Université paris 7 - Denis diderot
Ufr d'informatique

Thèse de Doctorat
Informatique

par
Jean-Loup Guillaume

sujet de la thèse

Analyse statistique et modélisation des grands réseaux d'interactions

Thèse dirigée par
Matthieu Latapy

soutenue le 20 décembre 2004 devant le jury constitué de

Stéphane Boucheron (Examinateur)
Serge Fdida (Rapporteur)
Éric Fleury (Examinateur)
Pierre Fraigniaud (Rapporteur)
Michel Habib (rapporteur)
Matthieu Latapy (directeur de thèse)

Remerciements

Si les trois années et quelques mois qu'a duré ma thèse se sont déroulés dans d'excellentes conditions, c'est le fait de plusieurs personnes auxquelles je souhaite ici adresser quelques mots pour qu'elles sachent, si ce n'est pas déjà le cas, tout ce que je leur dois.

Matthieu Latapy a été, depuis un peu plus de deux ans, mon directeur de thèse et c'est un honneur pour moi d'avoir été son premier thésard. C'est avec un réel plaisir que je repense à ces années durant lesquelles nos incessantes réunions ont fait progresser pas à pas une certaine vision des grands réseaux d'interactions. J'espère que cette collaboration restera fructueuse dans les années à venir.

Et même s'il n'apprécie pas particulièrement les grandes phrases et les éloges, je souhaite le remercier pour ses conseils, sa disponibilité et surtout son amitié. Il m'a toujours aidé dans les moments difficiles et m'a guidé dans la bonne direction dans mon travail. Je lui dois beaucoup.

Michel Morvan et Laurent Viennot m'ont encadré durant mon stage de DEA et ma première année de thèse. Ils m'ont fait découvrir le monde de la recherche et fait connaître le travail d'équipe. Ils m'ont aussi permis de trouver ma voie dans ce qui était à l'époque une première étape vers l'étude des grands réseaux d'interactions et qui a depuis bien progressé. Mais beaucoup reste à faire et je suis certain que nos routes se croiseront à nouveaux sur ce domaine ou un autre.

Je remercie aussi tout particulièrement Serge Fdida, Pierre Fraigniaud et Michel Habib qui ont accepté la lourde tache de relire mon mémoire de thèse et de me faire part de leur opinion la concernant. Je suis conscient du travail que cela représente. Je souhaite aussi remercier Stéphane Boucheron et Eric Fleury qui ont bien voulu faire partie de mon jury.

Au rang des collègues de bureau, je me dois de distinguer Clémence Magnien que j'ai côtoyée pendant toute la durée de ma thèse. Clémence a relu tous mes articles, thèse comprise, et m'a beaucoup aidé dans mon travail. Je lui en suis énormément redevable. Mais elle est bien plus qu'une simple collègue et je dois aussi la remercier pour sa gentillesse et les multiples viennoises au chocolat qu'elle n'a pas hésité à partager en échange d'un simple café.

De nombreuses autres personnes ont partagé mon bureau durant ces quelques années sur de plus ou moins longues périodes et ont contribué à perpétuer la tradition de convivialité

qui anime le bureau 6A51. Espérons que cela continue de la sorte. Dans le désordre, je remercie Mohssen Abboud, Frédéric Chavanon, Arnaud Dartois, Mahendra Mariadasson, Pascal Pons, Fabien Viger et Jérôme Waldispush.

Noëlle Delgado, secrétaire efficace s'il en est, a non seulement résolu tous mes problèmes administratifs avec le sourire tout au long de ma thèse, mais a fait bien plus que cela, partageant de nombreux déjeuners et autres moments.

J'ai aussi eu la chance durant ma thèse d'encadrer, au moins partiellement, de nombreux étudiants. En particulier, Stevens Le-Blond, Samuel Leboeuf et Jérôme Wrede. Leur présence m'a beaucoup apporté et m'a permis d'appréhender l'autre versant de ce type de collaboration.

Enfin je souhaite remercier tous les chercheurs du LIAFA et ceux que j'ai pu rencontrer à l'occasion de diverses conférences ou séminaires et qui m'ont toujours considéré comme un pair dès les premiers jours.

J'ai naturellement oublié de citer de nombreuses personnes, je les prie de m'en excuser. Même si leur nom n'apparaît pas dans cette thèse, qu'elles soient assurées de toute ma reconnaissance.

Enfin, merci à celle qui occupe l'autre moitié de ma vie pour son soutien, pour son aide et pour tout le reste...

Table des matières

Introduction

Depuis quelques années, on a observé une véritable explosion du nombre de travaux centrés sur l'étude des grands réseaux d'interactions, domaine qui était jusque là complètement anonyme, a été observée. Ces études visent à expliquer les interactions entre les différents individus d'un réseau par le biais des grandes lois le gouvernant, d'une part, et à comprendre les divers phénomènes pouvant se produire sur ces réseaux, tels la propagation d'une maladie dans un réseau social ou une tentative de destruction d'un réseau informatique, d'autre part.

Le récent engouement pour ce domaine est dû à la découverte de propriétés non triviales communes à un certain nombre de réseaux qui n'ont, *a priori*, rien en commun. Il semblerait, par exemple, peu naturel d'étudier dans un même article la manière dont les pages Web sont reliées entre elles et les interactions entre protéines à l'intérieur du corps humain. Pourtant, ces deux types de relations sont plus semblables qu'on ne pourrait le croire.

Les grands réseaux d'interactions regroupent une large variété d'objets. Parmi les plus typiques, en plus des interactions entre protéines et des liens entre pages Web, on peut citer le réseau de l'Internet (des ordinateurs reliés par des câbles) et les réseaux ferroviaires, routiers ou de distribution d'électricité. Divers types de réseaux sociaux sont aussi étudiés, selon le type de relation entre individus que l'on considère : amitié, relations de travail, correspondance par courrier, etc. Ces exemples, et bien d'autres, sont étudiés dans de nombreuses disciplines, de l'informatique aux sciences sociales, en passant par la biologie, la physique, l'économie ou encore la linguistique.

Comprendre comment sont structurés les grands réseaux d'interactions, comment ils évoluent et quels sont les phénomènes agissant dessus sont donc les points centraux de ce domaine de recherche. Ces divers points sont structurés en quatre grands axes de recherche sur les grands réseaux d'interactions : la métrologie, l'analyse, la modélisation et l'algorithmique de ces réseaux.

Métrologie, analyse, modélisation et algorithmique

La plupart des grands réseaux d'interactions ne sont pas accessibles directement, mais ne sont observables qu'au travers d'une opération de mesure. Celle-ci est généralement complexe et constitue en soi une problématique de recherche. Au-delà de la mesure d'un réseau particulier, la **métrologie** vise à déterminer, étant donné un processus de mesure

et son résultat, ce que celui-ci nous apprend sur l'objet réel. La vision obtenue est généralement *partielle* et même *biaisée*. L'évaluation de la représentativité de l'échantillon obtenu, l'identification des biais et la mise au point de méthodes de mesure plus efficaces sont au cœur de cette problématique.

Une fois que l'on dispose de données que l'on estime suffisamment représentatives, l'**analyse** a pour objet de déterminer les principales caractéristiques de ces données pour obtenir une description fine du réseau. Il s'agit également d'évaluer l'impact des propriétés de ces réseaux sur divers phénomènes comme, par exemple, les phénomènes de diffusion (rumeurs, maladies, attaques, etc.). C'est généralement au niveau applicatif que l'on détermine quelles sont les propriétés pertinentes à considérer. L'étude de cas particuliers pose souvent des questions originales qui peuvent ensuite être généralisées ou utilisées pour d'autres cas particuliers.

Une fois que l'on dispose d'une description suffisamment fine de l'objet étudié, on peut entreprendre une phase de **modélisation**. Cette étape consiste principalement à pouvoir *produire des objets artificiels similaires à l'objet réel*, c'est-à-dire ayant les même propriétés et/ou le même comportement. La modélisation permet notamment de quantifier l'influence des différents paramètres dans le comportement des réseaux, que ce soit par des simulations ou dans des analyses formelles. L'effort nécessaire pour proposer un modèle pertinent se traduit, de plus, par une meilleure compréhension de l'objet, on peut donc considérer la définition d'un modèle comme un but en soi.

Un autre aspect de la modélisation concerne la *modélisation des phénomènes*. Les phénomènes de diffusion (de rumeurs, d'informations, de maladies, d'innovations, etc.), et bien d'autres phénomènes qu'on peut souhaiter étudier sur des réseaux, posent ainsi des problèmes de modélisation qui font partie de nos préoccupations et seront abordés plus loin.

Finalement, un domaine peu exploré jusqu'à présent concerne le développement d'une **algorithmique** dédiée aux grands réseaux d'interactions. Leur taille étant généralement très importante, de quelques milliers à quelques milliards de sommets, il n'est pas toujours possible d'effectuer des calculs exacts sur ces réseaux, et l'on est souvent obligé de faire des approximations. Or, l'analyse des grands réseaux d'interactions fournit une description de ces réseaux, et notamment de leur propriétés statistiques. Celles-ci peuvent être prises en compte pour définir des algorithmes efficaces sur ces types de réseaux.

Méthodologie

Dans ce domaine majoritairement étudié par en physique statistique, nous avons privilégié une approche basée sur plusieurs problèmes pratiques et essayant de mettre en avant les méthodes et outils informatiques.

Les réseaux sur lesquels nous travaillons sont issus de problèmes appliqués ; nous avons donc interagi fortement avec des personnes impliquées dans ces applications, que ce soit en réseaux informatiques, en sciences humaines et sociales ou en physique. Ceci nous a

permis à la fois d'obtenir des données sur lesquelles travailler et d'évaluer la pertinence de nos résultats.

Les principaux outils que nous avons utilisés sont :

- la modélisation probabiliste, qui apparaît à plusieurs endroits dans nos travaux et joue un rôle particulièrement important pour le réalisme et la pertinence de nos résultats ;
- l'algorithmique, et notamment l'algorithmique des graphes, qui fournit un ensemble de méthodes et de notions qui sont au cœur de tous nos travaux ;
- la simulation, qui joue dans notre contexte un rôle central en permettant de se faire une intuition de certains résultats qui seraient, sans ce recours, difficiles à appréhender ;
- l'analyse statistique, utilisée pour l'analyse de données et l'analyse des résultats de simulation. Il s'agit, par exemple, d'être capable d'évaluer la représentativité d'un échantillon, de détecter des phénomènes d'échelle ou de corréler divers phénomènes.

