réseaux d'influence politique [Bol88, HH95]. Par exemple dans une étude maintenant bien connue, Padgett *et al.* justifient l'accession au pouvoir de la famille Médicis au début du XV$^{\text{ème}}$ siècle notamment à l'aide des différentes formes de centralité [PA93] dans le réseau des alliances socio-commerciales entre les grandes familles florentines. Par ailleurs, c'est la mesure de centralité d'intermédiarité qui guide la construction de l'algorithme de détection communautaire de Girvan et Newman évoqué en 1.2.3.f. [GN02].

Contraintes d'utilisation. Cette définition caractérise la situation d'un nœud vis-à-vis de l'ensemble du réseau, ce qui suggère que son application à de grands graphes pose des difficultés algorithmiques. $\mathbf{C_I}$ nécessite non seulement de déterminer la distance entre tous les couples de nœuds mais également de lister les géodésiques.

Ces problèmes usuels d'énumération des plus courts chemins font l'objet d'une importante littérature, et les différents algorithmes proposés sont d'une efficacité variable selon la topologie des graphes parcourus. Mais les plus classiques sur la question (e.g. Floyd-Warshall) ont une complexité en $\mathcal{O}(N^3)$, qui limite leur utilisation aux petits graphes (typiquement, quelques heures de calcul pour un millier de nœuds sur une machine standard). On trouve dans [Bra01] la proposition d'un algorithme en $\mathcal{O}(NL)$ pour calculer la centralité d'intermédiarité, ce qui en autorise l'utilisation pratique sur de grands graphes [Bar04].

Une autre solution consiste à recourir à une mesure plus locale (qualifiée de centralité égocentrée) qui ne prend en compte que le proche environnement du nœud considéré et dont la corrélation avec la centralité sur l'ensemble du réseau peut être suffisante pour justifier son utilisation [Mar02, EB05].

Diffusion sur les réseaux, force des liens. La centralité d'intermédiarité définit indirectement l'importance des nœuds selon leur capacité à agir sur la circulation de ce qui diffuse sur le réseau: information, biens matériels - des "objets mobiles".

Le sociologue Mark Granovetter a construit la **théorie des liens faibles** au cours des années 70 en s'inscrivant dans cette logique [Gra73, Gra83]. Celle-ci repose sur une certaine description du réseau social autour d'un individu: il est constitué d'une part de proches avec qui on établit des liens forts (passe beaucoup de temps ensemble, partage de nombreuses activités etc.) et des relations plus lointaines - les liens faibles.

L'idée essentielle est que ces liens faibles jouent un rôle fondamental de pont entre les communautés du réseau. Cela se traduirait en particulier sur les phénomènes de diffusion car les objets mobiles importants transitent relativement peu au travers des liens forts; "importants" dans le sens où le fonctionnement du réseau ne peut être

compris sans analyser ces canaux particuliers de circulation des objets mobiles. Ainsi dans un réseau commercial, si un acteur n'est approvisionné pour une marchandise que par un producteur, ce canal lui est essentiel, cela ne sera plus le cas s'il est lié à beaucoup de producteurs ou d'intermédiaires lui permettant d'être ravitaillé par d'autres chemins. Cette image permet de comprendre que la notion de liens faibles peut être perçue d'un point de vue stratégique pour le contrôle de la circulation de l'information dans le réseau[6].

Cette hypothèse a été mise à l'épreuve empiriquement: dans [LD72], une étude sur l'adoption d'innovation appuie en effet l'idée que la diffusion est assurée par des relations éloignées et que les environnement hétérogènes favorisent la circulation de l'information. D'autres travaux analysent le type de liens par lesquels passent des individus recherchant un emploi [Lan77]; ils aboutissent notamment à la conclusion que les liens faibles servent effectivement d'intermédiaires préférentiels pour atteindre les agents du réseau ayant un statut élevé [LEV81, Gra83].

Du point de vue graphique, comment pourrait alors être interprétée la notion de force? Une relation étroite pourrait se traduire par de nombreuses connaissances communes dans le réseau, on s'appuie alors sur la cohésion locale pour déterminer la force des liens, et selon la mesure de cohésion choisie (c.f. 1.2.3.e.), l'estimation de la force du lien varie. Des mesures sur la circulation de l'information à l'aide de ce formalisme fournissent des évaluations expérimentales quantitatives à l'appui de la théorie des liens faibles [CR09].

Cela s'accorde avec l'image qualitative des réseaux sociaux en tant que petits mondes: des groupes cohérents dans lesquels l'information circule facilement, reliés entre eux par quelques rares ponts, associant des nœuds de forte centralité.

1.2.5 Bilan

Nous avons fait dans cette partie un inventaire des mesures génériques couramment employées dans la littérature des graphes de réseaux complexes. En étudiant leurs applications, nous avons cherché à mettre en évidence d'une part ce qu'elles nous indiquent de la structure du réseau et d'autre part quel est leur prix algorithmique.

Cet exposé n'est pas exhaustif, on trouve plusieurs revues qui proposent d'autres mesures (e.g. [BLM+06, CRTB07]); nous n'avons par exemple pas évoqué les mesures spectrales, classiques (c.f. [Par98, FDBV01, SR03]) mais dont nous ne ferons pas usage par la suite en raison de leur coût trop élevé.

En bref, nous retiendrons que les mesures adoptées sont le résultat d'un compromis entre leur pertinence et les difficultés algorithmiques à les mettre en œuvre. Nous essaierons, aux cours des exemples traités, d'expliquer certains choix, mais il faut

[6]Cette idée est développée par la théorie du "trou structurel" [Bur95].

garder à l'esprit cette limitation pratique qui nous impose de mener les études avec un nombre restreint d'observables le plus souvent locales.

1.3 Graphes de réseaux sociaux

Des objets très divers sont regroupés sous le qualificatif de "réseaux complexes", parce que leur étude appelle l'utilisation des mêmes outils et qu'ils partagent certaines caractéristiques qualitatives (par exemple l'allure de la distribution de degré). Néanmoins, une analyse plus poussée nécessite de préciser certaines caractéristiques propres aux réseaux étudiés.

Ce travail est essentiellement consacré à l'analyse de réseaux sociaux, et nous avons vu à plusieurs reprises que les graphes représentant des données sociologiques se distinguent des autres: leur clustering est particulièrement élevé, leur assortativité souvent positive etc. Nous nous penchons donc maintenant plus précisément sur les caractéristiques propres à ces données: collecte, hypothèses de travail et mesures spécifiques.

1.3.1 L'échelle de l'étude

Une étude effectuée dans la tradition de la sociologie des petits réseaux, telle que celle menée par Padgett *et al.* [PA93] sur les familles florentines, examine les relations entre acteurs sur plusieurs niveaux. Dans cet exemple, on considère les transactions et accords commerciaux, les alliances maritales, les prêts, les mécénats; les explications proposées sur les événements et les évolutions des interactions s'appuient sur la répartition du pouvoir conjointement dans chacun de ces plans. On peut utiliser différents types d'interaction, en chercher d'autres quand certaines font défaut, modifier en conséquence la représentation du système - en l'occurrence le tissu social de l'élite florentine. C'est-à-dire que comprendre la mécanique d'un petit réseau nécessite un aller-retour constant entre la modélisation et l'interprétation des données: énoncer ce que représente une relation précisément et la valeur que les agents peuvent leur accorder, individualiser les nœuds dont les décisions ne sont pas nécessairement rationnelles etc.

Mais à l'échelle à laquelle nous travaillons, le rapport aux données est plus anonyme: la collecte et le traitement sont en grande partie automatisés, et ne permettent pas une telle précision dans la définition des caractéristiques des liens et des nœuds.

1.3.2 Collecte et traitement

Il existe de nombreuses typologies du large domaine des réseaux sociaux - e.g. [Bur80, WF94, BMBL09], en fonction de la nature des nœuds (individus, communautés, institutions) ou des liens (interactions, rôles relatifs, transferts matériels); les frontières

entre ces classifications sont souvent floues.

Nous exposons ici une approche pragmatique du choix des réseaux de travail: nous décrivons les données employées, les procédures de collectes et de traitement associées, puis expliquons quel est le cadre d'hypothèses sur leur interprétation dans lequel nous nous inscrivons. Pour des précisions techniques sur les données utilisées, on se référera à nouveau à l'Annexe A.

a. Collecte des données

Les grandes bases de données qui ont permis le développement d'une approche statistique des réseaux sociaux n'ont pas été construites spécifiquement pour en faire une analyse de ce type. Il s'agit le plus souvent de bibliothèques numériques, c'est le cas par exemple des bases de collaborations scientifiques qui sont extraites de collections bibliographiques.

Les modalités de collecte ne sont donc pas optimales: nous ne disposons pas nécessairement des informations qui permettraient des spécifications utiles sur les relations (nature, durée, force)[7]. Actuellement, les campagnes de collectes se multiplient dans le but d'analyser les données à l'aide d'outils de théorie des graphes, cependant nous n'avons pas effectué la collecte des données brutes: notre travail commence à la traduction de celles-ci en graphe.

b. Hypothèses de la modélisation

Les conditions de collecte et de traitement des données nous amènent à réfléchir aux hypothèses sous-jacentes à la modélisation en graphe du réseau social.

Représenter les données à l'aide du seul graphe revient à dire que nous n'aurons aucune précision sur le contenu des interactions, seulement l'indication de leur existence (accessoirement de leur orientation ou de leur poids). Pour qu'un tel modèle puisse effectivement être informatif, voire autorise des prédictions, il est nécessaire que les liens et les nœuds aient tous sensiblement la même signification, qu'ils représentent une même réalité. Nous ferons référence à cette condition comme une **hypothèse d'homogénéité** sur les données.

Nous ne pouvons évidemment pas respecter exactement cette condition: un échange d'information n'est jamais strictement identique à un autre et il n'existe pas de mesures qui restitue l'intégralité de sa signification.

Cependant, nous sélectionnons les bases de données parmi celles à disposition dans le but de nous approcher au mieux de cette situation idéale. Et si nous cherchions à définir ce que signifient les liens entre agents dans les représentations que nous étudions,

[7]Par analogie, on est davantage dans une situation où l'on chercherait à repérer des régularités d'un phénomène naturel observé, plutôt que dans celle d'un expérimentateur dont le protocole est développé dans le but de détecter un événement attendu.

il s'agirait de **l'existence d'un canal d'échange entre les deux agents**, la nature de l'échange étant spécifiée par le type de réseau (biens matériels, données etc.). Nous excluons volontairement les autres dimensions de l'interaction sociale pour nous ramener à une description unimodale, plus conforme à la représentation en graphe.

Cette interprétation donne un certain éclairage à l'hypothèse fondamentale de la description en graphe des réseaux: **la structure de leur environnement suffit à expliquer certains comportements des acteurs**. Cette hypothèse est déjà forte puisqu'elle suppose qu'on peut se passer de toute autre représentation du comportement individuel. Elle l'est pourtant moins que celle énoncée par Mizruchi [Miz94]: "the primary tenet of network analysis is that the structure of social relations determines the content of these relations"[8], où l'idée de détermination suppose que *toute* l'information sur la relation soit contenue dans la structure, point de vue que nous ne partageons pas. À l'opposé, certains défendent que le formalisme des réseaux est trop appauvri en information et ne prend pas suffisamment en compte l'environnement culturel et son influence sur les choix des agents (e.g. [EG94]).

Nous ne discuterons pas ces différents points de vue longuement débattus entre sociologues, et nous nous satisferons d'une position pratique sur la question. Le graphe est pour nous **un mode de projection d'une information réelle très complexe, dont il ne saisit certainement pas l'intégralité des mécanismes**. Toutefois l'information projetée, quelque partielle qu'elle soit, peut porter la trace du phénomène qui lui donne naissance et nous devons donc chercher des outils qui permettraient de l'identifier.

c. Bases décrites

Le déficit d'information sur les contenus transférés nous a donc amené à une définition minimale de la notion d'interaction: elle y représente l'existence d'un possible transfert (matériel ou immatériel) entre les agents du système.

Nous pouvons proposer un classement des différentes bases utilisées en fonction de ce que nous y voyons de commun, mais cela reste très arbitraire. Les catégories ci-dessous - non exclusives les unes des autres - examinent les réseaux selon trois points de vue différents, en fonction de la méthode de collecte utilisée en amont.

- **Selon la finalité du réseau.** Le réseau est constitué d'individus s'associant dans le but de réaliser un objectif commun: scientifiques coauteurs d'un article, artistes participant à un même projet, contributeurs d'une même page sur un site collaboratif. Nous les dénommerons par l'expression générique de **réseaux de collaborations**.

[8] le principe de base de l'analyse de réseau est que la structure des relations sociales détermine leur contenu

Nous supposons que ce type de réseau soit un substrat privilégié pour la circulation d'informations spécialisées (e.g. un concept scientifique).

- **Selon le support du réseau.** Dans ce cas ce n'est pas le but qui justifie la construction du réseau, mais le moyen de communication qui a permis la collecte des données: il peut s'agir de communications téléphoniques, d'e-mail, de posts sur un forum ou un blog.

- **Selon la quantité transférée.** Ce point de vue est privilégié lorsque l'interaction est caractérisée par un échange physique, il est alors possible de définir une unité de la quantité circulant et une structure matérielle supportant son acheminement. Un réseau de circulation des individus, des échanges commerciaux ou encore Internet (en tant que structure physique servant au transfert de paquets d'une machine à une autre) peuvent être considérés comme tels.

  remarque : Ce dernier type ne met en jeu qu'indirectement les relations entre individus, et il ne s'agit pas à proprement parler de réseaux sociaux. Nous les incluons toutefois dans notre cadre d'étude car ils nous semblent être un intermédiaire intéressant entre des réseaux dont la decription physique contient la quasi-intégralité de l'information (e.g. réseau électrique), et ceux qui - comme les réseaux sociaux - supposent des hypothèses fortes sur la capacité de ce formalisme à saisir le contenu complexe des interactions.

d. Filtrage

Dans une perspective pratique, la transformation en graphes de ces bases exige un **filtrage** des données brutes. Il s'agira d'une part d'éliminer les défauts inhérents à la collecte automatique des données, et surtout, conformément à l'hypothèse que nous énoncions précédemment, de rendre le plus uniforme possible les interactions représentées.

Étant donnée l'échelle de l'étude, il n'est possible d'examiner le contenu des interactions que très ponctuellement. Cela implique des vérifications qu'il faut adapter au contexte, et pour lesquelles on ne peut pas énoncer de règles méthodologiques générales. Afin de comprendre concrètement ce que traduit le terme de filtrage, prenons quelques exemples:

– La communauté des rédacteurs de *Wikipédia** peut être subdivisée en sous-catégories: administrateurs, utilisateurs enregistrés ou anonymes, programmes circulant de page en page (ou "robots") entre autres catégories plus ou moins formelles.
  Les comportements observés correspondants sont très différents: le nombre de contributions d'un robot peut être extrêmement élevé alors qu'il n'interagit pas

à proprement parler avec les autres contributeurs, on dervait donc les éliminer lorsque l'on cherche à mesurer la communication entre rédacteurs du réseau[9].

– Dans le cadre de collaborations scientifiques, on peut trouver des publications dont le nombre d'auteurs se comptent en centaines. C'est une caractéristique typique de la physique expérimentale des hautes énergies (entre autres), où il est d'usage de faire figurer tous les participants à la construction du dispositif. Il est bien évident que la relation entre deux acteurs d'un tel projet n'est pas comparable à celle de deux collaborateurs qui interagissent quotidiennement.
Parmi les moyens de contourner ce problème, le filtrage thématique (avec des mot-clefs par exemple) est simple et les articles sélectionnés devraient avoir une relative uniformité méthodologique. Mais c'est une solution arbitraire puisque, par principe, les participations de collaborateurs à un projet ne sont jamais équivalentes.

La notion de filtrage peut paraître secondaire, mais elle tient un rôle important car on y trace les frontières des données, elle exige d'énoncer explicitement la définition des liens. Nous voyons que cette question est liée à la spécification des contenus, il n'y a donc pas de solution universelle à ces problèmes qui dépendent de l'objectif poursuivi. Selon le contexte, on doit imaginer des solutions *ad hoc* nous permettant de respecter au mieux la condition d'homogénéité.

1.3.3 Nature bipartie

Les réseaux de collaboration que nous nous proposons d'étudier se prêtent à une représentation bipartie qui semble être un choix plus judicieux que la projection - de toutes façons plus pauvre en information. En effet, les interactions se produisent à l'occasion d'événements, que nous pouvons décrire comme les atomes, les "briques de base" du réseau: réunions, discussions, plate-formes collaboratives communes etc.

Dans une certaine mesure, les réseaux de communication peuvent être regardés de la même manière, les conversations téléphoniques ou autres événements ne faisant intervenir que deux acteurs pouvant être compris comme un cas limite d'une situation plus générale. Mais l'unité de l'échange d'information reste l'événement alors que l'existence d'un lien entre deux acteurs de la projection en est une image cumulée sur la période de capture des données.

Si l'usage de la représentation projetée est très répandue, cela tient d'une part au fait que les outils d'analyse sont généralement conçus pour des graphes monopartis, d'autre part à ce que la structure événementielle n'est pas toujours accessible. Nous

[9]D'autant plus que les *hubs* ont en général un poids considérable dans les statistiques [XZSS09]: ils "rapprochent" les nœuds, affectent lourdement les corrélations de degré, génèrent de nouveaux chemins, de nouveaux motifs etc.

allons donc examiner dans la suite de cette partie les conséquences induites par la nature bipartie des données sociales.

a. Mesures spécifiques

Même si les graphes bipartis peuvent être compris comme un sous-ensemble des graphes monopartis, la majorité des mesures décrites précédemment ne sont pas adaptées à des modélisations dans lesquelles les nœuds peuvent représenter deux entités différentes. Par exemple, les mesures de clustering usuelles ne contiennent aucune information sur ces graphes, puisqu'elles y sont nulles par construction.

Pour contourner ce problème, nous pouvons effectuer les mesures non pas directement sur les graphes bipartis mais sur leurs projections. Cela n'est pas contradictoire avec l'importance de la nature bipartie que nous soulignions précédemment car les modèles de réseaux prendront en compte les événements. En général, une projection semble plus naturelle que l'autre: on peut choisir de connecter entre eux les individus ou au contraire les événements, mais il est logique de considérer plutôt le réseau des acteurs, qui est l'objet usuel d'analyse monopartie.

b. Formes généralisées de clustering

On trouve dans la littérature plusieurs suggestions de généralisation du clustering applicables aux graphes bipartis (e.g. [LGH05]). Mais aucune ne s'impose naturellement à nos yeux, car autant la notion de triade ou plus généralement de cycle est significative dans le cas monoparti, autant un clustering biparti ne semble pas être indispensable. En effet, l'information relative serait la répétition d'interactions entre acteurs, qui peut-être saisie plus directement à l'aide de la mesure suivante.

c. Répétitions d'interactions et redondance

Pour évaluer à quel point un acteur tend à réitérer ses interactions avec un autre, nous pouvons mesurer le **nombre d'interactions répétées (ir)** d'un nœud y_1 (en pratique un événement) est défini comme le nombre de paires de voisins dans le graphe biparti qui sont associés entre eux par un autre nœud que y_1.

Soit formellement:

$$\mathbf{ir}(y_1) = |\,\{x_1, x_2\} \subset \mathcal{V}(y_1) \;\text{ tq }\; x_1 \neq x_2 \text{ et } \exists\; y_2 \neq y_1 \text{ et } (x_1, y_2) \in E \text{ et } (x_2, y_2) \in E\,|$$

Où $\mathcal{V}(y_1)$ désigne l'ensemble des voisins (dans le graphe biparti) de y_1.

Cette définition est en fait très proche du **coefficient de redondance** proposé dans [LMDV08], la différence étant que ce dernier est normalisé au nombre total de paires de voisins de y_1. On peut aussi interpréter ces grandeurs comme une mesure de la perte d'information entre les données biparties et leurs projections monoparties.

La base de données *Notre-Dame* précédemment évoquée a une structure bipartie sous-jacente (acteurs et films dans lesquels ils figurent), nous traçons dans la Figure 1.12 la distribution des répétitions d'interactions associée, dont nous pouvons remarquer le caractère *heavy-tailed*.

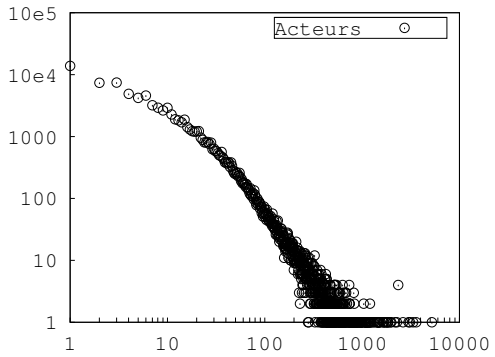


FIG. 1.12: Allure de la distribution des **ir** par événement pour la base *Notre-Dame*.

remarque : Cela peut paraître contradictoire d'examiner les évenements pour mesurer des répétitions d'interactions entre acteurs: il serait en effet possible de dénombrer par exemple les voisins d'un acteur dans la projection pour accéder à ce type d'information. La raison de ce choix est que la mesure du degré dans la projection nécessite d'établir la liste des évenements auxquels participent un acteur et de déterminer tous les autres acteurs qui y prennent part, cette opération est coûteuse algorithmiquement. À l'inverse, $\mathbf{ir}(y)$ ne demande que d'examiner les événements des acteurs participant à y, la complexité est donc réduite d'un facteur $\overline{\Delta}$, avec Δ le nombre d'événements auquel participe un acteur.

1.3.4 Mesures sur les projections

La plupart des mesures usuelles étant menées sur des graphes monopartis, il est légitime de les mettre en œuvre sur les projections et en particulier celle des acteurs. Mais la structure d'événements sous-jacente induit certaines spécificités que nous décrivons ici.

a. Motifs séquentiels ou structurels

Par construction, la projection monopartie d'un réseau social présente des cliques en grandes quantités, par exemple un événement rassemblant dix agents crée de manière mécanique autant de triangles qu'il y a de combinaisons à 3 éléments parmi 10 soit $C_3^{10} = 120$. Dans [New03c, NP03], les auteurs justifient par un argument de ce type les taux de clustering très élevés observés dans les réseaux sociaux.

Il semble alors naturel de distinguer les motifs produits par cet effet mécanique de ceux qui proviennent de corrélations à plus longue distance. On trouve d'ailleurs dans quelques articles cette distinction dans le cas des triangles (e.g. [LA05]), nous la généralisons à tout motif cyclique [TCR08]: un cycle sera dit **structurel** s'il provient d'un événement unique, **séquentiel** s'il est créé par une suite d'événements.

Un triangle peut être ainsi produit par deux types de combinaisons d'événements que nous représentons sous leur forme hypergraphique:

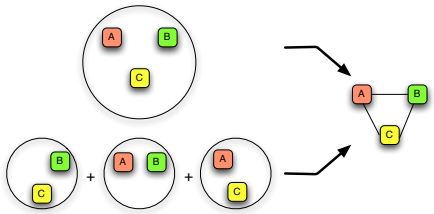


FIG. 1.13: Séquences d'événements produisant des triangles dans la projection.

Et les combinaisons possibles sont d'autant plus nombreuses que la taille des cycles augmente, par exemple à quatre nœuds:

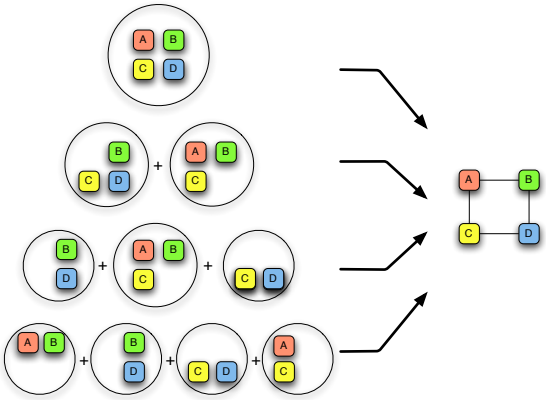


FIG. 1.14: Séquences d'événements produisant des cycles de taille 4 dans la projection.

Les définitions précédentes n'étant pas *a priori* exclusives l'une de l'autre, nous choisissons la convention où un cycle qui peut être généré par un événement unique sera seulement structurel.

Les réseaux à structure d'événements ont des caractéristiques qui peuvent être différentes vis-à-vis de cette propriété, même au sein d'une "famille" de graphe, comme le montrent quelques mesures sur les bases de données collaboratives suivantes* (c.f. Table 1.3):

– *AIO Europe* et *arXiv* sont des graphes de collaborations scientifiques,

– *TheyRule* et *DucthElite* regroupent les membres d'institutions économiques ou politiques,
– *Wiki(d)* est un graphe de discussions entre collaborateurs de *Wikipédia*, *Debian* un forum.

	arXiv	*AIO Europe*	*TheyRule*	*DutchElite*	*Wiki (d)*	*Debian*
△	$17{,}8 \cdot 10^3$	5.365	$110{,}5 \cdot 10^3$	$200{,}0 \cdot 10^3$	$69{,}0 \cdot 10^3$	$234{,}0 \cdot 10^3$
dont structurels	$16{,}3 \cdot 10^3$	5.318	$110{,}3 \cdot 10^3$	$200{,}0 \cdot 10^3$	$66{,}1 \cdot 10^3$	$102{,}2 \cdot 10^3$
proportion	0,92	0,99	1,00	1,00	0,95	0,44
◇	$43{,}5 \cdot 10^3$	8.979	$931 \cdot 10^3$	$608 \cdot 10^3$	$2{,}83 \cdot 10^6$	$17{,}3 \cdot 10^6$
dont structurels	$15{,}4 \cdot 10^3$	7.509	$905 \cdot 10^3$	$601 \cdot 10^3$	$2{,}45 \cdot 10^6$	$1{,}23 \cdot 10^6$
proportion	0,35	0,84	0,97	0,99	0,87	0,07
⬠	$159{,}8 \cdot 10^3$	16.747	$8{,}70 \cdot 10^6$	$277 \cdot 10^6$	$134 \cdot 10^6$	$1.391 \cdot 10^6$
dont structurels	$13{,}0 \cdot 10^3$	10.512	$8{,}10 \cdot 10^6$	$273 \cdot 10^6$	$105 \cdot 10^6$	$20.7 \cdot 10^6$
proportion	0,08	0,63	0,93	0.99	0,78	0,014

TAB. 1.3: Proportion des cycles d'origine structurelle dans des bases de données réelles.

Si la grande majorité des triangles sont structurels, les effets de corrélations à plus longues portées peuvent être très peu visibles ou au contraire dominants selon les graphes: nous le mesurons sur les autres motifs cycliques (mais une distinction équivalente peut bien sûr être employée sur d'autres types de motifs locaux).

b. Corrélations de degré dans les projections de graphes bipartis

Newman et Park [NP03] suggèrent que la structure de groupe qui sous-tend le réseau social puisse rendre compte des corrélations de degrés observées entre agents. En effet, si on lie avec une forte probabilité les acteurs d'une même communauté, cela induit mécaniquement une certaine homogénéité du degré dans celle-ci.

Dans le même ordre d'idée, l'effet de *rich-club* (c.f. 1.2.3.b.) pourrait être compris comme un artefact de la structure bipartie sous-jacente. Prenons un exemple extrême pour observer qualitativement un effet de ce type: la base de données de collaborations scientifiques *CERN** présente la particularité de contenir un petit nombre d'événements de très grandes tailles. Nous procédons alors comme dans [ZM03]:
– les acteurs sont classés par degré décroissant dans la projection; en cas d'égalité l'ordre est choisi aléatoirement,
– on évalue la fonction de *rich-club* $\phi(x)$ en fonction du classement (ramené au nombre de nœuds): le nombre de liens partagés par le nœud de classement x avec les nœuds de classement inférieur, que l'on normalise au nombre total de liens possibles entre l'ensemble de nœuds considérés, $\phi(x)$ est donc d'autant plus proche de 1 que les acteurs de degrés supérieurs à la valeur associée à x sont connectés entre eux.

Et nous comparons alors les courbes obtenues pour la base de données complète et cette même base à laquelle on retire arbitrairement les 25 plus grands événements (ce qui laisse dans ce cas précis le nombre total d'acteurs inchangé):

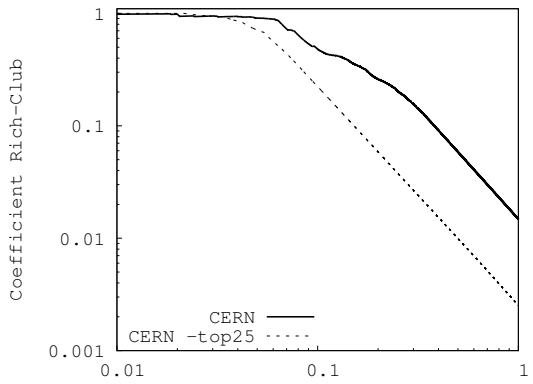


FIG. 1.15: Cœfficient de rich-club en fonction du rang normalisé pour la base *CERN*.

Nous constatons (Fig. 1.15) que cette modification affecte considérablement l'allure de la fonction de *rich-club*: entre les deux courbes, le dénominateur est identique, alors que le numérateur sera plus grand d'un facteur quasiment 10 aux grands classements[10].

1.4 Conclusion et définition des objectifs

Jusqu'à maintenant, nous avons présenté la gamme d'outils dont nous disposons pour analyser les réseaux d'interactions et ce pour quoi ils sont employés, notamment:

- identifier des caractéristiques typiques des réseaux,
- puis tester des mécanismes d'interactions qui en rendraient compte.

Réaliser ces tâches serait une avancée importante dans la compréhension de la structure du réseau social.

Cet objectif, aussi fascinant soit-il, restera pourtant une perspective à long terme dans tout ce travail. En effet, les graphes que nous observons sont le résultat d'un enchevêtrement de décisions individuelles ou collectives, d'événements "accidentels", de rétroactions de la communauté sur les acteurs par le biais d'institutions, de règlements et plus généralement de l'ensemble des conditions sociales qui constituent l'environnement de chacun. Bref, la nature extrêmement complexe des systèmes que nous étudions dissuade de proposer des mécanismes explicatifs sans plus de méthode.

[10]On mesure en effet que le nombre de liens dans la projection des acteurs est de l'ordre de 10^6 pour *CERN*, contre 10^5 pour la version sans les événements géants.

Cependant, il n'est pas exclu de trouver un ordre, une théorie sous-jacente à ces comportements d'apparence chaotique. Notre hypothèse de travail est qu'un nombre réduit d'éléments suffit à rendre compte des traits saillants du réseau observé. En effet, si à l'échelle de l'individu l'ensemble des éléments qui gouvernent les processus d'interactions est immense, en moyennant sur une population suffisamment importante, on perd de vue ces finesses de comportements au profit de grandes tendances qui émergent globalement et que nous cherchons à identifier.

Nous nous proposons alors de poser le problème méthodiquement: le réseau réel apparaît comme un ensemble de mesures topologiques interdépendantes, dont nous pensons que quelques caractéristiques seulement sont significatives. Il faut donc identifier les corrélations: voir quel est l'effet de la modification d'une variable sur le reste de la topologie. Nous avons donc besoin de graphes de référence, vérifiant un ensemble de propriétés choisies, auxquels nous pourrions comparer le graphe réel.

Autrement dit, nous souhaitons situer le réseau relativement à des points de repère dont les caractéristiques sont bien identifiées, voire dont on connaît des processus de génération. En conclusion, nous voulons **disposer d'un moyen simple et efficace de réaliser un balisage de l'environnement du réseau réel** dans l'espace des graphes.

Chapitre 2

Une méthode de génération de graphes synthétiques

Sommaire

Pour réaliser un balisage de l'espace, nous cherchons à produire des graphes d'une manière qui soit à la fois **modulable**, **efficace** et **statistiquement rigoureuse** et comparer leur structure à celle du réseau réel. Dans ce but, nous effectuons d'abord un tour d'horizon des familles de graphes décrites dans la littérature, particulièrement celles qui se rapprochent au plus de nos préoccupations et les procédures employées pour les générer. Nous proposons ensuite une généralisation méthodologique qui sera au cœur de la thèse. Dans un premier temps, nous en décrivons le principe, les potentialités et les limites. Ensuite, nous donnons quelques exemples élémentaires d'applications à la génération de graphes synthétiques.

2.1 Familles traditionnelles

2.1.1 Graphes d'Erdős-Rényi

La classe des graphes d'Erdős-Rényi est probablement celle sur laquelle la littérature est la plus abondante, à commencer par les articles originaux: [ER59, ER60, ER61]; dans [Bol01], l'auteur couvre une partie importante de leurs propriétés. Ils sont définis comme des graphes simples dont le nombre de nœuds et de liens sont fixés, et les liens sont répartis aléatoirement entre les nœuds; on peut donc les produire numériquement sans difficulté et les étudier analytiquement.

Mais dans le contexte qui nous occupe, la comparaison aux graphes d'Erdős-Rényi s'avère rapidement limitée car ils ne présentent que peu de points communs avec les graphes de réseaux réels: le diamètre des composantes est certes faible (typiquement en $ln(N)/ln(\overline{\delta})$), mais la distribution de degré suit une loi binomiale, et n'est donc pas *heavy-tailed*. Aux faibles densités qui correspondent aux cas réels, le graphe n'atteint pas le seuil qui permettrait de voir apparaître une composante connexe géante dans le graphe; leur clustering (en $\overline{\delta}/N$) est souvent très inférieur à celui des graphes observés etc. En bref, ils ne rendent compte d'aucune propriété résultant de corrélations entre les agents du graphe.

2.1.2 Vers des graphes de réseaux complexes

L'expression "réseau complexe" se répand à la fin des années 90, lorsque des chercheurs proposent les premiers modèles présentant des caractéristiques communes avec les graphes de réseau observés dans la nature.

a. Les petits mondes revisités

Watts et Strogatz [WS98] introduisent ainsi une famille de modèles simples, qui permettent de retrouver en ordre de grandeurs des propriétés de petits mondes. Le principe en est le suivant (c.f. Fig. 2.1):

- le graphe de départ est régulier, avec des nombres de nœuds et de liens fixés,
- avec une probabilité p, on déconnecte une extrêmité d'un lien pour la reconnecter à un nœud choisi aléatoirement,
- on examine les caractéristiques du graphe en faisant varier le paramètre p de 0 à 1 (entre les cas extrêmes: régulier et aléatoire).

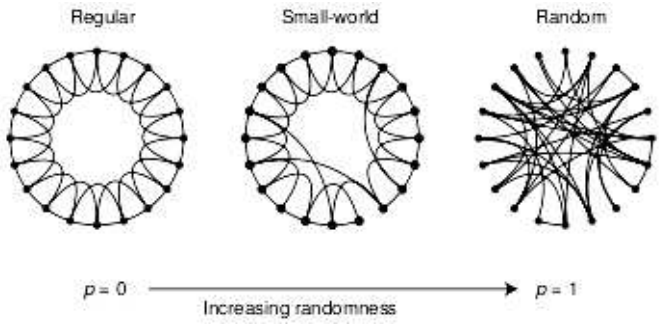


FIG. 2.1: Modèle petit monde de Watts et Strogatz. Extrait de [WS98].

Cela a notamment pour effet de diminuer la distance moyenne entre deux nœuds aux valeurs intermédiaires de p, tout en conservant une partie du clustering de la structure régulière, on simule ainsi le petit monde tel que décrit dans 1.2.4.b. Ce modèle ne rend toutefois pas compte d'autres caractéristiques présentées par les réseaux réels, telle que l'allure de la distribution de degré.

b. Attachement préférentiel

Peu après, Barabási et Albert [BA99] présentent un mécanisme dynamique de croissance du réseau basé essentiellement sur l'idée que les nouveaux entrants s'attachent préférentiellement aux nœuds existants les plus connectés.

Ce modèle est inspiré de ce que l'on qualifie dans le contexte des réseaux sociaux "d'effet Matthieu". Cette idée introduite par Merton [Mer68][1], visait à résumer l'observation selon laquelle, dans les milieux scientifiques, l'attention et la reconnaissance accordées croissent avec la notoriété de l'individu qui en fait l'objet. Ici, cela signifierait que de nouveaux agents du réseau s'attachent préférentiellement à ceux

[1]Par allusion au passage de l'Évangile selon saint Matthieu: "Car on donnera à celui qui a, et il sera dans l'abondance, mais à celui qui n'a pas on ôtera même ce qu'il a."

étant déjà fortement connectés; cette tendance a d'ailleurs effectivement été mesurée sur Internet [PSVV01].

En pratique, ce modèle (que nous noterons BA) est très simple à mettre en œuvre:
- initialement le graphe compte un petit nombre de nœuds,
- à chaque itération, on ajoute de nouveaux agents en les liant à un nombre fixe de nœuds existants, choisis selon une loi de probabilité proportionnelle au degré.

Cette procédure permet de produire des graphes dont la distribution de degré suit une loi de puissance.

Une telle propriété est en fait établie de longue date pour expliquer la banalité des distributions statistiques *scale-free* que nous discutions en 1.2.3.a.; l'hypothèse se ramène en effet au modèle général de Simon [Sim56] où la probabilité d'un phénomène - ici l'attachement d'un nœud - est proportionnelle au nombre de ses occurrences passées.

c. Prolongements et limites

Ces deux familles ont donné lieu a un grand nombre de développements: études analytiques [BW00, DM03], élargissements aux graphes dirigés [ACL01] et bipartis [PCMG07]. Des variations sur le concept d'attachement préférentiel suggèrent des mécanismes reposant non pas selon le critère du degré, mais sur d'autres caractéristiques du réseau, topologiques ou non: évaluation du caractère attractif de l'acteur [DMS00], relations communes [JGN01], âge des nœuds [HAB07], proximité sémantique des contenus éventuels [RC09] etc. L'article de revue [AB02] décrit les premiers développements du modèle BA.

Cependant, ces classes de graphes ne sont pas suffisantes pour justifier les détails topologiques d'un réseau réel, en particulier pour expliquer des mesures aussi complexes que l'abondance des motifs, les structures communautaires ou l'assortativité.