En pratique, l'étude d'un réseau passe par plusieurs étapes : au minimum une phase de mesure, suivie d'une phase d'analyse, et parfois une étape de modélisation. Nous verrons plus loin que ces étapes, liées aux quatre points définis précédemment, sont très imbriquées et ne s'effectuent pas dans un ordre linéaire. Ainsi, la métrologie est naturellement reliée à la mesure, mais définir un protocole de mesure efficace nécessite de connaître les propriétés de l'objet mesuré et éventuellement d'avoir testé ce protocole sur des graphes aléatoires issus de modèles pour tester son efficacité.

Ces différents axes généraux sont fortement liés entre eux, mais aussi avec les applications. Tout au long de cette thèse, et en particulier dans la troisième partie, nous montrerons comment une description des grands réseaux d'interactions est nécessaire afin d'obtenir une modélisation correcte et comment ces modèles peuvent être utilisés pour mieux comprendre certains phénomènes ou pour définir des procédures de mesure efficaces.

Organisation de la thèse

Les résultats présentés dans cette thèse sont articulés en trois parties axées respectivement sur l'analyse des grands réseaux d'interactions, leur modélisation, et enfin une mise en pratique de ces deux aspects qui éclaire leur apports respectifs et ouvre de nouvelles voies.

La première partie est consacrée à l'analyse des grands réseaux d'interactions et consiste essentiellement en un travail de description des réseaux étudiés mettant en avant leurs propriétés les plus caractéristiques. Cette partie est principalement une partie introductive faisant le point sur les réseaux et les propriétés étudiés. Mais nous nous attacherons aussi à discuter plus précisément certains points utiles pour la suite.

Le premier chapitre définit certains grands réseaux d'interactions représentatifs de ceux actuellement étudiés. Nous verrons dans ce chapitre quels sont les graphes les plus étudiés et les raisons pour lesquelles on les étudie. Nous présenterons aussi les méthodes utilisées

pour obtenir ces graphes, l'impact que cela peut avoir en termes de qualité de la mesure et quelques résultats obtenus sur ces graphes. Certains graphes particuliers feront l'objet d'une description plus poussée, notamment ceux reliés à l'Internet : le graphe de l'Internet lui-même (routeurs reliés par des liens physiques), le graphe du Web (pages Web reliées par des liens hypertextes) et des graphes d'échanges pair-à-pair. Nous évoquerons aussi des réseaux sociaux, biologiques et provenant d'autres domaines encore. Ces graphes, et notamment un ensemble représentatif de six cas, serviront de bases aux chapitres ultérieurs.

Le deuxième chapitre présente les principales propriétés utilisées pour l'analyse de ces graphes. Certaines propriétés sont partagées par l'ensemble des grands réseaux d'interactions rencontrés en pratique, notamment la faible distance entre les individus (effet *petit-monde*), le fait que les connaissances d'un individu soient très liées entre elles (*clustering*), ou encore la diversité du nombre de contacts des individus (graphes *sans-échelle*). Nous présenterons aussi diverses corrélations entre ces propriétés et d'autres notions plus subtiles. Nous illustrerons toutes ces notions sur les six exemples introduits dans le Chapitre 1, et en discuterons les implications.

La seconde partie constitue le cœur de cette thèse. Elle est consacrée à la modélisation des grands réseaux d'interactions, c'est-à-dire la construction de graphes artificiels similaires aux graphes rencontrés en pratique (au sens de diverses propriétés statistiques).

Le troisième chapitre présente un état de l'art passant en revue les principaux modèles actuellement utilisés : le modèle de graphes aléatoires d'Erdös et Renyi [46], les travaux fondateurs de Watts et Strogatz [134] sur le clustering et ceux d' Albert et Barabási sur les distributions des degré en loi puissance [8], ainsi que divers autres. Dans ce chapitre, nous distinguerons notamment deux grandes classes de modèles : la première est basée sur le tirage d'un graphe aléatoire ayant certaines propriétés imposées, la seconde repose sur la définition d'un processus de construction inspiré de la réalité.

Le quatrième chapitre met en lumière le fait que certaines propriétés non triviales permettent de proposer un modèle pour les grands réseaux d'interactions qui soit à la fois suffisamment simple pour que l'on puisse étudier formellement ses propriétés et malgré tout réaliste. Le modèle présenté dans ce chapitre est basé sur l'observation de plusieurs grands réseaux d'interactions et de la façon dont ils sont construits en pratique mais s'applique finalement à tous les grands réseaux d'interactions. Les résultats de ce chapitre ont été publiés dans [60, 61].

Le cinquième chapitre présente une extension du modèle du Chapitre 4. Ce dernier est en effet encore imparfait : il ne prend pas en compte de nombreuses corrélations qui sont présentes dans les graphes originaux. Le modèle présenté dans ce chapitre arrive à capturer la plupart des propriétés de base, mais aussi des propriétés plus complexes. Il est le modèle le plus performant actuellement disponible par les propriétés qu'il capture, son réalisme et sa simplicité. Un article le décrivant est actuellement en cours de rédaction.

Enfin, **la troisième partie** présente la mise en pratique des parties précédentes et l'apport qui en découle. Cette partie est donc applicative mais nous la verrons principalement sous l'angle *méthodologique* : outre les façons diverses d'aborder l'étude des grands réseaux d'interactions, elle montre que c'est souvent lors de l'étude de réseaux ou de problèmes

particuliers que les concepts fondamentaux les plus pertinents émergent.

Le sixième chapitre est voué à l'étude d'un réseau d'échange de données pair-à-pair. L'analyse des échanges entre les différents utilisateurs d'un tel réseau permet d'obtenir des informations nouvelles (comportement des utilisateurs, par exemple), mais est aussi la base de travaux plus généraux. Ce chapitre présente en particulier des résultats sur l'évolution du graphe des échanges au cours du temps, qui ouvre des pistes prometteuses pour l'analyse de la dynamique des graphes en général. Une partie des résultats de ce chapitre a été publiée dans [57, 65].

Le septième chapitre s'attache à l'étude de la résistance d'un réseau aux pannes et aux attaques. On étudie ainsi l'impact de la suppression d'un individu dans un réseau sur la capacité à communiquer des individus restants. Tout au long de ce chapitre, nous verrons que les propriétés du réseau lui-même ont un impact fondamental sur sa robustesse. Nous verrons aussi comment la modélisation peut être utilisée pour comprendre ces phénomènes et pour les étudier formellement. Certains résultats de ce chapitre ont fait l'objet d'une publication dans [66].

Le huitième et dernier chapitre utilise la modélisation des graphes pour étudier, à l'aide de simulations, l'impact de leur structure sur leur exploration. Dans ce chapitre, nous montrerons notamment qu'obtenir une carte précise de l'Internet est non seulement très complexe mais que la façon de procéder est très dépendante du réseau lui-même : en fonction des propriétés du graphe, et de la façon de le mesurer, les propriétés ne sont pas toutes aussi bien capturées. Ce chapitre montre comment la modélisation de grands réseaux d'interactions peut être utilisée pour mener des simulations apportant un éclairage précis sur une question concrète. Les résultats de ce chapitre ont été partiellement publiés dans [64].

D'autres études effectuées pendant cette thèse ont fait l'objet de publications dans des conférences ou des revues internationales ou nationales. C'est le cas de [58, 59, 63] qui présentent respectivement de courts états de l'art sur la mesure de l'Internet, la modélisation des grands réseaux d'interactions et le graphe du Web. Le contenu de ces trois courts articles est disséminé tout au long de la thèse, notamment dans les deux premières parties.

Première partie

Analyse statistique

Introduction

Jusqu'à la fin des années 1990, les grands réseaux d'interactions étaient analysés de manière isolée. Leur étude était souvent liée à un problème spécifique posant des questions originales, mais ces problématiques n'étaient pas étendues à d'autres contextes. Il y a quelques années, l'analyse des grands réseaux d'interactions a mis en évidence des similarités entre des réseaux très divers. L'observation de ces similarités s'est alors généralisée, de nombreux réseaux jusque-là complètement ignorés devenant soudain l'objet d'études.

Dans cette optique, l'analyse des grands réseaux d'interactions a pour objet de déterminer les principales caractéristiques des réseaux considérés afin de les décrire de la façon la plus pertinente possible. Ceci passe par la définition de propriétés, souvent statistiques, des réseaux. On mesure alors ces propriétés sur les cas considérés, dont on obtient ainsi une description statistique. Toute la difficulté réside dans l'identification et la définition de propriétés pertinentes, capturant une information importante sur le réseau. On peut ainsi détecter les propriétés partagées par tous les grands réseaux d'interactions et celles qui les différencient les uns des autres.

L'analyse se situe aussi en aval d'autres travaux, notamment de modélisation ou de simulation : il s'agit alors d'interpréter les résultats obtenus et d'en tirer des conclusions utiles d'un point de vue applicatif ou fondamental.

Dans de nombreux cas, il est aussi essentiel de disposer d'informations sur les phénomènes qui ont lieu sur ces réseaux, tels la propagation de virus. Les phénomènes et les réseaux sur lesquels ils agissent sont généralement très liés, la structure des réseaux influant souvent sur ces phénomènes (et parfois l'inverse).

Dans cette partie, nous présenterons tout d'abord plusieurs grands réseaux d'interactions de natures diverses. L'Internet, par sa nature numérique et en ligne offre notamment de nombreux réseaux à étudier, du réseau physique reliant ordinateurs et routeurs par des câbles, aux réseaux purement virtuels tels que le Web ou les réseaux d'échanges de fichiers avec le pair-à-pair. L'objectif de ce premier chapitre est de présenter quelques objets d'étude typiques, qui nous servirons de base dans toute la suite.

Ensuite, nous présenterons dans le Chapitre 2 un certain nombre de paramètres identifiés sur ces réseaux. Ces propriétés permettent de mieux comprendre les points communs entre les grands réseaux d'interactions ainsi que leurs différences, et seront utilisées principalement dans cette optique dans cette thèse.