2.2 Graphes à distribution de degré fixée

Si l'on souhaite imposer "un peu plus" que les contraintes d'Erdős-Rényi, un choix raisonnable consiste à reproduire la distribution de degré des réseaux réels - ce qui est d'ailleurs l'intention du modèle BA [BA99], même si celui-ci a également une portée explicative. Mais les distributions de degré réelles n'étant pas systématiquement assimilables à des lois de puissance, on voudrait disposer de graphes satisfaisant d'autres types de distributions. Par la suite nous nous référerons aux graphes vérifiant une certaine distribution comme satisfaisant $\mathbf{C_{min}}$ (pour contrainte minimum), et l'ensemble de ces graphes sera noté $\mathcal{E}_{\mathbf{C_{min}}}$.

D'ailleurs, des recherches récentes sur les graphes de réseaux complexes utilisent largement les graphes synthétiques ayant une distribution de degré fixée comme modèle de référence, que ce soit pour comparer à la structure du réseau réel [NSW01, NP03],

ou pour simuler et mener des études analytiques sur des processus s'y produisant [CEBAH00, MPSV02, BBPSV04].

Cette partie décrit les méthodes existant dans la littérature pour produire des graphes en partant de la distribution réelle des degrés, en insistant sur le cas particulièrement étudié des graphes simples. Nous évoquons brièvement les plus courantes, puis nous détaillons la méthode d'échange utilisée par la suite.

2.2.1 Quelques méthodes usuelles

a. S'ajuster à la distribution de degré

Dans cette classe de modèles, on suit une procédure d'assemblage stochastique qui génère un graphe dont le degré tend vers une certaine distribution de probabilité, on parle alors de distribution de degré attendue (par opposition à une distribution de degré exacte). On trouve dans [CL02] un modèle très standard de ce type.

Il existe dans la lignée de BA d'autres modèles de graphes dynamiques, utilisant des mécanismes de croissance ou de réarrangements des liens, par exemple pour les cas spécifiques du web [KRR$^+$00] ou de réseaux de collaborations [CCP04]. On trouve un exemple d'un genre un peu différent dans [BDML06]: en construisant un graphe orienté dont on élimine l'orientation puis fusionne les liens multiples, on obtient une distribution choisie convoluée à une loi de Poisson.

Après tirage d'une suite de degré reproduisant la distribution attendue, la génération peut être effectuée en affectant un degré de cette suite à chaque nœud puis en réalisant un assemblage qui respecte les degrés attribués.

b. Reproduire la suite de degré

Nous nous intéressons maintenant au cas suivant: la suite des entiers $\{\delta_i\}$ telle que δ_i est le degré du nœud d'index i est supposée connue, et nous cherchons à la reproduire exactement.

> **remarque :** Dans cette situation, il est habituellement nécessaire de savoir si la suite de degré examinée est graphique ou non, autrement dit existe-t-il au moins un graphe réalisant cette suite? Ce n'est par exemple pas le cas de la suite $\{3, 2, 2\}$ dans le contexte des graphes simples. Il existe dans la littérature des théorèmes pour trancher la question de l'existence (en particulier, le théorème d'Erdős-Gallai [EG60]) ou permettant de générer un tel graphe (th. d'Havel-Hakimi [Hak62]). Dans la perspective de notre travail, cela ne sera pas nécessaire d'y recourir: nous reprendrons les caractéristiques d'un graphe réel G_0, la suite de degré sera donc trivialement graphique.

Un algorithme populaire pour produire des graphes selon une distribution de degré précise consiste à attribuer une "moitié" de lien *(stub)* à un nœud pour chacun des

liens auquel il participe, puis à joindre les moitiés de liens aléatoirement; il s'agit donc d'un processus d'association (on parlera de *matching* ou *filling algorithm*). Celui-ci, connu depuis les années 70 [BC78] puis affiné et réactualisé [MR95, Bol01], est désigné en général par le terme de ***configuration model*** [New03a].

Il pose néanmoins quelques difficultés: un tel processus crée naturellement des boucles ou des liens multiples - en particulier avec des distributions de degré *heavy-tailed*; dans le cas très répandu des graphes simples, on doit contourner le problème moyennant des évolutions du modèle qui éliminent les structures de ce type (e.g. [CBPS05, BDML06]). L'algorithme proposé précédemment par McKay et Wormald [MW90], bien que nécessitant des conditions plus strictes sur la distribution de degré, préfigure les solutions de ce type.

Parmi d'autres algorithmes de génération de graphes simples à distribution de degré fixée, *"go with the winners"* [MKI+03] ou récemment celui proposé dans [DGKTB10], sont des méthodes permettant d'obtenir un échantillon où chaque graphe apparaît avec la même probabilité moyennant une pondération des graphes produits.

Toute les méthodes présentées jusqu'ici sont dites de *matching*: elles cherchent à générer le graphe satisfaisant une certaine distribution de degré "à partir de rien", dans le sens où l'on propose une procédure d'assemblage à partir de nœuds déconnectés les uns des autres. Nous adoptons maintenant la posture inverse: libérer progressivement des contraintes à partir d'un graphe réel afin de nous en éloigner de manière contrôlée.

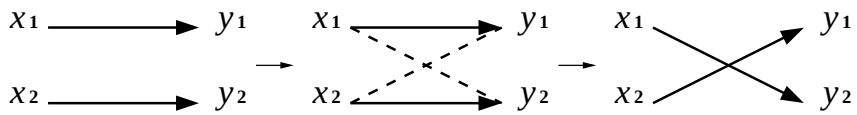
2.2.2 Méthode d'échange

Il existe une littérature importante autour des **méthodes d'échange** (*switching*, *rewiring* ou *swapping methods*), qui font partie de la grande famille des méthodes Monte-Carlo à base de chaînes de Markov (MCMC). Nous allons étudier en détails cette classe d'algorithmes: ses principes, le vocabulaire associé et son utilisation pratique; d'une part parce qu'elle est employée et discutée pour générer des graphes dont la distribution de degré est fixée (entre autres: [RJB96, Rob00, New02, MKI+03]), d'autre part parce qu'elle est à l'origine de celle que nous proposons par la suite.

a. Principe

Processus markovien. Le processus que nous décrivons est une **chaîne de Markov**: un processus stochastique discret $\mathcal{M}$ dont l'état à l'instant $t+1$ ne dépend que de l'état à l'instant t. L'étape élémentaire du processus est un échange entre les extrêmités de deux liens du graphe. Nous la représentons schématiquement dans le cas orienté[2]:

[2]Ce cas est plus explicite que le non-orienté car il distingue naturellement les deux extrêmités des arcs, mais le principe est identique pour des liens.



Et du point de vue de la matrice d'adjacence (ou d'affiliation) du graphe:

- pour les graphes non-pondérés, cette étape élémentaire peut être vue comme une permutation d'éléments, placés en configuration de "rectangle alterné" *(check-boarder units)*:

$$\begin{pmatrix} & \vdots & & \vdots & \\ \cdots & 1 & \cdots & 0 & \cdots \\ & \vdots & & \vdots & \\ \cdots & 0 & \cdots & 1 & \cdots \\ & \vdots & & \vdots & \end{pmatrix} \longrightarrow \begin{pmatrix} & \vdots & & \vdots & \\ \cdots & 0 & \cdots & 1 & \cdots \\ & \vdots & & \vdots & \\ \cdots & 1 & \cdots & 0 & \cdots \\ & \vdots & & \vdots & \end{pmatrix}$$

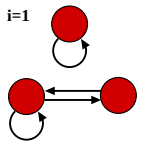
- pour les multigraphes, on ajoute ou retranche une unité aux éléments, par exemple:

$$\begin{pmatrix} & \vdots & & \vdots & \\ \cdots & 4 & \cdots & 1 & \cdots \\ & \vdots & & \vdots & \\ \cdots & 0 & \cdots & 7 & \cdots \\ & \vdots & & \vdots & \end{pmatrix} \longrightarrow \begin{pmatrix} & \vdots & & \vdots & \\ \cdots & 3 & \cdots & 2 & \cdots \\ & \vdots & & \vdots & \\ \cdots & 1 & \cdots & 6 & \cdots \\ & \vdots & & \vdots & \end{pmatrix}$$

La chaîne $\mathcal{M}$ est l'itération de cette étape élémentaire. Nous choisissons conventionnellement de ne pas considérer l'identité comme un échange de ce type. Par ailleurs, nous constatons qu'une telle méthode **conserve nécessairement la distribution de degré**, puisque le degré de chaque nœud n'est jamais modifié.

Description matricielle. De tels processus markoviens discrets sont représentés à l'aide de leur **matrice de transition M**, dont les éléments m_{ij} sont la probabilité de passer de l'état i à l'état j au cours d'un pas de la marche. Chaque ligne (ou colonne) représente alors un des éléments de l'ensemble dans lequel la processus est réalisé, en l'occurrence un graphe (ou la matrice d'adjacence qui lui est associée).

Illustrons cette idée sur l'exemple suivant d'un graphe simple orienté, acceptant les boucles, et dont le degré de chaque nœud, écrit sous la forme (degré entrant , degré sortant), satisfait la distribution $\{(1,1);(1,1);(2,2)\}$:



Nous effectuons des échanges entre les extrêmités des arcs permettant de rester dans $\mathcal{E}_{\mathbf{C_{min}}}$ (l'ensemble des graphes satisfaisant la distribution). Nous qualifions d'échanges

effectifs ceux qui respectent cette condition autres que l'identité, par exemple l'échange représenté sur la Figure 2.2.

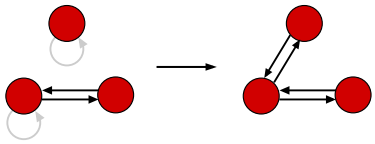
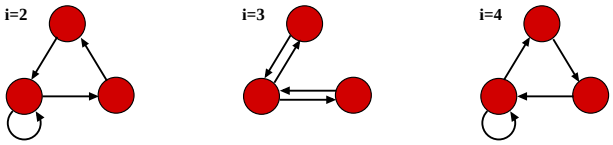
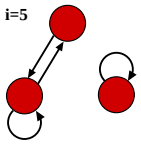


FIG. 2.2: L'échange des extrêmités des arcs grisés de $i = 1$ produit le graphe de droite ($i = 3$). Le graphe obtenu respectant la contrainte $\mathbf{C_{min}}$, l'échange est effectif.

Les trois échanges effectifs possibles amènent, chacun des trois graphes suivants (dont la probabilité sera alors de 1/3):



En effectuant des échanges à partir l'un de ces trois graphes, on découvrirait l'existence d'un cinquième élément dans cet ensemble $\mathcal{E}_{\mathbf{C_{min}}}$:



Et une analyse plus poussée révélerait qu'il n'y en a pas d'autre. D'où la première ligne de la matrice de transition $\mathbf{M}$: $\begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}$. Cela traduit qu'en effectuant un échange sur $i = 1$ entre deux arcs choisis aléatoirement dans l'ensemble E des arcs du graphe, on a une probabilité 1/3 d'atteindre les graphes $i = 2, 3$ ou 4 et nulle d'atteindre 5.

b. Convergence de la chaîne

Mesure stationnaire. Toutes les chaînes de Markov que nous considérons, et en particulier celle des échanges simples que nous venons de décrire, vérifierons les deux propriétés suivantes:

- **Irréductible** (*irreducible*) ou **ergodique**, signifie que tout élément de l'ensemble est accessible par le processus (depuis un point quelconque de l'ensemble).

  Plus formellement, un élément x_j est dit accessible depuis x_i si:

$$\exists \tau \geq 0 \text{ tel que } m_{ij}^{(\tau)} > 0$$

où $m_{ij}^{(\tau)}$ désigne l'élément (i, j) de la matrice $\mathbf{M}^\tau$. Comme m_{ij} représente la probabilité de transiter de x_i vers x_j en une itération, la composante (i, j) de la matrice $\mathbf{M}^\tau$ est la probabilité d'être passé de x_i à x_j après τ itérations. Par construction, cette propriété sera toujours satisfaite pour l'ensemble des graphes accessibles par $\mathcal{M}$ depuis G_0.

- **Récurrente positive** *(positive recurrent)*: cette propriété signifie que l'espérance du temps de premier retour à chaque état est non-nulle et finie.

  Le temps T_i de premier retour à un état x_i est la borne inférieure des valeurs de $\tau \geq 1$, telle que après τ itérations de la chaîne, on soit revenu en x_i en partant de cet état (si une telle valeur n'existe pas, $T_i = \infty$).

  Dans le cas qui nous occupe, elle est vérifiée puisque l'ensemble décrit lui-même est fini et que chacun de ses éléments est effectivement accessible.

Une chaîne qui vérifie ces propriétés admet une **mesure stationnaire** π unique:

$$\exists!\ \pi \text{ tel que } \Sigma_i \pi_i = 1 \text{ et } \pi\mathbf{M} = \pi$$

Les composantes de ce vecteur normalisé s'interprètent comme la fraction du temps passé asymptotiquement dans chacun des états de l'ensemble (e.g. [Yca02]).

Apériodicité. Bien que cette condition ne soit pas indispensable pour définir la mesure stationnaire, on comprend mieux cette propriété en regardant le cas particulier des chaînes **apériodiques** (*aperiodic*).

La période de l'état x_i sera définit comme le PGCD de l'ensemble Λ_{ii}, avec:

$$\Lambda_{ij} = \{\tau \geq 0 \text{ tel que } m_{ij}^{(\tau)} > 0\}$$

Si celle-ci vaut 1, la chaîne sera dite apériodique, cette propriété peut donc être facilement garantie dans les chaînes que nous construisons en autorisant avec une probabilité non-nulle de rester dans le même état.

Une chaîne apériodique, irréductible et récurrente positive est telle que le produit matriciel $\mathbf{M}^\tau$ converge:

$$\exists \mathbf{W} \text{ tel que } \lim_{\tau \to \infty} \mathbf{M}^\tau = \mathbf{W}$$

On montre alors que chacune des lignes de la matrice $\mathbf{W}$ s'identifie à π. La composante w_{ij} sera la probabilité de tendre vers l'état x_j partant de l'état x_i, et celle-ci sera donc la même quel que soit i.

De cette manière, en itérant la procédure, nous allons peu à peu perdre la mémoire du graphe de départ, pour tendre vers un état stationnaire où le graphe x_j de l'ensemble a une probabilité π_j d'être atteint.

c. Irréductibilité: l'ergodicité de la chaîne

Une partie importante de notre réflexion tient à la distinction entre *ensemble décrit* et *ensemble à décrire*, idée qui est au centre de la partie 2.4. En d'autres mots, est-ce que le processus de Markov nous permet effectivement d'atteindre tout élément de l'ensemble dont nous souhaitons réaliser un échantillon?

Or, pour un graphe non-orienté (graphe simple ou multigraphe, avec ou sans boucle), la littérature nous apprend que l'ensemble des graphes satisfaisant une distribution de degré arbitraire est effectivement ergodique au sens de la chaîne de Markov définie précédemment [Egg73, EH78, Tay80].

En revanche, les démonstrations proposées dans ces références ne sont valides que dans le cadre de cette contrainte, leur généralisation à d'autres cas ne semblent pas évidentes, ce qui aura de l'importance dans la suite de notre travail.

d. Garantir l'uniformité

Nous ne voulons pas seulement obtenir un échantillon de graphes de l'ensemble, nous attendons également de celui-ci qu'il soit uniformément aléatoire. Le terme "uniformément" traduit ici l'équidistribution: **tous les graphes de l'ensemble doivent avoir la même probabilité d'être représentés dans l'échantillon**.

Si nous procédons par une simple itération d'échanges entre les destinations d'arcs, l'état stationnaire de la chaîne n'est pas nécessairement équidistribué: l'équidistribution se traduirait par un vecteur π dont toutes les composantes sont égales, ce qui n'a aucune raison d'être le cas en général.

Dans [MP04], les auteurs expliquent comment nous pouvons garantir simplement l'équidistribution à l'aide d'une procédure très voisine qualifiée de méthode *trial-swap*. Il s'agit de ne pas fixer le nombre d'échanges effectifs mais le nombre de tentatives d'échange. On trouve d'ailleurs cette idée sous divers noms, dont *switching & holding* [ARS05]; en effet, on choisit deux liens au hasard, on cherche à effectuer l'échange de destination, si celui-ci ne permet pas d'obtenir un élément de l'ensemble à décrire, le graphe reste dans le même état (*hold*).

Nous donnons dans la sous-partie suivante un argument pour en comprendre la signification, et une démonstration plus formelle en Annexe C. L'algorithme correspondant à ce processus pour des graphes simples orientés est en Annexe D.

e. Métagraphe de l'ensemble des graphes

Définition. Pour en comprendre les enjeux, nous décrivons ici la procédure à l'aide d'un mode de représentation que nous pensons plus intuitif et que nous dénommerons **métagraphe**. Le principe en est simple: le métagraphe représente l'ensemble à décrire, associé à la chaîne de Markov $\mathcal{M}$ considérée. Un (méta)nœud représente un élément de

l'ensemble, un (méta)lien l'existence d'un échange permettant de transiter d'un élément à un autre. On peut alors comprendre simplement la chaîne de Markov comme **une marche aléatoire** dans le métagraphe.

> **remarque :** Il s'agit en fait d'une alternative à la représentation graphique classique des processus de Markov (qui provient de la matrice de transition), la différence étant que nous utilisons des métaliens - éventuellement multiples - plutôt que d'affecter une valeur de probabilité à chaque lien.

Dans le cas qui nous occupe (et tous ceux dont il sera question ensuite), le processus est réversible: s'il existe un lien de G_a vers G_b, il existe un lien de G_b vers G_a. La Figure 2.3 en est l'illustration pour la contrainte $\mathbf{C_{min}}$ sur l'ensemble des graphes de distribution de degré (entrant, sortant): $\{(1,1);(1,1);(2,2)\}$.

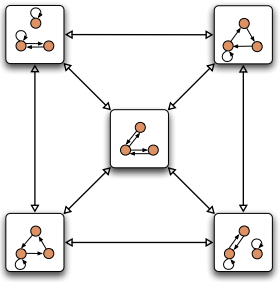


FIG. 2.3: Métagraphe associé à la procédure d'échanges simples.

Interprétation métagraphique du *trial-swap*. Passer des échanges simples aux tentatives d'échanges revient à créer des boucles dans le métagraphe, c'est-à-dire à considérer l'identité comme la réalisation d'une étape élémentaire du processus de Markov.

Or, le nombre de tentatives d'échanges possibles est strictement identique pour tous les graphes de l'ensemble: il s'agit de deux fois le nombre de combinaisons possibles à deux liens dans l'ensemble des liens (le facteur 2 venant du fait qu'on peut réaliser deux échanges différents avec un couple de liens). Comme un échec dans la tentative d'échange est considéré comme un lien du métanœud vers lui-même, le métadegré entrant est le même pour tous - et idem pour le métadegré sortant.

Cela signifie que la probabilité d'arriver ou partir d'un certain graphe au cours de la marche est identique pour tous, Une telle chaîne de Markov correspond donc effectivement à une distribution uniforme, comme on l'a affirmé en 2.2.2.d. Nous illustrons cette idée sur la Figure 2.4.

Dans le cas du processus d'échange simple (métagraphe de gauche), l'état stationnaire π est déterminé par $\pi\mathbf{M} = \pi$ avec

$$\mathbf{M} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$$

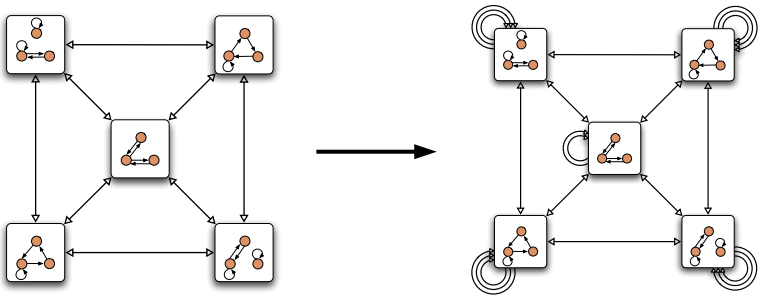


FIG. 2.4: Métagraphes associés aux échanges simples et aux tentatives d'échanges.

On peut vérifier que $\pi = \left(\frac{3}{16}; \frac{3}{16}; \frac{4}{16}; \frac{3}{16}; \frac{3}{16}\right)$ satisfait $\pi\mathbf{M} = \pi$, et l'état stationnaire n'est donc pas équidistribué. En passant au processus de tentatives d'échanges (métagraphe de droite), la matrice de transition devient:

$$\mathbf{M}' = \begin{pmatrix} 3/6 & 1/6 & 1/6 & 1/6 & 0 \\ 1/6 & 3/6 & 1/6 & 0 & 1/6 \\ 1/6 & 1/6 & 2/6 & 1/6 & 1/6 \\ 1/6 & 0 & 1/6 & 3/6 & 1/6 \\ 0 & 1/6 & 1/6 & 1/6 & 3/6 \end{pmatrix}$$

dont l'état stationnaire est équidistribué.

Interprétation métagraphique de l'ergodicité. Le langage intuitif des métagraphes permet de donner un nouveau point de vue sur la propriété d'irréductibilité de la chaîne: elle se traduit par un métagraphe connexe. Le graphe de départ G_0 appartient toujours à une certaine métacomposante connexe, mais est-ce que celle-ci se confond avec l'ensemble à décrire? Nous savons par théorème que la réponse est oui dans le cas de la seule distribution de degré pour les graphes non-orientés, mais l'élargissement à d'autres contraintes ne nous garantit pas de conserver cette propriété.

f. Étiquette des nœuds

Définition. Si l'on souhaite expliciter la description mathématique de la procédure, il faut distinguer graphes étiquetés (*labelised*) et non-étiquetés. Dans le premier cas, chaque noeud est étiqueté de manière à ce que l'on puisse le différencier de tout autre noeud du graphe; dans le second, on ne peut pas distinguer deux nœuds ayant le même environnement topologique. Formulé autrement, les nœuds sont discernables ou non.

Cette définition peut sembler secondaire mais elle importe du point de vue du dénombrement de l'ensemble. En effet, les ensembles étiquetés et non-étiquetés se rapportant à la même distribution de degré ne sont pas statistiquement identiques. On entend par "statistiquement identiques" que toute mesure menée sur l'un ou l'autre ensemble de graphes produit la même distribution de résultats.

En effet, les graphes comme ceux représentés en 2.5 (appartenant à l'ensemble de la figure 2.4), sont équivalents si on retire leur étiquette[3]:

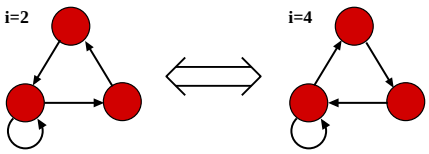


FIG. 2.5: Avec des nœuds indiscernables, les deux graphes sont identiques; cela n'est plus le cas si on attribue à chacun une étiquette.

Parcourir un ou plusieurs étiquetages? Un examen attentif de la procédure d'échange nous indique que nous ne parcourons en fait qu'un certain étiquetage de l'ensemble: le degré attaché à un certain nœud reste identique au cours de la procédure. Mais cela n'a pas de conséquence sur le dénombrement effectué; en effet l'opération qui consiste à associer à un nœud une autre étiquette est trivialement bijective (c'est une permutation de l'ensemble des étiquettes dans lui-même).

Donc quelle que soit la propriété topologique évaluée, la mesure sur les éléments de l'ensemble parcouru est distribuée de la même manière lorsque l'on parcourt un ou plusieurs étiquetages. Et donc parcourir la sous-partie du métagraphe représentative d'un étiquetage est statistiquement identique à parcourir le métagraphe complet.

Dans les cas qui nous occupent, les agents sont individualisés. Nous n'aurons plus besoin par la suite de revenir à cette notion, mais nous garderons en tête l'idée que **tous les algorithmes dont il sera question traitent exclusivement de graphes étiquetés**.

g. Problème de la durée de convergence

L'inconvénient commun aux algorithmes de cette famille tient à la difficulté à définir un critère d'arrêt. Il s'agit de savoir au bout de combien d'itérations on peut considérer que l'état atteint est effectivement quelconque, ce à quoi nous répondons dans cette partie à l'aide d'une méthode expérimentale.

Critère de convergence. Une très vaste littérature est consacrée aux questions de vitesse de convergence des chaînes de Markov et en particulier aux conditions pour obtenir un "mélange rapide" (ou *rapid mixing*, c.f. [JS97, KTV97, BDX04]). Nous ne développons pas l'aspect théorique de cette question, notre réflexion sur ce sujet restera de nature expérimentale.

Nous utilisons alors un critère répandu pour évaluer la convergence [GMZ03, VL05]: on choisit des observables que l'on pense représentatives, puis on mesure au long du

[3]Dans la plupart des représentations, pour plus de lisibilité, l'étiquetage est tacite et attaché à la position du nœud sur le schéma.

processus comment celles-ci vont évoluer depuis les valeurs initiales jusqu'aux valeurs moyennes sur l'ensemble. Lorsque les mesures choisies marquent un plateau, on considère avoir perdu la mémoire du point initial de la chaîne de Markov.

> **remarque :** On peut estimer que l'on est sur le plateau à l'aide d'une variété de critères, qui contiendront une part d'arbitraire. Une possibilité serait que les fluctuations des mesures sur un nombre fixé n d'itérations restent comprises dans un intervalle de $\beta\sqrt{n}\Delta x$, Δx représentant une estimation de la variation élémentaire de la grandeur mesurée et β un facteur ajusté en fonction de la qualité souhaitée. Par exemple, si on dénombre des motifs triangulaires, la variation élémentaire associée serait d'une unité, on ajuste à 0.2 le facteur β; alors, si on n'enregistre les fluctuations durant 100 itérations et que celles-ci n'excèdent pas 2 unités on considère que le palier est atteint.

Exemple. Nous illustrons ce critère sur la génération d'un échantillon de graphes à partir d'un réseau réel de collaborations artistiques* extrait du site *Allmusic*, mettant en jeu 7079 musiciens ayant participé à des projets collectifs. Nous suivons sur la figure 2.6 la génération à l'aide de différentes mesures topologiques dont nous représentons la moyenne (sur 50 réalisations) au cours du processus.

Caractériser la convergence. Nous souhaiterions définir une vitesse de convergence du processus, mais nous allons voir sur l'exemple précédent que cela n'est pas trivial. On sait que la convergence d'une chaîne de Markov irréductible, apériodique, dans un espace des états fini converge géométriquement vers sa mesure stationnaire [Yca02], c'est-à-dire $\exists\ C > 0$ et $0 \leq \alpha < 1$ tels que $\forall t \in \mathbb{N}$,

$$|m_{ij} - \pi_j| < C\alpha^t$$

Il est donc légitime de chercher un modèle de convergence qui soit exponentiel, car dans ce cas on vérifie la propriété précédente. Nous explorons donc expérimentalement cette possibilité:

- Supposons que la convergence suive une loi du type:

  $$f(t) = a(1 - e^{-t/\tau}) + b$$

  la dérivée de la courbe en échelle semi-logarithmique est une droite dont la pente nous fournit le temps τ caractéristique de la convergence.

- Les conditions initiales peuvent jouer un rôle important sur l'allure du processus de marche aléatoire: on peut le voir par exemple sur l'encart de A (Fig. 2.6). C'est pourquoi il peut être nécessaire d'exclure les premières itérations pour décrire l'évolution.

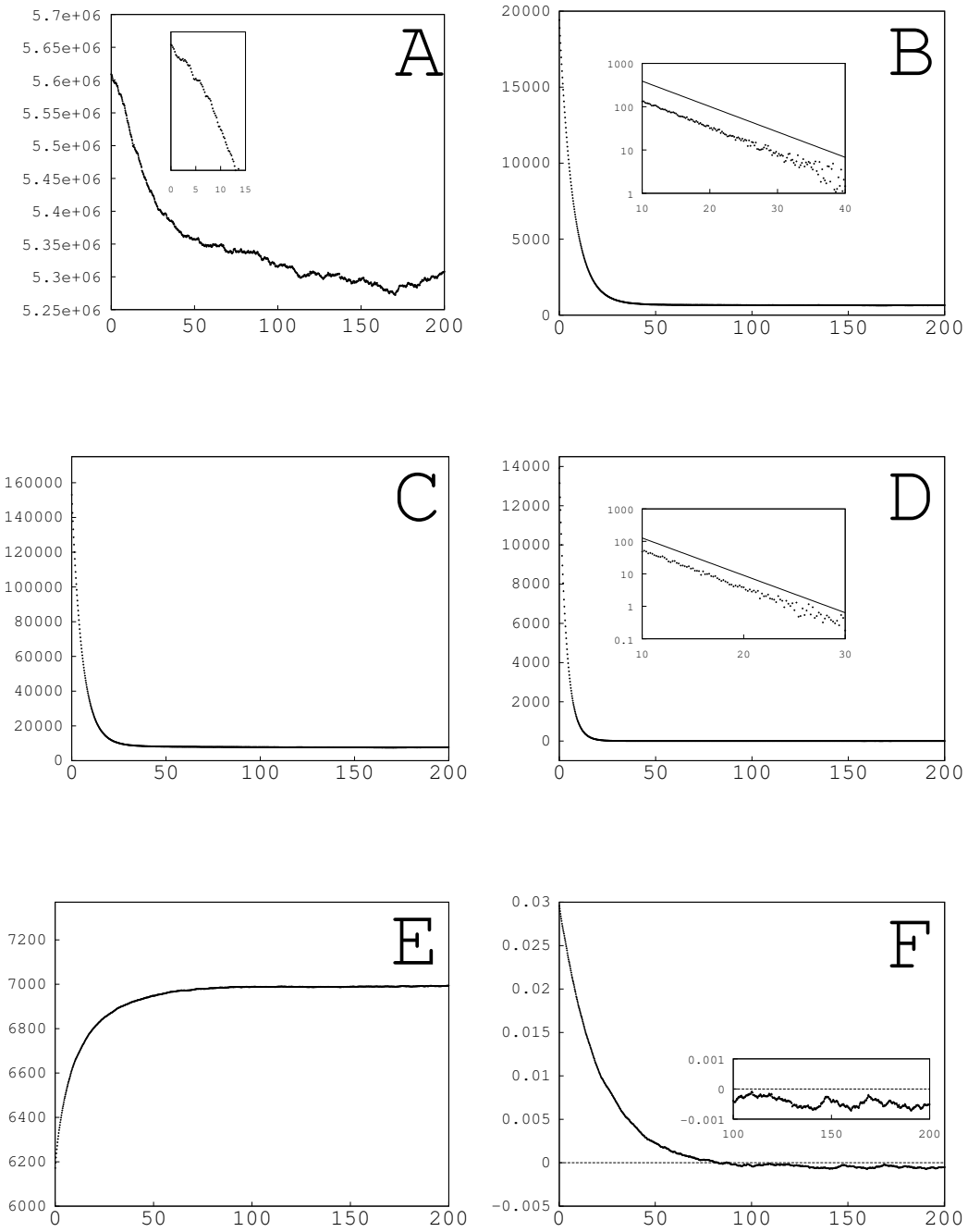


FIG. 2.6: Évolutions en fonction du nombre de tentatives d'échanges (en milliers) du nombre de chemins de longueur 3 (A); de cycles de taille 3 (B); de cycles de taille 4 (C); de cliques de taille 4 (D); de la taille de la composante géante (E); de l'assortativité (F).

- L'encart de la mesure B représente la valeur absolue de la dérivée de la mesure des cycles de taille 3, pour laquelle un modèle exponentiel semble satisfaisant, dans ce cas $\tau_{\mathrm{B}} \simeq 7000$ (tentatives d'échanges). Alors que la mesure D (cliques de taille

4) est aussi bien modélisée par une exponentielle, mais son temps caractéristique est inférieur $\tau_{\mathrm{D}} \simeq 4000$.

Le temps caractéristique de la convergence est donc variable en fonction des mesures considérées. En fait des courbes d'autres situations de convergence montreraient que le type de fonction qui décrit l'évolution varie aussi selon les contraintes et les graphes, peut-être même selon les mesures. Et le modèle exponentiel peut alors ne pas être approprié (nous verrons un exemple concret en 3.1.3).

Nous emploierons par la suite le terme de vitesse de convergence selon un point de vue pratique: il s'agira du **nombre caractéristique d'itérations nécessaire pour pouvoir affirmer que l'on a atteint le plateau de convergence**. Remarquons que celui-ci est également très variable: entre D et F par exemple, nous estimons un rapport de l'ordre de 10.

Fluctuations. D'autre part malgré le moyennage, certaines mesures restent très fluctuantes, c'est ici le cas du nombre de chemins de longueur 3 (courbe A). Une proposition pour remédier à ce problème, consiste à effectuer une moyenne temporelle cumulée de la mesure [AF01, GMZ03]: plutôt que de tracer la mesure $m(t)$, on trace alors $\frac{1}{t+1}\sum_{i=0}^{t} m(i)$.

Un inconvénient étant que l'observation de la convergence nécessite l'itération sur un nombre supérieur de pas de temps. Une solution intermédiaire consiste alors à effectuer une moyenne cumulée glissante dans le temps: pour un lissage sur g valeurs, on tracera au pas de temps t: $\frac{1}{g}\sum_{i=t-g}^{t} m(i)$ si $t \geq g$. Sur l'exemple de la mesure A on obtient la figure 2.7.

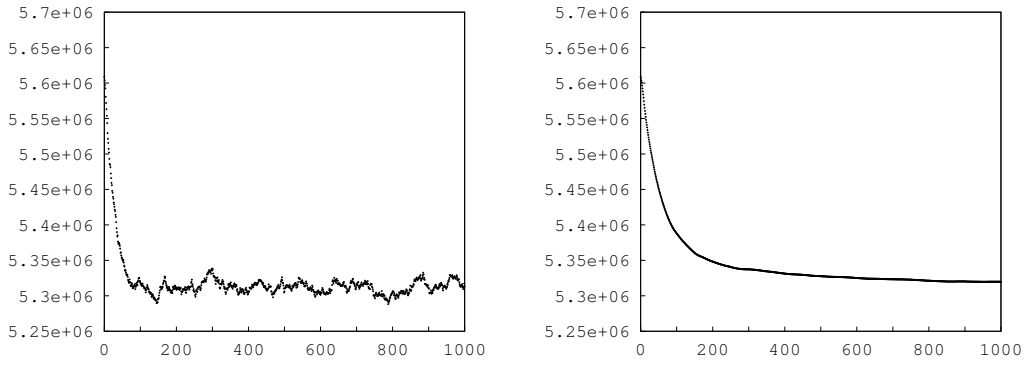


FIG. 2.7: Sur 50 réalisations, nombre de chemins de longueur 3 en fonction du nombre de tentatives d'échange (en milliers). Gauche: moyenne simple. Droite: moyenne cumulée.

remarque : Si nous examinons plus attentivement la mesure d'assortativité (F), nous remarquons que les graphes produits, non-corrélés par construction, ont une assortativité qui ne tend pas pour autant vers 0. Ici, cette mesure tend vers une valeur négative (certes, petite en valeur absolue devant 1). Cette observation est en accord avec l'idée d'abord conjecturée [NP03] puis démontrée dans [JTMM10], dont nous avons fait état en 1.2.3.b.: la moyenne statistique de l'assortativité sur cet ensemble est souvent négative.

2.3 Méthodes pour d'autres contraintes

Nous avons exposé des méthodes classiques pour générer des graphes vérifiant une distribution de degré déterminée. Est-il possible d'employer l'une d'entre elles pour réaliser notre objectif plus général: produire des graphes aléatoires satisfaisant des propriétés quelconques? Nous expliquons dans cette partie pourquoi les méthodes exposées précédemment ne suffisent plus.

2.3.1 Alternatives aux méthodes d'échange

Dans cet objectif, les algorithmes de *matching* ne sont - sauf exception - pas adaptés. D'abord parce qu'ils ne produiraient pas *a priori* un échantillon uniformément aléatoire; et surtout parce qu'ils ne correspondent pas à une procédure standardisée: il faut la réinventer pour chaque contrainte envisagée. Et, lorsque que cette dernière devient exigeante, il est difficile d'imaginer une procédure qui ne reste pas bloquée au cours de la construction du graphe.

Dans certains cas cependant, nous disposons de moyens: récemment, Newman a proposé une forme généralisée du *configuration model* qui permet de fixer également le clustering global [New09], elle serait potentiellement utilisable pour imposer d'autres motifs.

D'autre part, les méthodes d'ajustement sur une distribution de probabilité du type décrit en 2.2.1 (e.g. [CL02]) peuvent être généralisée dans une certaine mesure, par exemple pour générer des graphes satisfaisant asymptotiquement une distribution de corrélations de degré donnée [Dor03].

La classe des modèles à "variables cachées" *(hidden variables)* est également utilisable dans ce but [CCDLRM02, BPS03]; elle consiste à attribuer à chaque nœud une certaine variable distribuée de manière arbitraire - souvent appelée *fitness*, puis à connecter les nœuds entre eux selon la valeur de celle-ci. Avec des choix adéquats des lois de distribution et d'attachement, il est possible de générer des graphes ayant les distributions de degrés et de corrélations souhaitées.

Nous insistons particulièrement sur la classe des graphes aléatoires exponentiels

(*exponential random graphs - ERG)*, car celle-ci a été construite dans le même objectif que celui que nous poursuivons [FS86, SPRH06, RPKL07]: générer des graphes aléatoires à partir de quelques éléments des données réelles jugés fondamentaux pour restituer l'essentiel de la topologie du graphe. C'est une étape fondamentale pour la détermination des mécanismes de construction du réseau. En effet si la géométrie peut être expliquée à l'aide - par exemple - de la distribution de degré et du taux de clustering, un processus qui restitue ces deux caractéristiques pourra être considéré comme un modèle possible de constitution du réseau.

Pour ce faire, la classe des *ERG* est construite en partant de l'hypothèse que le graphe réel est le produit le plus vraisemblable d'une loi de probabilité de la forme:

$$\mathcal{P}(\mathbf{A}) = \frac{1}{\kappa}.exp\left(\sum_a \eta_a \delta_a(\mathbf{A})\right)$$

où $\mathbf{A}$ désigne le graphe (ou sa matrice d'adjacence), a: une configuration locale présente dans le graphe initial, η_a: le paramètre qui lui est associé, $\delta_a(\mathbf{A}) = 1$ si $\mathbf{A}$ contient la configuration locale a et 0 sinon, et enfin κ est le coefficient de normalisation.

Cette définition autorise un nombre gigantesque de paramètres: un par configuration locale, sachant qu'une configuration fait ici référence à un sous-graphe où les nœuds sont étiquetés. Il est donc courant d'ajouter une hypothèse d'uniformité: imposer un unique paramètre par motif, quelles que soient les étiquettes. Il s'agit ensuite de rechercher le faisceau de paramètres η_a qui restitue au mieux le graphe réel, ce qui peut n'être fait en général que selon des méthodes approchées (e.g. *maximum pseudo-likelihood* - [SI90]). On pourrait alors utiliser ces paramètres pour générer des graphes proches du graphe réel au sens des configurations choisies.

Une telle méthode présente l'avantage d'une grande polyvalence, puisque tout type de motif peut être considéré, incluant éventuellement des informations externes au graphe. En revanche, elle induit également certains inconvénients:

- le modèle repose sur l'hypothèse arbitraire que le graphe réel est la réalisation la plus probable du processus stochastique décrit,
- la recherche de l'optimum est approchée et les écarts quadratiques sur des mesures pratiques suggèrent que celui-ci est évalué avec une forte incertitude, même sur de petits graphes [RPKL07].

Par la suite, nous chercherons également à créer un échantillon de graphes de comparaison, mais d'une manière que nous espérons plus contrôlée, dans le sens où nous aimerions savoir précisément l'ensemble statistique que nous décrivons.

2.3.2 Limites d'utilisation des tentatives d'échanges simples

L'ergodicité de la chaîne des tentatives d'échanges dont il a été question jusqu'à présent est démontrée dans un certain nombre de cas. Nous avons évoqué en 2.2.2.c. le théorème d'Eggleton qui l'affirme pour une distribution de degré fixée sur un graphe non-orienté. Mais il existe d'autres cas où celle-ci est établie: pour des graphes dépourvus de cycle (ou arbres) [Col77], des graphes connexes [Tay80] ou 2-connexes [Tay82], avec dans ces trois cas une distribution de degré fixée et un graphe non-orienté. Il est donc possible de générer par exemple des échantillons uniformément aléatoires de graphes connexes ayant une distribution de degré fixée, par la méthode d'échange simple telle que nous l'avons décrite, en étant assuré de l'ergodicité de la chaîne [VL05].

Néanmoins, les démonstrations sont très spécifiques à chacune de ces contraintes et rien ne montre que la génération par itération d'échanges, soit effectivement ergodique dans un ensemble quelconque. Sans plus de vérification, il est possible que le sous-ensemble décrit soit une sous-partie non représentative statistiquement de la totalité de l'ensemble.

D'ailleurs, il n'existe pas de théorème aussi générique pour le cas qui semble pourtant très voisin des graphes simples orientés et sans boucle, en effet dans [RJB96], les auteurs donnent l'exemple du couple de matrices d'adjacence:

$$\begin{pmatrix} 0^* & 1 & 0 \\ 0 & 0^* & 1 \\ 1 & 0 & 0^* \end{pmatrix} \nleftrightarrow \begin{pmatrix} 0^* & 0 & 1 \\ 1 & 0^* & 0 \\ 0 & 1 & 0^* \end{pmatrix}$$

dont on peut constater qu'il est impossible de passer de l'une à l'autre par une chaîne d'échanges sans sortir de l'ensemble[4].

C'est à ce type de problèmes que nous souhaitons remédier dans ce qui suit.

2.4 Généralisation de la méthode d'échange

Jusqu'à présent, nous n'avons fait état que de méthodes de génération de graphes qui existent dans la littérature, mais celles-ci sont limitées à des contraintes relativement simples. Dans ce qui suit nous proposons une méthode originale dans le but de générer des graphes obéissant à des contraintes plus variées, en étant le moins restrictif possible.

2.4.1 Contrainte minimale

Toutefois, nous allons dans toute la suite de ce travail faire le choix d'une contrainte minimale que respectent tous les graphes que nous générons. En effet, nous aurons pour objet d'étude un certain graphe traduisant les données réelles - noté G_0; les graphes

[4]Les 0^* indiquent que ces valeurs ne peuvent être modifiées: comme les boucles sont interdites, tous les éléments diagonaux doivent rester nuls.

auxquels nous proposons de comparer sa structure auront **la même distribution de degré**. Nous nous référerons à la contrainte minimale par la notation $\mathbf{C_{min}}$, et à l'ensemble des graphes la satisfaisant par $\mathcal{E}_{\mathbf{C_{min}}}$.

Ce choix impose évidemment la nature du graphe à produire (dirigé ou non, avec ou sans liens multiples), mais aussi le nombre de nœuds, de liens et donc la densité moyenne. Précisons également que la distribution de degré conservée dans le cas de graphes orientés sera la distribution des couples (degré entrant ; degré sortant) du nœud, et cette contrainte est plus exigeante que reproduire les distributions de degré entrant et sortant indépendamment l'une de l'autre.

2.4.2 Une marche aléatoire dans l'ensemble $\mathcal{E}_{\mathbf{C_{min}}}$?

Avant d'approfondir la discussion, précisons encore les notations que nous employons: $\mathcal{E}_{\mathbf{C_{min}}}$ étant l'ensemble des graphes obéissant à une distribution de degré déterminée; les graphes que nous voulons produire vérifient un ensemble de contraintes $\mathbf{C}$ tel que $\mathbf{C_{min}} \subset \mathbf{C}$, ces graphes appartiennent à un ensemble $\mathcal{F}_{\mathbf{C}}$ tel que:

$$\mathcal{F}_{\mathbf{C}} \subset \mathcal{E}_{\mathbf{C_{min}}}$$

Une première méthode que nous pouvons imaginer consisterait alors à réaliser un processus d'échange classique dans l'ensemble de référence $\mathcal{E}_{\mathbf{C_{min}}}$ qui contient l'ensemble $\mathcal{F}_{\mathbf{C}}$ à décrire, puis de vérifier régulièrement si l'état à l'instant t est un élément de $\mathcal{F}_{\mathbf{C}}$. Mais il faut ici prendre conscience du nombre d'éléments typiques des ensembles que nous sommes amenés à parcourir, qui empêche la mise en œuvre pratique d'une telle méthode. La taille de l'ensemble de référence relativement à celle de l'ensemble à décrire est tellemment disproportionnée qu'une marche aléatoire dans le premier mettrait un temps littéralement astronomique à revenir dans le second. Cela serait également le cas les situations traitées par la suite.

remarque : Pour en donner un aperçu à titre indicatif, nous sortons du cadre strict des graphes à distribution de degré fixée, pour faire un calcul d'ordre de grandeur des tailles relatives d'ensembles de graphes d'Erdős-Renyi. Avec $p = L/C_2^N$ la probabilité pour que deux nœuds soient associés, nous avons une probabilité $p' = p^3$ pour qu'un triplet de nœuds quelconque forme un triangle, et $T = C_3^N$ triplets de nœuds. Le nombre de triangles des graphes ER suit donc une distribution binomiale, que nous approchons à l'aide d'une loi de Poisson. Supposons que nous souhaitions comparer la taille de l'ensemble des graphes ER de densité fixée, à celle de l'ensemble des graphes de même densité comprenant un nombre de triangles déterminé; puis prenons des tailles typiques auxquelles nous pourrions être confrontés: 1000 nœuds et 5000 liens. L'espérance du nombre de triangles est $E = T.p' \simeq 17$; avec une densité de probabilité $\mathcal{P}(\alpha)$ en $\frac{e^{-E}E^{\alpha}}{\alpha!}$, celle-ci sera de l'ordre de 10^{-4} pour un graphe comportant $2E$ triangles, 10^{-6} pour $3E$ et 10^{-20} pour $4E$. En résumé, il faudrait visiter de l'ordre de 10^{20} états dans cet ensemble de graphes ER pour trouver un élément qui compte $4E$ triangles...

Nous pouvons alors penser à biaiser cette marche aléatoire, de manière à chercher plus spécifiquement des graphes appartenant au sous-espace. Malheureusement, cela ne garantit plus d'obtenir un échantillon uniformément aléatoire de l'ensemble. Nous reviendrons toutefois à ce principe de ciblage en 3.3, où nous le mettrons à profit dans un objectif différent.

2.4.3 Principe du k-échange

La solution que nous proposons [TRC10] consiste à modifier l'étape élémentaire du processus markovien, de manière à ce que celui-ci nous permette d'accéder à un plus grand nombre d'éléments de l'ensemble.

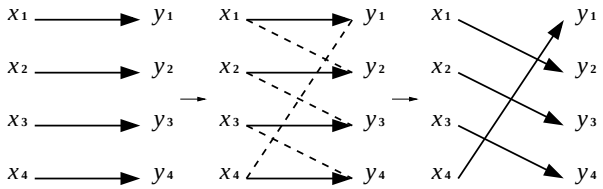
Cas orienté. Nous allons tout d'abord décrire le processus dans le cas des graphes orientés, où le principe est plus explicite. L'échange classique peut être compris comme une permutation de l'ensemble des destinations d'arcs. Ainsi l'échange entre $u_1 = (x_1, y_1)$ et $u_2 = (x_2, y_2)$ revient à effectuer:

$$\begin{cases} y_1 \rightarrow y_2 \\ y_2 \rightarrow y_1 \end{cases}$$

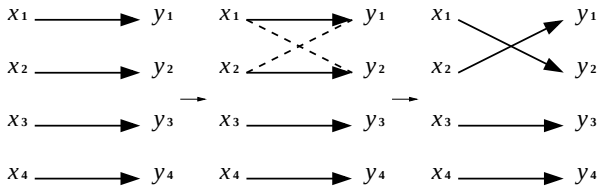
Choisissons maintenant un ensemble de k arcs, puis effectuons une permutation aléatoire de l'ensemble des destinations. Prenons deux exemples de telles permutations, pour $k = 4$ et donc un ensemble de quatre arcs $\{(x_1, y_1)\,;(x_2, y_2)\,;(x_3, y_3)\,;(x_4, y_4)\}$:

$$\begin{cases} y_1 \to y_2 \\ y_2 \to y_3 \\ y_3 \to y_4 \\ y_4 \to y_1 \end{cases} \quad \text{et} \quad \begin{cases} y_1 \to y_2 \\ y_2 \to y_1 \\ y_3 \to y_3 \\ y_4 \to y_4 \end{cases}$$

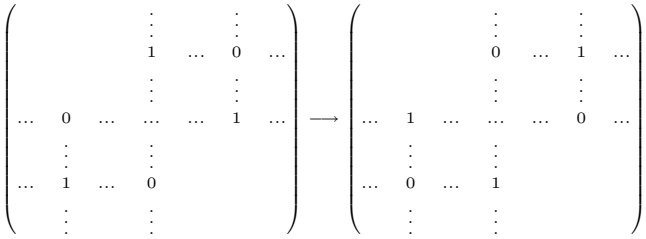
Nous sommes en train de réaliser une permutation circulaire de taille au plus k, nous parlerons d'effet d'avalanche, car la première permutation pourrait être décrite pratiquement par "y_2 remplace y_1, qui remplace y_4 etc.", on la schématiserait alors par:



Mais comme nous effectuons une permutation sur les destinations des arcs, il est possible qu'une destination soit envoyée sur elle-même, la taille de la permutation circulaire peut donc être inférieure à k, c'est le cas du second exemple:



De cette manière l'ensemble des permutations autorisées pour $k = k_1$ est également permise par toute valeur de $k \geq k_1$. Dans le cas où l'avalanche met systématiquement en jeu exactement 2 arcs, nous revenons aux échanges simples décrits en 2.2.2. Si elle met en jeu exactement 3 arcs, la configuration correspond aux *alternating hexagons* dont il est question dans [RJB96]. On donne ici une image matricielle de ce second cas:



Pour $k = L$, nous savons que le processus doit être ergodique dans l'ensemble à décrire. En effet, à partir du graphe $G = (\{x_1; ...; x_L\}; \{(x_1, y_1); ...; (x_L, y_L)\})$, il est

toujours possible d'accéder au graphe $G' = (\{x_1; ...; x_L\}; \{(x_1, y'_1); ...; (x_L, y'_L)\})$, on dit qu'il existe un **chemin canonique** de G à G', selon le vocabulaire en usage (ou *canonical path* - c.f. [JS97, KTV97]). Le nombre d'occurrences de chaque nœud est le même dans les deux graphes puisque la distribution de degré est conservée, et la transformation:

$$\begin{array}{rccl} & y_1 & \rightarrow & y'_1 \\ y'_1 = & y_i & \rightarrow & y'_i \\ y'_i = & y_j & \rightarrow & y'_j \\ & \vdots & & \\ & y_{L'} & \rightarrow & y'_{L'} \quad = y_1 \end{array}$$

qui permet de passer de G à G' et où $L' \leq L$, appartient bien à l'ensemble des L-échanges. Cette observation a surtout une valeur théorique car - sauf cas exceptionnels - nous serons pratiquement limités à des valeurs de k de l'ordre de quelques unités, et donc petites devant L.

Cas non-orienté. Le cas non-orienté n'est pas fondamentalement différent: moyennant une réécriture on peut se ramener à une description identique. En effet, on peut écrire un lien associant x à y comme deux arcs "miroirs" l'un de l'autre: (x, y) et (y, x), sur lesquels on effectue les permutations de manière ordinaire, hormis qu'un arc et son miroir ne peuvent être impliqués dans le même échange et que la modification de l'un implique celle de l'autre.

Uniformité. Nous garantissons l'uniformité de l'échantillon généré selon une méthode exactement analogue à celle développée en 2.2.2.d.: en effectuant une chaîne dont le nombre de tentatives d'échanges - plutôt que de réussites - est fixé. Quelles que soient les manières d'accéder à un métanœud, son degré dans le métagraphe n'est alors fonction que du nombre de combinaisons à k arcs parmi L, qui est le même pour tout les graphes de l'ensemble, induisant l'équidistribution de l'état stationnaire (c.f. Annexe C).

2.4.4 Point de vue métagraphique

L'ensemble des graphes considérés, associé à un processus de Markov, peut donner lieu à une représentation métagraphique. En effet, comme nous avons figuré par un métalien la transformation d'un graphe en un autre par un échange, nous pouvons représenter par un métalien la possibilité de générer un graphe à partir d'un autre au moyen d'un k-échange.

Rappelons qu'au cours d'un k-échange, le nombre de liens effectivement échangé est **inférieur ou égal** à k. De cette manière, nous nous assurons que s'il existe un métalien

entre deux des métanœuds associés au processus $k = k_1$, il en existe aussi un entre les mêmes métanœuds pour un k-échange où $k \geq k_1$. Le nombre de composantes connexes du métagraphe associé décroît lorsque k augmente, et pour $k = L$, le métagraphe est nécessairement connexe.

Nous illustrons cette idée sur le cas de graphes bipartis auxquels on impose simultanément la distribution de degré bipartie et celle de la projection monopartie des acteurs (contrainte $\mathbf{GHC_{min}}$). Dans cet exemple, les distributions imposées sont:

– pour le graphe biparti: $\{1, 1, 2, 2, 2\}$ (acteurs) et $\{1, 2, 2, 3\}$ (événements),

– pour la projection des acteurs: $\{1, 1, 2, 2, 2\}$,

satisfaites par exemple par le graphe de la figure 2.8.

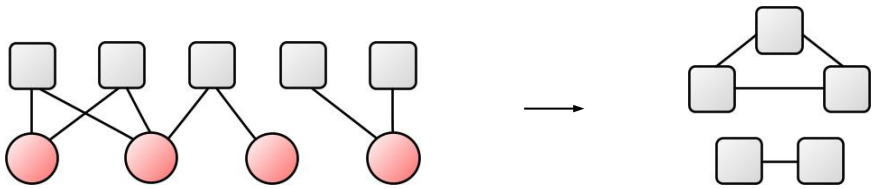


FIG. 2.8: Exemple de graphe biparti vérifiant les distributions de degré $\{1, 1, 2, 2, 2\}$ (nœuds gris) et $\{1, 2, 2, 3\}$ (nœuds rouges) et dont la projection des nœuds gris vérifie la distribution $\{1, 1, 2, 2, 2\}$.

Le métagraphe associé est alors décrit par la figure 2.9: chaque métanœud représente un des graphes de l'ensemble associé à la contrainte précédente et chaque métalien est la possibilité de passer d'un élément à un autre à l'aide d'un k-échange.

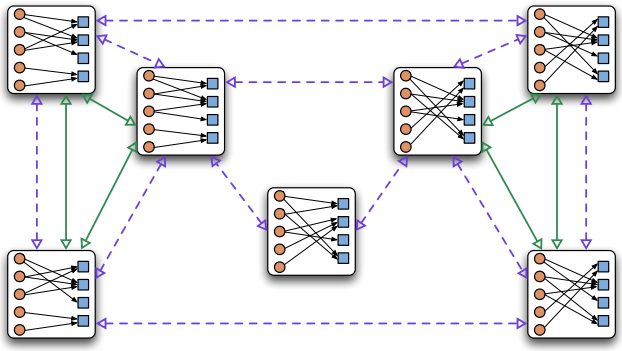


FIG. 2.9: En traits pleins verts: métaliens associés au processus pour $k = 2$ et $k = 3$; en traits pointillés bleus: métaliens supplémentaires pour $k = 4$. Pour une question de clarté, nous avons supprimer les métaboucles et métaliens multiples.

remarque : La représentation du métagraphe n'est ici possible que parce que le graphe considéré est de très petite taille, et le nombre d'éléments de l'ensemble à décrire suffisament faible. Dans les cas pratiques que nous envisageons par la suite, il sera totalement exclu de réaliser de telles images de l'ensemble.

2.4.5 Procédure pratique

En résumé, la méthode pratique que nous nous proposons d'appliquer de manière systématique est la suivante:

1. À k fixé, en commençant par $k = 2$, nous effectuons des k-échanges jusqu'à ce que l'on estime avoir atteint un palier de convergence pour un ensemble de mesures-test choisies (τ_k itérations).

2. Parmi les valeurs de k testées, on doit observe à partir d'une valeur minimum k^* que nous qualifierons de **seuil**, que toutes les mesures-test convergent vers une même limite quel que soit $k \geq k^*$. En pratique, nous estimons qu'il faut s'assurer de la stabilité sur 3 ou 4 valeurs de k pour que le seuil soit effectivement atteint.

3. Nous générons alors l'échantillon en effectuant τ_k itérations de la chaîne de k-échanges la plus rapide à atteindre l'état stationnaire et telle que $k \geq k^*$ (il s'agit en général de $k = k^*$).

2.5 Illustrations

2.5.1 Mise en pratique sur des "modèles-jouets"

Pour justifier l'intérêt de cette généralisation, nous allons mettre en évidence le fait qu'une chaîne d'échanges simples peut ne pas être suffisante pour décrire la totalité de l'ensemble sur des exemples artificiels, mais que nous pensons explicites.

a. Composantes connexes rigides

Graphe initial, contraintes. Imaginons le graphe simple non-orienté représenté figure 2.10, constitué de deux composantes connexes (une est indexée à l'aide de chiffres et l'autre de lettres).
Nous imposons - en plus de la distribution de degré - que le graphe ait toujours strictement deux composantes connexes, nous souhaitons donc décrire l'ensemble des graphes satisfaisant ces deux contraintes.

Échanges autorisés. Il est impossible avec $k = 2$ d'effectuer des échanges de nœuds entre les composantes connexes. En effet, pour lier une lettre et un chiffre à l'aide d'un

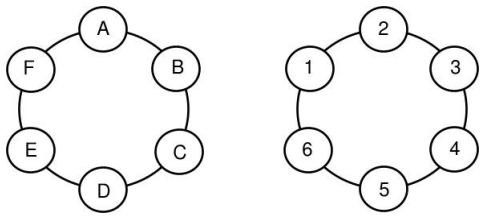


FIG. 2.10: Le graphe initial, comportant deux composantes connexes.

échange simple, il faut briser exactement un lien de chaque composante et en créer deux entre elles, le graphe entier est alors connexe et nous sortons de l'ensemble à décrire. À partir de $k = 3$, en revanche, il est possible de faire des échanges permettant à une composante d'absorber un nœud de l'autre. De cette manière, on peut produire les autres graphes de cet ensemble (c.f. Fig. 2.11).

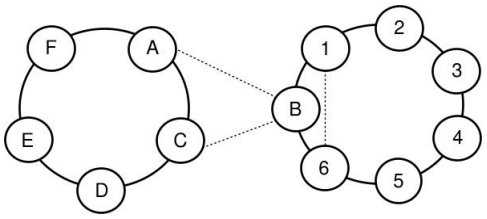


FIG. 2.11: Résultat d'un 3-échange sur le graphe de départ permettant de modifier les tailles des composantes connexes. En pointillés: les liens supprimés.

Discussion. Cet exemple met en évidence une configuration topologique qui est "gelée" lorsqu'elle est associée à une certaine contrainte. Ici, on accède soit à une partie très limitée de l'ensemble à décrire (pour $k = 2$), soit à sa totalité (pour $k > 2$). C'est un indice d'une caractéristique qualitative de l'algorithme: sa "force" augmente rapidement avec k. Métaphoriquement, augmenter k nous donne la possibilité de franchir les "barrières de potentiel" que sont les configurations gelées. L'exemple suivant montre toutefois que la transition vers l'ergodicité en fonction de k peut être progressive.

b. Triangles colorés

Graphe initial, contraintes. Considérons les graphes constitués exclusivement de triangles orientés, comportant donc $N = 3.N'$ nœuds, pour lesquels chacun dispose d'un arc entrant et d'un autre sortant. Par ailleurs, nous attribuons à chaque nœud une des trois couleurs rouge, vert ou bleu, de manière à ce que le graphe initial soit exclusivement constitué de triangles "rouge → vert → bleu" (R,V,B). Les contraintes que nous imposons sont:

- $\mathbf{C_{min}}$: la distribution de degré (entrant, sortant),
- le nombre de triangles orientés est également invariant.

De cette manière, d'un graphe à un autre de l'ensemble, seules varient les associations de couleurs dans les triangles. Nous allons donc les dénombrer afin de savoir quelle partie de l'ensemble est décrite selon la valeur de k.

Échanges autorisés.

- Pour $k = 2$, aucun échange n'est autorisé car tous briseraient un triangle au moins, sans en créer. La marche est piégée dans l'état initial.
- Pour $k = 3$, il est possible d'effectuer des permutations intra-triangulaires comme celle schématisée en 2.12. Celles-ci permettent de produire des triangles orientés dans le sens inverse aux triangles de départ.

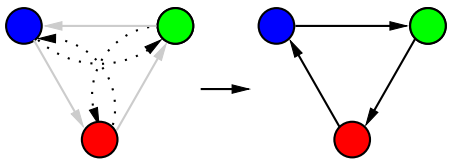


FIG. 2.12: Permutation intra-triangulaire: arcs supprimés en gris, créés en pointillés.

- Pour $k \geq 4$, il est possible de réaliser des permutations extra-triangulaires, comme sur la figure 2.13, et d'accéder alors à toutes les combinaisons de couleurs possibles.

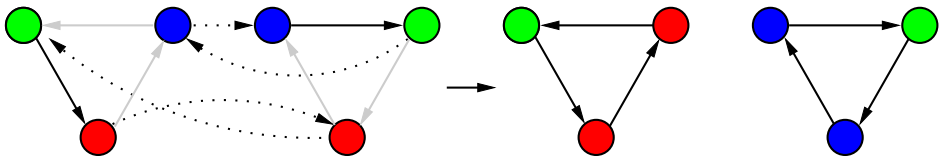


FIG. 2.13: Permutation extra-triangulaire: arcs supprimés en gris, créés en pointillés.

Résultats et discussion. Dans le cadre de cet exemple, il est relativement facile de prévoir la composition statistique de l'ensemble à décrire: le tirage d'un triangle quelconque de l'ensemble complet revient à tirer aléatoirement - en prenant en compte l'ordre du tirage - trois nœuds colorés parmi les $3.N'$ nœuds, dont N' de chaque couleur. Le calcul théorique est cohérent avec la mise en œuvre de l'algorithme de k-échanges sur un graphe de 180 nœuds dont les résultats sont compilés dans le tableau 2.1 et la figure 2.14.

Nous constatons qu'il est nécessaire de prendre $k \geq 4$ pour obtenir une mesure statistiquement satisfaisante de la composition de l'ensemble complet. Strictement parlant, cela ne démontre pas que tout élément de l'ensemble soit effectivement accessible par le processus à $k = 4$ (même si pour ce problème précis, c'est le cas), mais

	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	Théorique
R-R-R	0.	0.	0.036	0.036	0.036	0.036
V-V-V	0.	0.	0.036	0.036	0.036	0.036
B-B-B	0.	0.	0.036	0.036	0.036	0.036
R-V-V	0.	0.	0.111	0.111	0.111	0.111
R-B-B	0.	0.	0.111	0.111	0.111	0.111
V-V-B	0.	0.	0.111	0.111	0.111	0.111
V-B-B	0.	0.	0.111	0.111	0.111	0.111
R-R-B	0.	0.	0.111	0.111	0.111	0.111
R-R-V	0.	0.	0.111	0.111	0.111	0.111
R-B-V	0.	0.500	0.113	0.113	0,113	0.113
R-V-B	1.000	0.500	0.113	0.113	0,113	0.113
Succès	0	997 ± 74	2643 ± 108	2067 ± 132	936 ± 55	-

TAB. 2.1: Proportion de triangles de chaque combinaison de couleurs dans l'état final en fonction de k, moyenné sur 10000 réalisations d'une chaîne de Markov de 10^8 tentatives.

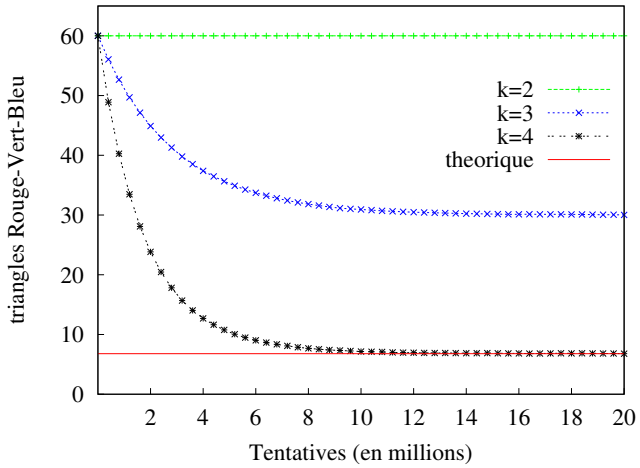


FIG. 2.14: Évolution du nombre de triangles orientés de type (Rouge, Vert, Bleu) pour $k = \{2, 3, 4\}$ en fonction du nombre d'essais. Moyenne sur 10000 réalisations.

seulement que parmi les mesures proposées, aucune ne permet de distinguer l'ensemble décrit pour $k = 4$ et les ensembles décrits par des valeurs de k supérieures.

remarque : Nous pouvons dès maintenant observer un problème qui sera récurrent dans l'utilisation de cette technique: lorsque la contrainte à satisfaire est exigeante, le nombre de combinaisons de liens permettant de réussir un échange est petit face au nombre total de combinaisons, le taux de réussite (échanges effectués / échanges tentés) est donc petit devant 1 (ici, de l'ordre de 10^{-5}) et par conséquent la durée pour observer la convergence de la chaîne de Markov peut être longue.

2.5.2 Quelques applications réalistes

Dans cette partie, nous voulons mettre en pratique la méthode sur des exemples de graphes et de contraintes qui soient plus proches de ce que nous souhaiterons étudier concrètement. Il ne s'agit pas ici de creuser la signification des mesures observées, mais seulement de tester les k-échanges en "conditions réelles". Cela nous permettra de mettre en évidence les difficultés pratiques rencontrées, et les situations où leur utilisation est nécessaire.

a. Contrainte des composantes connexes

En suivant l'idée que les composantes connexes d'un graphe puissent représenter des modules de fonctionnalité de celui-ci, on peut vouloir produire des graphes aléatoires dont **la distribution des tailles des composantes connexes** serait identique à celle d'un graphe réel (contrainte $\mathbf{C_{compo}}$).

Mise en œuvre. En pratique, la vérification de la distribution des composantes connexes à chaque itération limite la vitesse de l'algorithme. L'inconvénient d'une telle contrainte tient à son caractère global: la taille d'une composante n'est pas déterminée par le proche environnement du nœud. Une possibilité rudimentaire consiste à donner, pour tous les nœuds impliqués dans le k-échange, leur composante d'appartenance avant et après le k-échange, et vérifier que les tailles de ces composantes permettent de conserver une distribution identique.

Avec les topologies usuelles des graphes de réseaux complexes, où il existe presque toujours une composante géante, la probabilité pour qu'un des liens impliqués dans l'échange appartienne à celle-ci est proche de 1. Cela signifie que la stratégie précédente impose à chaque étape de passer en revue la majorité des nœuds du graphe. Même s'il est possible d'optimiser ce parcours, cette méthode reste peu efficace en temps et il ne sera possible de l'utiliser que sur des graphes de petit N. L'algorithme est décrit en Annexe D.

Exemple: résultats, discussion. Nous considérons cette contrainte sur un graphe d'interactions entre protéines* (*pathways*): il s'agit des voies métaboliques recensées chez l'homme dans la base *Ecocyc*, qui comporte 679 nœuds et 11.030 liens, et dont les composantes connexes sont distribuées selon:

- une composante géante (440 nœuds),
- six composantes de tailles "moyennes" (37, 35, 30, 27, 11 et 10),
- vingt-sept composantes de tailles ≤ 10 (1×7, 3×6, 3×5, 4×4, 5×3, 7×2, 4×1).

Pour estimer la convergence de la chaîne, la mesure utilisée doit être effectuée régulièrement au cours du processus, ce qui suggère de la choisir légère: le dénombrement

des motifs locaux présente l'avantage d'être à la fois simple et rapide. Nous employons le nombre de chemins de longueur 3 et effectuons 50 simulations de 200.000 tentatives d'échanges, puis reportons l'évolution en fonction de celles-ci dans la figure 2.15.

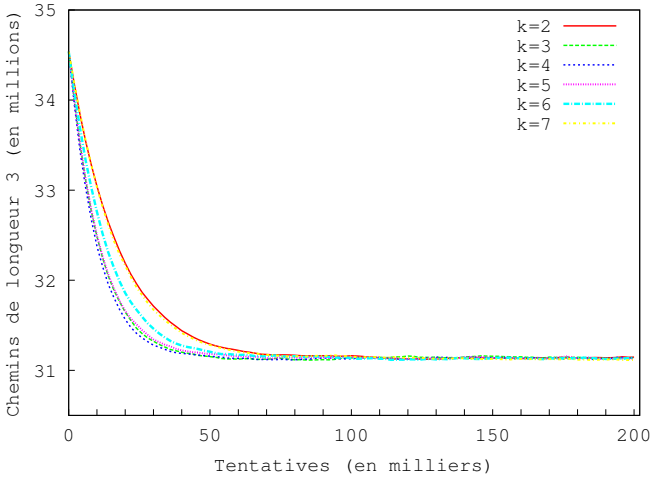


FIG. 2.15: Nombre de chemins de longueur 3 (en millions) selon le nombre de tentatives de k-échanges, $k \in [\![2, 7]\!]$.

Dans la table 2.2, nous résumons les mesures obtenues pour l'état stationnaire avec $k \in \{2, 3, 4, 5, 6, 7\}$, d'une part sur les motifs locaux de taille 3 et 4, mais également sur des caractéristiques plus globales: la moyenne $\overline{d}$ et l'écart quadratique σ_d de la distribution des distances dans la composante géante.

Observables	départ: G_0	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
△	$161,3 \cdot 10^3$	$51,7 \cdot 10^3$	$51,7 \cdot 10^3$	$51,7 \cdot 10^3$	$51,7 \cdot 10^3$	$51,7 \cdot 10^3$	$51,7 \cdot 10^3$
⟠	$2070 \cdot 10^3$	$178 \cdot 10^3$	$178 \cdot 10^3$	$178 \cdot 10^3$	$178 \cdot 10^3$	$178 \cdot 10^3$	$177 \cdot 10^3$
⟠	$12,95 \cdot 10^6$	$1,60 \cdot 10^6$	$1,60 \cdot 10^6$	$1,60 \cdot 10^6$	$1,60 \cdot 10^6$	$1,60 \cdot 10^6$	$1,59 \cdot 10^6$
◇	$6,74 \cdot 10^6$	$1,81 \cdot 10^6$	$1,81 \cdot 10^6$	$1,81 \cdot 10^6$	$1,80 \cdot 10^6$	$1,81 \cdot 10^6$	$1,80 \cdot 10^6$
◇	$34,5 \cdot 10^6$	$31,1 \cdot 10^6$	$31,1 \cdot 10^6$	$31,1 \cdot 10^6$	$31,1 \cdot 10^6$	$31,1 \cdot 10^6$	$31,1 \cdot 10^6$
$\overline{d}$	3,511	2,058	2,057	2,057	2,057	2,057	2,057
σ_d	1,435	0,490	0,489	0,489	0,488	0,489	0,489
Succès	-	42,300	60,400	52,800	38,100	25,500	20,400

TAB. 2.2: Moyenne sur 50 simulations des mesures après 200.000 tentatives de k-échanges depuis G_0 et nombre de réussite. L'écart quadratique est de l'ordre du pourcent.

Nous constatons sur cet exemple que les mesures testées correspondant à différents k convergent toutes vers une unique valeur stationnaire. Dans ce cas, la méthode des

k-échanges peut être ramenée à celle des tentatives d'échanges simples (équivalente à $k = 2$), et elle contribue à nous assurer qu'un échantillon généré de cette manière est effectivement uniformément aléatoire sur tout l'ensemble $\mathcal{F}_{\mathbf{C_{compo}}}$.

b. Contrainte du nombre de triangles

La génération de graphes reproduisant le taux de clustering en plus de la distribution de degré des réseaux réels a fait l'objet de travaux récents (c.f. 2.3.1). C'est ce que nous nous proposons de faire dans cet exemple: générer des échantillons de graphes simples, non-orientés et dont **le nombre de triangles reste invariant** (ensemble de contraintes $\mathbf{C_{tri}}$).[5]

Mise en œuvre. À l'inverse du cas précédent, la contrainte peut ici être testée de manière locale: les triangles créés ou détruits à chaque pas de l'algorithme mettent en jeu au moins un des liens impliqués dans le k-échange. De cette manière il nous suffit de vérifier que les nombres de triangles créés ou détruits se compensent à chaque étape de la chaîne pour nous assurer que la marche reste dans l'ensemble $\mathcal{F}_{\mathbf{C_{tri}}}$. Le test spécifique à cette contrainte est en Annexe D.

Exemple: résultats, discussion. Nous l'appliquons sur le graphe monoparti des collaborations scientifiques* extrait de l'*Anthropological Index Online*, dans la section archéologie scandinave, au cours des années 2000-2009. Le graphe comporte 273 individus et 280 liens, il présente par ailleurs la particularité de ne pas avoir de composante géante.

La petite taille du graphe induit de fortes fluctuations sur la plupart des mesures, c'est pourquoi nous étudions la convergence à l'aide de la mesure cumulée des motifs cycliques de taille 4 pour $k \in \{2, 3, 4, 5\}$ (Fig 2.16). En revanche, on remarquera que cette petite taille n'induit pas une convergence rapide.

Pour nous assurer de la qualité de l'échantillon produit, nous élargissons le faisceau de mesures dans l'état stationaire en fonction de k, en y ajoutant le nombre d'autres motifs de taille 4 et la taille moyenne des composantes connexes du graphe $\overline{\text{cc}}$ - qui caractérise le graphe à une échelle plus globale - (Tab 2.3).

Contrairement à l'exemple précédent, un échantillon créé à l'aide d'une chaîne de 2-échanges n'est pas satisfaisant, ici il est nécessaire d'avoir recours au minimum aux 3-échanges pour générer un échantillon uniformément aléatoire sur l'ensemble $\mathcal{F}_{\mathbf{C_{tri}}}$ associé à ce graphe. Au-delà, la dispersion des mesures ne permet pas de distinguer les

[5] La distribution de degré imposant le nombre de chemins de longueur 2 dans le graphe, un nœud de degré δ est au centre de $\frac{\delta(\delta-1)}{2}$ chemins de longueur 2, cette contrainte conserve aussi le clustering global du graphe.

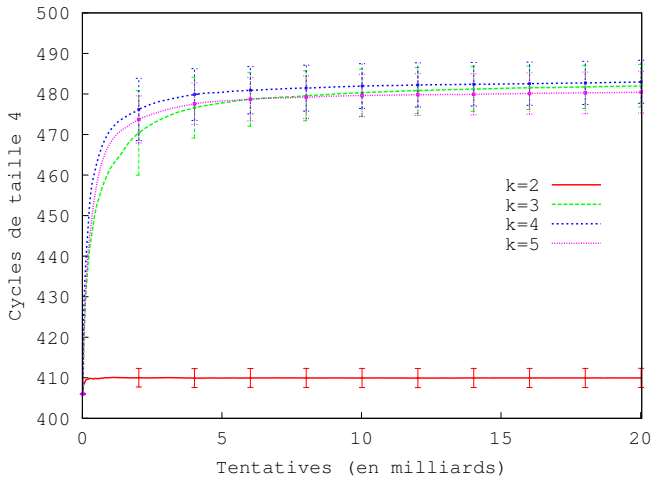


FIG. 2.16: Moyenne cumulée des cycles de taille 4 selon le nombre de tentatives de k-échanges, $k \in [\![2, 5]\!]$.

Observables	départ: G_0	$k = 2$	$k = 3$	$k = 4$	$k = 5$
⟐	108	108 ± 0	121,0 ± 1,7	121,1 ± 2,3	119,9 ± 2,0
⟡	730	733,6 ± 2,9	844,9 ± 8,8	847,0 ± 9,8	841,4 ± 8,7
◇	406	409,6 ± 2,9	481,9 ± 6,3	483,8 ± 4,2	481,8 ± 5,9
◿	2794	2798 ± 4	2904 ± 49	2935 ± 39	2896 ± 55
$\overline{cc}$	1,98	1,98 ± 0	2,11 ± 0,02	2,12 ± 0,02	2,10 ± 0,01
Succès	-	157 $\cdot 10^6$	430 $\cdot 10^6$	235 $\cdot 10^6$	84 $\cdot 10^6$

TAB. 2.3: Moyenne sur 10 simulations des mesures après $15 \cdot 10^9$ tentatives de k-échanges depuis G_0 et nombre de réussites (l'incertitude indiquée correspond à l'écart quadratique de la mesure).