Chapitre 1

Grands réseaux d'interactions étudiés

Dans ce chapitre, nous allons présenter des objets d'étude typiques de la discipline. La plupart de ces réseaux peuvent être classés dans trois catégories : les réseaux sociaux, les réseaux d'infrastructures et les réseaux biologiques. Tous ont fait l'objet d'études poussées dont nous citerons quelques résultats sans toutefois entrer dans les détails. Notre objectif ici est d'introduire un ensemble d'exemples qui nous seront utiles dans toute la suite et qui illustreront bien la diversité des champs d'application.

Nous allons ainsi présenter plus précisément plusieurs réseaux en insistant sur les problèmes spécifiques qu'ils posent, concernant notamment leur acquisition, et certaines solutions qui y ont été apportées. La Section 1.1 est dédiée aux réseaux qu'on peut trouver sur l'Internet, à la fois au niveau physique (comment sont connectés les ordinateurs) et au niveau des réseaux engendrés par les applications (Web, courrier électronique, pair-à-pair, etc.). Ensuite, les Sections 1.2, 1.3 et 1.4 décrivent plusieurs réseaux sociaux, biologiques et provenant d'autres domaines qui seront récurrents dans toute cette thèse : ils nous serviront de cas d'étude pour les problèmes que nous allons aborder. Nous insisterons en particulier sur les aspects qui vont nous être utiles dans la suite.

1.1 Différentes visions de l'Internet

L'Internet est constitué d'un ensemble de câbles reliés à des interfaces (cartes réseaux). Les interfaces sont physiquement placées dans des routeurs qui servent de carrefours pour faire transiter l'information. Ces routeurs appartiennent à leur tour à des entités nommées systèmes autonomes qui les gèrent. Ceci définit trois réseaux pour l'Internet physique : interfaces, routeurs et systèmes autonomes.

Mais on peut aussi s'intéresser aux applications qui utilisent l'Internet comme support pour faire circuler de l'information. Le Web, typiquement, constitue un réseau de pages reliées les unes aux autres par des liens hypertexte.

Par ailleurs, les usages que font les utilisateurs d'autres applications, comme par exemple les échanges de courriers électroniques ou les échanges de données sur des réseaux de type pair-à-pair, forment un troisième type de réseaux. Ceux-ci ont toutefois un caractère prin-

cipalement socio-culturel et seront donc traités dans la section suivante, dédiée à ce type de réseaux.

1.1.1 L'Internet au niveau physique

Au niveau le plus bas, l'Internet est constitué de routeurs reliés entre eux par des câbles (fibres optiques ou autres). Chaque routeur contient un ensemble d'interfaces et c'est à ces dernières que les câbles sont reliés. À chaque interface est associée une adresse IP qui permet de l'identifier de manière unique (voir Figure 1.1).

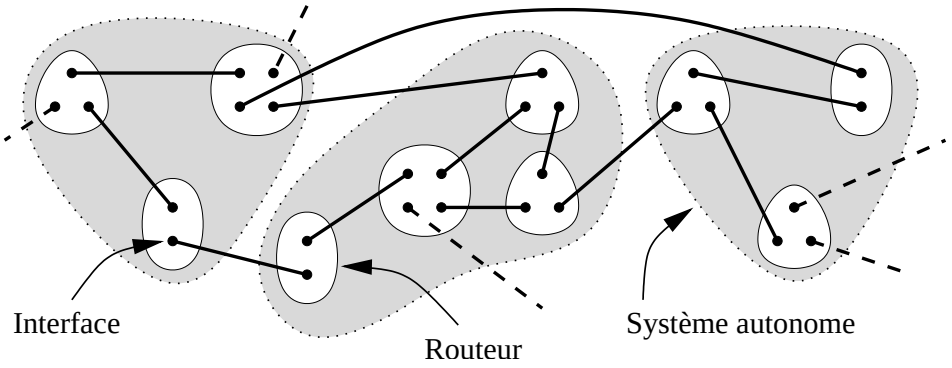

FIG. 1.1 – *Les trois niveaux de l'Internet : interfaces, routeurs et systèmes autonomes.*

Les routeurs forment le cœur du réseau : ils peuvent transmettre de l'information reçue par une interface à une autre interface et font donc transiter l'information. Cette information transite donc de routeur en routeur depuis une source jusqu'à une destination. Le travail principal des routeurs consiste surtout à router l'information dans la bonne direction.

Obtenir une carte de l'Internet

L'administration des routeurs n'étant pas centralisée, il n'est pas possible d'obtenir une vue complète du réseau directement, ce qui serait pourtant très utile, par exemple, pour router les messages efficacement, détecter les pannes ou savoir comment modifier le réseau pour le rendre plus performant. Il n'est pas non plus possible de contacter les routeurs pour obtenir une liste de leurs interfaces et savoir à qui elles sont reliées (certains routeurs l'autorisent mais ils sont très minoritaires).

La solution la plus utilisée pour pallier cette absence de carte du réseau repose sur une utilisation intensive de mesures. Le protocole de mesure consiste principalement à envoyer des messages depuis une machine source vers une autre et regarder par où ils passent [48, 50, 124, 132]. En utilisant l'outil `traceroute` [87], il est en effet possible de connaître la succession de routeurs par lesquels passe un message.

Le fonctionnement de cet outil repose sur l'utilisation de TTL (*Time To Live*) : ce champ contenu dans les messages indique le nombre de routeurs maximum que le message

peut traverser, ce qui permet d'éviter qu'il circule indéfiniment. À chaque fois qu'un routeur transfère un paquet, il décrémente ce champ. Dès que cette valeur atteint 0, le message est détruit et le routeur qui le détruit en informe l'expéditeur. En envoyant un message vers une destination quelconque avec un TTL de 1, il est donc possible de connaître le premier routeur (en fait une interface de ce routeur) traversé par le message, avec un TTL de 2 le deuxième routeur, etc. On peut ainsi reconstituer le chemin suivi par le paquet envoyé (Figure 1.2). Voir [87] pour plus de détails.

```
1  border-gw.utfsm.cl (200.1.20.193)              189.933 ms  247.787 ms  271.905 ms
2  12.145.193.89 (12.145.193.89)                  139.820 ms  140.628 ms  133.082 ms
3  nw2nj31ck4.pos2-1.ip.att.net (12.119.12.5)     140.198 ms  139.378 ms  151.884 ms
4  tbr2-p012402.n54ny.ip.att.net (12.123.219.133) 152.273 ms  138.339 ms  147.824 ms
5  ggr2-p3120.n54ny.ip.att.net (12.123.3.109)     182.826 ms  146.080 ms  137.283 ms
6  att-gw.nyc.opentransit.net (192.205.32.138)    137.205 ms  158.516 ms  150.851 ms
```

FIG. 1.2 – *Un exemple de* `traceroute` *: les différents routeurs traversés sont indiqués ainsi que leur adresse IP, et le temps mis pour effectuer l'aller-retour (l'opération est répétée trois fois).*

Des méthodes plus complexes utilisent le *source-routing* (ou LSRR : *loose source record route*) qui permet de trouver un chemin vers une destination en passant par un certain nombre de routeurs intermédiaires choisis [56, 68]. Cette technique peut être utilisée pour détecter des anomalies dans le réseau ou éviter de surcharger certains liens, mais elle peut aussi servir pour le piratage informatique et n'est maintenant plus utilisable, un grand nombre de routeurs détruisant les messages faisant appel à cette fonction.

En effectuant un grand nombre de `traceroute`, avec ou sans source-routing, depuis diverses sources vers diverses destinations, il est possible de reconstituer une partie non négligeable du réseau des interfaces. Ce réseau est cependant très rarement étudié, et on lui préfère le réseau des routeurs, plus facile à interpréter. Obtenir le réseau des routeurs à partir de celui des interfaces consiste à identifier les différentes interfaces associées à un routeur [56, 124]. Cette identification est toutefois un problème délicat, pour lequel Il n'y a pas de méthode sûre, mais plusieurs heuristiques.

Une méthode consiste à utiliser le fait que les routeurs qui répondent à un message doivent s'identifier et le font souvent en donnant l'adresse de l'interface par laquelle ils répondent **??**. Ainsi, en envoyant un message vers une interface X d'un routeur[1], on peut obtenir une réponse par l'interface Y, ce qui signifie que X et Y appartiennent au même routeur. Ces opérations doivent être répétées souvent car les routes peuvent changer et donc l'interface utilisée par le routeur aussi. Il faut aussi noter que certains routeurs ne répondent pas à ces messages. D'autres méthodes ont été introduites pour tenter de prédire les interfaces probables d'un même routeur en considérant, parmi d'autres hypothèses, que

1. En pratique, un message UDP est envoyé vers un port invalide pour obtenir une réponse ICMP *port unreachable*.

les adresses IP de ces interfaces sont généralement proches, ou que les temps mis pour atteindre deux interfaces d'un même routeur sont similaires [124].

Une fois identifiées les interfaces de tous les routeurs, on peut définir le réseau des routeurs en considérant que deux routeurs sont reliés si deux de leurs interfaces le sont.

Problèmes liés à la mesure

Comme nous venons de le voir, l'obtention d'une carte de l'Internet passe par plusieurs étapes qui posent de gros problèmes de fiabilité :

- la mesure repose tout d'abord sur l'utilisation de `traceroute` à partir d'un certain nombre de sources. Or, pour pouvoir utiliser `traceroute`, il faut pouvoir se connecter sur la machine source. C'est la raison pour laquelle les cartes actuelles sont construites avec très peu de sources, de l'ordre de quelques dizaines ;
- le nombre de destinations est *a priori* illimité puisqu'il suffit de connaître l'adresse IP d'une machine pour lancer un `traceroute` vers elle. Cependant, on ne connaît pas l'ensemble des adresses réellement utilisées et le nombre d'adresses IP existantes est trop important pour envisager une approche exhaustive. En pratique, on se restreint donc à un nombre limité de destinations (quelques centaines de milliers) ;
- à chaque étape, `traceroute` donne l'adresse d'un routeur à une certaine distance de la source, ce qui ne veut pas dire que deux routeurs retournés successivement soient effectivement reliés. En effet, les routeurs peuvent choisir de ne pas toujours faire transiter les messages dans la même direction, notamment pour répartir la charge. Deux messages successifs peuvent donc emprunter des chemins distincts, ce qui fausse le résultat de `traceroute` ;
- les routeurs devraient répondre aux messages qui expirent avec une adresse correcte, mais certains ne répondent pas ou donnent de fausses informations afin de dissimuler la structure du réseau, ou parce qu'ils sont mal configurés[2] ;
- le réseau contient des liens utilisés uniquement en cas de problèmes et donc rarement empruntés. De tel liens sont par conséquent très difficiles à capturer en utilisant `traceroute` ;
- l'évolution du réseau est permanente et plus rapide que le temps qu'il faut pour faire une mesure (plusieurs jours). On obtient donc des clichés de petits morceaux du réseau pris à divers instants, que l'on tente ensuite de recoller pour obtenir une seule image.