échantillons, et selon ce critère $k^* = 3$.

2.6 Signification statistique de la comparaison

Cette partie est autonome, elle n'est pas liée à la technique de génération de graphes utilisée: quel que soit le graphe considéré et l'ensemble dans lequel il est situé, il faut s'interroger sur la taille de cet ensemble. Le principe des échanges va nous donner des outils pour examiner le problème.

Quel sens donner aux écarts entre les mesures effectuées sur un graphe réel et celles sur un échantillon synthétique de référence? En effet, si l'ensemble que nous décrivons est très lourdement contraint, le métagraphe sera relativement petit et la distance

séparant un élément quelconque du graphe réel G_0 suffisamment faible pour que leurs topologies soient proches, indépendamment de la signification des contraintes.

Nous cherchons ici à distinguer cette situation du cas où l'ensemble des contraintes choisies permet d'obtenir une topologie proche de G_0 en raison, non pas du cardinal de l'ensemble, mais de la nature des propriétés choisies pour le définir.

Ce qui suit est difficile d'accès, car cette partie repose davantage sur la compréhension qualitative de la méthode que sur un formalisme indiscutable. Comme elle n'est pas indispensable pour la compréhension de la suite du travail, nous recommandons de la passer en première lecture.

2.6.1 Méthode expérimentale de validation

Il s'agit alors de donner une estimation des tailles relatives d'ensembles, c'est un problème dit d'***approximate counting*** [JS97], qui pour être résolu de manière précise nécessiterait d'importants développements. Nous nous limiterons donc à une méthode expérimentale de validation, qui - à notre connaissance - n'a pas d'équivalent dans la littérature.

a. Principe

Les méthodes d'échanges nous suggérent une estimation plus simple à mettre en œuvre mais assez grossière, que nous décrivons ici pour $k = 2$:

1. Nous mesurons **le nombre n_e d'échanges autorisés** dans le graphe G_0 pour la contrainte à tester, ainsi que **le nombre de liens distincts** L_e impliqués dans ces échanges.

2. Nous tirons alors au hasard L_e liens dans G_0; puis parmi les paires possibles entre ces liens, nous en sélectionnons n_e aléatoirement que nous qualifierons "d'échangeables".

3. Nous effectuons une procédure usuelle d'échanges entre les paires sélectionnées.

> **remarque :** Les paires évoluent au cours du processus, si bien que le statut de paire échangeable est en fait attaché à un des deux nœuds de la paire (par exemple le nœud-source pour un arc).

De cette manière, nous espérons générer un échantillon de graphes appartenant à un ensemble aléatoire $\mathcal{F}_{\mathbf{vérif}}$, contenant G_0, **dont le cardinal soit du même ordre que celui de l'ensemble à décrire $\mathcal{F}_{\mathbf{C}}$**. En effet, reprendre n_e revient à autoriser autant de transitions vers des destinations quelconques de $\mathcal{E}_{\mathbf{C_{min}}}$, situées à distance 1 de G_0 au sens des 2-échanges, qu'il y en a autour de G_0 dans $\mathcal{F}_{\mathbf{C}}$. Nous pouvons donc

comprendre cette démarche comme la construction d'un métagraphe artificiel pour lequel le métanœud G_0 a autant de voisins que dans le métagraphe réel.

Nous comparerons alors les mesures topologiques obtenues sur cet échantillon de $\mathcal{F}_{\mathbf{vérif}}$ à celles effectuées sur l'ensemble contraint $\mathcal{F}_{\mathbf{C}}$ et sur le graphe G_0. Nous avons alors une indication sur la signification des résultats: s'ils sont le produit d'un simple effet statistique, les écarts des mesures entre $\mathcal{F}_{\mathbf{vérif}}$ et G_0 seraient du même ordre que les écarts des mesures entre $\mathcal{F}_{\mathbf{C}}$ et G_0.

b. Hypothèses sous-jacentes

Au premier voisin. Affirmer que la reconstruction de l'environnement local du métanœud G_0 soit suffisante pour obtenir un ensemble comparable en taille à $\mathcal{F}_{\mathbf{C}}$, suppose tacitement que l'environnement proche de G_0 soit comparable à celui d'un métanœud quelconque de $\mathcal{F}_{\mathbf{C}}$. Cela impliquerait que la proportion de succès de la chaîne soit approximativement stable au cours du processus (i.e. le nombre de voisins d'un métanœud est stable dans le métagraphe).

Cette condition est assez bien vérifiée sur les exemples traités expérimentalement. Sur les cas examinés précédemment, on dénombre n_e sur les graphes de départ et quelques graphes choisis au hasard dans l'échantillon. Dans le cas de $\mathcal{F}_{\mathbf{C_{compo}}}$, où le nombre de tentatives possibles d'échange est élevé, nous nous limitons à une estimation du taux $t_e = n_e/n_{tot}$ réalisée sur 10^5 tentatives d'échanges. Dans la table 2.4, nous constatons que les fluctuations sont dans les deux cas inférieures à 10%.

$\mathbf{C_{compo}}$	G_0	G_1	G_2	G_3
t_e	$0,240$	$0,211$	$0,213$	$0,211$

$\mathbf{C_{tri}}$	G_0	G_1	G_2	G_3
n_e	1.227	1.227	1.227	1.227

TAB. 2.4: Taux ou nombre de réussites (t_e, n_e) de quelques graphes (dont les graphes réels) des ensembles $\mathcal{F}_{\mathbf{C_{compo}}}$ et $\mathcal{F}_{\mathbf{C_{tri}}}$.

Au second voisin et plus. Cependant, la procédure repose également sur d'autres hypothèses plus difficiles à vérifier. Même si le nombre de voisins dans le métagraphe est identique pour tous les métanœuds, cela n'assure pas que le nombre de voisins à distance 2, 3 ou p soient le même dans $\mathcal{F}_{\mathbf{vérif}}$ et $\mathcal{F}_{\mathbf{C}}$; et donc que les tailles de ces deux ensembles soient effectivement comparables.

Fixer L_e pour générer l'échantillon aléatoire est un moyen indirect de chercher à satisfaire cette condition. En effet, à n_e fixé, le nombre de liens susceptibles d'être échangés conditionne la diversité des structures de graphe auxquelles on peut accéder. Et donc imposer L_e revient à réduire le nombre de voisins à distance 2 (ou plus) dans le métagraphe. De sorte qu'en prenant cette précaution additionnelle, l'ensemble $\mathcal{F}_{\mathbf{vérif}}$ aura un cardinal plus proche de celui des graphes de $\mathcal{F}_{\mathbf{C}}$. La mise en œuvre de

vérifications plus précises alourdit la procédure au point que nous nous contenterons des validations décrites ci-dessus.

c. Amélioration: élargissement à k quelconque

La procédure présentée ci-dessus pour les échanges simples peut être mise en œuvre de la même manière sur des k-échanges, on dénombrera alors non les paires de liens mais les ensembles de k liens mis en jeu, ce qui peut être long lorsque k croît mais ne diffère pas sur le principe.

La vérification statistique sera appliquée à k^*: la valeur de k minimum pour laquelle nous considérons que l'échantillon de l'ensemble $\mathcal{F}_{\mathbf{C}}$ obtenu est uniformément aléatoire. En effet, pour les valeurs inférieures, l'échantillon n'est de toutes façons pas significatif; pour les valeurs supérieures, n_e et L_e sont nécessairement plus élevés, et donc l'échantillon $\mathcal{F}_{\mathbf{vérif}}$ plus grand que celui associé à k^*, comme nous voulons nous assurer précisément que $\mathcal{F}_{\mathbf{C}}$ n'est pas "trop petit", on le compare au $\mathcal{F}_{\mathbf{vérif}}$ de taille minimum.

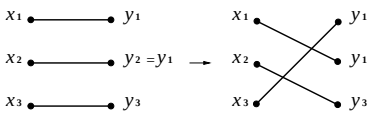
d. Amélioration: problème des isomorphismes

D'autre part, parmi les échanges autorisés dans le graphe réel, certains sont des isomorphismes, i.e. des réétiquetages des nœuds, mais qui ne vont modifier en rien la structure du graphe obtenu. On peut identifier deux cas simples (pour $k = 2$):

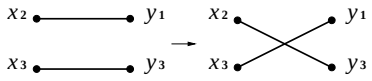
- dans l'échange de (x_a, y_a) avec (x_b, y_b), si les deux extrêmités x (ou les deux y) sont de degré 1, l'échange produit un graphe dont la topologie est identique,
- c'est également le cas, si l'une des extrêmités est commune aux deux liens échangés.

remarque : Pour $k > 2$, on considérera comme des isomorphismes les cas suivants:

- si les extrêmités x_i (ou y_i) de tous les liens échangés sont de degré 1,
- puis, supposons que nous effectuons le 3-échange:



Il ne s'agit pas d'un isomorphisme mais le 2-échange:



nous amènerait à un résultat identique. Pour ne pas les dénombrer plusieurs fois, ce type de k-échanges sera traité comme des isomorphismes.

Les échanges de ce type peuvent être très majoritaires parmi les échanges réussis pour l'ensemble $\mathcal{F}_{\mathbf{C}}$ car ils sont toujours autorisés: en effet si les deux graphes obtenus sont identiques aux étiquettes près, et que l'un respecte la contrainte, l'autre aussi.

L'analyse s'en trouverait biaisée puisque nous comparons à un ensemble construite à l'aide d'échanges choisis de manière aléatoire. Cela signifie que les cardinaux de $\mathcal{F}_{\mathbf{vérif}}$ et $\mathcal{F}_{\mathbf{C}}$ seraient effectivement comparables, mais que les éléments de $\mathcal{F}_{\mathbf{C}}$ ont une plus forte probabilité d'être des réétiquetages les uns des autres. Alors une alternative envisageable à la vérification consiste à ne compter **que les permutations effectives (non-isomorphes) et les liens associés**, nous noterons n_e^{eff} et L_e^{eff} les valeurs correspondantes.

Ce point attire notre attention sur un autre biais d'interprétation possible aux résultats: la topologie des éléments de $\mathcal{F}_{\mathbf{C}}$ peut être proche de celle de G_0 parce que $\mathcal{F}_{\mathbf{C}}$ contient un grand nombre de graphes isomorphes les uns aux autres, induisant une faible diversité topologique effective dans l'ensemble.

2.6.2 Illustration

Nous reprenons le graphe de collaborations archéologiques étudié en 2.5.2 avec la même contrainte $\mathbf{C_{tri}}$, sur lequel nous effectuons la vérification statistique proposée pour diverses valeurs de k. Même si l'échantillon n'est uniformément aléatoire que pour $k \geq 3 = k^*$ nous l'appliquons pour $k \in \{2; 3; 4\}$ dans le but d'en illustrer le principe.

a. Estimations de n_e et L_e

Nous mesurons n_e (nombre d'échanges autorisés) et L_e (nombre de liens distincts impliqués) pour chaque valeur de k. Lorsque k augmente, cela peut devenir très long de tester toutes les combinaisons d'échanges possibles dans le graphe initial. On peut se satisfaire d'une estimation de n_e en extrapolant à partir du taux de réussite t_e et du nombre de liens différents impliqués évalués sur un échantillon de k-échanges.

Nous avons pour cela besoin du nombre total n_{tot} de tentatives d'échanges. Sachant que pour un graphe non-orienté de L liens, il y a C_k^L ensembles de k liens, et que chacun peut amener à $2k.(2k-2)...(2k-2(k-1))/(2.k)$ permutations mettant en jeu exactement k liens, on arrive à la conclusion que le nombre n_{tot} de tentatives d'échanges où k liens sont effectivement modifiés est:

$$n_{tot} = \frac{2^{k-1}}{k} L(L-1)...(L-(k-1))$$

En ce qui concerne L_e, on atteint la valeur limite au bout d'un nombre de pas de temps relativement faible, ce qui permet d'en donner une borne inférieure proche de la valeur exacte, c'est ce qui est mis en pratique ici (résultats table 2.5).
Remarquons la différence d'ordre de grandeur manifeste selon que l'on élimine ou non les isomorphismes, ce qui nous incite à ne considérer que le cas où ceux-ci sont éliminés.

		$k=2$	$k=3$	$k=4$
	n_{tot}	$78{\cdot}10^3$	$29{\cdot}10^6$	$12{\cdot}10^9$
avec isomorphismes	n_e	1.227	$\simeq 13{\cdot}10^3$	$\simeq 190{\cdot}10^3$
	L_e	280	280	280
sans isomorphisme	n_e^{eff}	5	263	$\simeq 14{\cdot}10^3$
	L_e^{eff}	4	26	$\gtrsim 74$

TAB. 2.5: Nombre de tentatives possibles de k-échanges (n_{tot}), valeurs des nombres de k-échanges (n_e) et k-échanges effectifs (n_e^{eff}) possibles et nombre de liens associés (L_e, L_e^{eff}) pour $k \in \{2;3;4\}$, dans le contexte de la contrainte $\mathbf{C_{tri}}$ sur le graphe *AIO: Scandinavie*.

b. Mesures sur les ensembles $\mathcal{F}_{\mathbf{vérif}}$

Conformément à la procédure que nous décrivions, nous générons des échantillons de 50 graphes des ensembles aléatoires correspondants, puis comparons les valeurs stationnaires aux valeurs réelles - résultats table 2.6.

		$k=2$		$k=3=k^*$		$k=4$	
Observables	départ: G_0	$\mathcal{F}_{\mathbf{C_{tri}}}$	$\mathcal{F}_{\mathbf{vérif}}$	$\mathcal{F}_{\mathbf{C_{tri}}}$	$\mathcal{F}_{\mathbf{vérif}}$	$\mathcal{F}_{\mathbf{C_{tri}}}$	$\mathcal{F}_{\mathbf{vérif}}$
	108	108	101 ± 5	121	62 ± 8	121	21 ± 5
	730	734	689 ± 27	845	468 ± 47	847	191 ± 31
	406	410	388 ± 12	482	287 ± 24	484	145 ± 15
	2794	2798	2811 ± 17	2904	2915 ± 42	2935	3059 ± 53

TAB. 2.6: Mesures comparées des motifs de taille 4 sur le graphe réel, l'ensemble contraint $\mathcal{F}_{\mathbf{C_{tri}}}$ et l'ensemble de vérification $\mathcal{F}_{\mathbf{vérif}}$, pour différentes valeurs de $k \in \{2;3;4\}$.

Pour $k \geq k^*$, même si nous ne pouvons pas énoncer de critère universel pour trancher cette question, en comparant les mesures m sur les deux ensembles, nous constatons que $(m_{\mathbf{C_{tri}}} - m_{G_0})$ et $(m_{\mathbf{vérif}} - m_{G_0})$ ne sont pas du même ordre. Donc les caractéristiques topologiques des échantillons obtenus ne sont manifestement pas le résultat d'un effet de petite taille, mais bien corrélées au choix de la contrainte.

Par ailleurs, l'échantillon généré à l'aide de $k = 2$, même s'il n'est pas significatif de l'ensemble, mérite un court commentaire. Supposons qu'on ait observé les mêmes résultats et que $k^* = 2$. Si tel était le cas, et malgré la proximité topologique de G_0 et de l'échantillon $\mathcal{F}_{\mathbf{C_{tri}}}$ correspondant, on ne pourrait pas prétendre avoir isolé les éléments topologiques suffisants pour expliquer la structure du réseau réel. En effet, si nous comparons par exemple $|m_{\mathbf{C_{tri}}} - m_{G_0}|$ et $|m_{\mathbf{vérif}} - m_{G_0}|$, l'écart serait insuffisant pour que l'ensemble $\mathcal{F}_{\mathbf{C_{tri}}}$ puisse être de toutes façons très différent de G_0. Nous seri-

ons donc dans la situation où la correspondance des mesures sur les données réelles et l'échantillon de comparaison dérive d'un effet statistique.

Nous fermons ici la parenthèse sur la signification statistique des mesures obtenues, certes importante pour l'interprétation des résultats, mais secondaire pour ce qui est de la description de la méthode.

2.7 Limites d'utilisation

Nous avons mis en avant la polyvalence de la famille d'algorithmes décrite, qui est son principal intérêt. Mais cette qualité a des contreparties que nous allons chercher à évaluer ici. À nouveau, cette partie peut être difficile en première lecture car elle ne s'appuie pas seulement sur les mesures expérimentales ou des formules théoriques, mais aussi sur des arguments plus discutables qui relèvent de l'usage que l'on fait des algorithmes d'échange. Cependant, elle est fondamentale à nos yeux, car elle permet de séparer les problèmes importants à résoudre pour rendre la méthode opérationnelle de ceux qui sont en fait assez secondaires.

2.7.1 Limite fondamentale

Une limite forte de cette méthode tient à sa nature intrinsèquement expérimentale, ce qui se traduit au travers de deux aspects.

a. Atteindre l'uniformité

Comme pour l'algorithme à base d'échanges simples (c.f. 2.2.2.g.), il est difficile de prévoir à l'avance le nombre de pas nécessaires - à valeur de k fixée - pour assurer la convergence. Dans le cas de l'ensemble $\mathbf{C_{min}}$, l'expérience montre que l'on peut considèrer que la chaîne a convergé lorsque le nombre d'échanges effectivement réalisé est de l'ordre de L [GMZ03]. Mais cette observation ne tient pas en augmentant les contraintes: sur $\mathbf{C_{tri}}$, il fallait réaliser plusieurs centaines de millions d'échanges pour observer le palier avec un graphe où $L < 300$.

Nous avons donc choisi d'évaluer la vitesse de convergence à l'aide d'un ensemble de mesures-tests, mais expérimentalement, on constate que celle-ci est très variable et on peut observer jusqu'à plusieurs ordres de grandeur de différence en fonction de la mesure (toutes choses égales par ailleurs).

Donc, même si le processus choisi est ergodique, nous ne pouvons affirmer avec certitude qu'il n'existe pas une mesure en-dehors de notre ensemble-test dont la convergence ne soit pas achevée au pas de temps d'arrêt. C'est pourquoi il est recommandé de prendre plus que le nombre de pas *a priori* nécessaires à la convergence: si on pense

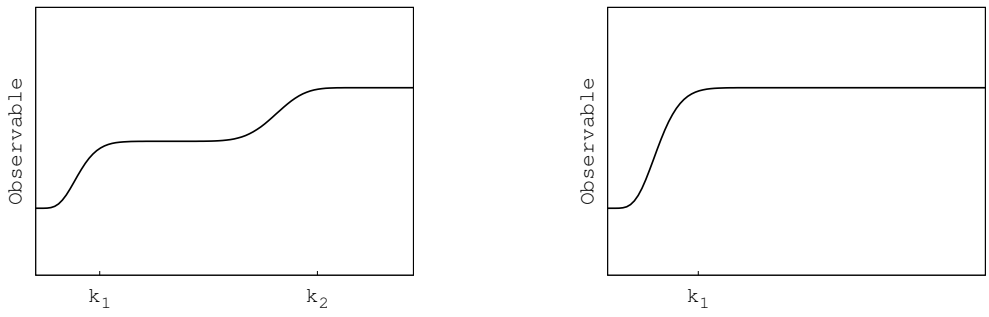
avoir atteint un palier à t tentatives pour toutes les mesures tests, on pourrait prendre par exemple $10.t$ itérations.

Nous pouvons aussi contourner ce problème en précisant l'objectif dans lequel nous travaillons: les graphes produits sont souvent destinés à être comparés aux réseaux réels, une telle comparaison est réalisée sur la base de mesures topologiques. Nous nous assurerons donc que ces mesures de comparaison ont effectivement convergées dans l'échantillon de graphes synthétiques. Et il faudrait alors parler de convergence par rapport à une certaine mesure. Cela peut sembler simple à réaliser mais on peut rencontrer quelques difficultés pratiques: en particulier, nous avons essentiellement envisagé des mesures scalaires, mais si nous souhaitions tester une mesure vectorielle, ou encore le comportement du graphe vis-à-vis d'un processus complexe, la notion de convergence de la mesure est moins triviale et il faudra donc en discuter le cas échéant.

En bref, l'objectif d'obtenir un échantillon uniformément aléatoire de graphes est accessible, mais nous n'avons pas de moyen incontestable de savoir si un tel objectif est effectivement atteint, nous pouvons seulement améliorer notre degré de confiance en **élargissant la gamme de mesures et le nombre d'itérations**.

b. Atteindre l'ergodicité

Nous évaluons également à l'aide de critères expérimentaux l'identité des états stationnaires en fonction de k. Il n'est pas possible - sauf cas particuliers - de tester la valeur $k = L$ dont nous sommes certains qu'elle converge vers l'état uniformément aléatoire sur tout l'ensemble à décrire. En pratique, nous serons bien souvent cantonnés à de "petites" valeurs de k (quelques unités). Par conséquent, il n'est pas totalement exclu de pouvoir observer des situations comme celle représentée à gauche, plutôt que le schéma supposé usuel de convergence en fonction de k, à droite:



Mais, outre le fait que nous n'avons pas observé de situations de ce type en pratique, nous avons des raisons de croire qu'elles seraient tout à fait exceptionnelles, et donc que **l'étude de petites valeurs de k est suffisante**. En effet, s'il existe deux valeurs k_1 et k_2 comme sur la figure, nous pouvons l'interpréter comme l'existence

d'un nombre caractéristique de l'ensemble de graphes contraints considéré: il existe des structures gelées pour une taille k_1 qui ne le sont plus à k_2, ce qui suppose que ces structures mettent en jeu k_2 liens ou moins. D'ailleurs, c'est ce que nous observions sur l'exemple des modèles-jouets (2.5.1): la contrainte gèle les triangles colorés totalement pour $k = 2$, partiellement pour $k = 3$, mais pas au-delà. En bref, il serait contre-intuitif qu'il puisse exister une taille caractéristique radicalement différente de celle introduite par la contrainte - excepté dans certains cas "pathologiques".

Un autre élément en faveur de cette idée consiste à imaginer sur les quelques exemples vus (comme la figure 2.9), le nombre considérable de métaliens créés par l'incrémentation de k. D'un nombre de l'ordre de C_k^N combinaisons de liens possibles, on passe à C_{k+1}^N combinaisons, soit $\frac{N-k}{k+1}$ fois plus[6], alors que le nombre de métanœuds (i.e. le cardinal de l'ensemble de graphes) reste lui inchangé. Bien sûr, cet argument reste très qualitatif et discutable puisque beaucoup de ces nouveaux métaliens sont des boucles, et rien n'assure qu'ils permettent de rendre connexe le métagraphe.

Une fois que nous avons conscience des hypothèses qui sous-tendent l'utilisation de cette méthode, nous ne la modifions pas, mais nous pouvons améliorer le degré de confiance que nous avons en son application. Cela signifie en pratique: élargir l'ensemble de mesures-tests et augmenter le nombre de pas temps pour le problème de l'uniformité; accroître la gamme des k testés pour celui de l'ergodicité.

2.7.2 Limite pratique: vitesse, complexité

La réalisation de ces améliorations est concrètement conditionnée par la vitesse de l'algorithme, qui est le problème pratique majeur auquel nous sommes confrontés - ce qui explique d'ailleurs la petite taille des échantillons que nous sommes parfois amenés à considérer.

Cette question est d'autant plus délicate à discuter que nous construisons une famille d'algorithmes sur laquelle nous ne pouvons pas donner de classes de complexité générale, d'une part parce que celle-ci dépend de la contrainte examinée, d'autre part parce qu'elle ne dépend pas simplement des caractéristiques du graphe.

Ainsi, si on a généré un échantillon en τ itérations pour une contrainte à partir d'un graphe G_0, on ne sait rien du nombre d'itérations nécessaires pour générer un échantillon depuis un autre graphe G'_0 pour la même contrainte. Par exemple, il faut approximativement le même nombre d'itérations de l'algorithme pour observer la convergence avec la contrainte $\mathbf{C_{tri}}$ avec le graphe *AIO: Scandinavie*, qui compte quelques centaines de nœuds et de liens, et le graphe *arXiv*, qui en a plusieurs dizaines de milliers*, nous développons cet exemple en 3.2.3.e.

[6]Pour l'expression exacte, c.f. 2.6.1.

Nous discutons ici quelques aspects de ce problème:

- **Complexité en temps.** En suivant la description de l'algorithme (c.f. Annexe D) et si nous choisissons N comme variables pour décrire sa complexité, une étape élémentaire consiste en un tirage aléatoire de liens ($\mathcal{O}(\log N)$), un test des contraintes, puis éventuellement un échange ($\mathcal{O}(1)$).

  La classe de complexité du test des contraintes additionelles dépend évidemment de la nature de celles-ci, mais dans les exemples que nous traitons, il est rare de pouvoir la décrire avec la seule variable N. Cependant, la vitesse d'une étape élémentaire est presque toujours déterminée par ce test et c'est donc sur celui-ci que sont concentrés les efforts d'optimisation.

- **Taux de réussites.** Une fois la chaîne de Markov (la valeur de k) choisie, le nombre d'étapes nécessaire à l'obtention d'un échantillon uniformément aléatoire ne dépend plus de notre volonté, mais de la nature de la contrainte. Pour $\mathbf{C_{min}}$, le taux de réussite des tentatives d'échanges est maximum (par rapport à toutes les autres contraintes que nous envisageons). Pour des contraintes "fortes", ce taux décroît considérablement, comme l'indique le nombre de succès mesurés dans les exemples de 2.5.

  L'augmentation de la valeur de k ne conduit pas nécessairement à une baisse du taux de réussites, on le voit par exemple sur $\mathbf{C_{compo}}$ où le taux maximum de réussites correspond à $k = 3$, pour un état stationnaire identique (selon notre critère expérimental). Mais sur les exemples vus - et à venir - on observe la même allure: un maximum atteint pour une petite valeur de k, puis une décroissance rapide.

- **Taux de mélange.** Parallèlement, plus la valeur de k est élevée, plus le k-échange modifie la structure du graphe, et donc du point de vue de la marche aléatoire dans le métagraphe, plus elle nous éloigne du point de départ. Nous dirons alors que nous augmentons le "taux de mélange" de la chaîne de Markov.

  La vitesse de calcul dépend de ces trois éléments: la complexité de l'étape élémentaire (qui croît avec k), les taux de réussites et de mélange du processus markovien.

- **Complexité en espace.** En ce qui concerne la mémoire, l'espace occupé consiste essentiellement en l'espace nécessaire au stockage du graphe, que l'on modifie au fur et à mesure. Le type de données tableau de listes tire parti du caractère épars du graphe, ce n'est pas vrai du type matrice (c.f. 1.2.2).

  En fonction de la contrainte, nous travaillons sur des graphes dont la taille peut varier de la centaine à quelques dizaines de milliers de nœuds; au-delà les algorithmes convergent trop lentement pour devenir utilisables pratiquement, on

utilise donc au plus quelques Mo en type tableau de liste. La complexité en espace n'est donc pas problématique pour cette méthode tant que nous utilisons ce type.

Dans le chapitre suivant, nous allons mettre en œuvre notre méthode dans des contextes pratiques où elle peut être très utile car les contraintes imposées seront fortes. Nous serons alors confrontés aux difficultés dont nous faisions ici état, il faudra alors proposer et mettre en œuvre des heuristiques pour les contourner.

Chapitre 3

Applications pratiques

Sommaire

La famille d'algorithmes exposée dans le chapitre précédent a été conçue dans l'idée d'être appliquée à des réseaux sociaux collaboratifs. Or, comme nous l'avons vu précédemment, les classes usuelles de modèles ne décrivent pas précisément les caractéristiques géométriques de tels réseaux. En permettant de produire de nouvelles références dans l'espace des graphes, la génération de graphes aléatoires satisfaisant des contraintes flexibles permet de répondre à deux besoins au moins. D'abord elle permet de déterminer des éléments géométriques essentiels du graphe: nous recherchons un petit ensemble de caractéristiques suffisantes pour expliquer l'ensemble des propriétés du réseau réel. En effet, un modèle permettant de restituer ces éléments essentiels pourrait être une proposition de mécanisme régissant la constitution du réseau social. Les parties 3.1 et 3.2 portent sur des applications de ce type. Le premier cas abordé n'est pas d'ordre sociologique mais plutôt méthodologique: il met en parallèle notre approche et une étude menée récemment sur la structure de l'Internet. Nous revenons ensuite à une application centrée sur les réseaux collaboratifs, en cherchant à explorer la notion d'activité au travers d'éléments topologiques. D'autre part, les réseaux sociaux sont le siège d'une variété de processus de diffusion qui font l'objet de beaucoup d'attention. Notre méthode permet de synthétiser des graphes aux propriétés ajustables, nous pourrons alors estimer comment les caractéristiques topologiques affectent ces processus. Nous étudierons ce problème complexe de l'influence des éléments géométriques sur le transport dans un réseau d'échanges commerciaux en 3.3.

3.1 Génération de *"dK-graphs"*

Nous avons souligné qu'une lacune de la plupart des modèles classiques de réseaux complexes tient à leur incapacité à prendre en compte les corrélations entre les éléments de topologie du graphe. Les corrélations de degrés entre voisins en sont la forme la plus simple et traduisent des phénomènes intuitifs comme la tendance à l'assortativité dans les réseaux sociaux.

Dans cette partie nous appliquons notre méthode à l'analyse des corrélations telle qu'elle a été réalisée pour étudier la structure physique de l'Internet: nous résumons d'abord la méthode et les résultats d'une étude de la littérature, puis nous mettons en œuvre notre méthodologie dans un contexte similaire et comparons les résulats obtenus, enfin nous appliquons la technique de validation statistique proposée précédemment pour discuter les conclusions de l'article original.

3.1.1 Contexte

Dans un article déjà cité [MKFV06], Mahadevan *et al.* proposent d'étudier les corrélations de degré à l'aide de la famille des *"dK random graphs"*. Il s'agit d'une suite d'ensembles

de graphes simples inclus les uns dans les autres, construite à partir d'un graphe G_0 de manière à ce que le $d^{\text{ème}}$ ensemble en conserve les corrélations jusqu'à la distance d; ainsi elle converge vers le singleton $\{G_0\}$ (et l'atteint pour une valeur de d finie), ce que nous représentons sur le schéma 3.1.

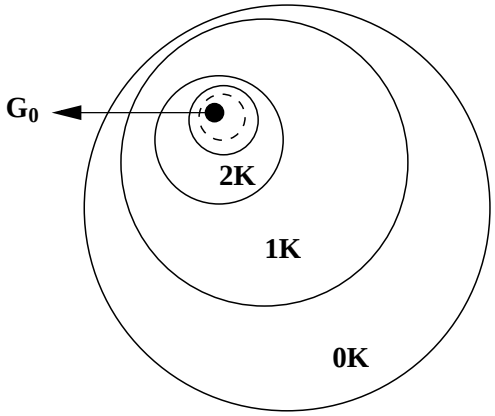


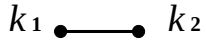
FIG. 3.1: Représentation schématique de la suite d'ensembles dK. Adaptée de [MKFV06].

Nous faisons ici une rapide présentation de cette étude qui nous sert de toile de fond, en décrivant sa méthodologie et ses conclusions.

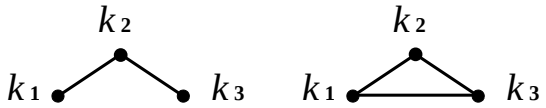
a. Définitions

Reconstruire les corrélations à distance d signifie ici que les distributions des degrés de toutes les structures connexes à d éléments restent identiques. En pratique, la suite d'ensembles considérée se limite à:

- $d = 0$, ensemble des graphes aléatoires de même N et L que le graphe réel (ce sont donc des graphes d'Erdős-Rényi).
- $d = 1$, ensemble $\mathcal{F}_{1K} = \mathcal{E}_{\mathbf{C_{min}}}$ des graphes aléatoires de même distribution de degré que G_0.
- $d = 2$, ensemble $\mathcal{F}_{2K} \subset \mathcal{F}_{1K}$ des graphes aléatoires reproduisant la distribution des motifs à deux nœuds de G_0:



- $d = 3$, ensemble $\mathcal{F}_{3K} \subset \mathcal{F}_{2K} \subset \mathcal{F}_{1K}$ des graphes aléatoires reproduisant la distribution des motifs à trois nœuds de G_0:



b. Bases de données

Les auteurs examinent la structure de l'Internet* à différents niveaux hiérarchiques: à l'échelle des connexions entre systèmes autonomes (AS) et à celle des routeurs (R). Pour chacun de ces deux niveaux, l'étude se concentre sur deux sources:

- Au niveau AS : ils étudient une acquisition sur un mois de *traceroute* (graphe *skitter*).
- Au niveau R : il existe une grande variété de topologies selon la fonction de la sous-partie de l'Internet examinée, ils emploient ici un modèle synthétique proposé dans [LAWD04]: *Heuristically Optimal Topology* (HOT), qui serait une topologie adaptée pour un fournisseur d'accès.

Les auteurs souhaitent générer les ensembles dK dérivés de ces graphes, puis repérer la topologie vis-à-vis de ces références, leur permettant ainsi d'émettre des hypothèses sur la profondeur à laquelle se structure l'Internet à chacune de ces échelles. Une autre perspective de l'étude serait d'utiliser ces graphes artificiels pour simuler sur des topologies réalistes des phénomènes réels tels que le routage des paquets d'information, l'adaptation du réseau à des attaques ciblées etc.

c. Comparaison aux graphes réels

Les auteurs choisissent un ensemble de mesures pour estimer la proximité de leurs reconstructions vis-à-vis du réseau réel - qui seront également pour nous des tests pour la convergence des algorithmes d'échanges. Parmi celles-ci, nous retenons: l'assortativité (r), le *clustering* local moyen ($\overline{c}_{3-l}$), la centralité d'intermédiarité (C_I), ainsi que les premier ($\overline{d}$) et second (σ_d) moments de la distribution des distances entre paires de nœuds de la composante géante - les autres mesures de l'article n'apportant pas de conclusions fondamentalement différentes pour ce qui va suivre.

d. Génération des graphes synthétiques

Pour les petites valeurs de d, il est possible de recourir aux algorithmes de *matching*[1]: modèles stochastiques ou *configuration model* que les auteurs adaptent jusqu'au cas $d = 2$. En revanche au-delà, et conformément à ce que nous faisions remarquer en

[1]Dans [MKFV06], le terme de *matching* se réfère à un cas particulier de la famille que nous regroupons sous cette dénomination.

2.4, il faut recourir aux procédures d'échange. Mahadevan *et al.* utilisent alors deux techniques:

- Le *3K-randomizing rewiring*, qui correspond à une chaîne classique d'échanges simples. D'après la discussion du chapitre précédent, on ne peut pas affirmer que cette méthode soit uniforme pour ces contraintes.

- Le *3K-targeting rewiring*, que nous allons développer dans la partie 3.3. Il s'agit, partant d'un élément de $\mathcal{F}_{dK}$, de "viser" un élément de $\mathcal{F}_{d+1K}$ à l'aide d'une procédure d'échanges. L'idée essentielle consiste à employer une mesure pour évaluer la distance entre le graphe de $\mathcal{F}_{dK}$ et l'ensemble $\mathcal{F}_{d+1K}$ et à n'effectuer un échange que si celui-ci diminue cette distance. Cette procédure présente l'avantage de ne pas nécessiter un élément de l'ensemble. Nous n'en détaillons pas plus le principe pour l'instant, mais elle n'est *a priori* ni ergodique, ni uniforme.

e. Résultats et conclusions

Lorsque plusieurs types d'algorithmes sont possibles, tous produisent des résultats cohérents[2] (les écarts de valeurs n'excèdent pas 10%). Cette observation est importante dans notre perspective: cela tend à montrer que la stratégie d'échanges simples adoptée pourrait être suffisante dans ce contexte pour générer des échantillons dont les propriétés sont proches de l'uniformément aléatoire.

Par ailleurs, les mesures menées sur les graphes synthétiques convergent vers celles du graphe réel lorsque d croît. Les structures semblent proches pour $d = 2$ au niveau AS, pour $d = 3$ au niveau R, selon l'appréciation des auteurs. Nous reprenons dans la table 3.1 et la figure 3.2 ci-dessous certains des résultats obtenus au niveau R.

Observables	$0K$	$1K$	$2K$	$3K$	HOT (réel)
$\overline{\delta}$	2,47	2,59	2,18	2,10	2,10
r	-0,05	-0,14	-0,23	-0,22	-0,22
$\overline{c}_{3-l}$	0,002	0,009	0,001	0	0
$\overline{d}$	8,48	4,41	6,32	6,55	6,81
σ_d	1,23	0,72	0,71	0,84	0,57

TAB. 3.1: Mesures topologiques sur HOT et les modèles dK. Extraites de [MKFV06].

Nous pouvons faire quelques remarques sur ces mesures:
– Certaines sont par construction restituées par les modèles contraints: $\overline{\delta}$ par tous, r pour 2K et plus. Ce n'est qu'approximativement le cas ici, alors que la méthode d'échanges telle que nous l'avons décrite permet de conserver strictement ces caractéristiques. Cela indique soit que les graphes sont générés par *matching*,

[2]Hormis les méthodes stochastiques, qui ne sont pas fiables sur ces topologies.

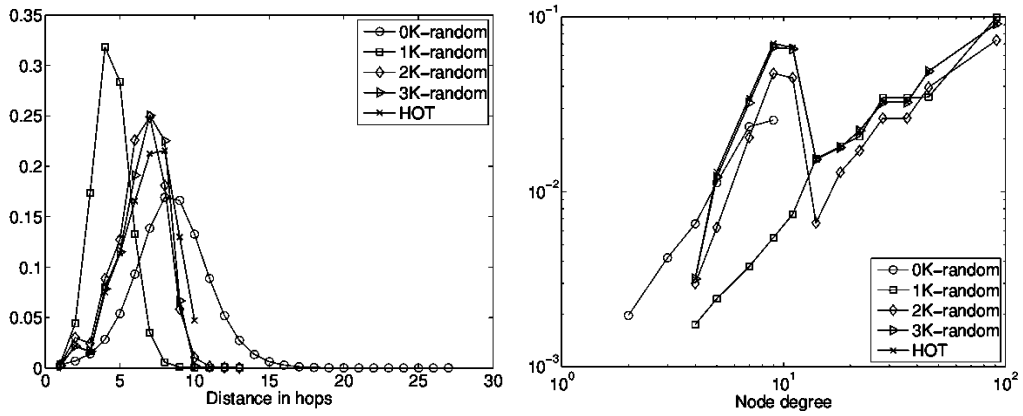


FIG. 3.2: Distribution des distances (gauche) et de la centralité d'intermédiarité en fonction du degré (droite) pour le graphe *HOT* et les différents modèles dK. Extrait de [MKFV06].

soit que les échanges sont menés non pas en imposant exactement la valeur des contraintes, mais un encadrement de celles-ci[3].