Avec ces remarques, on se rend compte qu'il n'est pas possible d'obtenir une vue fiable du réseau, car toute carte est non seulement partielle mais certainement biaisée par le protocole de mesure. Plusieurs études récentes [31, 62, 67, 76, 112, 118] ont effectivement montré que cette façon de faire peut influencer les visions qu'on en a. Le dernier chapitre de cette thèse sera entièrement consacré à cette problématique.

2. On rencontre, par exemple, une grande quantité d'adresses privées de type $192.xxx.xxx.xxx$ ou $10.xxx.xxx.xxx$.

Nous utiliserons dans la suite plusieurs mesures du réseau de l'Internet au niveau des routeurs [29, 56], et principalement une mesure de Mercator de 75 000 sommets [55]. Ces mesures sont toutes partielles et biaisées, mais elles sont aujourd'hui notre seul point d'entrée pour étudier l'Internet et sont largement utilisées par la communauté scientifique.

1.1.2 Au niveau des systèmes autonomes

Le niveau physique décrit précédemment ne prend pas du tout en compte le fait que les routeurs appartiennent à des entreprises, institutions ou autres. Ces groupes sont appelés *systèmes autonomes* (*autonomous systems* ou *AS* en anglais) et gèrent leurs routeurs, en particulier les mécanismes de routage entre ces routeurs, comme ils le souhaitent. Chaque système autonome possèdent un numéro unique permettant de l'identifier.

Parmi les routeurs d'un système autonome, certains sont internes au système et sont chargés de router les messages d'un point à un autre de celui-ci, alors que d'autres, appelés *routeurs de bordure*, sont aux portes du système autonome : ils sont directement reliés à des routeurs (de bordure) d'autres systèmes autonomes.

S'il est très difficile d'obtenir une vue satisfaisante des liens entre routeurs, les liens entre systèmes autonomes sont eux plus faciles à obtenir. Les tables de routage des routeurs de bordure utilisent en effet un protocole de routage inter systèmes autonomes spécifique qu'il est possible d'interroger : BGP, pour *Border Gateway Protocol*. Les tables de routage BGP permettent de savoir à qui transférer un message, étant donnée l'adresse du destinataire. Elles permettent aussi de connaître les liens entre systèmes autonomes.

L'exemple de la Figure 1.3 illustre le contenu classique d'une telle table. À partir de l'adresse de la destination, le routeur détermine dans quel réseau (champ *Network*) elle se trouve, puis vers quel réseau l'envoyer (champ *Next Hop*) en fonction du coût associé (champ *Metric*). L'estimation du coût est basée sur de nombreux paramètres : temps de transfert du message, surcharge du réseau, politique commerciale avec les systèmes autonomes voisins, politique entre les nations (notamment pour éviter de traverser certaines zones sensibles), etc.

Les tables de routage sont mises à jour en fonction des problèmes dans le réseau et échangées entre les routeurs de bordure : le champ *Path* contient la liste des systèmes autonomes qu'une mise à jour a traversé avant d'arriver au routeur. Si un routeur reçoit une nouvelle mise à jour, il vérifie que son numéro de système autonome n'est pas dans la liste, met sa table à jour, se rajoute dans la liste et transmet l'information.

Avec le contenu des tables BGP, on arrive à avoir des informations très précises sur les connexions entre systèmes autonomes. Une table BGP comme celle de l'exemple peut contenir plusieurs millions d'entrées (7 105 749 dans ce cas), vers de très nombreux réseaux (168 045). Si l'augmentation de la taille des tables de routage pose de nombreux problèmes d'un point de vue réseau, elle a de notre point de vue l'avantage d'apporter beaucoup d'information sur les liaisons entre systèmes autonomes.

```
   Network    Next Hop            Metric LocPrf Weight Path
*  3.0.0.0    207.126.96.1            20             0 6461 701 80 i
*             204.42.253.253                         0 267 2914 701 80 i
*             216.140.8.63             3             0 6395 701 80 i
*             216.140.14.127        1596             0 6395 701 80 i
*             4.0.0.2               2115             0 1 701 80 i
*>            12.127.0.249                           0 7018 701 80 i
*             204.70.4.89                            0 3561 701 80 i
*  4.0.0.0    207.126.96.1             0             0 6461 1 1 1 1 i
*             203.62.248.4                           0 1221 16779 1 i
*             203.62.252.21                          0 1221 16779 1 i
```

FIG. 1.3 – *Le contenu typique d'une table BGP d'un routeur Cisco tiré de* `www.routeviews.org`.

1.1.3 Le World Wide Web

Le Web[3] est un ensemble distribué de documents qui a vu le jour en 1989 dans le but de pouvoir partager facilement de l'information. Tout le monde peut créer des documents et les mettre en ligne simplement. Ces documents sont alors accessibles *via* l'Internet. Chaque document a un nom sur le Web, son URL (pour *Uniform Ressource Locator*) comme, par exemple, `http://www.liafa.jussieu.fr`, et contient non seulement de l'information, mais aussi des liens (liens hypertexte, ou hyperliens) vers d'autres documents.

Cette notion de pages Web (ou plutôt d'URL) reliées par des liens hypertextes définit naturellement une structure de graphe [25, 74]. Le contenu des pages est bien évidemment très important, mais la structure elle-même apporte beaucoup d'information : le fait de créer un lien hypertexte de sa page vers une autre, par exemple, n'est pas anodin et confère à la page référencée un certain crédit.

La taille du graphe du Web était estimée à 4 milliards de pages il y a quelques années. Google recense actuellement $4,3$ milliards de pages Web, mais ceci ne représente qu'une petite partie du Web : la plus grande partie, *le Web caché*, n'est accessible qu'en remplissant des formulaires. Sa taille est donc très difficile à estimer.

De nombreuses études ont été menées sur le graphe du Web, la plus célèbre [25] étant celle dite du *nœud-papillon*. Cette étude porte sur la structure du graphe du Web et cherche à savoir quelles pages on peut atteindre en cliquant sur des liens hypertexte (voir Figure 1.4).

D'après cette étude, le Web est composé de quatre parties distinctes :

- un *Cœur* dans lequel on peut accéder à toutes les pages le composant à partir de n'importe laquelle uniquement en suivant les liens hypertexte ;
- une zone *In* que l'on ne peut pas atteindre depuis le *Cœur*, mais à partir de laquelle le *Cœur* est accessible ;

3. Aussi appelé WWW, le World Wide Web, W3, la Toile et, moins couramment : Cybertoile, Hypertoile, Infocentre.

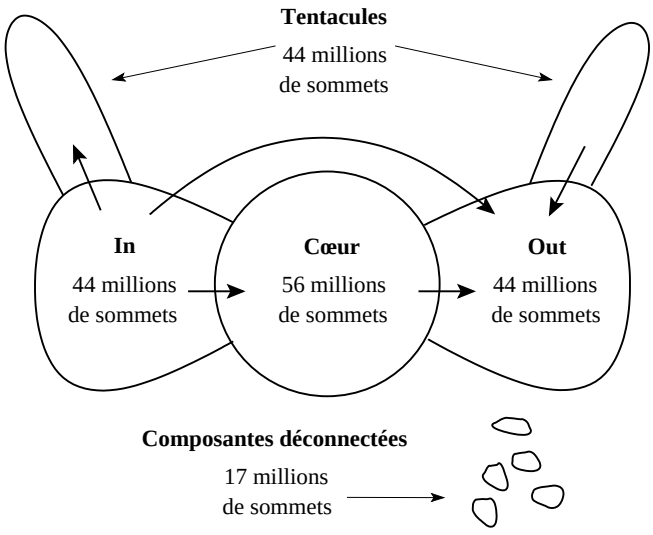

FIG. 1.4 – *Le modèle du nœud papillon.*

- une zone *Out*, symétrique de *In*, qui est accessible depuis le *Cœur* mais depuis laquelle on ne peut pas l'atteindre ;
- enfin, toutes les autres pages sont isolées du *Cœur* et appartiennent soit à des *Tentacules*, soit à des zones complètement déconnectées.

Tout graphe orienté peut, bien sûr, être décomposé en quatre parties de ce type. L'information pertinente réside dans le fait que, pour le Web, ces zones ont des tailles similaires. Aucune explication n'a été apportée à ce jour concernant cette caractéristique. Les pages de *In* sont souvent des pages personnelles mal référencées et difficiles à trouver. Les pages de *Out*, quant à elle, appartiennent souvent à des entreprises qui mettent rarement des liens vers l'extérieur (*capture* du visiteur).

De nombreuses applications prennent en compte la structure du graphe du Web, la plus importante étant sans doute les moteurs de recherche. Le moteur de recherche Google a en effet été le premier à utiliser non seulement le contenu des pages, mais aussi la structure du graphe pour classer les pages par ordre d'importance. Cette technique, connue sous le nom de PageRank [24], repose sur l'intuition selon laquelle un lien hypertexte confère de l'importance à la page Web qui le reçoit. Les pages reçoivent et redistribuent de l'autorité et celles qui en ont le plus sont les "meilleures". Une page qui a beaucoup d'autorité va en redistribuer plus aux pages vers lesquelles elle pointe. Ainsi, une page recevant peu de liens, mais ceux-ci venant de pages importantes sera plus importante qu'une page recevant beaucoup de liens de pages inconnues. Les pages les plus importantes apparaissent en premier lors des réponses aux requêtes entrées. C'est cette méthode qui a donné à Google le succès qu'on lui connaît aujourd'hui.

D'autres applications utilisent cette structure comme, par exemple, la recherche de communautés ou la détection de phénomènes émergeants **??**. Ces études reposent sur la recherche de structures particulières dans le graphe du Web (une communauté est, par

exemple, un ensemble de pages très liées les unes aux autres) ou sur l'évolution du graphe lui-même (certaines zones changent plus vite que d'autres).