– La distribution des distances tracées dans l'article ne correspond pas aux valeurs définies usuellement comme l'écart-type σ_d. On s'affranchit par la suite de cette mesure, dans le doute sur l'origine du problème.

Enfin, la question de la signification statistique des résultats est soulevée dans l'étude. En effet, le graphe *HOT* comportant moins d'un millier de nœuds, les auteurs constatent qu'à partir des contraintes $d = 3$, le nombre d'échanges effectivement possibles dans le graphe G_0 deviendrait très faible.

3.1.2 Démarche

a. Objectif

Nous nous posons la question de la validité méthodologique de la démarche suivie dans [MKFV06]. Lorsque les échantillons sont générés exclusivement par des algorithmes de *switching* (en pratique pour $d = 3$), il faut s'interroger sur l'interprétation des résultats. Comme nous l'avons discuté en 2.6.1, nous en distinguons trois possibles:

1. L'ensemble de contraintes suffit à produire des graphes artificiels proches du graphe réel parce que les éléments topologiques choisis contiennent l'essentiel de l'information expliquant la structure du graphe. La démarche est alors **valide et concluante**.
2. L'ensemble de contraintes suffit à produire des graphes artificiels proches du graphe réel par un simple effet statistique: l'ensemble à décrire est trop petit (ou

[3]Ce genre de procédure permet d'accélérer la convergence, ce que nous discuterons par la suite.

contient trop de graphes isomorphes les uns des autres) pour que ses éléments soient très différents de G_0. La démarche est alors **valide mais non concluante**.
3. La marche aléatoire à base d'échanges simples n'est pas ergodique mais cantonnée à un sous-ensemble statistiquement non-significatif. La démarche **n'est pas valide**.

Afin de trancher entre ces trois possibilités, nous mettons en œuvre les k-échanges dans un cadre qui est le plus proche possible de celui de cette étude afin de s'assurer que les graphes produits par des procédures de *dK-randomizing rewiring* sont effectivement utilisables dans ce contexte.

b. Données, mesures

La question de la validité se pose surtout au niveau R, où la petite taille du graphe et le fort niveau de contraintes peuvent compromettre l'ergodicité. Nous chercherons à produire jusqu'aux graphes 3K sur ce niveau.

Comme nous ne disposons pas exactement des données employées dans [MKFV06], nous reprenons le réseau tel qu'il figure dans l'article dont il est adapté [LAWD04], dont la structure est proche (c.f. Tab. 3.2). On y fera référence par le nom *HOT'*.

	N	L	$\overline{\delta}$	r	$\overline{c}_{3-l}$	$\overline{d}$
HOT	939	988	2,10	-0,22	0	6,81
HOT'	830	878	2,12	-0.19	0	6,54

TAB. 3.2: Comparaison entre les structures des graphes *HOT* et *HOT'*.

c. Ensembles décrits

La famille 1K ne pose pas de problème puisque la chaîne est irréductible par théorème, il ne peut donc pas y avoir de problème de validité. Nous produisons donc des échantillons de graphes 2K et 3K.

> **remarque :** On peut noter que même si 3K ne pose pas de problème d'ergodicité avec des échanges simples (ce qui reste à prouver), cela ne garantit pas que la chaîne relative à 2K soit elle aussi irréductible. Plus généralement, il n'est pas exclu qu'avec des contraintes $\mathbf{C_A} \subset \mathbf{C_B}$, le processus soit ergodique pour $\mathbf{C_B}$ et non pour $\mathbf{C_A}$. Autrement dit, en augmentant les contraintes, il n'est pas impossible - même si cela semble peu probable - que le métagraphe devienne connexe.

Les algorithmes associés suivent une logique analogue à celle de la contrainte $\mathbf{C_{tri}}$: à chaque pas, on dénombre les structures locales détruites ou créées dans le voisinage des nœuds impliqués dans l'échange:

– pour 2K: les paires de nœud (avec le degré de chacun),
– pour 3K: les triangles et les chemins de longueur 2 (là encore avec le degré de chaque sommet).

Pour plus détails, on peut se reporter à l'Annexe D.

3.1.3 Résultats

a. Ensemble 2K

Nous testons la convergence à l'aide de diverses mesures de motifs et de la distance moyenne. La détermination des distances entre les nœuds est une opération coûteuse algorithmiquement, elle est ici possible en raison de la petite taille du graphe. Comme le montrent les courbes de $\overline{d}$ (Fig. 3.3), les mesures dans l'état stationnaire produisent des résultats indiscernables pour toutes les valeurs de k testées. Nous estimons donc que $k^* = 2$ est suffisant pour créer un échantillon considéré comme uniformément aléatoire.

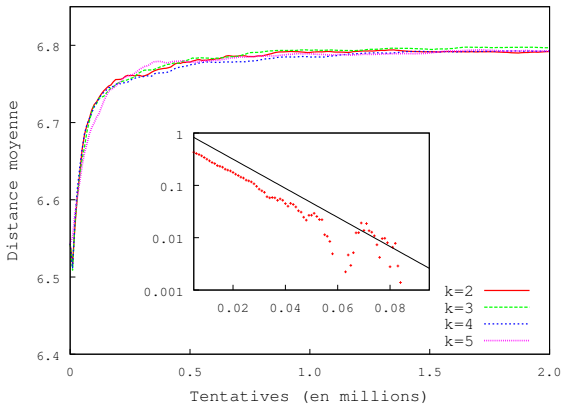


FIG. 3.3: Modèle 2K: moyenne cumulée de $\overline{d}$ selon le nombre de tentatives de k-échanges, $k \in [\![2, 5]\!]$ (échantillons de 50 simulations - encart: vitesse sur 500 simulations pour $k = 2$).

b. Ensemble 3K

Le protocole est identique pour 3K et les conclusions du point de vue de la méthode également: comme les mesures tendent vers la même valeur pour tous les k testés, selon notre critère le seuil $k^* = 2$.

Cela justifie *a posteriori* l'usage d'algorithmes d'échanges simples dans [MKFV06] et écarte alors la troisième hypothèse de a.: on parcourt bien une partie significative du métagraphe, la méthode d'échanges suivie est valide pour 2K comme pour 3K.

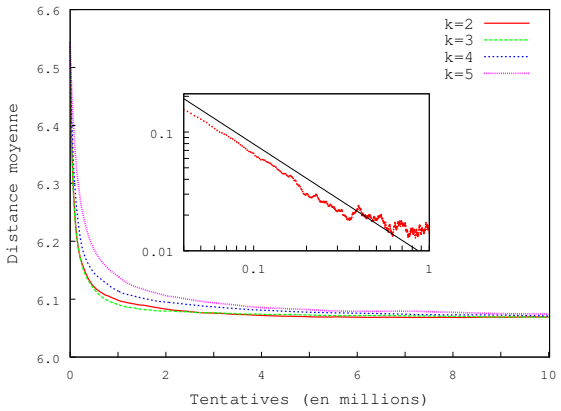


FIG. 3.4: Modèle 3K: moyenne cumulée de $\overline{d}$ selon le nombre de tentatives de k-échanges, $k \in [\![2, 5]\!]$ (échantillons de 50 simulations - encart: vitesse sur 500 simulations pour $k = 2$).

c. Remarque sur la vitesse de convergence

Nous faisons ici une parenthèse relative à la vitesse de convergence telle que définie en 2.2.2.g.: nous avions vu dans cette partie qu'avec une contrainte simple (en l'occurrence $\mathbf{C_{min}}$), un modèle exponentiel de convergence pouvait être satisfaisant, mais que selon la mesure utilisée, le temps caractéristique variait.

Les encarts des figures 3.3 et 3.4 sont les tracés, pour les contraintes 2K et 3K, de $|\overline{d}(x) - \overline{d}(\infty)|$ avec x le nombre d'itérations, pour $k = 2$ et en moyennant sur 500 réalisations sur la partie significative de la convergence, au-delà le signal est très bruité. On constate que si pour la contrainte 2K (fig. 3.3), un modèle exponentiel est relativement satisfaisant, en revanche, ce n'est pas le cas pour 3K (fig. 3.4), où la convergence est plus proche d'une loi de puissance (d'exposant $\simeq 0,96$).

Nous voyons donc que le comportement asymptotique d'allure exponentielle prévu théoriquement n'est pas nécessairement visible expérimentalement, mais dépend de la contrainte et du graphe.

d. Comparaison des résultats

Résultats. Nous regroupons les résultats des modèles 2K et 3K relativement au graphe *HOT'* pour les mesures d'intérêt (c.f. table 3.3 et figure 3.5).

Observables	*HOT'*	2K	3K
$\overline{\delta}$ (degré moyen)	2,12	2,12	2,12
r (assortativité)	-0,21	-0,21	-0,21
$\overline{d}$ (distance moyenne)	6,54	6,80	6,06

TAB. 3.3: Mesures sur les modèles 2K et 3K, obtenues depuis le graphe *HOT'*.

Pour effectuer la comparaison avec l'article original en s'affranchissant du fait que les graphes de référence ne sont pas tout à fait identiques dans les deux cas, nous employons pour chaque mesure $m_{d\mathrm{K}}$ sur le modèle dK, la grandeur normalisée:

$$Z_{dK} = \frac{m_{d\mathrm{K}} - m_{\mathrm{ref}}}{m_{\mathrm{ref}}}$$

m_{ref} désignant selon le cas la mesure sur *HOT* ou *HOT'* (c.f. Tab. 3.4).

	Cas *HOT'* (notre méthode)		Cas *HOT* (dans [MKFV06])	
Observables	$Z_{2\mathrm{K}}$	$Z_{3\mathrm{K}}$	$Z_{2\mathrm{K}}$	$Z_{3\mathrm{K}}$
$\overline{\delta}$	0,00	0,00	0,04	0,00
r	0,00	0,00	0,04	0,00
$\overline{d}$	0,04	-0,07	-0,07	-0,04

TAB. 3.4: Résultats normalisés comparés à ceux obtenus dans [MKFV06] sur le graphe *HOT*.

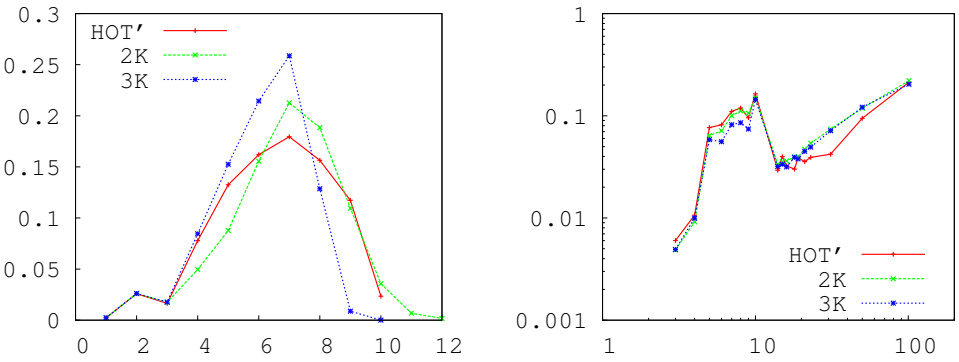


FIG. 3.5: Distribution des distances (gauche) et centralité en fonction du degré (droite).

Discussion. Les mesures $\overline{\delta}$ et r produisent les résultats attendus: les écarts des mesures de distance sur *HOT'* sont cohérents avec ceux sur *HOT*. Mais nous pouvons observer sur les courbes de distribution des distances et de centralité d'intermédiarité (Fig. 3.5) que les écarts semblent plus importants ici qu'ils ne le sont dans l'étude originale (Fig. 3.2). Surtout, la conclusion selon laquelle les graphes 3K reconstruisent mieux le graphe initial que 2K reste discutable, les écarts entre chacun des modèles et *HOT'* étant du même ordre selon toutes les mesures.

Cette dernière observation peut étonner: selon les indications de Mahadevan *et al.* on pouvait s'attendre en effet à ce que les graphes 3K soient plus proches de *HOT'* par un effet de petite taille de l'ensemble.

3.1.4 Validation statistique

Nous entreprenons alors de refaire le dénombrement des permutations autorisées pour k^* selon la méthode du 2.6.1.

a. Nombre d'échanges autorisés

Le dénombrement amène la table de valeurs 3.5, à comparer aux valeurs sur *HOT*. On rappelle que n_e désigne le nombre d'échanges possibles depuis le graphe de départ, L_e le nombre de liens distincts impliqués dans ces échanges; n_e^{eff} et L_e^{eff} sont les valeurs correspondantes pour les échanges qui ne sont pas des isomorphismes. Comme les valeurs de L_e et L_e^{eff} ne figure pas dans [MKFV06], nous ne rapportons que n_e et n_e^{eff}.

HOT'	2K	3K
n_e	$\simeq$ 390.000	$\simeq$ 110.000
n_e^{eff}	24.939	2.793

HOT	2K	3K
n_e	326.409	146
n_e^{eff}	268.871	44

TAB. 3.5: Valeurs du nombre d'échanges n_e et du nombre d'échanges effectifs n_e^{eff} (non-isomorphes) dans le graphe de départ selon les processus des deux modèles. Le cas *HOT'* est évalué par notre méthode, les valeurs pour *HOT* sont celles figurant dans [MKFV06].

Les résultats comparés des dénombrements sont très surprenants: on attendrait que pour un même modèle, les dénombrements sur *HOT* et *HOT'* soient du même ordre de grandeur. De tels écarts sont injustifiables par les seules différences topologiques entre les deux graphes, alors comment peut-on les expliquer?

- La seule mesure correspondante - en ordre de grandeur - est celle de n_e pour le modèle 2K, on peut donc supposer que les deux méthodes autorisent les mêmes permutations. Alors, l'écart sur la mesure n_e^{eff}(2K) doit simplement venir du fait que certains isomorphismes que nous excluons sont conservés dans l'article original.

- Les résultats sur le modèle 3K sont plus problématiques: soit nous surestimons le nombre d'échanges autorisés, soit l'article original les sous-estiment. Mais dans le premier cas, nous pourrions très facilement le constater puisque le graphe généré présenterait des distributions de corrélations à trois nœuds différentes. Par conséquent, l'explication la plus cohérente à nos yeux serait que certains échanges qui devraient être autorisés ne sont pas permis par l'algorithme de [MKFV06]

Cela explique que dans l'article original, les propriétés des graphes 3K soient si proches de celles du graphe *HOT*: en diminuant drastiquement le nombre de permutations autorisées, l'échantillon de graphes obtenu décrit un ensemble très restreint autour du graphe d'origine, mais pas l'ensemble recherché $\mathcal{F}_{\mathbf{3K}}$.

b. Comparaisons à l'ensemble de vérification

Par ailleurs, nous générons un ensemble de vérification de 50 graphes à partir des valeurs mesurées pour le modèle 3K. Nous rappelons que l'objectif de cette procédure est d'identifier si les mesures observées pourraient être un simple artefact statistique, en conséquence du fait que l'ensemble décrit est trop petit. Les échantillons produits ont des caractéristiques sensiblement différentes de celles de *HOT'*, par exemple sur les chemins de longueur 4 (c.f. Table 3.6), ce qui nous permet d'écarter la seconde hypothèse du a.: il ne s'agit pas d'un cas non-concluant.

	HOT'	$\mathcal{F}_{\mathbf{3K}}$	$\mathcal{F}_{\mathbf{vérif}}$
⬠	38.297	23.900 ± 1.400	76.000 ± 14.000

TAB. 3.6: Mesure du nombre de chemins de longueur 4 pour l'ensemble $\mathcal{F}_{\mathbf{3K}}$ comparée à une moyenne sur des ensembles de vérification $\mathcal{F}_{\mathbf{vérif}}$.

Cela modifie l'interprétation des résultats: dans [MKFV06], les auteurs suggèrent que le modèle 3K reconstruit la topologie au niveau routeur, probablement en raison d'un effet de petite taille de l'ensemble (on serait donc dans le cas d'une procédure non concluante). Notre conclusion est qu'il n'y a pas de tel effet, mais que le modèle 3K ne saisit pas sensiblement mieux la topologie routeur que ne le faisait déjà le 2K.

3.1.5 Conclusion

En résumé, nous avons montré dans cette partie que la démarche suivie dans [MKFV06] est *a posteriori* valide: les échanges simples permettent effectivement de générer un échantillon uniformément aléatoire représentatif de l'ensemble. Le fait que les caractéristiques topologiques de ces graphes soient proches du réseau réel n'est pas attribuable à la taille de l'ensemble décrit, en revanche, la démarche proposée n'est pas correctement mise en œuvre, ce qui amène une erreur quant à l'interprétation des résultats.

La méthodologie suivie dans cette étude nous semble intéressante car elle est utilisable de manière générique. Il s'agit d'identifier un ensemble de graphes qui reproduise la topologie du réseau réel et qui soit un élément d'une suite d'ensembles inclus les uns dans les autres. Pour que cela soit efficace, la famille de contraintes doit être définie par les caractéristiques que nous pensons fondamentales. C'est ce modèle de démarche que nous allons adopter pour l'appliquer aux réseaux sociaux.

3.2 Contraintes de connectivité dans les réseaux de collaborations

Dans les modèles d'attachement préférentiel et plus généralement les diverses propositions de mécanismes de constitution des réseaux, il est tacitement accepté que les comportements locaux des acteurs suffisent à expliquer la structure globale. Cette idée paraît vraisemblable dans les réseaux sociaux où les agents ont souvent une connaissance très partielle de l'ensemble de la structure, si bien que la création du réseau semble guidée par des interactions locales.

La rétroaction de l'environnement sur les agents existe, bien qu'elle n'apparaisse qu'indirectement dans le graphe: ceux-ci agissent en fonction de réglements, d'usages ou de contraintes matérielles. Nous nous efforcons d'en rechercher la trace au travers de diverses mesures. Dans cette partie, on explore à quel point le graphe est contraint par la connectivité des nœuds qui le constituent; cela serait une manière simple de traduire la limite d'activité des acteurs du réseau.

3.2.1 Données d'exploration

Partant de l'hypothèse que la limite d'activité soit effectivement une contrainte primordiale pour rendre compte de la structure du réseau, nous recherchons des contextes où la participation à un événement représente un investissement pour les agents du réseau (sous quelque forme que ce soit: financièrement, en temps, en réflexion). On suppose que le degré serait une bonne évaluation de ce coût[4] (en moyennant sur l'ensemble de la population). Cela n'est pas nécessairement le cas: on peut par exemple douter de l'exactitude de ceci pour les réseaux de commentaires sur des sites de mise en commun de contenus (photos, vidéos, blogs), où la participation est presque gratuite.

Il est très difficile de satisfaire - et même seulement de tester - une telle hypothèse avec les données à disposition. Nous nous proposons donc de travailler sur divers types de collaborations*, où circule une information spécialisée qui réclame une forme d'expertise, et donc un certain coût pour les agents:

- des publications d'articles scientifiques dans diverses disciplines (*arXiv, SW*),
- des cercles de décision politiques ou économiques (*DutchElite*),
- des forum ou des pages de discussion sur Internet (*Wiki*).

Soulignons que nous ne prétendons pas produire une analyse en profondeur du contenu de ces formes de collaboration: plutôt que de répondre à la question "quels sont les phénomènes expliquant les topologies de ces réseaux à cette échelle?", nous

[4]Nous pouvons comprendre cette supposition comme la traduction de l'hypothèse d'homogénéité dans ce contexte.

apportons sur un exemple concret des éléments pour comprendre comment il serait possible de rechercher ces phénomènes.

3.2.2 Choix du modèle

Nous évoquons les modèles existants pour expliquer le choix de l'ensemble sur lequel nous allons concentrer cette étude.

a. Lien avec $\mathbf{C_{min}}$

Notre intention est de mesurer l'effet des limites de connectivité dans le réseau social, nous allons donc faire usage des modèles associés à la contrainte $\mathbf{C_{min}}$ (i.e. la contrainte de distribution de degré). Plus précisément, nous distinguons:

- Le modèle que nous dénommerons $\mathbf{GC_{min}}$ (**G** pour *"Graphic"*), qui reprend la distribution de degré de la projection monopartie entre acteurs du graphe, rendant ainsi compte des contraintes sur le nombre d'acteurs avec qui chacun peut effectivement interagir. Nous utiliserons la notation $\mathbf{GC_{min}}$ pour qualifier indifféremment le modèle ou la contrainte qui lui est associée.

  Mais ainsi que nous l'indiquent par exemple les résultats de simulations sur le graphe de collaborations artistiques *Allmusic* (c.f. 2.2.2.g.), les écarts sont importants entre des mesures topologiques basiques sur un échantillon de graphes de $\mathbf{GC_{min}}$ et les valeurs réelles. Ce modèle s'avère donc insuffisant pour rendre compte de l'essentiel des éléments topologiques de réseaux intrinsèquement bipartis.

- Cette observation amenait Newman *et al.* [NSW01] à utiliser le modèle $\mathbf{HC_{min}}$ (**H** pour *Hypergraphic*), où l'on conserve la structure d'hyperliens sous-jacente.

  Mais dans bon nombre de cas, cela ne s'avère pas non plus satisfaisant: rien n'impose les corrélations entre acteurs de différents événements, alors que l'on sait par exemple que la répétition d'interactions est fréquente dans les réseaux collaboratifs [TCR10].

  En conséquence, le degré des acteurs et la taille des événéments restant identiques, on observe typiquement une surestimation du nombre de voisins de la projection monopartie. Une rapide mesure le montre: la densité du graphe projeté des acteurs d'*arXiv* est de 3,60 contre 4,88 pour un graphe quelconque de l'ensemble $\mathcal{F}_{\mathbf{HC_{min}}}$.

b. Contrainte de redondance $\mathbf{C_{Red}}$

Face aux insuffisances des modèles précédents, on souhaiterait contraindre au moins la forme la plus élémentaire de corrélations entre les hyperliens: la répétition des interactions entre acteurs. Or c'est précisément ce que cherche à estimer la redondance définie en 1.3.3.c., nous allons alors imposer à cette grandeur une certaine valeur moyenne sur le graphe.

c. Point de vue des répétitions d'interactions

En imposant la redondance globale, on contraint les répétitions d'interactions entre agents à l'échelle du graphe entier. Ce n'est pas équivalent à imposer le nombre exact de répétitions[5], mais peut-être plus significatif, car la répétition d'interaction au cours de grands événements (ce qui a peu d'influence sur la redondance) a plus de chances d'être accidentelle qu'une répétition au cours de petits.

En revanche un tel modèle ne saisit pas *a priori* des structures séquentielles plus complexes (c.f. 1.3.4.a.), mais celles-ci pourraient être le produit indirect de la combinaison des contraintes mises en jeu. L'étude de $\mathbf{C_{Red}}$ peut donc nous aider à savoir si les structures de ce type sont "accidentelles" ou au contraire la trace de corrélations à plus longues distances entre les nœuds du graphe.

3.2.3 Accélérations de l'algorithme

La mise en pratique de la contrainte $\mathbf{C_{Red}}$ va poser des difficultés de vitesse de convergence. C'est pour nous l'occasion de discuter de manière plus générale l'accélération de l'algorithme et diverses heuristiques utilisables dans cet objectif.

Des développements théoriques rigoureux seraient très délicats car la vitesse de l'algorithme dépend simultanément de la complexité des tests, du taux de réussite et du taux de mélange (voir 2.7.2). Nos arguments resteront donc essentiellement qualitatifs, le critère essentiel en faveur d'une heuristique étant le temps nécessaire pour observer la convergence.

a. Remarques générales

Nous faisons ici quelques remarques pratiques élémentaires communes à tous les algorithmes de cette famille pour optimiser leur vitesse d'exécution.

- Comme le plus souvent l'étape qui limite la vitesse de l'algorithme est le test de la (ou des) contrainte(s) additionnelle(s) - c'est-à-dire celles qui ne relèvent

[5]Des mesures indiquent même que cela peut-être très différent: *arXiv* a une fraction d'interactions répétées de $0,35$, alors qu'un graphe de l'ensemble $\mathcal{F}_{\mathbf{C_{Red}}}$ qui lui est associé a en moyenne une fraction de l'ordre de $0,10$.

pas de la méthode d'échange usuelle, celui-ci doit être efficace, en particulier on privilégie les tests locaux aux globaux. Par exemple, pour les contraintes dK (c.f. 3.1), on ne produira pas la totalité des distributions de corrélations à chaque test, mais on examine uniquement la sous-partie de la distribution susceptible d'être modifiée.

- Par ailleurs, il est moins coûteux de détecter les isomorphismes les plus simples (tels qu'on les a définis en 2.6.1.d.) et de les considérer comme des échecs plutôt que de réaliser un échange qui corresponde à un réétiquetage et n'a donc strictement aucune incidence sur les propriétés topologiques du graphe.

- Autre remarque élémentaire: lorsqu'il faut tester plusieurs contraintes successivement, on ordonne les tests de manière à ce que ceux de complexité moindre soient réalisés en premier, cela évite de tester inutilement les contraintes lourdes en cas d'échec. Par exemple, lorsque l'on doit tester que le graphe obtenu est simple (pas de boucle, ni de lien multiple) - ce qui peut être fait en $\mathcal{O}(1)$ - on le fera avant un test de contrainte du type nombre de triangles ($\mathbf{C_{tri}}$), qui serait en $\mathcal{O}(\overline{\delta}^4)$ - c.f. Ann. D.

b. Fenêtre temporelle, problème du taux de réussite

Dans la littérature sur l'amélioration de la vitesse de convergence des processus dans le cadre d'échanges simples, [GMZ03, VL05] utilisent le fait qu'il n'est pas nécessaire de tester les contraintes à chaque itération.

La méthode proposée consiste alors à utiliser une fenêtre τ, et à tester la contrainte seulement toutes les τ itérations: en cas de réussite de l'échange on augmente τ, et on la diminue en cas d'échec. De cette manière τ oscille dans un intervalle de valeurs qui permet de tester le moins souvent possible la contrainte tout en produisant des graphes majoritairement dans l'ensemble à décrire.

Malheureusement, il n'est que rarement possible d'employer cette technique dans le contexte qui nous occupe. En effet, la méthode n'est utilisable que si le taux de réussite du test est proche de 1 et en tous cas supérieur à $0,5$: au-dessous, les échecs sont plus nombreux que les succès et il n'y aura pas de gain. Or sur des contraintes complexes, le taux de réussite sera presque toujours inférieur à ce seuil - excepté dans de rares cas tels $\mathbf{C_{compo}}$ évoqué précédemment.

c. Compromis réussite-mélange

Une fois atteint la valeur seuil k^* des k-échanges, toutes les chaînes convergent vers le même état stationnaire quelle que soit la valeur de k. Il n'est pas nécessaire (bien que cela soit souvent le cas), que la valeur seuil permette d'atteindre l'état stationnaire plus rapidement qu'une valeur de k plus élevée. En effet si le taux de succès des k-échanges

a plutôt tendance à décroître une fois le palier atteint, en revanche le taux de mélange croît nécessairement: chaque itération réussie modifie davantage la structure du graphe que ne le ferait une valeur inférieure de k.

Mais à nouveau, cette méthode ne permet un gain que pour une contrainte dont le taux de réussite est élevé, car dans le cas contraire, le passage de $k(\geq k^*)$ à $k+1$ s'accompagne d'une telle décroissance du taux de réussite que le meilleur compromis réussite-mélange correspondra presque certainement à $k = k^*$.

La contrainte $\mathbf{HC_{min}}$ a en général un taux de réussite proche de 1, nous examinons alors la convergence sur celle-ci appliquée au graphe *arXiv*. La figure 3.6 montre effectivement que les chaînes de Markov à $k = 3$ ou $k = 4$ convergent plus rapidement que celle de $k = 2$ - les abscisses représentent ici le temps (et non pas le nombre d'itérations).

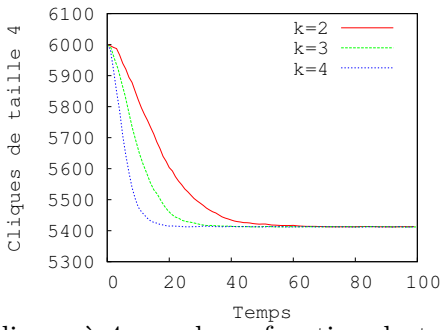


FIG. 3.6: Nombre de cliques à 4 nœuds en fonction du temps (unité arbitraire).

d. Utiliser plusieurs racines

En partant de racines différentes, on peut intuitivement supposer que le mélange est déjà en partie effectué et qu'il faut donc un nombre de pas inférieur pour obtenir un échantillon de graphes quelconques de l'ensemble. Nous en donnons une illustration dans le cas du modèle 3K sur la Figure 3.7, sur laquelle nous constatons que l'état stationnaire est effectivement atteint plus rapidement que ce qui avait été observé en 3.1.3 (environ 2 fois plus vite).

> **remarque :** Cette manière de procéder peut être utile lorsque l'on cherche à évaluer k^*: après une première série d'itérations pour $k = 2$, on a généré un échantillon de graphes dont on peut se servir comme racines pour la convergence à $k = 3$ et ainsi de suite.

e. Relaxation de contrainte

Enfin, et bien qu'il ne s'agisse pas à proprement parler d'une accélération de la même procédure puisque l'échantillon décrit aura des caractéristiques différentes, une solution

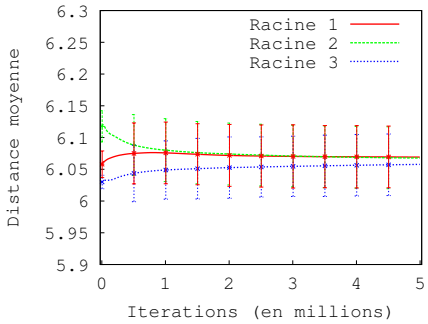


FIG. 3.7: Illustration de la convergence depuis plusieurs racines sur le modèle 3K: on génère 25 graphes avec $k = 3$ depuis 3 sources différentes produites avec $k = 2$.

peut être d'alléger les contraintes imposées. Nous voulons en fait insister ici sur le fait que lever - même très peu - les contraintes peut avoir une incidence considérable sur la vitesse de convergence.

Il n'y a bien sûr pas de méthode systématique pour appliquer cette idée qui dépend de l'interprétation de la contrainte. Nous étudions sa mise en pratique sur l'exemple particulier de $\mathbf{C_{tri}}$ appliquée à *arXiv*.

- On impose d'abord strictement la valeur du nombre de triangles du graphe, comme en 2.5.2. La dérivée du nombre de motifs en fonction du nombre d'itération est tracée en Figure 3.8 : nous constatons qu'on modèle exponentiel avec un temps caractéristique de convergence $\tau = 140.10^6$ tentatives est satisfaisant lorsque nous suivons la convergence avec pour observable le motif à 4 noeuds:  , que nous qualifierons de "double-triangle".

- Nous autorisons maintenant un nombre de triangles situé dans un intervalle de ± 1 autour de la valeur mesurée sur le graphe initial. La Figure 3.8 nous montre que pour la même observable et le même modèle, on peut estimer un temps caractéristique de $\tau = 70.10^6$, soit deux fois plus court. D'autres observables et d'autres critères rendent compte d'une accélération du même ordre.

La relaxation de contrainte produit donc une accélération efficace du processus sur cet exemple. Comment comprendre cette observation? Relâcher les contraintes implique une augmentation de la taille de l'ensemble dans lequel on se déplace, mais il devient également possible d'emprunter de nouvelles voies pour se déplacer d'un point à un autre du métagraphe. De manière imagée, on pourrait comprendre l'accélération globale observée comme la création de raccourcis dans le métagraphe.

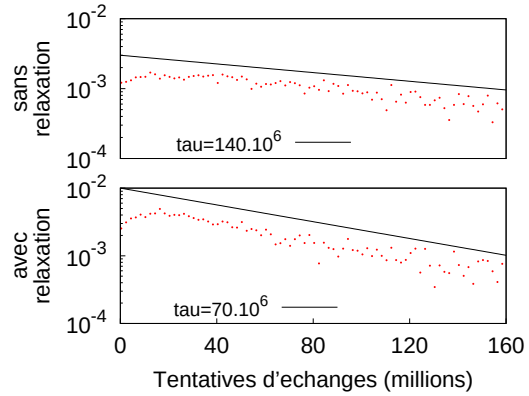


FIG. 3.8: Dérivée de la moyenne des motifs double-triangle au cours de la convergence du graphe *arXiv*. En haut: on fixe strictement le nombre de triangles, en bas: on fixe le nombre de triangles à ± 1.

remarque : Par ailleurs, l'observation du palier sur cet exemple nécessite plusieurs dizaines de milliards d'itérations. C'est-à-dire que le nombre d'itérations nécessaires à la convergence de l'algorithme est du même ordre que ce que l'on avait observé pour le graphe *AIO Scandinavie* (c.f. 2.5.2), de densité comparable mais dont le nombre de nœuds est 60 fois inférieur. **Cela met en évidence que le temps de convergence n'est pas une fonction simple de la taille des graphes**.

f. Conclusion sur la relaxation de contrainte

Nous avons vu au cours de cette partie un certain nombre de moyens dont l'utilisateur dispose afin de contourner la principale limite pratique de nos algorithmes: leur vitesse de convergence. Pour qu'un tel outil puisse être utilisé dans une vaste gamme de domaines, il serait très bénéfique de pousser plus avant les réflexions relatives à la complexité algorithmique et la vitesse de mélange. Nous refermons maintenant cette parenthèse pour reprendre le fil de notre application.

3.2.4 Protocole

À partir de **la plus grande composante connexe**[6] des bases explorées, nous produisons des échantillons de 25 graphes avec une précision à deux chiffres significatifs sur la redondance. La contrainte se traduit alors, sur les quatre bases dont nous allons détailler les résultats par:

Données	*arXiv*	*SW*	*DutchElite*	*Wiki(d)*
$\overline{\text{rc}} \in$	[0,40 ; 0,41]	[0,58 ; 0,59]	[0,017 ; 0,018]	[0,26 ; 0,27]

[6]C'est pourquoi les résultats de cette partie diffèrent des mesures de 1.3.4.a. qui sont réalisées sur la base complète.

Nous comparons ensuite certaines caractéristiques topologiques des projections de graphes produits à ce que l'on obtient avec les modèles usuels de graphes de réseaux sociaux $\mathbf{GC_{min}}$ et $\mathbf{HC_{min}}$:

- le nombre de motifs locaux de la projection,
- puis plus particulièrement la nature de ces motifs: séquentielle ou structurelle, puisque le modèle ne contient pas *a priori* cette information sur les données.

Pour reprendre une image analogue à celle de [MKFV06] dans le cas des graphes dK, nous pourrions représenter les ensembles relatifs aux contraintes de la manière suivante:

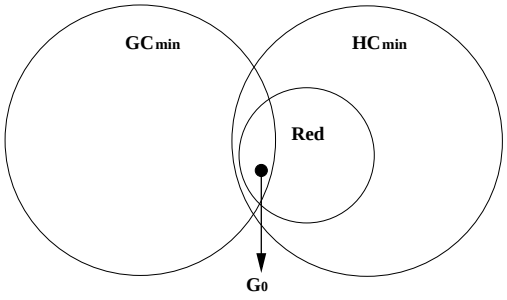


FIG. 3.9: Représentation schématique des ensembles contraints et du graphe réel.

3.2.5 Résultats de mesures toplogiques

a. Motifs locaux

Les résultats des mesures brutes sont rassemblés dans le tableau 3.7, nous nous focalisons sur les chemins et les motifs cycliques, qui sont les plus intéressants pour l'analyse que nous allons faire de ces résultats. Afin d'en faciliter la lecture, les résultats en orange correspondent à des reconstructions à un facteur f tel que $2 > f > 1,5$, et pour les résultats en rouge: $f > 2$.

On observe d'abord que le modèle $\mathbf{GC_{min}}$ restitue par définition le nombre de chemins de longueur 2. Conformément à ce que l'on attendait, en dehors de cette qualité, il reconstruit de manière insuffisante les caractéristiques des graphes (en particulier les motifs cycliques), c'est pourquoi nous ne considérerons plus ce modèle par la suite.

$\mathbf{HC_{min}}$ reproduit les caractéristiques du graphe *DutchElite*, ce qui avait déjà été observé dans des contextes comparables et semble donc être une caractéristique générique à certains graphes d'affiliations à des institutions politiques ou économiques [NSW01].