Dans ce contexte complètement distribué, le problème de l'acquisition du graphe se pose aussi. La méthode générale utilisée pour explorer le graphe est la suivante : on commence avec une première page dans laquelle on cherche tous les liens hypertexte. Chacun de ces liens pointe sur une page. On recommence alors l'opération à partir de chacune de ces pages en allant chercher tous les liens qu'elles contiennent, et ainsi de suite. Le processus s'arrête quand on ne trouve plus de nouvelles pages. Ce processus automatisé est effectué par un logiciel appelé couramment *robot* ou *crawler*.

Avec cette méthode, même si l'on pouvait visiter mille pages par seconde (dans la réalité, il est difficile d'en visiter plus de 100), il faudrait plus d'un mois pour connaître toutes les pages accessibles depuis la page initiale, et ce, quelle que soit la puissance des machines utilisées. Pendant ce temps, plusieurs centaines de millions de pages auront vu le jour, et de nombreuses autres auront été supprimées ou déplacées.

Il y a un grand nombre d'autres problèmes qui rendent la récupération du graphe du Web impossible en pratique. Il faut notamment éviter de trop solliciter un serveur donné (pas plus de quelques pages par minute) dont l'objectif est avant tout de servir les utilisateurs du site. Les très gros serveurs sont donc rarement explorés entièrement (cela dépend des crawlers). D'autre part, si une page n'est pointée par aucune autre, elle ne sera jamais découverte avec cette méthode et, de façon plus générale, si une page n'est pas atteignable à partir de la page initiale, on ne la trouvera jamais[4]. De plus, de nombreuses autres pages ne sont accessibles qu'en remplissant des formulaires sur des sites Web (encyclopédies, par exemple). Les pages obtenues de la sorte sont difficilement accessibles de manière automatique, même si certains crawlers sont maintenant capables de remplir des formulaires.

Ces problèmes montrent que la définition même du "graphe du Web" n'est pas claire : on peut considérer toutes les URLs valides, toutes les pages accessibles par un crawl, inclure ou non les pages dynamiques, ou se restreindre aux pages intéressantes selon un critère à définir. À cet égard, la page http://www.liafa.jussieu.fr/~latapy/alltheweb/arbre.php montre bien les problèmes posés par ce type de définition. En effet, sur cette page, l'utilisateur peut créer une URL caractère par caractère et, à chaque étape, il peut cliquer sur l'URL en cours de construction. Depuis cette page, on peut donc atteindre n'importe quelle URL en la construisant petit à petit. Il convient donc de préciser la définition retenue lorsque l'on parle du Web. Dans l'étude du nœud-papillon présentée ci-dessus, par exemple, les auteurs considéraient l'ensemble des pages connues à ce moment-là par Altavista.

Dans la suite, nous utiliserons un échantillon du graphe du Web correspondant au site complet de l'Université de Notre-Dame (Indiana) qui est disponible en ligne [8, 38] et qui, de part sa (relativement) petite taille (un peu plus de 300 000 sommets), nous permet d'effectuer des calculs. C'est bien évidemment une toute petite portion du graphe.

4. Certains diront que les pages non accessibles ne sont pas intéressantes et ne méritent donc pas d'être découvertes.

1.2 Réseaux sociaux

Le terme *réseaux sociaux* regroupe les réseaux constitués d'individus reliés entre eux par tous types de relations : relations de connaissance proche, de travail, sexuelles, etc.[5] De nombreux réseaux sociaux ont été étudiés mais, jusqu'à récemment, leur étude reposait surtout sur des enquêtes avec des formulaires à remplir ou des entretiens avec les personnes. Ce type d'approches, lourdes à mettre en œuvre et à analyser, à naturellement limité les progès dans ce domaine, notamment en les limitant à des études locales des réseaux sociaux.

Une étude à plus grande échelle du réseau des connaissances proches a toutefois été menée en 1967 par Stanley Milgram [75, 90, 91]. Dans cette expérience, une personne choisie au hasard dans le Nebraska devait faire parvenir une lettre à un agent de change de Boston bien précis qu'elle ne connaissait pas. Pour faire transiter la lettre, cette personne devait la donner à une de ses connaissances qui, à son tour, la donnait à une de ses connaissances, jusqu'à arriver à destination. Cette expérience a été effectuée avec un relativement grand nombre de personnes (environ 300).

Parmi toutes les lettres, environ un quart sont arrivées à destination et ce avec 5 intermédiaires en moyenne. Diverses critiques ont été formulées concernant l'expérimentation elle-même, notamment sur le choix des personnes sources et sur le nombre de lettres comptabilisées [75]. Malgré tout, le résultat est significatif : il est étonnant qu'autant de lettres soient arrivées à destination, surtout avec seulement 5 intermédiaires. De plus, il est probable que les lettres auraient pu arriver plus vite à destination (les chemins suivis n'étant pas forcément optimaux).

Cette étude a été suivie par de nombreuses autres avec divers outils (téléphone, courriers électroniques, etc.) et, récemment, des travaux à la frontière entre informatique et sociologie ont vu le jour [17, 73, 134] pour tenter de comprendre l'existence de ces chemins courts.

Alors que cette étude avait été faite manuellement, l'avènement de l'informatique et de l'Internet ouvre de nouvelles voies pour l'étude des réseaux sociaux, en offrant une quantité phénoménale de données riches et disponibles instantanément. Parmi ces réseaux, certains reposent sur des applications de l'Internet, comme les réseaux d'échanges de courriers électroniques reliant les personnes qui s'en envoient, alors que d'autres réseaux sont simplement disponibles dans des bases de données, comme le graphe des acteurs dans lequel les acteurs ayant joué dans un même film sont reliés.

Nous allons par la suite présenter les réseaux issus d'applications liées à l'Internet, puis un certain nombre de réseaux de collaboration (qui travaille avec qui) ayant des caractéristiques particulières que nous utiliserons dans la suite.

1.2.1 Réseaux sociaux de l'Internet

Parmi les nombreuses applications reposant sur l'Internet et le Web, beaucoup impliquent des interactions entre utilisateurs. Celles-ci peuvent, par exemple, être des dia-

5. Il existe d'autres réseaux sociaux dont les sommets sont des institutions, des entreprises et autres, mais nous n'en rencontrerons pas dans la suite.

logues en direct (chat, jeux en ligne), en différé (courriers électroniques, forums de discussions, blogs) ou encore des échanges de données (pair-à-pair, ftp), etc.

Parmi tout ces réseaux, très peu ont été étudiés en profondeur et cela pour plusieurs raisons. Tout d'abord, les données ne sont pas toujours faciles à obtenir en grande quantité et, d'autre part, certains réseaux commencent tout juste à faire leur apparition (le concept de blog est, par exemple, très récent). Mais de plus en plus de chercheurs prennent la mesure du potentiel contenu dans ces données riches, numériques et en-ligne, ce qui entraîne actuellement une véritable explosion du nombre d'études sur ces objets. Nous allons détailler ci-dessous deux exemples : les échanges de fichiers dans un système pair-à-pair et les échanges de courriers électroniques.

Échanges de fichiers dans un système pair-à-pair

Un système pair-à-pair (P2P) est composé d'un grand nombre d'utilisateurs qui veulent rendre des données (musique, films, logiciels, etc.) disponibles pour d'autres utilisateurs. Tous ces participants jouent *a priori* le même rôle : ce sont des pairs. L'objectif du système est alors de permettre à un utilisateur qui désire une donnée mise à disposition par un autre utilisateur de la récupérer, généralement en mettant en contact les deux utilisateurs. Un système P2P doit être aussi distribué que possible, sans autorité centrale, pour être à la fois robuste aux pannes pour pouvoir gérer de très grands nombres de pairs et de données, et pour préserver l'anonymat des utilisateurs. En pratique, toutefois, les systèmes P2P sont parfois semi-centralisés : certains utilisateurs ont alors un rôle prépondérant.

Dans le cas décentralisé, les utilisateurs sont connectés entre eux pour maintenir un réseau virtuel (*overlay network*). Dans ce réseau virtuel, chaque utilisateur a un certain nombre de contacts qu'il peut interroger pour trouver un fichier. Si ses contacts ne possèdent pas le fichier, ils vont relayer la question à leurs contacts et, de proche en proche, le fichier sera trouvé s'il existe. Ce réseau est dynamique : quand un individu veut rejoindre le réseau, il doit trouver des contacts, qu'il perdra en se déconnectant.

De nombreuses études visent à définir une bonne topologie pour ce type de réseau, la qualité étant généralement mesurée par le temps nécessaire pour trouver une donnée dans le réseau [52, 86, 115, 127]. Ces topologies cherchent notamment à s'assurer que les individus les plus éloignés dans le réseau ne le soient pas trop et que les pairs n'aient pas besoin d'avoir trop de contacts.

Obtenir une carte de ce réseau virtuel est, par contre, assez complexe et très dépendant du protocole utilisé. Avec Gnutella (v0.4), pour entrer dans le réseau, il faut trouver un client déjà présent. Une procédure autorise cela et permet en plus de connaître tous les contacts de ce client. En simulant une procédure d'entrée vers un client, puis vers ses contacts, les contacts de ses contacts, et ainsi de suite, il est en théorie possible de découvrir tous les clients de proche en proche [119]. Cette procédure est très similaire à un *crawl* du Web et a reçu le même nom. Dans [120], un crawl typique de Gnutella dure 2 minutes durant lesquelles on estime que 25% à 50% des utilisateurs sont recensés.

Malgré tout, ce type de *crawl* ne peut pas se faire sur tous les systèmes P2P et génère beaucoup de trafic. Les problèmes liés à la dynamique du réseau (connexions et décon-

nexions très fréquentes) influent aussi sur la mesure : une mesure trop courte donne une vision incomplète, une mesure trop longue est faussée par l'évolution du réseau.

Le réseau virtuel défini précédemment n'est pas vraiment social, les liens entre les pairs étant gérés par le protocole. Un deuxième réseau toutefois peut être défini en reliant deux individus s'ils échangent des données via le système. Ce réseau apporte beaucoup d'informations sur les usages des individus et le type de données qu'ils fournissent. Il peut aussi permettre d'identifier des communautés d'utilisateurs.