$\mathbf{HC_{min}}$ reproduit également *Wiki (d)* avec une qualité moindre. Ce n'est en revanche pas le cas des bases de collaborations scientifiques dans lesquelles le nombre de

		∧	△	◇	◇	⬠	⬠
arXiv	**Réel**	**225 $\cdot 10^3$**	**16,7 $\cdot 10^3$**	**2,05 $\cdot 10^6$**	**42,5 $\cdot 10^3$**	**20,3$\cdot 10^6$**	**159 $\cdot 10^3$**
	$\mathbf{GC_{min}}$	225 $\cdot 10^3$	102	1,92 $\cdot 10^6$	672	16,4 $\cdot 10^7$	4,55 $\cdot 10^3$
	$\mathbf{HC_{min}}$	556 $\cdot 10^3$	17,7 $\cdot 10^3$	8,48 $\cdot 10^6$	22,0 $\cdot 10^3$	129 $\cdot 10^6$	94,5 $\cdot 10^3$
	Red	354 $\cdot 10^3$	17,4 $\cdot 10^3$	3,86 $\cdot 10^6$	17,4 $\cdot 10^3$	42,1 $\cdot 10^6$	28,9 $\cdot 10^3$
SW	**Réel**	**72,2 $\cdot 10^3$**	**20,7 $\cdot 10^3$**	**2,76 $\cdot 10^6$**	**646 $\cdot 10^3$**	**120 $\cdot 10^6$**	**23,0 $\cdot 10^6$**
	$\mathbf{GC_{min}}$	72,2 $\cdot 10^3$	1,9 $\cdot 10^3$	1,59 $\cdot 10^6$	30,6 $\cdot 10^3$	34,7 $\cdot 10^6$	0,53 $\cdot 10^6$
	$\mathbf{HC_{min}}$	129 $\cdot 10^3$	21,9 $\cdot 10^3$	5,22 $\cdot 10^6$	681 $\cdot 10^3$	246 $\cdot 10^6$	23,0 $\cdot 10^6$
	Red	99,9 $\cdot 10^3$	21,2 $\cdot 10^3$	4,26 $\cdot 10^6$	661 $\cdot 10^3$	193 $\cdot 10^6$	23,8 $\cdot 10^6$
DutchElite	**Réel**	**810 $\cdot 10^3$**	**188 $\cdot 10^3$**	**36,3 $\cdot 10^6$**	**5,97 $\cdot 10^6$**	**2,01 $\cdot 10^9$**	**276 $\cdot 10^6$**
	$\mathbf{GC_{min}}$	810 $\cdot 10^3$	4,84 $\cdot 10^3$	24,9 $\cdot 10^6$	0,11 $\cdot 10^6$	0,76 $\cdot 10^9$	2,73 $\cdot 10^6$
	$\mathbf{HC_{min}}$	868 $\cdot 10^3$	188 $\cdot 10^3$	39,4 $\cdot 10^6$	5,93 $\cdot 10^6$	2,21 $\cdot 10^9$	273 $\cdot 10^6$
	Red	872 $\cdot 10^3$	188 $\cdot 10^3$	40,4 $\cdot 10^6$	5,93 $\cdot 10^6$	2,30 $\cdot 10^9$	274 $\cdot 10^6$
Wiki (d)	**Réel**	**383 $\cdot 10^3$**	**69,0 $\cdot 10^3$**	**23,4 $\cdot 10^6$**	**2,83 $\cdot 10^6$**	**1,48 $\cdot 10^9$**	**134 $\cdot 10^6$**
	$\mathbf{GC_{min}}$	383 $\cdot 10^3$	25,2 $\cdot 10^3$	19,9 $\cdot 10^6$	0,99 $\cdot 10^6$	1,04 $\cdot 10^9$	41,1 $\cdot 10^6$
	$\mathbf{HC_{min}}$	438 $\cdot 10^3$	73,3 $\cdot 10^3$	28,0 $\cdot 10^6$	3,12 $\cdot 10^6$	1,83 $\cdot 10^9$	154 $\cdot 10^6$
	Red	441 $\cdot 10^3$	73,6 $\cdot 10^3$	29,2 $\cdot 10^6$	3,27 $\cdot 10^6$	1,67 $\cdot 10^9$	197 $\cdot 10^6$

TAB. 3.7: Mesures de motifs de tailles 3, 4 et 5 sur les différentes bases de données. La dispersion des mesures varie selon le motif et le graphe, mais n'excède jamais 5% excepté pour les mesures de motifs cycliques du modèle $\mathbf{GC_{min}}$ (qui peuvent atteindre 10%).

chemins est surévalué et les motifs cycliques au contraire sous-évalués.

On vérifie que la contrainte $\mathbf{C_{Red}}(\supset \mathbf{HC_{min}})$ produit une reconstruction de qualité comparable des graphes d'affiliations institutionnelles et de discussions en ligne, ce nouveau modèle apporte donc peu d'éléments nouveaux vis-à-vis de ces bases. En revanche, il permet une reconstruction améliorée du graphe *SW* en diminuant le nombre de chemins. L'effet existe aussi pour *arXiv* mais sans permettre de produire un échantillon dont les caractéristiques soient du même ordre que celles du graphe réel. On note d'ailleurs que du point de vue des motifs cycliques, le modèle **Red** n'apporte pas d'amélioration vis-à-vis de $\mathbf{HC_{min}}$.

b. Structure globale

En plus de ces mesures élémentaires de motifs, nous examinons la distribution de distance de la composante géante de chacun de ces graphes (de taille N_c) pour saisir

sa structure de manière plus globale[7].

	arXiv			*SW*			*DutchElite*			*Wiki (d)*		
	Réel	**HC$_{\mathbf{min}}$**	**Red**	**Réel**	**HC$_{\mathbf{min}}$**	**Red**	**Réel**	**HC$_{\mathbf{min}}$**	**Red**	**Réel**	**HC$_{\mathbf{min}}$**	**Red**
N_c	**11.654**	11.053	10.909	**765**	730	704	**3.620**	3.405	3.405	**889**	697	677
$\overline{d}$	**7,18**	4,61	5,04	**7,55**	3,22	3,73	**4.47**	3,89	3,89	**3,22**	2,80	2,81
σ_d	**1,87**	0,92	0,99	**3,03**	0,84	1,11	**1.25**	0,92	0,92	**0,96**	0,78	0,80

TAB. 3.8: Premier et second moments de la distribution de distances (resp. $\overline{d}$ et σ_d).

Nous constatons que dans les modèles, les nœuds sont systématiquement plus "proches" que dans le cas réel et que la distribution est plus piquée. C'est particulièrement visible sur *arXiv* et *SW*, les modèles **HC$_{\mathbf{min}}$** et **Red** utilisent des hypothèses locales qui ne restituent pas la structure à longue portée des graphes lorsque celle-ci est élaborée (collaborations scientifiques).

c. Corrélations entre hyperliens

Nous cherchons maintenant l'origine de ces observations élémentaires en effectuant des mesures plus fines. Comme nous l'avons dit, le modèle **Red** cherche à contraindre la forme la plus simple de corrélations entre hyperliens: les répétitions d'interactions; en revanche, il ne contient pas *a priori* de structures séquentielles plus subtiles, nous mesurons donc le nombre de motifs structurels (c.f. Tab. 3.9) et vérifions que les deux modèles **HC$_{\mathbf{min}}$** et **Red** produisent des valeurs comparables aux valeurs réelles dans tous les cas.

À la lumière de ces résultats il possible d'interpréter les résultats au niveau de la structure graphique, c'est-à-dire de comprendre en quoi les contraintes induisent ou non la présence de certains types de motifs, et l'influence que cela aura sur les mesures. En effet on constate que les nombres d'éléments structurels sont reconstruits par les deux modèles, et c'est donc aux éléments séquentiels que tiennent les écarts observés:

- Concernant les graphes *DutchElite* et *Wiki*, l'un et l'autre sont reconstruits par le modèle **HC$_{\mathbf{min}}$** de manière correcte, mais pour des raisons différentes[8]:
  - Dans *DutchElite*, il y a peu de corrélations entre les hyperliens: la redondance est faible et les structures séquentielles inexistantes.
  - En revanche ces effets existent dans *Wiki*, mais sont approximativement restitués par les distributions de degré biparties.

[7]Nous pouvons effectuer cette comparaison parce que la plus grande composante connexe des modèles est de taille voisine de celle du graphe initial (à 10 % environ).

[8]On observe d'ailleurs que la base *TheyRule* (conseils d'administration de sociétés) produit des résultats analogues à *DutchElite* et *Debian* (forum) à *Wiki*.

		△ *str*	◇ *str*	⬠ *str*
arXiv	**Réel**	**15,2 $\cdot 10^3$**	**14,6 $\cdot 10^3$**	**12,5 $\cdot 10^3$**
	$\mathbf{HC_{min}}$	17,3 $\cdot 10^3$	15,4 $\cdot 10^3$	12,8 $\cdot 10^3$
	Red	17,3 $\cdot 10^3$	15,4 $\cdot 10^3$	12,8 $\cdot 10^3$
SW	**Réel**	**20,7 $\cdot 10^3$**	**642 $\cdot 10^3$**	**22,9 $\cdot 10^6$**
	$\mathbf{HC_{min}}$	21,1 $\cdot 10^3$	642 $\cdot 10^3$	22,9 $\cdot 10^6$
	Red	21,0 $\cdot 10^3$	642 $\cdot 10^3$	22,9 $\cdot 10^6$
DutchElite	**Réel**	**188 $\cdot 10^3$**	**5,90 $\cdot 10^6$**	**271 $\cdot 10^6$**
	$\mathbf{HC_{min}}$	188 $\cdot 10^3$	5,90 $\cdot 10^6$	271 $\cdot 10^6$
	Red	188 $\cdot 10^3$	5,90 $\cdot 10^6$	271 $\cdot 10^6$
Wiki (d)	**Réel**	**66,1 $\cdot 10^3$**	**2,45 $\cdot 10^6$**	**105 $\cdot 10^6$**
	$\mathbf{HC_{min}}$	66,5 $\cdot 10^3$	2,45 $\cdot 10^6$	105 $\cdot 10^6$
	Red	66,4 $\cdot 10^3$	2,45 $\cdot 10^6$	105 $\cdot 10^6$

TAB. 3.9: Nombre de cycles structurels de taille 3, 4 et 5, proches des valeurs réelles pour les deux modèles $\mathbf{HC_{min}}$ et **Red**.

- Concernant les bases de données de collaborations scientifiques, non seulement les corrélations sont fortes, mais l'information n'est pas contenue dans la contrainte $\mathbf{HC_{min}}$.
  - La reconstruction d'*arXiv* reste très imprécise: la contrainte $\mathbf{C_{Red}}$ limite en partie le nombre de répétitions, mais les effets séquentiels sont tels que le nombre de motifs cycliques est largement sous-estimé.
  - Pour *SW*, l'amélioration est plus sensible: les effets séquentiels existent mais sont suffisamment faibles pour que le nombre de motifs cycliques soient proches de la valeur réelle, l'effet de multiplication des chemins est limité par la contrainte de redondance.

remarque : On peut d'ailleurs comprendre le fait que la reconstruction des motifs cycliques soit meilleure avec le modèle $\mathbf{HC_{min}}$, c'est le résultat d'une "compensation d'erreurs": celui-ci sous-estime les motifs séquentiels mais la multiplication des chemins fait qu'accidentellement, certains produiront des cycles.

3.2.6 Tests de modèles diffusifs

Nous avons insisté sur le rôle du réseau social en tant que support à la transmission d'information; pour évaluer la qualité des modèles, nous pouvons tester leur comportement sur des simulations de phénomènes de ce type.

Dans ce but, il est fréquent de recourir à des modèles stylisés de propagations épidémiques, qui font l'objet d'une importante littérature (pour une revue: [Het00]).

L'idée commune est de répartir la population en groupes (sains, infectés, immunisés) et d'autoriser au moyen de taux de transition le passage d'un groupe à un autre. La topologie du graphe vient se superposer à cette description en ne rendant possible le contact qu'entre nœuds adjacents du réseau.

Les travaux existant sur cette question mettent en évidence que les topologies synthétiques de graphes sont souvent déficientes et amènent des résultats très différents des topologies réelles (c.f. [CR07]).

a. Processus dynamique

Ici, nous cherchons uniquement à mesurer l'influence de la structure sur le processus de contamination. C'est pourquoi, nous procédons à nouveau comme dans [CR07], où les auteurs choisissent une dynamique la plus simple possible pour comparer les effets des différentes topologies. Celle-ci est conçue de la manière suivante:

- chaque agent à deux états possibles: S (sain) ou I (infecté),
- initialement une fraction λ $(= 0,01)$ des nœuds est dans l'état I,
- à chaque pas de temps, on choisit un nœud et un de ses voisins aléatoirement,
- si un des deux agents est I, son voisin devient ou reste I.

Et nous mesurerons alors au cours du processus la fraction de nœuds dans l'état I.

b. Résultats

Après une moyenne sur 100 simulations pour chaque graphe des échantillons, nous observons les courbes de contamination regroupées dans la figure 3.10 (les encarts représentent la dérivée de la courbe).

Nous constatons que le modèle **Red** n'amène pas d'amélioration significative sur la simulation de ces processus, même pour les graphes *arXiv* et *SW* où l'on pourrait penser que la limitation du nombre de chemins locaux permettrait de "ralentir" la diffusion. Une interprétation possible serait que le processus de diffusion simulé est fortement dépendant de la structure globale du graphe.

3.2.7 Conclusion

Cet exemple nous semble intéressant non pas tant par ses résultats - car le modèle **Red** ne fournit qu'une amélioration très relative via-à-vis de $\mathbf{HC_{min}}$ - que par la manière avec laquelle on cherche à identifier comment une contrainte influence les propriétés topologiques du graphe.

Cela nous a permis ici de distinguer des classes de bases de données où "la structure à longue portée" est peu élaborée et dans ce cas les éléments locaux suffisent à saisir l'essentiel de la topologie; et des bases où ces modèles s'avèrent insuffisants et dont la

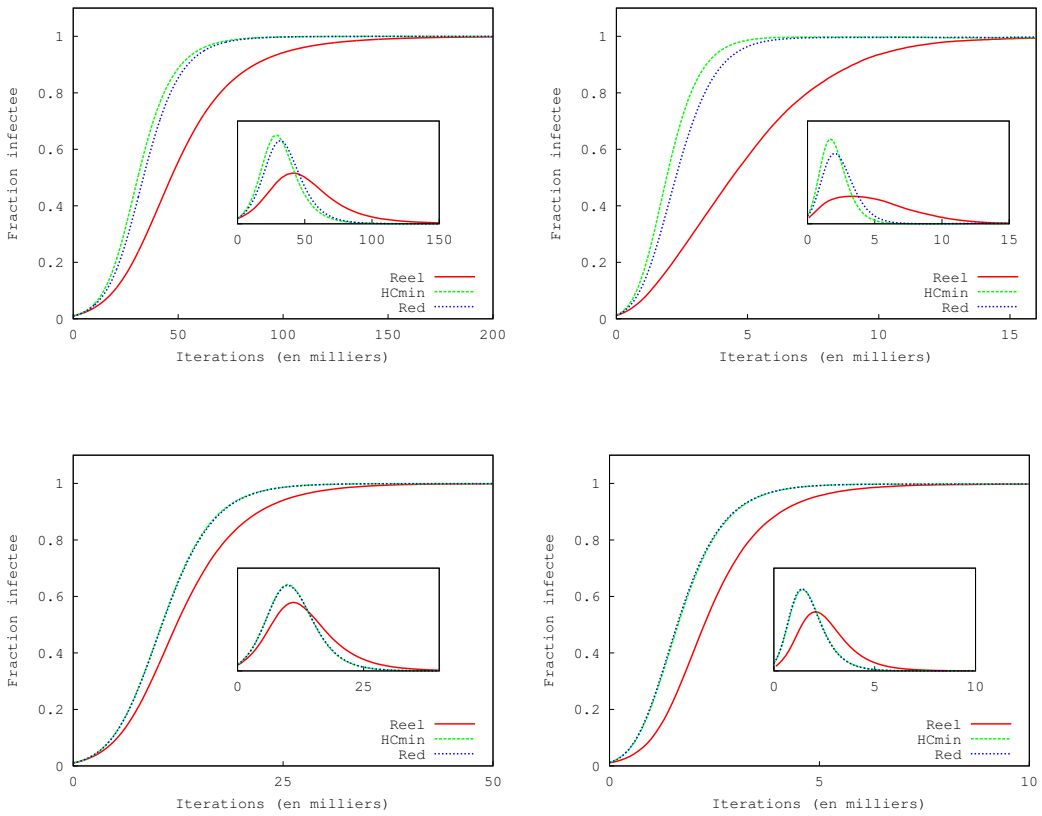


FIG. 3.10: Fraction des nœuds contaminés (et dérivée en encart) au cours de la simulation du processus sur le graphe réel et les modèles $\mathbf{HC_{min}}$, **Red**. En haut (de gauche à droite): *arXiv*, *SW*; en bas: *DutchElite*, *Wiki*.

structure du graphe contient plus d'information sur les mécanismes de constitution du réseau.

Un prolongement de cette exploration consisterait à imposer une contrainte locale plus élaborée qui reconstruise le nombre de chemins avec une meilleure précision; et à plus long terme on pourrait chercher à restituer les corrélations entre liens de l'hypergraphe à longue distance.

3.3 Méthodologie de ciblage

Au cours des parties 3.1 et 3.2, nous avons décrit des applications dont le but est de déterminer les éléments géométriques essentiels du réseau, c'est-à-dire des caractéristiques suffisantes pour reconstruire la topologie de celui-ci. L'objectif poursuivi ici est différent: nous supposons disposer d'un certain réseau sur lequel se produit un

processus de propagation, et nous voulons alors estimer comment la modification des caractéristiques topologiques du réseau affecte le processus.

Il faudrait alors être en mesure de générer un graphe ayant des propriétés fixées à partir d'un graphe qui ne les a pas. Nous proposons une modification des procédures d'échange qui permet de rechercher dans l'ensemble des graphes un sous-ensemble vérifiant des propriétés particulières[9]. Nous qualifierons cette pratique de **ciblage**, puisque contrairement aux cas précédents, nous allons d'abord effectuer une marche aléatoire biaisée afin d'atteindre la cible visée.

Nous allons appliquer cette variante des méthodes d'échanges sur une étude de cas spécifique: la propagation d'une perturbation dans un réseau commercial. Mais la portée du ciblage est en fait très vaste: elle permet (en théorie) de générer un graphe aléatoire vérifiant un ensemble de propriétés choisies quelconques.

3.3.1 Principe du ciblage

Jusqu'à présent nous utilisions un élément de l'ensemble dont nous voulons générer un échantillon aléatoire comme état initial de la chaîne de Markov. Supposons que nous ne disposions pas d'un tel graphe, mais d'un graphe ne satisfaisant qu'un sous-ensemble $\mathbf{C}'$ des contraintes souhaitées (avec $\mathbf{C_{min}} \subseteq \mathbf{C}' \subset \mathbf{C}$), ce graphe est donc un élément de l'ensemble $\mathcal{F}_{\mathbf{C}'}$ des graphes satisfaisant $\mathbf{C}'$, et $\mathcal{F}_{\mathbf{C}} \subset \mathcal{F}_{\mathbf{C}'}$.

Le problème consiste alors à chercher dans $\mathcal{F}_{\mathbf{C}'}$ un élément de $\mathcal{F}_{\mathbf{C}}$. Pour ce faire nous pourrions en principe effectuer une série de permutations simples ($k = 2$), jusqu'à atteindre un graphe de l'ensemble. En effet, si nous prenons $\mathbf{C}' = \mathbf{C_{min}}$ (pour des graphes simples, non-orientés) et effectuons une marche aléatoire classique, nous savons que la marche est exhaustive [Egg73]; par conséquent nous allons nécessairement visiter un élément de $\mathcal{F}_{\mathbf{C}}$ en un temps fini.

Mais nous avons vu en 2.4.2 qu'une telle marche n'a qu'une probabilité infime de toucher un élément de $\mathcal{F}_{\mathbf{C}}$ en un temps raisonnable. Il sera donc souvent beaucoup plus efficace de "guider" les permutations simples. Le principe général est le suivant: on attribue un score permettant de mesurer à quel point le graphe est proche du sous-ensemble souhaité. **Si la permutation permet d'approcher le score de la valeur souhaitée, elle est effectivement réalisée.** Ainsi, supposons que nous cherchions à générer des graphes ayant une certaine valeur d'assortativité r_0, il est naturel de choisir $r - r_0$ comme mesure de l'écart au sous-ensemble.

> **remarque :** En revanche, si l'ensemble recherché ne peut pas être simplement défini à l'aide d'un scalaire, il nous faut choisir une mesure adaptée. Par exemple, si nous cherchons à reproduire une certaine distribution, on choisira un test caractérisant l'écart entre deux jeux de valeurs discrètes (type χ^2, Kolmogorov-Smirnov ou autre, selon le type de la distribution).

[9] Selon le même principe que 3.2.6.

Le processus d'échange devient alors assimilable à un algorithme d'optimisation. Cela induit également que nous ne savons pas *a priori* si l'objectif est effectivement accessible, et la procédure d'optimisation peut nous piéger dans un minimum local de la mesure d'écart. Cependant, si nous atteignons effectivement un élément de l'ensemble, les conditions nécessaires sont réunies pour appliquer la procédure de k-échanges standard. À nouveau, comme nous cherchons à décrire un ensemble de graphes aux propriétés complexes, celle-ci s'impose pour réaliser un échantillon de graphes uniformément aléatoires satisfaisant les contraintes que nous ciblons.

3.3.2 Application pratique

Nous développons dans cette partie un exemple de mise en pratique du ciblage. Cela nécessite des explications sur le contexte décorrélées du problème initial. Nous décrivons dans un premier temps le modèle dynamique utilisé, qui est indépendant de la question du ciblage. Puis nous présentons les caractéristiques topologiques ciblées, et la réalisation pratique de cette opération. Enfin nous détaillons les résultats obtenus lors de l'implantation du modèle sur les graphes produits.

a. Modèle dynamique

Contexte. Nous employons dans cette partie le modèle *ARIO-network*[10], développé par Fanny Henriet et Stéphane Hallegatte au CIRED[11]. Nous n'en décrivons ici que le principe et les éléments techniques utiles à la compréhension des simulations menées. Pour plus de détails et une vue plus complète sur les enjeux et les perspectives de ces travaux, on se référera à [HH09].

Le modèle ARIO décrit un système économique à l'échelle régionale en quantifiant les échanges commerciaux entre les différentes unités de production. Son objectif est de caractériser la résistance du système économique à une perturbation majeure affectant la production localement, e.g. une catastrophe naturelle. Le cyclone Katrina qui a touché la Louisiane en 2005 servira de base pour comparer les caractéristiques du modèle à celle d'une situation réelle.

Les estimations actuelles des coûts engendrés par les catastrophes naturelles sont déficientes, notamment parce que les évaluations se fondant sur les seuls dégâts directs sont insuffisantes pour évaluer le prix à long terme. L'intérêt et l'orginalité de cette approche est justement de proposer un formalisme pour évaluer comment se propagent les dommages dans le système économique.

Mais la conception de modèles cherchant à décrire des situations aussi complexes posent des difficultés pratiques; en particulier, il existe un grand nombre de paramètres

[10] *Adaptative Regional Input-Output*

[11] Centre International de Recherche sur l'Environnement et le Développement

à prendre en compte dont certains sont mal renseignés par les données. Le modèle est alors construit en fonction des mesures disponibles et à l'aide d'hypothèses simples lorsque les données font défaut.

Description des éléments statiques. Nous décrivons ici le réseau qui supporte la circulation des marchandises dans le système économique régional. Celui-ci est modélisé par un graphe orienté et pondéré, dont les noeuds sont les unités de production (U.P.) - une usine par exemple. Les liens entrant représentent le flux de marchandises (ou de services) consommées pour la production; les liens sortant, les marchandises produites.

Comme on ne dispose pas des données de flux circulant d'U.P. à U.P., celles-ci sont regroupées en 15 secteurs d'activité, selon la nomenclature du *Bureau of Economics Analysis* (e.g. l'agriculture, la manufacture ou les transports). Le volume total des échanges de secteur à secteur pour la Louisiane en 2004 est connu, ainsi que le nombre d'unités de production par secteur. Le graphe décrivant les échanges de marchandises est alors construit à l'aide des hypothèses simplificatrices suivantes:

- Toutes les U.P. d'un même secteur produisent les mêmes marchandises (en pratique, cela signifie qu'elles sont interchangeables dans le réseau).
- Les U.P. d'un même secteur sont de même taille, par conséquent le volume de marchandises entrantes et sortantes est sensiblement le même pour toutes les U.P. d'un secteur.

Description du fonctionnement dynamique. Le modèle évolue de manière discrète, chaque pas de temps correspondant à un jour. Le fonctionnement de chaque unité de production est construit à l'aide d'hypothèses *ad hoc* les plus simples possibles en accord avec le comportement attendu. À nouveau, nous renvoyons à [HH09] pour la description et la discussion exhaustives de ces hypothèses et les équations qui leurs sont associées, dont nous ne faisons ici qu'un résumé.

- *Modélisation des stocks:* La notion de stock joue un rôle considérable dans la vitesse de dégradation du niveau de production, c'est pourquoi celle-ci est intégrée de la manière suivante:
  - On suppose que chaque unité dispose d'une reserve de produits pour chacun de ses fournisseurs, celle-ci lui permettant de produire au rythme normal durant une période fixe (paramètre ajustable) τ_-.
  - Si une U.P. fonctionne à un niveau inférieur à sa capacité de production, elle peut reconstituer ses stocks selon un temps caractéristique τ_+.
  - Certains secteurs produisent des marchandises non-stockables (les transports en particulier) et il n'est pas possible de disposer de reserves de ce type.

- *Modélisation de la demande:* La demande totale au fournisseur j prend en compte d'une part les demandes internes des diverses U.P. du réseau, et d'autre part la demande finale, supposée invariante (les consommateurs extérieurs au réseau).

  Pour intégrer les demandes liées à la constitution de stocks d'une durée fixe, on ajoute à la demande totale un terme ajusté d'un jour sur l'autre selon l'état des stocks, depuis la production initiale (avant la catastrophe).

- *Modélisation de la production:* Chaque U.P. a une limite intrinsèque de production, correspondant à son niveau de production normal, mais celle-ci peut être réduite d'un certain facteur en raison des dégâts directement causés par la catastrophe. Par ailleurs, la production peut être inférieure encore à cette limite si les stocks sont insuffisants. La production effective sera donc le minimum des productions maximum permises par ces limitations.

- *Élements non-considérés:*

  – Le système est fermé: il n'y a pas d'importation ni d'exportation.

  – On ne permet pas aux U.P. affectées de se reconstituer après l'événement initial, ni l'adaptation des acheteurs qui pourraient choisir des fournisseurs dont une partie des capacités de production serait inutilisée.

  – Les facteurs géographiques - comme le coût d'acheminement d'une production d'une unité à une autre - ne sont pas non plus considérés.

  – On ne prend pas en compte les arrêts de production pour des raisons financières (même si ceux-ci ont joué un rôle important dans le cas de Katrina).

  De tels éléments ne sont pas intégrés au modèle car ils sont d'une nature différente de la stricte circulation des marchandises dans le réseau régional sur laquelle se focalise la description. Leur prise en considération nécessiterait alors un faisceau d'hypothèses et de paramètres supplémentaires. Elles pourraient néanmoins faire l'objet de développements du modèle.

À partir de maintenant, le modèle dynamique sera pour nous une "boîte noire" et nous nous concentrons sur la structure du réseau sur lequel il est employé.

b. Objectifs et protocole

Nous voulons identifier - au moins qualitativement - quelle peut être l'influence des caractéristiques topologiques du réseau sur sa résistance aux événements catastrophiques.

Topologie du réseau. Nous décrivons ici quels sont les différents graphes statiques dont nous allons tester le comportement relativement au modèle; en expliquant d'abord comment est construit le graphe racine de la procédure, puis sur quels paramètres nous allons mettre en œuvre le ciblage.

- *Matrice racine*

  La taille du réseau testé est limitée à 532 U.P. (contre environ 10^5 en Louisiane), en raison des difficultés algorithmiques à gérer un réseau de plus grande taille dans le modèle dynamique d'évolution.

  Dans la suite, nous nommons bloc XY la sous-partie de la matrice d'adjacence $\mathbf{A}$ dont les destinations des arcs (clients) sont les noeuds du secteur X et les sources (fournisseurs) ceux de Y. D'après ce qui précède, les contraintes imposées au réseau ($\mathbf{C_{ARIO}}$) se traduiraient par:

  – Le volume total $\mathcal{W}_{XY}$ d'échanges dans le bloc XY est fixé.

  – Le total de chaque ligne du bloc est fixé et ce total est le même pour chacune des x lignes (toutes les unités de X reçoivent le même volume produit par le secteur Y).

  – Le total de chaque colonne du bloc est fixé et ce total est le même pour chacune des y colonnes (toutes les unités de Y produisent le même volume destiné au secteur X).

  Ces conditions ne sont pas toujours strictement compatibles si $\mathcal{W}_{XY}/x$ ou $\mathcal{W}_{XY}/y$ n'est pas entier; on fera alors en sorte que le reste soit distribué entre le maximum de noeuds du secteur. Par exemple pour un volume total de 10 unités à répartir entre 6 fournisseurs et 4 clients, on accorde aux producteurs un volume de 2, 2, 2, 2, 1, 1 et aux acheteurs: 3, 3, 2 et 2.

  Dans [HH09], les auteurs proposent un graphe artificiel satisfaisant ces contraintes à l'aide d'une procédure de remplissage par bande de la matrice qui produirait pour l'exemple précédent:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

  Dans le cas étudié cela ne sera pas nécessaire: on arrondit les flux de secteurs à secteurs pour que la condition soit satisfaite. On procède ainsi sur chaque bloc XY puis on réunit ces blocs afin d'obtenir une matrice de l'ensemble $\mathcal{F}_{\mathbf{ARIO}}$.

  Dans l'annexe E nous donnons une représentation par blocs de la matrice utilisée, faisant figurer ses caractéristiques (taille des domaines, volume des échanges).

- *Matrice de G_0*

  On peut choisir d'utiliser la matrice-racine comme point de départ du protocole, ou un élément d'un échantillon uniformément aléatoire de matrices de l'ensemble $\mathcal{F}_{\mathbf{ARIO}}$, i.e. l'ensemble des matrices répondant aux contraintes énumérées dans la partie précédente.

  Construire un tel échantillon ne pose pas de difficulté nouvelle: comme chaque bloc satisfait $\mathbf{C_{min}}$, nous pouvons leur appliquer une série d'échanges simples, puis les réunir afin d'obtenir une matrice quelconque de $\mathcal{F}_{\mathbf{ARIO}}$.

- *Matrices ciblées*

  Nous décrivons ici quelles sont les particularités topologiques des matrices que nous souhaitons produire et ce qui motive le choix de ces propriétés.

  Clustering orienté. Il a été observé sur des modèles de diffusion épidémique qu'à densité égale, le clustering tendait à ralentir la propagation (e.g. [CR07]). Or le fonctionnement dynamique du modèle suggère que la diffusion de la baisse de productivité suive une logique analogue, nous souhaitons vérifier qu'il existe ici aussi un effet d'enfermement dans les zones denses.

  Nous utilisons ici une des généralisations possibles de la notion de clustering global définie en 1.2.3.c. dans le cas de graphes orientés. On substitue aux arcs multiples des arcs simples puis on évalue:

  $$c_{3-o} = 3.\frac{\text{  }}{\text{}}$$

  Concentration. Nous pouvons imaginer qu'une circulation des marchandises reposant sur quelques unités de production rend l'économie locale fortement dépendante d'un petit nombre de circuits d'approvisionnement. Dans le contexte du modèle, cela se traduirait par de fortes pondérations de quelques liens.

  Cette situation de concentration des échanges pourrait accélérer l'effondrement de la production car dès que ces circuits privilégiés sont touchés, on s'attend à ce que le système entier soit rapidement affecté; mais à l'inverse la production pourrait rester longtemps quasi-normale lorsque seuls des circuits secondaires de production sont affectés.

  Il nous faut donc estimer le degré de dispersion de la circulation des marchandises, et pour ce faire nous définissons la concentration $\mathcal{C}$ du réseau. Elle est construite selon le même principe que l'indice d'Herfindahl couramment utilisé par les économistes pour évaluer la concentration des marchés[12]:

[12]Employé par exemple dans le *Merger Guidelines* pour déterminer les situations de monopoles.

– Avec $w_{ij} = \mathbf{A}_{ij}$ le poids d'un fournisseur j (secteur Y) pour l'approvisionnement d'une unité de production i (secteur X), la somme sur l'ensemble du secteur des poids de la colonne est fixée par hypothèse: $\mathcal{W}_{iY} = \sum_{j \in Y} w_{ij}$.

– Nous définirons alors la concentration pour le secteur Y de l'acheteur i:

$$c_{iY} = \sum_{j \in Y} \left(\frac{w_{ij}}{\mathcal{W}_{iY}} \right)^2$$

– puis la concentration du bloc XY sera la somme pondérée des concentrations par acheteur:

$$c_{XY} = \frac{\sum_{i \in X} c_{iY}.\mathcal{W}_{iY}}{\sum_{i \in X} \mathcal{W}_{iY}}$$

– enfin la concentration totale $\mathcal{C}$ sera alors la somme pondérée des concentrations par bloc:

$$\mathcal{C} = \frac{\sum_{X,Y} c_{XY}.\mathcal{W}_{XY}}{\sum_{X,Y} \mathcal{W}_{XY}}$$

Avec cette définition, la concentration est une grandeur normalisée telle que $\mathcal{C} = 1$ correspond à un marché totalement concentré: un acheteur du secteur X s'approvisionne auprès d'un unique fournisseur dans le secteur Y. En revanche, si $\mathcal{C}$ est "faible"[13], le marché est très déconcentré et chaque acheteur dispose de plusieurs fournisseurs par secteur d'activité.

Protocole. Maintenant les paramètres à cibler fixés, nous mettons en œuvre la méthode elle-même, autrement dit, nous produisons un échantillon de matrices aléatoires ayant une valeur particulière du coefficient de concentration ou de clustering. Nous décrivons le ciblage pour $\mathcal{C}$, le principe est strictement identique pour le clustering orienté.

Ciblage. La mesure de l'écart choisie est le scalaire $\mathcal{C}_0 - \mathcal{C}$, avec $\mathcal{C}_0$ la valeur visée. On itère le processus à $k = 2$ en autorisant uniquement l'échange lorsque la quantité $|\mathcal{C}_0 - \mathcal{C}|$ s'approche de 0. Nous constatons que le taux d'échec augmente au cours du processus: lorsque nous arrivons à de "fortes" valeurs de $\mathcal{C}$, le nombre de graphes permettant de s'approcher de la cible diminue. On peut s'en rendre compte par exemple en traçant la valeur de $\mathcal{C}$ (et de c_{3-o}) en fonction du nombre de tentatives: la pente de la courbe décroît.

Uniformisation. Une fois $\mathcal{C}_0$ atteint, on procède à une séquence de k-échanges usuels en imposant, en plus de l'ensemble de contraintes $\mathbf{C_{ARIO}}$, d'avoir une valeur de $\mathcal{C}$ contenue entre deux bornes dont nous fixons arbitrairement l'écart à $\pm 0,005$ autour de

[13] La valeur minimum de $\mathcal{C}$ ($\simeq 0,358$) est celle de la matrice-racine pour laquelle tous les poids sont unitaires.

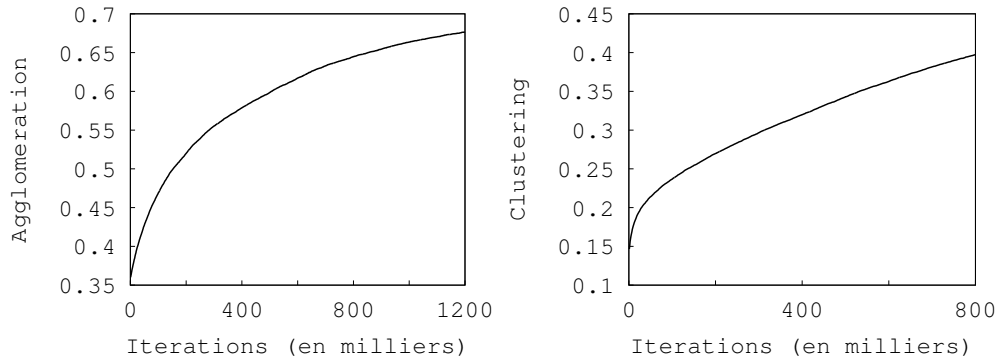


FIG. 3.11: Gauche: évolution en fonction du nombre d'itérations de $\mathcal{C}$ au cours de la convergence ciblée. Droite: *idem* pour c_{3-o}.

$\mathcal{C}_0$.

remarque : Nous pouvons faire ici une incise, de notre point de vue intéressante, car elle met en évidence le lien étroit existant entre ces algorithmes et l'*approximate counting*, i.e. le dénombrement approché de grands ensembles statistiques [JS97]: en principe, nous pourrions utiliser la mesure de $\mathcal{C}$ pour attester de l'uniformisation, mais ce n'est pas un bon choix.
En effet, nous décrivons des échantillons de graphes dont les valeurs de $\mathcal{C}$ sont très différentes de la valeur moyenne pour la contrainte $\mathbf{C_{ARIO}}$. Or, lorsque nous nous éloignons de cette valeur moyenne, le taux d'échec augmente. Cela traduit une diminution du nombre de graphes pour une valeur de $\mathcal{C}$ proche de la valeur atteinte. Donc si nous tracions la distribution de $\mathcal{C}$ des graphes de l'ensemble, nous constaterions sans doute que le poids du sous-ensemble dans lequel nous uniformisons est très petit devant le poids de tout l'ensemble et même devant le poids de sous-ensembles d'uniformisation correspondant à des intervalles plus proches de la valeur moyenne.
Ainsi en uniformisant l'échantillon, la mesure de $\mathcal{C}$ va rapidement tendre vers la borne la plus proche de la valeur moyenne sur l'ensemble $\mathbf{C_{ARIO}}$. C'est ce que traduit l'image 3.12: on uniformise dans l'intervalle $[0,40; 0,70]$, la borne inférieure est plus proche de la valeur moyenne ($\simeq 0,372$), nous constatons donc une rapide dérive vers 0,40. Comme ce phénomène se produit quelle que soit la valeur de k, il est très probable qu'on ne puisse pas discerner par cette mesure la valeur de k^*.

Nous testons alors l'uniformisation de $\mathcal{C}$ avec la mesure du nombre de triangles dirigés - les forts degrés des nœuds du graphe impose une mesure légère. On constate alors que $k^* = 2$ est suffisant pour obtenir un échantillon qui puisse être considéré comme représentatif de l'ensemble (c.f. Fig. 3.13).

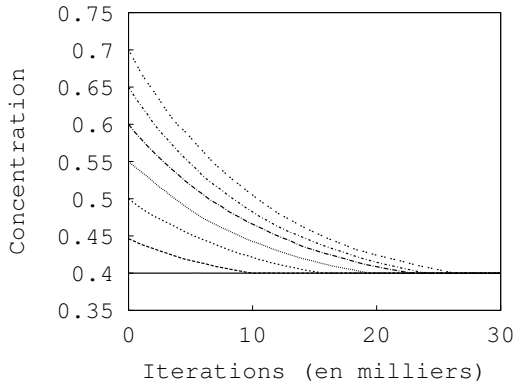


FIG. 3.12: Évolution de la concentration $\mathcal{C}$ en uniformisant dans l'intervalle $[0,40;0,70]$ depuis plusieurs racines - ici $k = 2$.

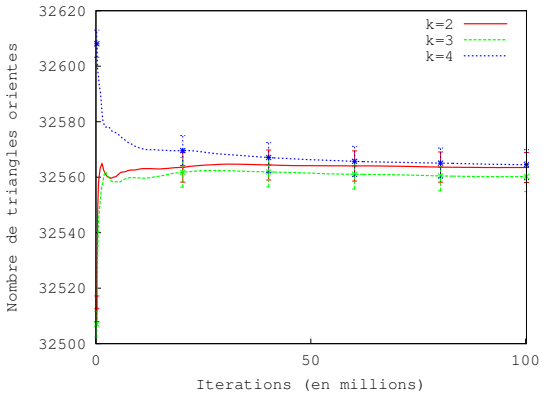


FIG. 3.13: Moyenne cumulée du nombre de triangles orientés pour différentes valeurs de k, en uniformisant dans l'intervalle $\mathcal{C} \in [0,40;0,70]$ depuis des racines distinctes.

3.3.3 Résultats

a. Conditions

Les valeurs des différents paramètres pour les simulations effectuées sont:

- $N = 532$ et $L = 22.282$ (pour la répartition des flux entre secteurs: c.f. Ann. E),
- les stocks permettent aux U.P. de produire normalement pendant une journée ($\tau_- = 1$), et la durée caractéristique de leur reconstitution est de 6 jours ($\tau_+ = 6$).
- la catastrophe initiale détruit à 99% une unique unité de production choisie dans le domaine de l'extraction minière.

Des séries de simulations sont réalisées sur des échantillons générées pour des valeurs de $\mathcal{C} \in [0,36;0,70]$ et de $c_{3-o} \in [0,01;0,40]$. Nous utilisons la matrice-racine pour point de départ (plutôt qu'une matrice de $\mathcal{F}_{\mathbf{ARIO}}$).

b. Mesure de l'impact

Comme nous ne laissons pas la possibilité au réseau de se reconstituer, le système économique s'effondre bien que la catastrophe n'affecte initialement qu'une U.P.: après cent jours, on mesure (sur l'ensemble des simulations) que la production totale $\mathcal{P}$ est située entre 3 et 5,5% de la production initiale.

L'effondrement se produit généralement de manière subite après un certain nombre d'itérations. Pour évaluer l'impact de l'événement, nous intégrons la fraction de la production assurée entre le premier et le centième jour suivant la catastrophe, soit:

$$I = \frac{1}{100} \sum_{t=0}^{100} \frac{\mathcal{P}(t)}{\mathcal{P}(0)}$$

Pour chacune des simulations, on trace l'impact en fonction de la caractéristique topologique ciblée sur les Figures 3.14 et 3.15.

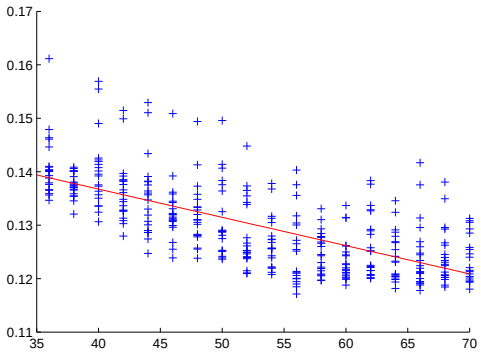


FIG. 3.14: Impact en fonction de $\mathcal{C}$: on observe une corrélation négative, comme le montre la meilleure régression linéaire (selon les moindres carrés).

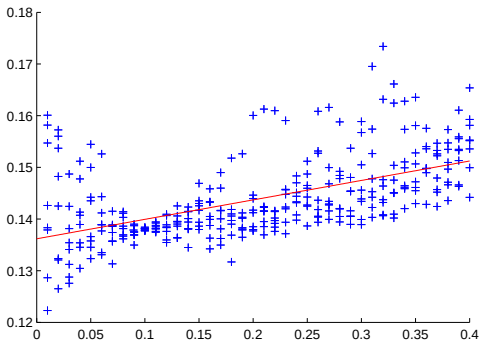


FIG. 3.15: Impact en fonction de c_{3-o}: on observe une corrélation positive.

c. Discussion

Il est difficile de donner un sens aux valeurs des fractions de la production mesurées, car ainsi que nous l'avons dit, le modèle n'est pas (et ne se veut pas) une reconstruction réaliste, mais les corrélations qualitatives sont instructives.

La Figure 3.14 montre que plus les clients ont de fournisseurs, plus le système est robuste. Or, ce résultat n'est pas aussi trivial que dans les tests de robustesse liés à la connexité, où lorsqu'on diminue le nombre de liens, le seuil de nœuds à supprimer pour déconnecter le graphe décroît aussi. Mais ici, augmenter le nombre de fournisseurs accroît aussi les chances pour un client d'être affecté plus vite par la crise. Il y a donc concurrence entre les deux effets antagonistes.

Concernant le clustering orienté, nous voyons que l'augmentation du taux de triangulation tend effectivement à ralentir la vitesse de propagation de la crise. On pouvait en douter car augmenter le taux de clustering se fait également en diminuant le nombre de chemins, ce qui peut se faire en augmentant l'agglomération[14].

Il serait alors utile de balayer le plus possible le champ des paramètres, notamment le nombre d'U.P. dans le système et les conditions initiales, pour voir comment ceux-ci influencent le comportement et en particulier si les effets mesurés peuvent être imputés à la taille du modèle.

Une difficulté manifeste d'un travail de modélisation de ce type consiste à intégrer les éléments réels qu'on pense indispensables à la compréhension des phénomènes essentiels (et leur mesure), en évitant autant que possible les hypothèses arbitraires. Pour être employé dans un objectif de prévisions réalistes des coûts, un tel modèle devrait d'abord intégrer tous les éléments exclus (la reconstruction, les apports extérieurs), puis être étalonné à des données réelles ou ajusté en fonction d'observations passées.

3.3.4 Autres applications du ciblage

Nous avons employé le ciblage pour examiner le comportement d'un modèle de propagation lorsque l'on modifie les propriétés topologiques du graphe. Nous pouvons en déduire des informations pour améliorer la conception des réseaux: les rendre plus résistants aux attaques, accroître la vitesse de propagation de l'information etc. Mais l'intérêt du ciblage est plus général: il permet de s'affranchir de la nécessité d'avoir un élément dans l'ensemble de départ pour générer un échantillon de graphes.

Nous pouvons alors en tirer de nouvelles applications, en particulier nous pouvons agrandir artificiellement des bases de données réelles. En effet, supposons que les données réelles puissent être reconstruites à l'aide de leur distribution de degré, et de diverses propriétés topologiques notées $\{\mathbf{P_i}\}_{i=1,2,...}$, nous pouvons alors générer un

[14] On observe d'ailleurs une corrélation positive entre $\mathcal{C}$ et c_{3-o} sur les graphes produits.

graphe selon cette distribution, puis viser successivement chacune des propriétés $\mathbf{P_i}$ pour synthétiser un graphe satisfaisant les contraintes $\mathbf{C_{min}} \cup \mathbf{C_{P_1}}$, puis $\mathbf{C_{min}} \cup \mathbf{C_{P_1}} \cup \mathbf{C_{P_2}}$, et ainsi de suite jusqu'à obtenir un élément de l'ensemble modélisant les données réelles.

3.4 Commentaires généraux sur les applications

Nous avons étudier dans ce chapitre deux types distincts d'applications de la méthode des k-échanges.

Dans l'exemple des graphes dK comme dans celui de l'analyse des limites de connectivité, nous cherchions à déterminer quelles caractéristiques topologiques peuvent être qualifiées d'élémentaires pour expliquer la structure du réseau réel, un problème important puisqu'il peut être considéré comme une première étape pour proposer un mécanisme d'assemblage du réseau réel.

La dernière application étudiait, non pas la constitution du réseau, mais un processus se déroulant sur celui-ci - en l'occurrence un phénomène de crise. Nous avons vu qu'une généralisation relativement simple du protocole permettait de prévoir comment les modifications de la topologie du réseau pouvait modifier les caractéristiques de la propagation.

Par ailleurs, nous constatons sur ces quelques exemples pratiques qu'on peut souvent se satisfaire d'échantillons générés à l'aide d'échanges simples (i.e. $k = 2$). Cela semble pourtant improbable que le processus soit effectivement ergodique dans des situations si variées et avec des contraintes si complexes, nous pensons plutôt qu'à un certain niveau de précision dans les mesures, il apparaît comme tel.

Nous tirons de ces exemples une image intuitive du métagraphe associé à un k-échange: si celui-ci présente plusieurs composantes connexes, nous pensons que l'une d'entre elles serait géante dans le sens où elle regrouperait la grande majorité des éléments de l'ensemble.

En augmentant la valeur de k nous souhaitons rendre connexe le métagraphe; mais d'après cette hypothèse, si de nouvelles composantes sont réunies avec la composante géante, celles-ci n'auraient pas un poids statistique suffisant pour différencier les mesures. Cela expliquerait que les mesures suivies au cours du processus semblent fréquemment converger vers une valeur identique pour $k = 2$ et des valeurs supérieures. Cette supposition n'a pas bien sûr de portée universelle, elle ne se fonde que sur l'observation de quelques cas spécifiques.

Quelle que soit le domaine de validité de cette remarque, pour s'assurer dans une certaine mesure de la légitimité d'une procédure de génération de graphes aléatoires

sous contrainte au moyen d'échanges, il faudrait l'accompagner de vérifications correspondant à des valeurs supérieures de k.

Chapitre 4

Perspectives

Sommaire

Nous développons dans ce court chapitre des prolongements possibles au travail exposé dans cette thèse, d'abord sur le plan des applications, ensuite sur celui de la méthodologie.

4.1 Intégration de caractéristiques externes

L'analyse en réseau des systèmes sociaux - et en particulier de leur représentation en graphe - donne des agents et de leurs modes d'interaction une image homogène: les nœuds et les liens sont tous identiques, ils n'ont pas de propriétés permettant de les distinguer autrement que par leur rôle dans la structure.

La capacité à saisir les mécanismes de systèmes aussi composites par une modélisation aussi simple a été souvent mise en question. Mais nous pouvons envisager deux façons de contourner cette lacune:

- **Isoler des systèmes qui soient suffisamment uniformes** et bien renseignés pour que la description en graphe soit autonome, c'est-à-dire qu'il ne serait pas

nécessaire dans ces cas de recourir à des arguments extérieurs pour comprendre la constitution ou le fonctionnement du réseau.

Les réseaux construits sur le flux d'une quantité transférée peuvent être considérés comme un terrain privilégié pour ce point de vue, car ce qui circule est quantifiable et donc l'équivalence entre un lien et une interaction n'est pas arbitraire. Parmi ceux-ci, nous pouvons évoquer les réseaux tels que l'Internet, les réseaux pair-à-pair et tous les réseaux sur lesquels circulent des paquets d'information. En contrepartie, l'information d'ordre sociologique dans ces réseaux est limitée.

- **Intégrer à la description graphique des informations externes**. Donner une "couleur", une communauté d'appartenance, un âge, des caractéristiques sémantiques à un nœud revient tacitement à enrichir le modèle de graphe.

  Nous procédons ainsi en 3.3 lorsque nous attribuons à une unité de production un secteur d'activité. Nous discutons ci-dessous d'autres cas où nous pourrions mettre à profit ce second point de vue.

4.1.1 Caractérisations sémantiques

Décrire la signification de l'information transmise amène à étudier le contenu sémantique des signaux circulant dans le réseau. À l'aide d'outils apropriés, il est en effet possible d'extraire d'un texte ses concepts-clefs et ainsi de reconstituer sous forme simplifiée un environnement sémantique des acteurs, des interactions ou des événements (e.g. [RB06b]).

Nous pourrions alors traiter ces éléments sémantiques comme des contraintes sur le réseau. Pour illustrer cette suggestion, nous décrivons une application possible aux collaborations scientifiques:

- on extrait, par exemple sur des bases de données d'articles, des mots-clefs (choisis dans un ensemble fini et supposés pertinents),
- puis on produit un échantillon aléatoire de graphes imposant aux acteurs de conserver le même environnement sémantique.

Si l'on choisit un lexique de mots-clef très large, chaque article utilise une combinaison unique de concepts, et le graphe produit sera le graphe initial. En réduisant cette base, les articles partageront de plus en plus de mots-clefs, pour le processus de génération cela revient à autoriser de plus en plus d'échanges. Selon le niveau de ressemblance entre l'échantillon synthétique et le graphe réel, on pourrait estimer la représentativité d'un concept dans la communauté décrite.

4.1.2 Le temps

Jusqu'à présent l'analyse menée considérait le réseau comme une image statique, or le temps joue certainement un rôle essentiel, pour des raisons de causalité, par exemple, si A appelle B puis B appelle C, l'information peut circuler de A vers C mais pas de C vers A. Ainsi les considérations sur l'activité d'un agent du réseau ou sur la diffusion d'information pourraient être nettement améliorées en y intégrant l'aspect dynamique des interactions.

Une façon d'analyser le problème dans la continuité de ce que nous avons fait jusqu'à présent serait de voir le graphe dynamique comme une succession d'images statiques correspondant aux données cumulées sur une période précise, c'est d'ailleurs une approche très usitée (entre autres: [BBY06]). Néanmoins, on souhaiterait disposer de moyens pour saisir les caractéristiques intrinsèquement dynamiques des réseaux complexes, de nouveaux cadres théoriques se développent d'ailleurs dans ce sens [SBF+08].

Les méthodes d'échanges pourraient permettre d'aborder cet aspect de manière originale. La variable de temps a un statut particulier, notamment parce qu'à la différence de celles évoquées jusqu'à présent, ce n'est pas une variable discrète. En imposant un échantillonage temporel, nous pouvons la traiter comme une variable externe usuelle, (il sera toutefois nécessaire de réfléchir à la signification de la granularité temporelle).

Supposons que nous autorisions les échanges de liens uniquement entre agents actifs au même moment à $\pm\Delta t$ (ce qui fixerait l'échantillonage temporel); les graphes produits selon ce principe présenteraient un avantage vis-à-vis de modèles de référence construits à partir de photographies successives: en effet, alors que dans ces derniers il n'y a pas de rapport entre une référence statique et la suivante, on rétablirait au moyen de ces "échanges dynamiques" une continuité dans les données artificielles de comparaison. Cela permettrait d'utiliser ces graphes dynamiques synthétiques par exemple pour la simulation de phénomènes eux-mêmes dynamiques (comme la diffusion).

4.2 Sonder l'espace des graphes

En plus du champ d'applications possibles à la génération de graphes sous contraintes, nous pensons que la méthodologie proposée peut-être employée dans un objectif fondamental de grande envergure qui serait la description de très grands ensembles statistiques.

4.2.1 Lien avec l'*approximate counting*

L'état stationnaire de la chaîne de Markov renseigne sur la composition de l'ensemble décrit. Nous pourrions par exemple nous servir des algorithmes d'échanges pour décrire

comment une certaine caractéristique (disons la mesure scalaire m) est distribuée dans l'ensemble.

Imaginons par exemple la procédure suivante:

- On divise l'ensemble en intervalles de m, puis on effectue des échanges afin d'obtenir un échantillon uniforme de graphes sur chacun de ces intervalles.
- En faisant l'hypothèse d'une distribution linéaire sur l'intervalle considéré, on pourrait déduire sa pente de la valeur moyenne de m sur l'ensemble délimité par l'intervalle - le nombre de graphes avec $m \in [m_{inf}; m_{moy}]$ devant être égal au nombre de graphes avec $m \in [m_{moy}; m_{sup}]$.
- Et de proche en proche on reconstituerait la distribution de la mesure m sur la totalité de l'ensemble, comme illustré schématiquement:

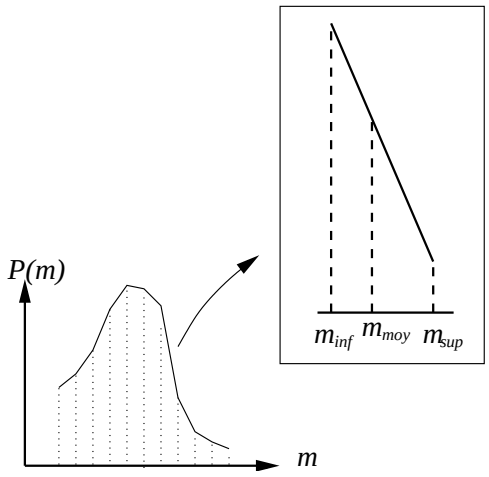


FIG. 4.1: Allure possible de la distribution d'une mesure m sur l'espace des graphes décrit.

Cette description reste théorique, et il faudrait sans doute faire face à un certain nombre de difficultés pratiques, dont deux au moins sont prévisibles:

– si l'intervalle est choisi trop petit, les échanges risquent de ne plus être ergodiques et il faudra alors ajuster la valeur de k;
– si la distribution est très inhomogène dans l'intervalle, nous verrons la mesure dériver vers une des bornes (comme en 3.3.2.b.), et l'estimation de la pente sera imprécise.

On peut même espérer déduire non pas seulement des probabilités de distribution, mais le cardinal d'un intervalle ou plus généralement d'espaces de graphes. En effet, dans des articles récents [BCPV08, Bia09], Bianconi *et al.* proposent des évaluations

approchées de la taille des ensembles de graphes sous diverses contraintes (dont $\mathbf{C_{min}}$) à l'aide de fonctions de partition.

Nous ne savons pas si ce type d'estimation pourraient être adaptées à des cas aussi complexes que ceux que nous cherchons à étudier, mais en connaissant le volume d'un ensemble relativement simple (comme $\mathcal{E}_{\mathbf{C_{min}}}$), on pourrait en déduire le nombre de graphes dans ses sous-parties.

4.2.2 Vitesse de convergence

Nous avons dit que le régime transitoire de la convergence était difficile à décrire car il dépendait non seulement de la contrainte associée au processus, mais également du graphe et de la mesure utilisée.

Cette caractéristique peut aussi indiquer que la convergence porte la trace de propriétés de l'espace parcouru. On pourrait alors entreprendre une analyse expérimentale recherchant des régularités dans ce régime transitoire: des mesures, des contraintes qui seraient associées à une certaine forme de convergence, avec l'espoir de maîtriser la vitesse de ces processus.

En disposant par ailleurs d'indications sur la taille de l'ensemble et sur le nombre de voisins au sens du processus de Markov, il serait possible de se faire une idée plus précise de la topologie du métagraphe.

4.2.3 Explorer les frontières en ciblant

Les exemples traités révèlent que dans beaucoup de cas, les échanges simples sont suffisants pour produire des échantillons qui soient représentatifs de l'ensemble et peuvent donc être utilisés pratiquement. À nouveau, cela ne signifie pas que le métagraphe associé soit connexe: les mesures choisies sont peut-être seulement trop imprécises pour permettre de détecter les écarts statistiques entre les différents ensembles.

Si maintenant nous voulons donner une borne inférieure à la valeur de k nécessaire pour rendre le métagraphe connexe, et ce avec une meilleure précision que notre estimation de k^*, on peut affiner l'approche expérimentale par ciblage.

En effet, si nous pensons que le métagraphe n'est pas connexe, il est vraisemblable que cela serait visible aux "frontières". On entend par frontières les régions de l'ensemble où certaines caractéristiques structurelles prennent des valeurs extrêmes.

Considérant la mesure scalaire m, nous pouvons chercher à atteindre le maximum accessible de m en effectuant des ciblages comme en 3.3. Le nombre de voisins d'un métanœud du métagraphe va diminuer, et la probabilité augmente pour qu'une certaine valeur k' de k produise des graphes qui seraient inaccessibles avec des valeurs inférieures à k'. On serait ainsi à la frontière du métagraphe et on y trouverait des sous-ensemble

encore inconnus.

Pour montrer de manière certaine que k' est une borne inférieure, on peut réaliser une uniformisation avec $k = k'$ dans un intervalle très étroit autour de la valeur extrême et constater que l'ensemble obtenu a des caractéristiques effectivement différentes de celles qu'il aurait en uniformisant avec $k' - 1$.

Au travers de ces dernières applications, les algorithmes d'échanges apparaissent non plus seulement comme un principe de génération de graphes aléatoires, mais comme un moyen expérimental pour sonder les ensembles de graphes. Et la représentation en métagraphe ne serait plus seulement un moyen de décrire le processus de Markov employé, mais aussi une manière de situer les éléments de l'ensemble les uns par rapport aux autres, ou autrement dit de munir l'ensemble des graphes d'une métrique et de le cartographier.

Conclusion

La modélisation en graphe des réseaux d'interactions sociales repose sur l'hypothèse fondamentale que ce mode simple de représentation du réel est suffisant pour en comprendre les mécanismes.

Il ne faut pas interpréter ce présupposé de manière stricte, car il est possible d'enrichir le modèle en superposant d'autres niveaux d'information à la structure de graphe: éléments dynamiques, contenus sémantiques ou autres variables externes qui permettent d'inclure les éléments que l'on pense nécessaires à la compréhension du fonctionnement du réseau.

Pour établir le degré de validité de cette hypothèse, il est nécessaire d'étudier d'une part le système modélisé et d'autre part la structure formelle qui le représente. C'est sur ce second point que nous avons concentré nos efforts au cours de ce travail. Il faut en effet pouvoir distinguer ce qui, dans le modèle, serait la trace d'un mécanisme de fonctionnement de ce qui serait accidentel.

Nous avons cherché à situer le réseau réel relativement à un environnement construit comme l'ensemble des graphes satisfaisants des caractéristiques particulières du cas réel, parmi lesquelles la distribution des degrés.

Dans ce but nous avons proposé une méthode algorithmique de génération de graphes aléatoires. Celle-ci permet de produire des échantillons dont on peut s'assurer de la qualité par des mesures expérimentales plus ou moins poussées. Elle utilise pour point de départ un élément de l'ensemble, mais il est possible de s'en affranchir moyennant une évolution de la procédure (ciblage). Son utilisation pratique nécessite toutefois des mesures exploratoires - pour en ajuster les paramètres - et éventuellement la mise en œuvre d'heuristiques d'accélération.

Au travers d'applications à des cas concrets, nous avons décrits comment effectuer la recherche d'éléments structurels suffisants à une reconstruction du réseau réel, éléments qui contiendraient donc l'information accessible par la représentation graphique.

Au fil de l'étude, une autre dimension des utilisations possibles de ces algorithmes apparaît: ils peuvent également être regardés comme des sondes des ensembles de graphes, informant sur leur taille ou la diversité de leurs éléments et dont on pense

qu'ils seraient utilisables pour examiner la distribution de caractéristiques des graphes dans l'ensemble.

L'exploration expérimentale de ces gigantesques espaces est une perspective de recherche passionnante car il s'agit d'appréhender autrement que par leur représentation des ensembles si vastes qu'il est impossible d'en avoir une vue exhaustive. Elle ouvre une quantité de possibilités en autorisant à étudier ces espaces synthétiques par des moyens comparables à ceux qu'utilisent les physiciens pour observer un système naturel.

Annexes

Annexe A

Familles de réseaux complexes

Nous présentons dans cette annexe un inventaire - non-exhaustif - des graphes de réseaux complexes usuels et donnons quelques détails sur ceux qui font l'objet de notre étude: caractéristiques élémentaires, accessibilité, définitions des nœuds, filtrages et précisions sur le contenus, ainsi que les parties du texte où l'on y fait référence - que ce soit à simple but illustratif ou pour une étude plus approfondie.

Les caractéristiques sont données avec les notations suivantes:

- N: nombre d'agents (d'acteurs),
- M: nombre d'événements si la base est bipartie,
- L: nombre de liens (ou d'arcs),
- L_b: nombre de liens événement-acteur si la base est bipartie,
- L_a: nombre de liens entre acteurs dans la projection si la base est bipartie.
- L_m: nombre de liens en cas de multigraphe.

A.1 Graphes de réseaux sociaux

Nous reprenons la nomenclature choisie en 1.3.2, mais les exemples concrets présentés ici vont mettre en évidence que les frontières entre ces catégories sont perméables.

A.1.1 Réseaux collaboratifs

a. Domaine scientifique

Les dépôts d'articles scientifiques en ligne sont très répandus et exhaustifs, ce qui en fait un domaine privilégié pour l'analyse des réseaux sociaux. De plus, des informations sémantiques sur les contenus sont souvent accessibles à l'aide d'outils de collecte adaptés.

La collecte peut être réalisée selon les options des moteurs de recherche bibliographiques: par auteur, date, mots-clefs, journal... Les données sont en général

modélisées de manière bipartie, les événements correspondant à des publications (articles, rapports techniques, actes de conférence...) et les acteurs aux scientifiques eux-mêmes.

- *AIO: Europe.*

  $N = 12.112$, $M = 17.869$, $L_b = 24.382$, $L_a = 9.090$.

  Extrait de l'*Anthropological Index Online* (aio.anthropology.org.uk/aiosearch/).

  Graphe de collaborations sur l'archéologie en Europe entre 1983 et 2009, sans restriction de langue. On élimine les articles qui n'ont pas d'auteur identifié (de l'ordre d'un pour mille dans l'*AIO*).

- *AIO: Scandinavie.*

  $N = 273$, $M = 185$, $L_b = 345$, $L_a = 280$.

  Sous-partie de la base précédente, portant exclusivement sur l'archéologie scandinave au cours de la décennie 2000-2009. La projection des acteurs ne présente pas de composante géante (la plus grande composante connexe compte 12 acteurs).

- *arXiv.*

  $N = 16.400$, $M = 19.885$, $L_b = 45.902$, $L_a = 29.552$.

  Extraite d'*arXiv: Condensed Matter* (arxiv.org/list/cond-mat).

  Graphe de collaborations dans le domaine de la physique de la matière condensée, extrait des versions pré-imprimées d'articles déposées sur le site entre 1995 et 1999.

- *SW.*

  $N = 2.764$, $M = 1.673$, $L_b = 4.827$, $L_a = 7.194$.

  Extraite de l'*ISI Web of Knowledge* (apps.isiknowledge.com).

  Graphe de collaborations dans le domaine des systèmes complexes, tous les articles compilés dans la base *WoK* sélectionnées avec le mot-clef *"Small-World"*, sans restriction de date (acquisition Mars 2007).

- *Medline.*

  $N = 65.211$, $M = 37.136$, $L_b = 126.312$, $L_a = 179.488$.

  Extraite de *PubMed* (www.ncbi.nlm.nih.gov/pubmed/).

  Graphe de collaborations dans le domaine de la biomédecine, tous les articles compilés dans la base *Medline* sélectionnés avec le mot-clef *"ligament"*, sans restriction de date (acquisition Mars 2007).

- *CERN.*

  $N = 11.033$, $M = 12.032$, $L_b = 35.625$, $L_a = 900.100$.

  Extraite du *CERN document server* (cdsweb.cern.ch).

  Graphe de collaborations dans le domaine de la physique des hautes énergies, tous les articles compilés dans la base avec le mot-clef *"string"*, sans restriction de date (acquisition courant 2008). On élimine tout type de documents autres que les articles publiés (rapports techniques, vidéos...). La base de données présente des événements géants (25 articles de tailles entre 100 et 550, pour une taille moyenne de 2.96).

b. Domaine artistique

- *Notre-Dame: Acteurs.*

  $N = 392.340$, $M = 127.823$, $L_b = 1.470.418$, $L_a = 15.038.088$.

  Accessible sur le site de l'Université Notre-Dame (www.nd.edu/~networks).

  Graphe de collaborations entre acteurs au cours de films, collecté sur l'*Internet Movie Database* (www.imdb.com), étudié dans [BA99].

- *Allmusic.*

  $N = 7.079$, $L = 20.922$.

  Extraite du site *Allmusic* (www.allmusic.com). Base monopartie uniquement.

  Graphes obtenus par l'intersection de collaborations entre musiciens et une classification d'experts par genre musical: on ne conserve les interactions de musiciens d'un même genre ayant joué ensemble. Le contenu des données est détaillé dans [TBCB08][1].

c. Domaine institutionnel et économique

Les typologies usuelles des réseaux sociaux les classeraient dans la catégorie des réseaux d'affiliations, ce terme générique recouvre les réseaux sociaux que l'on décrit par l'appartenance de membres à des groupes et décrivent donc indirectement les interactions. Ceux-ci se prêtent particulièrement à une modélisation hypergraphique, où les nœuds seraient les individus et les hyperliens: les groupes.

- *TheyRule.*

  $N = 4.300$, $M = 493$, $L_b = 5.530$, $L_a = 30.253$.

  Extraite du site They Rule (www.theyrule.net).

[1] Nous remercions T. Teitelbaum et ses collaborateurs d'avoir mis à notre disposition ces données.

Graphe réunissant les membres des conseils d'administrations des plus grandes sociétés américaines en 2004.

- *DutchElite.*

  $N = 936$, $M = 4.745$, $L_b = 6.154$, $L_a = 29.833$.

  Extraite du site de V. Batagelj (vlado.fmf.uni-lj.si/pub/networks/data/).

  Graphe réunissant les hauts-fonctionnaires des principales institutions politiques des Pays-Bas en 2006.

A.1.2 Réseaux de communication

Les réseaux de communication et d'échange d'information recouvrent les bases de données collectées à partir de leur support technologique: par téléphone, e-mail, blogs, messageries instantanées...

Une catégorie importante est représentée par les réseaux du *world wide web*, qui sert de support à une vaste gamme d'échanges d'information dont les données peuvent être accessibles. Une vue comparative de réseaux du *web* est proposée dans [MMG+07].

a. Réseaux collaboratifs en ligne

Si on les considère non en fonction de leur objectif mais de leur support, les projets tels que *Wikipedia* et plus généralement l'ensemble des *Wikis* et autres sites modifiables par les utilisateurs peuvent être classés dans cette catégorie.

À chaque *Wiki*, correspond une certaine structure de fonctionnalités, ainsi que des rôles et des droits des collaborateurs. Ces sites permettent alors des modélisations en graphe variées, pour *Wikipedia* quelques options possibles:

- en tant que sous-parties du *web*, on peut les réprésenter comme des pages liées entre elles par des liens hypertextes (e.g. [ZBŠD06]),
- on peut s'intéresser à l'aspect collaboratif en étudiant quels utilisateurs contribuent à quelles pages,
- ou encore à l'aspect interactif entre agents, en examinant les contributeurs participant aux pages de discussion.

- *Wiki (d).*

  $N = 1.398$, $M = 4.502$, $L_b = 6.507$, $L_a = 6.675$.

  Données collectées dans le cadre du projet *Autograph* (overcrowded.anoptique.org).

  Les événements sont les pages de discussion de *Wikipedia* (autres que les "discussions utilisateurs"), les acteurs en sont les contributeurs. Nous utilisons une sélection des modifications de la *Wikipédia* francophone entre le 13 et le 25 Juin 2004, en éliminant les robots et IP anonymes.

b. Forums

Organisés hiérarchiquement par thèmes, les forums de discussion en ligne permettent une modélisation bipartie, dont les événements sont les fils de discussion et les acteurs: les membres.

- *Debian.*

  $N = 1.997$, $M = 5.639$, $L_b = 17.517$, $L_a = 24.046$.

  Extraite du site *Debian-fr* (forum.debian-fr.org).

  Forums de discussions de la communauté des utilisateurs francophones du système d'exploitation *Debian*, à partir d'une collecte effectuée entre août 2003 et Juillet 2004. Les données sont étudiées dans [DLCA07][2].

  Ces données sont bruitées de par la liberté laissée aux utilisateurs de modifier les titres des fils de discussion, ou d'intervenir à l'aide de plusieurs identités caractérisées par des adresses électroniques, nous regroupons autant que possible les adresses associées au même utilisateur et supprimons les fils de discussion publicitaires.

c. Réseaux d'échanges de mails

- *Kiel.*

  $N = 59.912$, $L = 51.524$ (sans arcs multiples).

  Échanges de mails mettant en jeu les adresses d'étudiants de l'Université de Kiel entre le 29 Juillet et le 17 Novembre 2001. Tous les nœuds n'ont pas le même statut, 5.165 sont des étudiants, les autres ne sont pas membres de l'Université mais correspondent avec. La base est décrite dans [EMB02].

A.2 Réseaux à quantité transférée

Ces réseaux sont situés à la frontière du domaine social: sans en faire partie, ils informent indirectement sur leur fonctionnement et présentent l'avantage d'être caractérisables à l'aide de flux mesurables; il peut s'agir d'information (réseaux de partage de fichiers), de bien matériels, d'individus dans un réseau de transport, envisagés du point de vue de la quantité transférée, on peut aussi y inclure la plupart des réseaux de communications.

[2]Nous remercions R. Dorat et ses collaborateurs d'avoir mis à notre disposition ces données.

A.2.1 Internet

Nous nous référons ici à l'Internet en tant que structure physique de connexion entre les machines qui supportent le *web* - et plus généralement au transport d'information d'une machine à une autre au moyen de l'*Internet Protocol*.

Dans une structure aussi complexe, on dégage plusieurs niveaux de hiérarchie [PSV04] et son étude graphique peut alors se faire à différentes échelles et avec un degré de description des éléments plus ou moins fin. Notamment:

- Une description possible regarde l'Internet comme **un réseau de routeurs interconnectés**: des machines dont le rôle est de chercher dans leur environnement local, via une table de routage, quel est le chemin le plus adapté à l'acheminement des paquets d'information vers une destination identifiée.
- Les **systèmes autonomes** (*autonomous systems*) sont des sous-parties du réseaux administrées de manière autonome, et connectées entre elles par des passerelles (*gateways*). L'analyse des connexions entre AS correspond donc à une cartographie du réseau à plus large échelle que celle des routeurs.

- *Routeviews*.

  $N = 22.963$, $L = 48.936$.

  Données accessibles sur la page de M. Newman (www-personal.umich.edu/~mejn/).

  Image de l'Internet au niveau AS réalisée depuis les acquisitions du projet *Route Views*, à partir des tables *BGP* de routage entre les systèmes autonomes.

- *skitter*.

  $N = 9.204$, $L = 28.959$.

  Données collectées dans le cadre de *CAIDA* (www.caida.org/tools/measurement/).

  Image cumulée sur un mois des chemins de routage entre systèmes autonomes, réalisée à l'aide de *traceroute*. Cette base est décrite dans [MKFV06].

- *HOT / HOT'*.

  $N = 939/830$, $L = 988/878$.

  Données synthétiques extraites de [LAWD04].

  HOT est l'acronyme d'*Heuristically Optimal Topology*, un modèle de graphe artificiel représentant la topologie interne d'un système autonome au niveau de ses routeurs - typiquement un fournisseur d'accès.

A.2.2 Réseaux de circulation des marchandises

Ce type de réseaux représentent les flux de marchandises - ou éventuellement leur contrepartie financière - entre différentes unités économiques (sociétés, pays...). Le réseau synthétique *ARIO* décrit et utilisé dans la partie 3.3 en est un exemple.

A.3 Autres types de réseaux

A.3.1 Graphes de réseaux biologiques

Le volume de données disponibles dans ce domaine croissent spéctaculairement: on dispose aujourd'hui d'une large gamme de bases recensant les génomes et protéomes, depuis des organismes rudimentaires jusqu'à l'homme, pour lequel ces informations restent encore en grande partie à découvrir.

Parmi toutes les données accessibles, certaines se prêtent particulièrement à une modélisation en graphes, en particulier les réseaux de régulations génétiques où l'on examine comment une partie du génome interagit avec une autre par l'intermédiaire des protéines.

a. Interactions entre protéines

On représente les protéines qui interagissent dans un organisme comme les nœuds du réseau, qui sont connectés si le couple de protéines est effectivement susceptible d'interagir au cours d'une réaction biochimique. Un point de vue voisin se prêtant davantage à une modélisation bipartie du réseau consiste à regrouper dans un même ensemble les protéines intervenant dans une même voie métabolique.

- *Pathways.*

  $N = 679$, $L = 11.030$.

  Extraite de la base *Ecocyc* (ecocyc.org).

  Projection monopartie du réseaux des interactions entre protéines chez l'homme, regroupés par voie métabolique, tel qu'il est décrit dans les données *Ecocyc* (membre du réseau de banques de données *Biocyc* (www.biocyc.org)).

A.3.2 Réseaux linguistiques

Ces réseaux sont construits généralement sur le principe de la cooccurrence: la présence simultanée d'éléments au sein d'un même groupe. Dans une analyse lexicologique, les structures de textes d'une même langue peuvent être comparées via la cooccurrence des mots dans les phrases. Ce principe est aussi utilisable en phonologie: cooccurrence de phonèmes dans les mots d'une langue etc.

- *Wordnet.*

  $N = 82.670$, $L = 133.445$.

  Extraite du site de V. Batagelj (vlado.fmf.uni-lj.si/pub/networks/data/).

  Réseau de cooccurence de mots dans le dictionnaire *Word Net* (wordnet.princeton.edu), ici examiné comme un réseau de mots participant à une même définition. $N = 82.670$ correspond au nombre de mots du dictionnaire selon la source, seuls 67.539 apparaissent dans la base.

Annexe B

Dénombrement de motifs

Les algorithmes de dénombrement de motifs utilisés suivent une stratégie simple de parcours en largeur, que nous décrivons dans le cas non-orienté[1]. Supposons que nous souhaitions compter des motifs de taille p, à partir de chaque nœud i du graphe, nous procédons selon la procédure résumée par le schéma B.1 et décrite ci-dessous:

- On initialise l'ensemble $\mathcal{M}$ qui décrira les éventuels motifs à $\{i\}$.
- On construit l'ensemble des voisins $\mathcal{V}$ de i, tels que $j \geq i$ (pour éviter les dénombrements multiples). Ce parcours définit la profondeur 1 autour de i.
- On définit la profondeur 2 autour de i: pour chaque j , $\mathcal{M} \longleftarrow \mathcal{M} \cup j$.
- Et ainsi de suite jusqu'à la profondeur $p-1$, en ajoutant à l'ensemble des voisins un nœud seulement si celui-ci n'est pas déjà dans $\mathcal{M}$ (puisqu'il n'apparaîtra qu'une seule fois dans le motif).
- Enfin, pour chaque ensemble-motif $\mathcal{M} = \{i, j, k, ...\}$ obtenu, on vérifie le schéma de connectivité pour identifier de quel type de motifs à p nœuds il s'agit.