Comme précédemment, il est assez difficile d'obtenir des informations sur ce type de réseau en général. Mais, dans certains contextes tels que les systèmes P2P semi-centralisés, cela devient possible. Un système semi-centralisé comporte plusieurs serveurs utilisés pour référencer tous les fichiers partagés par les pairs qui s'y connectent. La recherche de fichiers est ensuite simplifiée puisqu'il suffit de demander un fournisseur au serveur.

Ainsi, en ayant accès à un serveur, on peut savoir qui partage quels fichiers et surtout qui les recherche. Si le serveur enregistre les requêtes des clients, il sait donc quels pairs vont entrer en contact les uns avec les autres [57]. Ceci donne par conséquent la vision complète des échanges entre les utilisateurs qui sont connectés sur le serveur. Le Chapitre 6 sera entièrement consacré à l'étude d'un tel réseau d'un point de vue statique et dynamique.

Échanges de courriers électroniques

Toujours dans le cadre des échanges sur l'Internet, il est possible d'étudier le réseau des échanges de courriers électroniques. Dans ce réseau, deux individus sont reliés s'ils correspondent par courrier électronique. Bien entendu, il n'est pas question d'étudier le réseau global des courriers électroniques mais, si l'on se fixe un cadre restreint (entreprise, laboratoire, etc.), cela devient possible.

Il existe plusieurs méthodes pour récupérer les échanges de courriers électroniques selon ce que l'on souhaite en faire. Dans [130], le réseau des échanges de courriers électroniques au sein d'une entreprise est étudié en récupérant tous les courriers électroniques qui ont circulé pendant presque 3 mois (environ 200 000) entre les 485 personnes de l'entreprise. L'objectif était de savoir s'il est possible de découvrir automatiquement des communautés et les personnes à responsabilité uniquement à partir des échanges. Les résultats ont été plutôt positifs dans ces deux domaines.

Avec le même protocole, mais à plus grande échelle (4 mois, 60 000 individus), le réseau d'échange d'une université a aussi été étudié [45], cette fois pour déterminer les différences de comportements entre utilisateurs. Cette étude a montré que la majorité des individus envoient très peu de mails alors qu'un petit nombre en envoient une énorme quantité.

Enfin, dans [137], une expérience a permis d'observer comment des informations se propagent par courrier électronique. Un programme analysait le contenu des messages, notamment les transferts de courriers électroniques. Les conclusions montrent que l'information se diffuse bien à l'intérieur des groupes d'utilisateurs, mais très peu entre les groupes.

1.2.2 Réseaux collaboratifs

Parmi les autres réseaux sociaux, on distingue en particulier les réseaux collaboratifs, reliant les individus qui travaillent ensemble sur un projet. La particularité de ces réseaux est que les individus ne travaillent en général pas par deux, mais en groupes de taille variable. Nous allons détailler deux cas particuliers, celui des acteurs qui collaborent en jouant dans des films et celui des scientifiques qui co-signent des articles.

Graphe des acteurs

Le graphe des acteurs représente les relations entre acteurs ayant joué dans un même film. La majorité des films ayant plus de 2 acteurs, la structure de ce réseau est donc très particulière avec, pour chaque film, de grands ensembles d'acteurs complètement connectés.

Ce réseau a été beaucoup étudié pour plusieurs raisons : il est très facilement récupérable grâce à l'*Internet Movie Database* [39, 134], il est très grand (plusieurs centaines de milliers d'acteurs et de films), et il évolue constamment avec le tournage de nouveaux films qui sont ajoutés très régulièrement dans la base de données. Ce graphe a aussi été étudié de manière ludique à travers le jeu de Kevin Bacon[6]. Le but de ce jeu est, étant donné un acteur quelconque, de trouver une suite d'acteurs le reliant à Kevin Bacon. Ceci doit se faire dans le graphe des acteurs, donc film par film[7].

Il faut noter que, contrairement à de nombreux autres réseaux sociaux, celui-ci peut être récupéré entièrement, bien que souffrant de quelques petits problèmes pouvant affecter sa qualité. Tout d'abord, il manque certains anciens films, ceux-ci étant rentrés dans la base petit à petit. D'autre part, certains acteurs n'apparaissent pas dans le casting des films et, enfin, certains acteurs sont dupliqués à cause d'erreurs de transcription dans leur nom, ou fusionnés par homonymie. Mais comparativement, la qualité de ces données est excellente.

Graphes de co-signature

De manière similaire au graphe des acteurs, les graphes de co-signature sont définis en reliant ensemble deux personnes qui ont co-signé un article [97, 98, 106]. Ce graphe est étudié depuis longtemps, notamment à travers le nombre d'Erdös qui définit, pour chaque scientifique, à quelle distance il se trouve de Paul Erdös. Les personnes ayant co-signé avec Erdös on un nombre de 1, ceux qui ont co-signé avec un co-signataire d'Erdös ont un nombre de 2, etc.

Plusieurs bases de données d'articles sont disponibles de la sorte pour diverses revues en ligne (Arxiv, Medline, Spires, Ncstrl, etc.) desquelles il est possible d'extraire les co-signatures. Dans la suite, nous utiliserons le réseau de co-signature d'Arxiv [13]. Ces réseaux de collaboration scientifique souffrent des mêmes problèmes que le graphe des acteurs et sont de plus largement incomplets.

6. http://www.cs.virginia.edu/oracle/

7. Par exemple, Richard Anconina a joué avec Harvey Keitel dans *Une pierre dans la bouche*, Harvey Keitel a joué dans *The Two Jakes* avec Eli Wallach, qui a joué dans *Mystic River* avec Kevin Bacon.

1.2.3 Autres réseaux sociaux

D'autres réseaux sociaux ont été étudiés dans le cadre des grands réseaux d'interactions. Le réseau des appels téléphoniques a été étudié en 1999 sur le réseau téléphonique d'AT&T [2, 117]. À l'époque, il y avait environ 250 millions d'appels par jour. À partir de ces données, un réseau a été construit sur 12 heures comprenant 53 millions de numéros de téléphones distincts. Parmi tous les utilisateurs recensés, 21 millions n'ont pas été appelés et 22 millions n'ont pas appelé. Les 10 millions restant ont à la fois passé et reçu des appels.

L'objectif de cette étude était de présenter un algorithme de recherche de groupes de personnes complètement connectés (*cliques*) sur de très grands réseaux. En l'occurrence, un groupe de 30 personnes a été identifié dans lequel tout le monde a téléphoné à tout le monde pendant ces 12 heures.

L'étude de ce genre de réseaux est très intéressante en raison de la quantité, la qualité et le caractère dynamique des données. En effet, pour chaque appel téléphonique, il est possible de connaître les deux numéros de téléphone impliqués ainsi que l'heure et la durée de l'appel. Il n'y a *a priori* aucune erreur de mesure dans ces données.

Citons comme dernier exemple le réseau des contacts sexuels. Dans [79, 80], ce réseau est étudié grâce à un questionnaire rempli par 2 810 personnes en 1996. Outre la variété de comportement sexuels, ces études montrent que des stratégies de prévention ciblées seraient efficaces pour stopper la propagation de maladies sexuellement transmissibles, en visant prioritairement les personnes ayant beaucoup de partenaires.

De nombreuses autres études ont été menées sur d'autres réseaux sociaux, notamment en sociologie ou en économie et ce, en ayant recours à d'autres outils que ceux utilisés pour les grands réseaux d'interactions.

1.3 Réseaux biologiques

De nombreux réseaux sont également étudiés en sciences du vivant : réseaux d'interactions entre protéines ou entre gènes, réseaux trophiques (qui mange qui) et bien d'autres. Nous allons les présenter succinctement car ils ne sont, dans cet ouvrage, que des exemples de grands réseaux d'interactions dont nous avons une connaissance limitée. Nous renvoyons donc aux références pour plus de détails.

1.3.1 Réseaux neuronaux

Diverses maladies neurologiques sont causées par la mort de cellules nerveuses. Pour la maladie de Parkinson, par exemple, la mort de cellules produisant de la dopamine entraîne une perte de contrôle des mouvements du corps. L'étude des réseaux neuronaux pourrait peut-être, à long terme, permettre de mieux comprendre les raisons qui font que ces cellules meurent ou faire en sorte que cela n'ait pas d'influence en faisant que les substances nécessaires soient produites par d'autres méthodes.

Il n'est pas pour l'instant question de mesurer le réseau neuronal d'un être humain, mais, pour certains animaux de très petite taille, les liaisons entre neurones sont parfaitement identifiées. En particulier, *Caenorhabditis elegans*, un petit ver, possède 282 neurones [134], chacun étant relié à 14 autres neurones en moyenne. Cela implique que les neurones sont tous très proches les uns des autres ; les informations peuvent donc se propager très vite. Il semble aussi que les neurones ne soient pas connectés de manière aléatoire.

1.3.2 Réseaux de réactions métaboliques

De nombreuses réactions métaboliques se produisent dans les organismes vivants afin de produire de l'énergie. Ces réactions impliquent divers constituants : protéines, ADN, ARN, et autres molécules. Ils sont généralement trop nombreux pour que le réseau puisse être étudié dans son ensemble, mis à part pour quelques organismes très simples où le nombre de constituants reste raisonnable (de l'ordre du millier).

Dans [70], les réseaux métaboliques de 43 organismes différents ont été étudiés, 25 d'entre eux étant complètement connus. Tous ces réseaux, pourtant de tailles variables, sont très similaires par de nombreux points. En particulier, les constituants qui apparaissent le plus fréquemment dans les réactions sont les mêmes pour toutes les espèces, alors que les constituants plus rares sont généralement spécifiques à une certaine espèce.

1.3.3 Interactions entre protéines

On peut aussi s'intéresser plus spécifiquement aux interactions entre protéines uniquement. Si les tâches spécifiques de chaque protéine sont maintenant relativement bien connues, la manière dont ces protéines interagissent de manière globale est encore peu connue. De nombreuses études sont menées sur le réseau dans lequel deux protéines sont reliées si elles réagissent l'une avec l'autre [70].

Parmi les réseaux disponibles, l'un des plus gros est celui de la levure qui contient 1 870 protéines distinctes reliées entre elles par 2 240 interactions. Les études sur ce réseau, et quelques autres, ont montré que certaines protéines ont de très nombreuses interactions[8] alors que la plupart en ont peu et, d'autre part, que ces quelques protéines très connectées n'interagissent pas entre elles. Cette structure très hétérogène assure à l'organisme une forte résistance aux mutations.

1.3.4 Réseaux trophiques

Les réseaux trophiques sont les réseaux de prédation entre espèces, dans lesquels deux espèces sont reliées si l'une se nourrit de l'autre. Ces réseaux sont souvent étudiés dans une zone restreinte en déterminant toutes les espèces présentes ainsi que les relations existant entre elles. De par l'origine diverse des réseaux étudiés, on dispose aujourd'hui de plusieurs

8. Chez l'homme, l'une des protéines les plus connectées à un rôle majeur pour empêcher le développement de tumeurs.

réseaux de type (terrestre, aquatique) et de taille variables. En pratique, ces réseaux sont souvent petits (une centaine d'espèces), l'obtention des données étant déjà très complexe sur ces petits réseaux.

Diverses études [44, 95, 136] ont été menées sur ces réseaux afin de déterminer leur structure ou la façon dont ils se créent. La robustesse de ces réseaux est aussi étudiée, afin de déterminer l'impact de la disparition d'une espèce sur l'écosystème.

1.4 Autres graphes étudiés

De très nombreux autres grands réseaux d'interactions ont été étudiés. Sans vouloir être exhaustifs, citons, par exemple, les réseaux suivants :

Cooccurrence de mots : on relie les mots s'ils apparaissent dans un même contexte. Ainsi, il est possible d'étudier les cooccurrences dans les requêtes faites à un moteur de recherche, les mots des phrases d'un livre, et bien d'autres. La cooccurrence de mots dans les phrases d'un livre a déjà été étudiée [49]. Il y a plusieurs façons de définir ce graphe de cooccurrence, soit en reliant les mots qui sont proches dans le texte (à un ou deux mots d'écart), soit en reliant tous les mots d'une même phrase. Avec le premier choix il est possible de perdre beaucoup de corrélations entre des mots couramment éloignés dans un texte ; le second choix, au contraire, peut entraîner des liaisons entre des mots très éloignés dans une phrase et donc créer des relations non pertinentes.

Nous avons choisi l'approche qui consiste à relier tous les mots d'une phrase, sans supprimer les mots de liaison comme cela est parfois fait, sur une version de la Bible disponible en ligne [133]. Ce graphe a, par sa construction même, une structure similaire à celle des graphes de collaboration présentés plus tôt : tous les mots d'une phrase sont reliés entre eux, comme les acteurs d'un même film.

Graphe du dictionnaire : on relie deux mots si l'un apparaît dans la définition de l'autre. Il est ensuite possible d'étudier les voisinages des mots et d'identifier automatiquement les mots qui ont un voisinage proche. Dans [20], une méthode est proposée pour extraire automatiquement des synonymes en se basant uniquement sur cette remarque. Bien que cette méthode soit moins performante que les bases de données de synonymes faites manuellement, elle propose néanmoins beaucoup plus de synonymes et apporte des informations pertinentes. Elle est utilisée comme outil d'aide au classement manuel.

Citations dans les articles scientifiques : de manière un peu similaire aux graphes de co-signature d'articles scientifiques, on peut relier deux articles si l'un des deux est cité par l'autre [116]. Chaque nouvel article contient un certain nombre de références vers des articles antérieurs et le choix des références est libre pour chacun. Il s'avère, malgré tout, que la majorité des articles sont peu ou pas cités, alors que certains le sont énormément.

Ce graphe a plusieurs particularités du fait de sa construction : on ne peut citer que des articles préexistants, sans possibilité de modifier *a posteriori* les liens créés. Ainsi, sauf cas particulier, deux articles ne peuvent pas se citer mutuellement et, de façon plus générale, il ne peut pas y avoir de cycles.

Réseaux de transports : divers réseaux de transports (routiers, aériens, d'électricité, etc.) ont aussi été étudiés [134]. Ces graphes ont des structures très variables dues à certaines contraintes : si un aéroport peut avoir autant de liaisons vers d'autres aéroports qu'il le souhaite, il n'en est pas de même pour un réseau autoroutier qui ne relie en général que des villes proches. Ces réseaux ont donc des structures assez diverses.

Conclusion

Les réseaux que nous avons présentés proviennent de contextes très variés. Les réseaux biologiques ont évolué depuis des millions d'années vers leur état actuel, de même que les réseaux linguistiques qui codent l'évolution du langage sur une plus courte durée. Les réseaux technologiques sont des créations plus ou moins concertées, parfois avec un objectif global (faire transiter de l'information efficacement sur l'Internet) et parfois non. Les réseaux sociaux enfin sont constitués d'individus ayant chacun un comportement spécifique.

L'étude de la plupart de ces réseaux est motivée par des applications pratiques : routage efficace sur l'Internet, définition de moteurs de recherche performants pour le Web, travaux en biologie ou en linguistique, etc. Leur étude en tant que graphe a déjà donné des résultats probants qui confirment l'intérêt de cette approche.

Plus de références sur les réseaux décrits sont disponibles dans [7, 42, 102]. Tout au long de cette thèse, nous utiliserons quelques grands réseaux d'interactions parmi ceux que nous avons présentés. Il s'agit d'un réseau d'interactions protéiques [70], d'une carte de l'Internet au niveau des routeurs [55], du graphe du Web de l'Université de Notre-Dame [8, 38], du graphe des Acteurs [39, 134], de la cooccurrence de mots dans la Bible [133, 49] et des relations de co-signature d'articles sur Arxiv [13]. Nous nommerons en général ces réseaux *Protéines*, *Internet*, *Web*, *Acteurs*, *Cooccurrence*, et *Co-signature* respectivement. Nous avons choisi cet ensemble car il est représentatif de l'ensemble des grands réseaux d'interactions étudiés dans la littérature : il contient des graphes de tailles diverses et d'origines variées.

Le point le plus singulier est qu'on s'attendait *a priori* à ce que ces réseaux exhibent des propriétés bien distinctes, liées à leur origine, leur processus de création ou l'objectif du réseau. Or, il s'avère qu'ils ont tous des propriétés non triviales en commun que nous détaillerons dans le Chapitre 2. Ces propriétés n'étaient pas attendues : ces réseaux "réels" forment donc un objet d'étude transdisciplinaire, qui s'est développé récemment sous la forme de l'étude des grands réseaux d'interactions en général.

Chapitre 2

Paramètres étudiés

De très nombreux paramètres ont été introduits pour l'étude des grands réseaux d'interactions, dans l'objectif d'en obtenir une description assez fine. Dans la suite de ce chapitre, nous allons détailler les paramètres élémentaires les plus utilisés, puis présenter des propriétés de graphes. À chaque fois, nous donnerons et discuterons les valeurs de ces paramètres pour les six exemples de grands réseaux d'interactions que nous utiliserons dans cette thèse.

On verra que certaines propriétés étaient attendues au sens où elles sont partagées par tous les réseaux, qu'ils soient réels ou aléatoires. D'autres, au contraire, sont plus surprenantes. Une caractéristique des grands réseaux d'interactions est qu'ils partagent des propriétés qui les différencient fortement des autres réseaux. D'autres propriétés, plus subtiles, les différencient entre eux.

2.1 Définitions de base

Un graphe G est un couple (V, E), où V est un ensemble (dont les éléments sont appelés sommets) et E est une partie de $V \times V$. On peut distinguer les graphes orientés pour lesquels les éléments de E sont des couples (ordonnés) de sommets et les graphes non orientés pour lesquels les éléments de E sont des paires (non ordonnées) de sommets. Les éléments de E sont appelés arcs dans le cas orienté et arêtes dans le cas non orienté. Dans la suite, on aura tendance à appeler les deux *liens*, le contexte permettant de faire la distinction.

Un graphe valué est un triplet (V, E, ω) où $\omega : E \to \mathbb{R}$ est une fonction associant à chaque lien un poids réel ou entier, selon les cas, et généralement positif. Dans la suite, la plupart des graphes seront non valués.

Le sous-graphe $G' = (V', E')$ de G est le graphe dont l'ensemble des sommets V' est un sous-ensemble de V et dont les liens sont ceux qui existent entre les sommets de V' dans G. Un sous-graphe dans lequel tous les sommets sont reliés deux à deux, $E' = V' \times V'$ est appelé une *clique*.

Dans un graphe orienté (resp. non orienté), un *chemin* (resp. une *chaîne*) entre deux sommets u et v est une suite de sommets reliés deux à deux par des arcs (resp. arêtes). On utilisera essentiellement le terme *chemin* dans la suite, même pour des graphes non

orientés. Si $u = v$, on parlera alors de *cycle*. La longueur d'un chemin est le nombre d'arcs qu'il contient. Un cycle de longueur 1 est un lien de u vers u et est appelé une *boucle*. Parmi tous les chemins reliant deux sommets, on peut distinguer les plus courts chemins qui sont ceux de longueur minimale si le graphe est non valué. La distance entre deux sommets est la longueur d'un plus court chemin les reliant.

Le *degré* d'un sommet u, noté $d(u)$, est le nombre d'arêtes qui y sont reliées. Dans le cas orienté, on parlera de degré entrant, noté $d^+(u)$, pour les arcs allant vers un sommet et de degré sortant pour les arcs qui en partent, et on le notera $d^-(u)$. Pour un sommet u, on note $N(u)$ l'ensemble des voisins de u.

Dans toute la suite, pour un graphe $G = (V, E)$, on notera $n = |V|$ et $m = |E|$.

2.2 Densité et degré moyen

La densité, notée δ, est définie comme le rapport du nombre de liens dans le graphe sur le nombre de liens maximal qu'il pourrait contenir :

$$\delta(G) = \frac{m}{\binom{n}{2}} \sim \frac{2 \cdot m}{n^2},$$

Cette approximation est valable quand $n >> 1$.

Le degré moyen du graphe G, noté k, vaut :

$$k = \frac{1}{n} \sum_{u \in V} d(u) = \frac{2 \cdot m}{n}$$

La densité du graphe peut alors être réécrite en fonction du degré moyen :

$$\delta(G) = k/(n-1) \sim k/n.$$

La plupart des grands réseaux d'interactions ont un nombre de liens du même ordre que leur nombre de sommets : la densité est donc peu élevée (de l'ordre de $1/n$), comme on peut le voir dans le Tableau 2.1.