Nous donnons maintenant une estimation de la complexité de l'algorithme avec $\overline{\delta} = 2L/N$, le degré moyen. Le parcours de la liste des voisins de i est en moyenne en $\mathcal{O}(\overline{\delta})$, comme la probabilité d'être voisin d'un nœud quelconque augmente linéairement avec le degré, le parcours de la liste des voisins d'un nœud j est en moyenne en $\mathcal{O}(\overline{\delta}^2)$, et la complexité serait donc multipliée par $\overline{\delta}^2$ lorsque l'on augmente la profondeur d'un rang. Il est nécessaire de parcourir jusqu'à une profondeur de $p-1$, donc la complexité totale de l'algorithme serait donc en $\mathcal{O}(N\overline{\delta}^{2p-3})$. Enfin il faut prendre en compte la vérification du schéma de connectivité, et il est alors avantageux de stocker également le graphe également sous sa forme de matrice d'adjacence, ce qui permet une vérification

[1]La méthode développée dans [Wer06], téléchargeable sur http://theinf1.informatik.uni-jena.de/motifs/ suit un principe analogue.

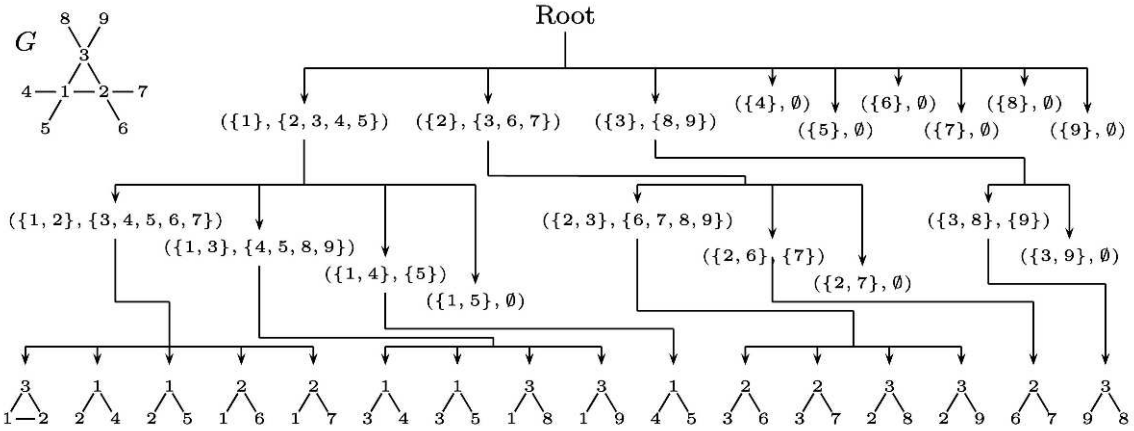


FIG. B.1: Dénombrement des motifs à trois nœuds sur du graphe G selon l'algorithme décrit. Extrait de [Wer06].

en $\mathcal{O}(1)$ - avec les tailles typiques que nous considérons, cela ne pose pas de problème de complexité spatiale.

Dans le cas orienté, un raisonnement analogue amène $\mathcal{O}(N\overline{\delta_s}^{p-1}\overline{\delta_e}^{p-2})$ en fonction des degrés entrant et sortant (respectivement δ_e et δ_s).

Cette stratégie n'est pas optimale, d'autres - utilisant le produit matriciel - peuvent être plus efficaces ([AYZ97, YZ04, Lat08]); mais elle présente les avantages d'être très simple à mettre en œuvre et à adapter à divers motifs; de plus, elle est efficace sur des graphes épars et n'utilise que peu de mémoire (au plus quelques Mo sur nos exemples).

Nous comparons ci-dessous les performances de l'algorithme à celui proposé par [Lat08][2], spécialement construit pour dénombrer les triangles de graphes épars, dont la distribution de degré est en loi de puissance (complexité typiquement $\subset \mathcal{O}(L^{3/2})$), puis à l'algorithme proposé par [MIK+04][3], au contraire très polyvalent - il peut dénombrer tous les motifs d'une taille fixée, orientés ou non - mais quasi-inutilisable sur de grands graphes. Les mesures sont menées sur les projections monoparties de graphes précédemment évoqués.

Graphes	N	L	$\triangle$	notre algorithme	algorithme [Lat08]	algorithme [MIK+04]
Medline	65.211	179.488	276.331	0,90 s	0,14 s	170 s
Notre-Dame	392.340	15.038.088	346.813.199	17,000 s	15 s	> 1.000.000 s

TAB. B.1: Temps de dénombrement de triangles à l'aide de trois algorithmes réalisés sur une machine standard (processeur 2×2.33 GHz; 2 Go de mémoire).

[2]Téléchargeable sur http://www-rp.lip6.fr/~latapy/index.php?item=programs&lang=fr .

[3]Téléchargeable sur http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html .

Annexe C

Démonstration de l'uniformité

Nous démontrons ici le caractère uniforme de la distribution d'un algorithme construit sur une chaîne de tentatives d'échanges. La preuve suit le modèle de celle proposée par Miklós et Podani [MP04], que l'on peut trouver sur esapubs.org/archive/ecol/E085/001/appendix-A.htm .

C.1 Algorithme de Metropolis

Supposons que nous souhaitions décrire un ensemble statistique discret X dont les éléments x_i seraient répartis selon une loi de probabilité $\pi(x_i)$. C'est possible avec un algorithme de Metropolis ([MRR+53]) i.e. de la forme:

1. Partant à l'itération t d'un élément x_i de l'ensemble X (soit $x(t) = x_i$), on tire un élément x_j selon une loi de transition $\mathcal{T}$, telle que $\sum_{x_i \in X} \mathcal{T}(x_i \to x_j) = 1$.
2. On tire un réel u de manière uniformément aléatoire sur l'ensemble $[0; 1]$,
   si $u < min\left(1\ ,\ \frac{\pi(x_j)\mathcal{T}(x_i \to x_j)}{\pi(x_i)\mathcal{T}(x_j \to x_i)}\right)$ alors $x(t+1) = x_j$,
   sinon $x(t+1) = x_i$.

La loi de transition $\mathcal{T}$ définit une chaîne de Markov, si celle-ci est apériodique et irréductible, l'algorithme converge vers π (pour une preuve, voir par exemple: [Liu08]).

C.2 Lien avec la chaîne de tentatives d'échanges

Qu'il s'agisse de tentatives d'échanges simples ou de k-échanges, il s'agit de montrer que la chaîne de Markov $\mathcal{M}$ employée dans les procédure d'échanges est un algorithme de Metropolis. Dans ce contexte:

- puisque nous souhaitons obtenir une distribution uniforme de l'ensemble $\mathcal{F}$,

$$\forall\ G_i \in \mathcal{F},\ \pi(G_i) = 1/|\mathcal{F}|$$

- on vérifie par définition du processus que $\sum_{G_i \in \mathcal{F}} \mathcal{M}(G_i \to G_j) = 1$,
- la chaîne de tentatives de k-échange est symétrique car il existe - à k fixé - exactement le même nombre de tentatives quel que soit l'élément de $\mathcal{F}$ considéré, donc:

$$\mathcal{M}(G_i \to G_j) = \mathcal{M}(G_j \to G_i)$$

Et donc nous avons:

$$min\left(1\ ,\ \frac{\pi(G_j)\mathcal{M}(G_i \to G_j)}{\pi(G_i)\mathcal{M}(G_j \to G_i)}\right) = \begin{cases} 1 & \text{si } G_j \text{ est voisin de } G_i \text{ par le processus,} \\ 0 & \text{sinon,} \end{cases}$$

Il n'est donc pas nécessaire de tirer u, et le processus de transition $\mathcal{M}$ est de la forme de $\mathcal{T}$. C'est un algorithme de Metropolis.

De plus

- **Apériodicité.** Pour qu'il n'existe strictement aucune configuration correspondant à l'échec d'une tentative d'échange, il faut qu'il n'apparaisse que des structures de la forme:

$$\begin{pmatrix} & \vdots & & \vdots & \\ \cdots & 1 & \cdots & 0 & \cdots \\ & \vdots & & \vdots & \\ \cdots & 0 & \cdots & 1 & \cdots \\ & \vdots & & \vdots & \end{pmatrix}$$

  ce qui n'est pas possible - sauf pour le cas particulier et sans intérêt de l'ensemble constitué par les deux graphes: $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ - or, si la probabilité de rester dans le même état est non nulle, le processus est apériodique.

- **Irréductibilité.** Par construction, nous raisonnons sur l'ensemble des graphes accessibles par le processus de Markov, donc par construction $\mathcal{M}$ est ergodique.

Nous réunissons toutes les conditions pour affirmer que le processus de tentatives de k-échanges est un algorithme Metropolis et converge vers une loi de probabilité uniforme sur l'ensemble $\mathcal{F}$ des graphes accessibles.

Annexe D

Algorithmes d'échanges

D.1 Algorithme de tentatives de k-échanges

D.1.1 Boucle principale: cas orienté, non-pondéré, sans boucle

La partie principale de tous les algorithmes de la famille ne varie pas (à l'orientation, la pondération et la présence de boucles près). Dans le cas des graphes simples et orientés:

Entrées:

- représentation éparse du graphe de départ: $G_0 = (V_0, E_0)$; E_0 est la liste des arcs connectant les nœuds de V_0
- nombre de tentatives d'échanges: n
- valeur de k

Sortie:

- graphe G produit au terme de n tentatives d'échanges

---

1. $G = (V, E) \longleftarrow G_0$ *// intialisation*
2. **pour** j **de** 1 **à** n
   (a) tirer k arcs différents au hasard: $\{(a_i, b_i)\}_{i \in I}$ de E ;
   (b) tirer une des $k!$ permutations de l'ensemble des indices I au hasard: σ ;
   (c) construire l'ensemble des arcs permutés $\left\{(a_i, b_{\sigma(i)})\right\}_{i \in I}$;
   (d) $E' \longleftarrow E \cup \left\{(a_i, b_{\sigma(i)})\right\} \setminus \{(a_i, b_i)\}$; $G' = (V, E')$; *// graphe produit par l'échange σ*
   (e) $\forall i \in I\ ,\ \mathcal{W}_i = \{b\ :\ \exists\ (a_i\ ,\ b) \in E\} \setminus \{b_i\}$; *// ensemble des voisins de a_i sauf b_i*
   si $\forall i\ ,\ a_i \neq b_{\sigma(i)}$ *// pas de boucles*
   et $\forall i\ ,\ b_{\sigma(i)} \notin \mathcal{W}_i$ *// pas de liens multiples*
   et ... *// contrainte complémentaire*
   alors $G \longleftarrow G'$;

D.1.2 Boucle principale: autres cas

Afin de traiter le cas non-orienté, on peut modifier l'algorithme précédent en substituant à la liste des liens une liste d'arcs doubles faisant figurer (a, b) et (b, a) si le lien non-orienté $(a, b) \in E$.

On procède ensuite de manière identique en ajoutant:

– **si** $(a, b) \in \{(a_i, b_i)\}_{i \in I}$ **alors** $(b, a) \notin \{(a_i, b_i)\}_{i \in I}$

– $E' \longleftarrow E \cup \left\{(a_i, b_{\sigma(i)}); (b_{\sigma(i)}, a_i)\right\}_{i \in I} \setminus \{(a_i, b_i); (b_i, a_i)\}_{i \in I}$

Les généralisations aux cas sans boucles ou pondérés se font sans difficultés particulières.

D.2 Conditions supplémentaires

Nous décrivons ici les cas traités dans la thèse avec une évaluation sommaire de leur complexité temporelle. Selon les différentes contraintes imposées, on ajoute des conditions supplémentaires à la boucle principale. En pratique il faut ajouter les modules décrits ici à l'étape (e), on écrira la contrainte $\mathbf{C}$ sous la forme $\mathbf{C} = \mathbf{c} \cup \mathbf{C_{min}}$, et $\mathcal{F}_{\mathbf{C}} = \mathcal{F}_{\mathbf{c}} \cap \mathcal{E}_{\mathbf{C_{min}}}$.

D.2.1 Contrainte $\mathbf{C_{compo}}$

Nous utilisons une méthode simple mais peu efficace pour traiter cette contrainte, consistant à déterminer pour chaque a_i la taille de sa composante avant et après l'échange, ce qui peut se faire selon la procédure de parcours en largeur (pour le nœud a_i dans le graphe G, $\mathcal{V}_G(x)$ désigne l'ensemble des voisins dans G du nœud x):

α. $\mathscr{Q}(a_i) \longleftarrow \{a_i\}$; *// initialisation de l'ensemble à parcourir*

β. $\mathscr{C}_G(a_i) \longleftarrow \{a_i\}$; *// initialisation de la composante connexe*

γ. **tant que** $\mathscr{Q}(a_i) \neq \emptyset$, $\forall q \in \mathscr{Q}(a_i)$

  si $q \notin \mathscr{C}(a_i)$

  alors $\mathscr{C}_G(a_i) \longleftarrow \mathscr{C}_G(a_i) \cup \{q\}$; $\mathscr{Q}(a_i) \longleftarrow \mathscr{Q}(a_i) \cup \mathcal{V}_G(q) \setminus \{q\}$;

  sinon $\mathscr{Q}(a_i) \longleftarrow \mathscr{Q}(a_i) \setminus \{q\}$;

L'ensemble des tailles des composantes est une sous-partie de la distribution, si cet ensemble est identique avant et après l'échange, ou formellement:

$$\{|\mathscr{C}_G(a_i)|\}_{i \in I} = \{|\mathscr{C}_{G'}(a_i)|\}_{i \in I}$$

la contrainte additionnelle $\mathbf{c}$ est satisfaite.

La probabilité pour que l'un des nœuds impliqués au moins appartienne à la composante géante étant proche de 1, on peut estimer grossièrement que l'on passe en revue tous les liens du graphe à chaque vérification de la contrainte et donc que la complexité de l'étape élémentaire serait en $\mathcal{O}(L)$. Il existe dans la littérature des méthodes visiblement beaucoup plus efficaces pour déterminer la connexité, qui serait à adapter à ce contexte (e.g. [HK99]).

D.2.2 Contrainte $\mathbf{C_{tri}}$

Avec les mêmes notations que précédemment, on évalue le nombre de triangles créés et détruits pendant l'échange puis on vérifie s'ils se compensent:

α. $\mathscr{T}_d(i) \longleftarrow \{\}$; *// initialisation de l'ensemble des triangles détruits*

β. $\mathscr{T}_c(i) \longleftarrow \{\}$; *// initialisation de l'ensemble des triangles créés*

γ. **pour** $i \in I$

- **si** $c \in \mathcal{V}_G(a_i)$ **et** $c \in \mathcal{V}_G(b_i)$
  alors $\mathscr{T}_d(i) \longleftarrow \mathscr{T}_d(i) \cup \{\{a_i, b_i, c\}\}$
- **si** $c \in \mathcal{V}_{G'}(a_i)$ **et** $c \in \mathcal{V}_{G'}(b_{\sigma(i)})$
  alors $\mathscr{T}_c(i) \longleftarrow \mathscr{T}_c(i) \cup \big\{\big\{a_i, b_{\sigma(i)}, c\big\}\big\}$

δ. **si** $\left|\bigcup_{i\in I} \mathscr{T}_d(i)\right| = \left|\bigcup_{i\in I} \mathscr{T}_c(i)\right|$ **alors** la contrainte **c** est satisfaite.

Il est pratique d'estimer ici la complexité de l'étape avec $\overline{\delta} = 2L/N$, le degré moyen. Avec la structure de données adoptée (tableau de liste), il est nécessaire de parcourir la liste des voisins de a_i et b_i. De plus lorsque nous tirons un lien au hasard le degré des extrêmités n'est pas en moyenne de $\overline{\delta}$ car la probabilité pour qu'un lien quelconque mène à un nœud de degré k croît linéairement avec k, ainsi le temps moyen de parcours de la liste des voisins de a_i ou b_i sera proportionnel à $\overline{\delta}^2$. Et l'étape élémentaire aurait donc une complexité en $\mathcal{O}(\overline{\delta}^4)$.

D.2.3 Contrainte $\mathbf{C_{dK}}$

On ne détaille pas l'algorithme correspondant à cette contrainte car il est du même type que le précédent: on dénombre les structures créés et détruites au cours de l'échange, si leur nombre se compensent, l'échange est réalisé.

La complexité dépend de la taille de la structure considérée.

- Pour la contrainte 2K, il suffit de construire un tableau indiquant le degré de chaque nœud - qui reste inchangé au cours du processus, sa consultation est en $\mathcal{O}(1)$.

– Pour la contrainte 3K, il s'agit des triangles et des chemins de longueur 2, ce qui se fait comme dans $\mathbf{C_{tri}}$ en $\mathcal{O}(\overline{\delta}^4)$, le calcul des degrés des sommets mis en jeu se faisant à nouveau en $\mathcal{O}(1)$.

D.2.4 Contrainte $\mathbf{C_{Red}}$

Pour chaque événement mis en jeu au cours de l'échange, on calcule le cœfficient de redondance avant et après. On en déduit quelle serait la redondance moyenne sur le graphe si l'échange est effectivement réalisé et donc si le graphe satisfait la contrainte **c**.

Pour calculer la redondance d'un événement de manière simple, on stocke les données d'une part dans le tableau de listes donnant les acteurs participant à un événement et d'autre part dans un tableau de listes des événements auxquels participe chaque acteur, qu'il faudra bien sûr mettre également à jour.

Si δ_e désigne la taille moyenne d'un évenement et δ_a: le nombre moyen d'événements auxquels participe un acteur, il faudra à chaque itération parcourir l'ensemble des couples d'acteurs de l'événement considéré (dont la taille est en δ_e^2). De plus en tirant un lien aléatoirement, il existe toujours un biais de probabilité proportionnel au degré, si bien qu'en moyenne le parcours des couples d'acteurs est en $\mathcal{O}(\overline{\delta_e}^3)$.

Pour chaque couple d'acteur, il faut parcourir la liste des événements auxquels chacun participe (taille en δ_a). Avec le biais du tirage aléatoire des liens, cela sera en moyenne en $\mathcal{O}(\overline{\delta_a}^2)$. La complexité de l'étape élémentaire serait donc au total en $\mathcal{O}(\overline{\delta_e}^3\overline{\delta_a}^2)$.

Annexe E

Matrice racine du modèle ARIO

Caractéristiques de la matrice racine modélisant les échanges commerciaux en Louisiane[1], base de l'étude 3.3 (lignes: clients, colonnes: fournisseurs, valeurs: volume financier des transactions):

	1. Agriculture	2. Extraction minière	3. Équipement	4. Construction	5. Manufacture	6. Commerce (gros)	7. Commerce (détail)	8. Transport	9. Information	10. Finance	11. Services (sociétés)	12. Éducation, santé	13. Arts, culture	14. Services (autres)	15. État	**TOTAL**
Nombre d'U.P.	**10**	**10**	**2**	**40**	**20**	**30**	**90**	**20**	**10**	**60**	**60**	**60**	**50**	**50**	**20**	**532**
1. Agriculture	20	20	10	80	40	60	180	40	20	120	120	120	100	100	40	1070
2. Extraction minière	10	10	10	40	20	30	90	20	10	60	60	60	50	50	20	540
3. Équipement	10	10	2	40	20	30	90	20	10	60	60	60	50	50	20	532
4. Construction	80	40	40	160	40	120	360	40	40	120	120	120	200	200	40	1720
5. Manufacture	40	20	20	40	40	60	180	20	20	60	60	60	100	100	20	840
6. Commerce (gros)	60	30	30	120	60	60	90	60	30	60	60	60	150	150	60	1080
7. Commerce (détail)	180	90	90	360	180	90	540	180	90	180	180	180	450	450	180	3420
8. Transport	40	20	20	40	20	60	180	40	20	60	60	60	100	100	20	840
9. Information	10	10	10	40	20	30	90	20	10	60	60	60	50	50	20	540
10. Finance	120	60	60	120	60	60	180	60	60	60	60	60	300	300	60	1620
11. Services (sociétés)	120	60	60	120	60	60	180	60	60	60	60	60	300	300	60	1620
12. Éducation, santé	120	60	60	120	60	60	180	60	60	60	60	60	300	300	60	1620
13. Arts, culture	100	50	50	200	100	150	450	100	50	300	300	300	500	500	100	3250
14. Services (autres)	100	50	50	200	100	150	450	100	50	300	300	300	250	250	100	2750
15. État	40	20	20	40	20	60	180	20	20	60	60	60	100	100	40	840
TOTAL	1050	550	532	1720	840	1080	3420	840	550	1620	1620	1620	3000	3000	840	22282

[1]Proportions du *Census Bureau* à l'échelle régionale (www.census.gov).

Bibliographie

[AA04] I. Albert and R. Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics-Oxford*, 20(18):3346–3352, 2004.

[AB02] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[ACL01] W. Aiello, F. Chung, and L. Lu. Random evolution in massive graphs. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science, 2001.*, pages 510–519, 2001.

[AF01] D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs. *Book in preparation*, 2001.

[AJB99] R. Albert, H. Jeong, and A.L. Barabási. Diameter of the World-Wide Web. *Nature(London)*, 401(6749):130–131, 1999.

[AJB00] R. Albert, H. Jeong, and A.L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[Alb73] R.D. Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1):113–126, 1973.

[ARS05] Y. Artzy-Randrup and L. Stone. Generating uniformly distributed random networks. *PRE*, 72(5):056708, 2005.

[ASBS00] L.A.N. Amaral, A. Scala, M. Barthélemy, and H.E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149, 2000.

[AYZ97] N. Alon, R. Yuster, and U. Zwick. Finding and counting given length cycles. *Algorithmica*, 17(3):209–223, 1997.

[BA99] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[Bar81] E.R. Barnes. An algorithm for partitioning the nodes of a graph. In *20th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes, 1981*, volume 20, 1981.

[Bar04] M. Barthélemy. Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):163–168, 2004.

[Bav47] A. Bavelas. A mathematical model for group structures. *Human Organization*, 7(3):16–30, 1947.

[Bav50] A. Bavelas. Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22:725–730, 1950.

[BBPSV04] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical Review Letters*, 92(17), 2004.

[BBY06] D. Braha and Y. Bar-Yam. From centrality to temporary fame: Dynamic centrality in complex networks. *Complexity*, 12(2):59–63, 2006.

[BC78] E.A. Bender and E.R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *J. Combin. Theory Ser. A*, 24(3):296–307, 1978.

[BCPV08] G. Bianconi, A.C.C. Coolen, and C.J. Perez Vicente. Entropies of complex networks with hierarchically constrained topologies. *Physical Review E*, 78(1):16114, 2008.

[BDML06] T. Britton, M. Deijfen, and A. Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6):1377–1397, 2006.

[BDX04] S. Boyd, P. Diaconis, and L. Xiao. Fastest Mixing Markov Chain on a Graph. *SIAM Review*, 46(4):667–689, 2004.

[Ben38] F. Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

[Ber70] C. Berge. Graphes et hypergraphes. 1970.

[BGLL08] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

[Bia09] G. Bianconi. Entropy of network ensembles. *Physical Review E*, 79(3), 2009.

[BLM+06] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[BMBL09] S.P. Borgatti, A. Mehra, D.J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916), 2009.

[Bol88] J.M. Bolland. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233–253, 1988.

[Bol01] B. Bollobás. *Random graphs*. Cambridge University Press, 2001.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[BPS03] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Physical Review E*, 68(3):36112, 2003.

[Bra01] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[Bur80] R.S. Burt. Models of network structure. *Annual Review of Sociology*, 6(1):79–141, 1980.

[Bur95] R.S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, 1995.

[BW00] A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[CBPS05] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Physical Review E*, 71(2), 2005.

[CCDLRM02] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(25):258702, 2002.

[CCP04] M. Catanzaro, G. Caldarelli, and L. Pietronero. Assortative model for social networks. *Physical Review E*, 70(3):37101, 2004.

[CEBAH00] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628, 2000.

[CFSV06] V. Colizza, A. Flammini, M.A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Arxiv preprint physics/0602134*, 2006.

[CH56] D. Cartwright and F. Harary. Structural balance: a generalization of Heider's theory. *Psychological Review*, 63(5):277–293, 1956.

[CL02] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

[CNM04] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), 2004.

[Col77] C.J. Colbourn. *Graph generation*. University of Waterloo, 1977.

[Col88] R. Collins. *Theoretical sociology*. Harcourt College Pub, 1988.

[CR05] M. Caesar and J. Rexford. BGP policies in ISP networks. *IEEE Network Magazine*, 19(6):5–11, 2005.

[CR07] J.P. Cointet and C. Roth. How realistic should knowledge diffusion models be? *Journal of Artificial Societies and Social Simulation*, 10(3):5, 2007.

[CR09] J.P. Cointet and C. Roth. Socio-semantic Dynamics in a Blog Network. In *Proceedings of the 2009 International Conference on Computational Science and Engineering-Volume 04*, pages 114–121. IEEE Computer Society, 2009.

[CRTB07] L.F. Costa, F.A. Rodrigues, G. Travieso, and P.R.V. Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[CSN10] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 2010.

[DA05] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):27104, 2005.

[Dav63] J.A. Davis. Structural balance, mechanical solidarity, and interpersonal relations. *American Journal of Sociology*, 68(4):444–462, 1963.

[DDGDA05] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, 2005.

[DGKTB10] C. Del Genio, H. Kim, Z. Toroczkai, and K. Bassler. Efficient sampling of graphs with arbitrary degree sequence. In *APS Meeting Abstracts*, 2010.

[DLCA07] R. Dorat, M. Latapy, B. Conein, and N. Auray. Multi-level analysis of an interaction network between individuals in a mailing-list. In *Annales des Télécommunications*, volume 62, page 325. Presses Polytechniques Romandes, 2007.

[DM03] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, USA, 2003.

[DMS00] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(21):4633–4636, 2000.

[Dor03] S.N. Dorogovtsev. Networks with given correlations. *Preprint*, 2003.

[EB05] M. Everett and S.P. Borgatti. Ego network betweenness. *Social Networks*, 27(1):31–38, 2005.

[EG60] P. Erdős and T. Gallai. Graphs with prescribed degrees of vertices. *Mat. Lapok*, 11:264–274, 1960.

[EG94] M. Emirbayer and J. Goodwin. Network analysis, culture, and the problem of agency. *American Journal of Sociology*, 99(6):1411, 1994.

[Egg73] R.B. Eggleton. Graphic sequences and graphic polynomials: a report. *Infinite and Finite Sets*, 1:385–392, 1973.

[EH78] R.B. Eggleton and D.A. Holton. The graph of type $(0,\infty,\infty)$ realizations of a graphic sequence. In *Proceedings of the 6th Australian Conference on Combinatorial Mathematics*, pages 41–54, 1978.

[EMB02] H. Ebel, L.I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(3):35103, 2002.

[ER59] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6(26):290–297, 1959.

[ER60] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[ER61] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267, 1961.

[FB07] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36, 2007.

[FDBV01] I.J. Farkas, I. Derenyi, A.L. Barabási, and T. Vicsek. Spectra of "real-world" graphs: Beyond the semicircle law. *Physical Review E*, 64(2):26704, 2001.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, page 262. ACM, 1999.

[For09] S. Fortunato. Community detection in graphs. 2009.

[Fre79] L.C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[Fri04] N.E. Friedkin. Social cohesion. 2004.

[FS86] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

[GMZ03] C. Gkantsidis, M. Mihail, and E.W. Zegura. The markov chain simulation method for generating connected power law random graphs. In *Proc. 5th Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2003.

[GN02] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

[Gra73] M.S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360, 1973.

[Gra83] M. Granovetter. The strength of weak ties: a network theory revisited. *Sociological Theory*, 1:201–233, 1983.

[GSPA04] R. Guimerà, M. Sales-Pardo, and L.A.N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):25101, 2004.

[HAB07] A. Herdağdelen, E. Aygün, and H. Bingol. Measuring preferential attachment. *Europhysics Letters*, 78:60007, 2007.

[Hak62] S.L. Hakimi. On the Realization of a Set of Integers as Degrees of the Vertices of a Graph. *J. SIAM Appl. Math*, 10:496–506, 1962.

[Het00] H.W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[HH95] P. Hage and F. Harary. Eccentricity and centrality in networks. *Social networks*, 17(1):57–63, 1995.

[HH09] S. Hallegatte and F. Henriet. Assessing the Consequences of Natural Disasters on Production Networks: A Disaggregated Approach. *Fondazione Eni Enrico Mattei Working Papers*, page 259, 2009.

[HHL+99] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, et al. From molecular to modular cell biology. *Nature*, 402(6761):47, 1999.

[HK99] M.R. Henzinger and V. King. Randomized fully dynamic graph algorithms with polylogarithmic time per operation. *Journal of the ACM*, 46(4):502–516, 1999.

[Jen06] P. Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E*, 74(3):35101, 2006.

[JGN01] E.M. Jin, M. Girvan, and MEJ Newman. Structure of growing social networks. *Physical Review E*, 64(4), 2001.

[JS97] M. Jerrum and A. Sinclair. The Markov chain Monte Carlo method: an approach to approximate counting and integration. *Approximation algorithms for NP-hard problems*, pages 482–520, 1997.

[JTMM10] S. Johnson, J.J. Torres, J. Marro, and M.A. Muñoz. The entropic origin of disassortativity in complex networks. *Physical Review Letters*, 104(10):108702, Mar 2010.

[KH07] D. Krackhardt and M. Handcock. Heider vs Simmel: Emergent features in dynamic structures. *Statistical Network Analysis: Models, Issues, and New Directions*, pages 14–27, 2007.

[KKR+99] J.M. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A.S. Tomkins. The web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science*, pages 1–17, 1999.

[KL70] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.

[Kle99] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[Kle02] J.S. Kleinfeld. The small world problem. *Society*, 39(2):61–66, 2002.

[KRR+00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 57–65, 2000.

[KTV97] R. Kannan, P. Tetali, and S. Vempala. Simple Markov-chain algorithms for generating bipartite graphs and tournaments. In *Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, pages 193–200. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1997.

[LA05] R. Lambiotte and M. Ausloos. N-body decomposition of bipartite author networks. *Physical Review E*, 72(6):66117, 2005.

[Lan77] S. Langlois. Les réseaux personnels et la diffusion des informations sur les emplois. *Recherches sociographiques*, 2:213–245, 1977.

[Lat08] M. Latapy. Practical algorithms for triangle computations in very large (sparse (power-law)) graphs. *J. Theoretical Computer Science*, 407:458–473, 2008.

[LAWD04] L. Li, D. Alderson, W. Willinger, and J. Doyle. A first-principles approach to understanding the internet's router-level topology. *ACM SIGCOMM Computer Communication Review*, 34(4):3–14, 2004.

[LBKT08] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[LC07] C.C. Leung and H.F. Chau. Weighted assortative and disassortative networks model. *Physica A: Statistical Mechanics and its Applications*, 378(2):591–602, 2007.

[LD72] W.T. Liu and R.W. Duff. The strength in weak ties. *Public Opinion Quarterly*, 36(3):361–366, 1972.

[LEV81] N. Lin, W.M. Ensel, and J.C. Vaughn. Social Resources and Strength of Ties: Structural Factors in Occupational Status Attainment. *American Sociological Review*, pages 393–405, 1981.

[LGH05] P.G. Lind, M.C. Gonzalez, and H.J. Herrmann. Cycles and clustering in bipartite networks. *Physical review E*, 72(5):56127, 2005.

[LH08] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. 2008.

[Liu08] J.S. Liu. *Monte Carlo strategies in scientific computing*, chapter 5: Metropolis algorithm and beyond. Springer Verlag, 2008.

[LMDV08] M. Latapy, C. Magnien, and N. Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

[Luc50] R.D. Luce. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15(2):169–190, 1950.

[Mar02] P.V. Marsden. Egocentric and sociocentric measures of network centrality. *Social Networks*, 24(4):407–422, 2002.

[Mer68] R.K. Merton. The Matthew effect in science: the reward and communication systems of science are considered. *science*, 159(3810):56, 1968.

[MGGP08] B. Mitra, N. Ganguly, S. Ghose, and F. Peruani. Generalized theory for node disruption in finite-size complex networks. *Physical Review E*, 78(2), 2008.

[MIK+04] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538, 2004.

[Mil67] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[Miz94] M.S. Mizruchi. Social network analysis: Recent achievements and current controversies. *Acta Sociologica*, 37(4):329, 1994.

[MKFV06] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat. Systematic topology analysis and generation using degree correlations. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, page 146. ACM, 2006.

[MKI+03] R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *Arxiv preprint cond-mat/0312028*, 2003.

[MMG+07] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, page 42. ACM, 2007.

[Mok79] R.J. Mokken. Cliques, clubs and clans. *Quality and Quantity*, 13(2):161–173, 1979.

[MP04] I. Miklós and J. Podani. Randomization of presence-absence matrices: comments and new algorithms. *Ecology*, 85(1):86–92, 2004.

[MPSV02] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

[MR95] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 161(6):161–179, 1995.

[MRR+53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087, 1953.

[MS02] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910, 2002.

[MSLC01] M. McPherson, L. Smith-Lovin, and J.M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[MSOI+02] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824, 2002.

[MW90] B.D. McKay and N.C. Wormald. Uniform generation of random regular graphs of moderate degree. *J. Algorithms*, 11(1):52–67, 1990.

[MW03] J. Moody and D.R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, pages 103–127, 2003.

[New02] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[New03a] M.E.J. Newman. *Handbook of graphs and networks: From the genome to the Internet*, chapter Random graphs as models of networks. Vch Verlagsgesellschaft Mbh, 2003.

[New03b] M.E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):26126, 2003.

[New03c] M.E.J. Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):26121, 2003.

[New03d] M.E.J. Newman. The structure and function of complex networks. *Arxiv preprint cond-mat/0303516*, 2003.

[New09] M.E.J. Newman. Random graphs with clustering. *Physical Review Letters*, 103(5):58701, 2009.

[NG04] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):26113, 2004.

[NP03] M.E.J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):36122, 2003.

[NSW01] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):26118, 2001.

[PA93] J.F. Padgett and C.K. Ansell. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology*, 98(6):1259, 1993.

[Par98] B.N. Parlett. *The symmetric eigenvalue problem*. Society for Industrial Mathematics, 1998.

[PCMG07] F. Peruani, M. Choudhury, A. Mukherjee, and N. Ganguly. Emergence of a non-scaling degree distribution in bipartite networks: A numerical and analytical study. *Europhysics Letters*, 79, 2007.

[PDFV05] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[PSV01] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.

[PSV04] R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet: A statistical physics approach.* Cambridge University Press, 2004.

[PSVV01] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the Internet. *Physical Review Letters*, 87(25):258701, 2001.

[RB06a] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):16110, 2006.

[RB06b] C. Roth and P. Bourgine. Lattice-based dynamic and overlapping taxonomies: The case of epistemic communities. *Scientometrics*, 69(2):429–447, 2006.

[RC09] C. Roth and J.P. Cointet. Social and semantic coevolution in knowledge networks. *Social Networks*, 2009.

[RCC+04] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658, 2004.

[RJB96] A.R. Rao, R. Jana, and S. Bandyopadhyay. A Markov chain Monte Carlo method for generating random (0, 1)-matrices with given marginals. *Sankhyā: The Indian Journal of Statistics, Series A*, 58(2):225–242, 1996.

[Rob00] J.M. Roberts. Simple methods for simulating sociomatrices with given marginal totals. *Social Networks*, 22(3):273–283, 2000.

[RPKL07] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, 2007.

[SBF+08] A. Scherrer, P. Borgnat, E. Fleury, J.L. Guillaume, and C. Robardet. Description and simulation of dynamic mobility networks. *Computer Networks*, 52(15):2842–2858, 2008.

[Sch09] S. Schnettler. A structured overview of 50 years of small-world research. *Social Networks*, 2009.

[Sei83] S.B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

[SI90] D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

[Sim56] H.A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, 1956.

[SMP90] Borgatti S.P., Everett M.G., and Shirey P.R. LS sets, lambda sets and other cohesive subsets. *Social Networks*, 12(4):337–357, 1990.

[SOMMA02] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

[SPRH06] T.A.B. Snijders, P.E. Pattison, G.L. Robins, and M.S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, 2006.

[SR03] A.J. Seary and W.D. Richards. Spectral methods for analyzing and visualizing networks: an introduction. In *Dynamic Social Network Modeling and Analysis: workshop summary and papers*, 2003.

[SR06] M.V. Simkin and V.P. Roychowdhury. Re-inventing Willis. *Arxiv preprint physics/0601192*, 2006.

[Tay80] R. Taylor. Constrained switchings in graphs. *Combinatorial Mathematics*, 8:314—336, 1980.

[Tay82] R. Taylor. Switchings constrained to 2-connectivity in simple graphs. *SIAM Journal on Algebraic and Discrete Methods*, 3:114, 1982.

[TBCB08] T. Teitelbaum, P. Balenzuela, P. Cano, and J.M. Buldú. Community structures and role detection in music networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18:043105, 2008.

[TCR08] L. Tabourier, J.P. Cointet, and C. Roth. Cycles in hypergraph-based networks: signal or noise, artefacts or processes? In *Proceedings of Algotel'08 10th "francophone summit on algorithms for telecommunications"*, 2008.

[TCR10] C. Taramasco, J.P. Cointet, and C. Roth. Academic team formation as evolving hypergraphs. *Scientometrics*, pages 1–20, 2010.

[TM69] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[TRC10] L. Tabourier, C. Roth, and J.P. Cointet. Generating constrained random graphs using multiple edge switches. *Arxiv preprint arXiv:1012.3023*, 2010.

[VL05] F. Viger and M. Latapy. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. *Lecture Notes in Computer Science*, 3595:440, 2005.

[Wer06] S. Wernicke. Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(4):347–359, 2006.

[WF94] S. Wasserman and K. Faust. Social network analysis: Methods and applications. 1994.

[WS98] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

[XZSS09] X.K. Xu, J. Zhang, J. Sun, and M. Small. Revising the simple measures of assortativity in complex networks. *Physical Review E*, 80(5):56106, 2009.

[Yca02] B. Ycart. *Modèles et algorithmes markoviens*, chapter 4: Exploration markovienne. Springer, 2002.

[YZ04] R. Yuster and U. Zwick. Detecting short directed cycles using rectangular matrix multiplication and dynamic programming. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 254–260. SIAM, 2004.

[ZBŠD06] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1):16115, 2006.

[Zip49] G.K. Zipf. Human Behavior and the Principle of Least Effort, 1949.

[ZM03] S. Zhou and R.J. Mondragon. The rich-club phenomenon in the Internet topology. *Arxiv preprint cs/0308036*, 2003.

[Zwi01] U. Zwick. Exact and approximate distances in graphs - a survey. *Lecture Notes in Computer Science*, pages 33–48, 2001.