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
n	75 885	325 729	392 340	16 401	9 297	2 113
m	357 317	1 090 108	15 038 083	29 552	392 066	2 203
k	9.42	6.69	76.66	3.60	84.34	2.09
δ	1.2e-4	2.1e-5	1.9e-4	2.2e-4	9.1e-3	9.9e-4

TAB. 2.1 – *Nombre de sommets, de liens, degré moyen et densité des grands réseaux d'interactions utilisés comme référence.*

2.3 Connexité

Dans un graphe $G = (V, E)$, une *composante connexe* (resp. composante fortement connexe) est un sous-ensemble maximal de sommets $V' \subset V$ tel qu'il existe une chaîne (resp. chemin) entre tout couple de sommets de V'. On dira qu'un graphe est connexe (resp. fortement connexe) si V forme une composante connexe (resp. fortement connexe). On peut parler de composante faiblement connexe pour un graphe orienté en ne tenant pas compte de l'orientation des arcs.

Dans des réseaux de communication, l'existence de composantes déconnectées indique que certains usagers ne communiquent pas (volontairement ou non) avec les autres. De manière similaire, une composante fortement connexe de pages Web est un ensemble de pages toutes accessibles les unes depuis les autres en cliquant sur des liens hypertextes.

La Figure 2.1 présente le nombre de composantes connexes de différentes taille pour nos exemples. Dans deux cas, *Internet* et *Web*, le graphe est connexe. La manière dont ils sont collectés (`traceroute` ou crawl à partir d'une source) ne pouvait que donner des échantillons connexes. En particulier, d'autres explorations du graphe du Web sur une plus grande échelle montrent l'existence de composantes déconnectées (*cf.* le modèle du nœud papillon, Section 1.1 [25]).

Sur les quatre autres réseaux, on remarque la présence d'un nombre non négligeable de petites composantes. En particulier, sur *Co-signature*, on peut identifier facilement un ensemble de 38 scientifiques "isolés" qui ont écrit les uns avec les autres, mais pas avec d'autres auteurs. Dans tous les cas, il existe une composante connexe contenant une large majorité de sommets.

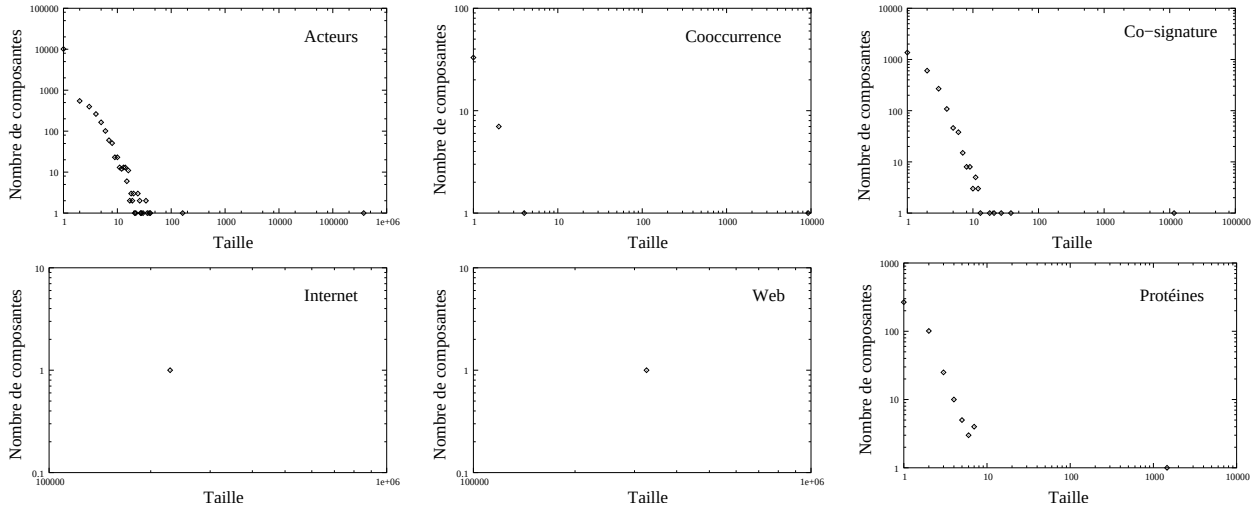

FIG. 2.1 – *Nombre de composantes connexes de chaque taille pour nos six exemples.*

2.4 Distance moyenne et diamètre

La *distance moyenne* est définie comme la moyenne sur tous les couples de sommets de la distance entre ces sommets. Le *diamètre*, quant à lui, est la distance maximale dans le graphe.

Les premières études sur la distance entre les sommets des grands réseaux d'interactions ont eu lieu sur des réseaux sociaux et ont engendré la notion de "six degrés de séparation". Ils ont été effectués par Stanley Milgram en 1967 [90] (voir Section 1.2). Ces résultats sont surprenants car ils montrent que les distances entre individus sont très faibles.

Il est apparu plus récemment que tous les grands réseaux d'interactions, et pas seulement les réseaux sociaux, ont une distance moyenne faible, variant typiquement comme le logarithme du nombre de sommets n. Il a été montré que cette propriété est en fait assez générale [22, 81, 105, 106].

En pratique, le calcul de la distance moyenne et du diamètre (Tableau 2.2) n'est pas toujours aisé pour les très grands graphes, la complexité étant en $\mathcal{O}(nm)$. On calcule donc généralement une valeur approchée de la distance moyenne en l'évaluant pour un certain nombre de couples de sommets, ce qui converge rapidement vers la valeur exacte. Pour le diamètre, il est plus difficile d'obtenir une valeur représentative sans effectuer le calcul exact.

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
n	75 885	325 729	392 340	16 401	9 297	2 113
d	5.80	7	3.6	7.18	2.13	6.74
diamètre	29	27	12	20	8	19

TAB. 2.2 – *Taille, distance moyenne et diamètre pour les principaux grands réseaux d'interactions. La distance moyenne est très faible et est considérée en général comme logarithmique en le nombre de sommets.*

Dans ce tableau, on peut aussi voir qu'il n'y a pas de lien *a priori* entre la distance moyenne et le diamètre, ce dernier pouvant être de 2 à 5 fois plus élevé que la distance moyenne.

La Figure 2.2 montre la distribution des distances, c'est-à-dire la proportion de couples de sommets à distance donnée l'un de l'autre. Ces distributions suivent approximativement des lois de Poisson :

$$P[distance = d] \sim e^{-\lambda} \cdot \frac{\lambda^d}{d!}.$$

Cela signifie qu'en choisissant deux sommets quelconques, la distance les séparant est, sinon égale, du moins très proche de la valeur moyenne λ. Ceci justifie le calcul approché de la distance moyenne en effectuant la moyenne sur un échantillon.

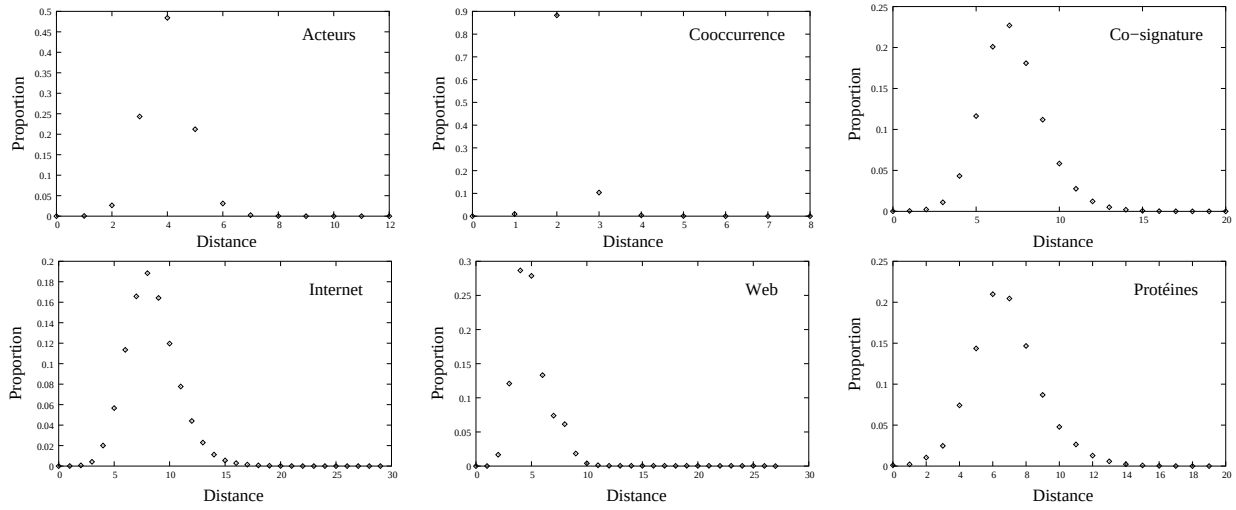

FIG. 2.2 – *Distribution des distances entre tous les couples de sommets pour nos six exemples.*

2.5 Distribution des degrés

La distribution des degrés d'un graphe donne, pour chaque entier d, le nombre de sommets de degré d. On aurait pu s'attendre à ce que les grands réseaux d'interactions exhibent une distribution en loi de Poisson similaire à celle présentée pour les distances. Une telle distribution aurait signifié que tous les sommets ont un degré à peu près identique. En fait, il n'en est rien. La distribution des degrés suit une loi de puissance :

$$P[degré = d] = C \cdot d^{-\alpha},$$

où C est une constante de normalisation. Cette loi a une décroissance polynômiale, ce qui implique que, bien qu'il y ait un grand nombre de sommets de faible degré, le nombre de sommets de très fort degré est non négligeable, comme le montre la Figure 2.3.

L'exposant α de la loi représente sa vitesse de décroissance. Plus α est grand et plus la probabilité d'obtenir des sommets de fort degré diminue.

On peut ainsi remarquer que, pour une telle loi, le degré moyen, qui est calculé comme suit :

$$\sum_{d=0}^{n} d \cdot P[degré = d] = C \cdot \sum_{d=0}^{n} d^{-\alpha+1},$$

diverge dès que α est inférieur ou égal à 2. Le degré moyen tend alors vers l'infini quand la taille des graphes tend vers l'infini. De même, si α est compris entre 2 et 3, le degré moyen est fini mais l'écart type est infini. Enfin, si α est supérieur à 3, ces deux valeurs sont finies. En pratique, pour la majorité des grands réseaux d'interactions étudiés, l'exposant est compris entre 2 et 3 (voir Tableau 2.3).

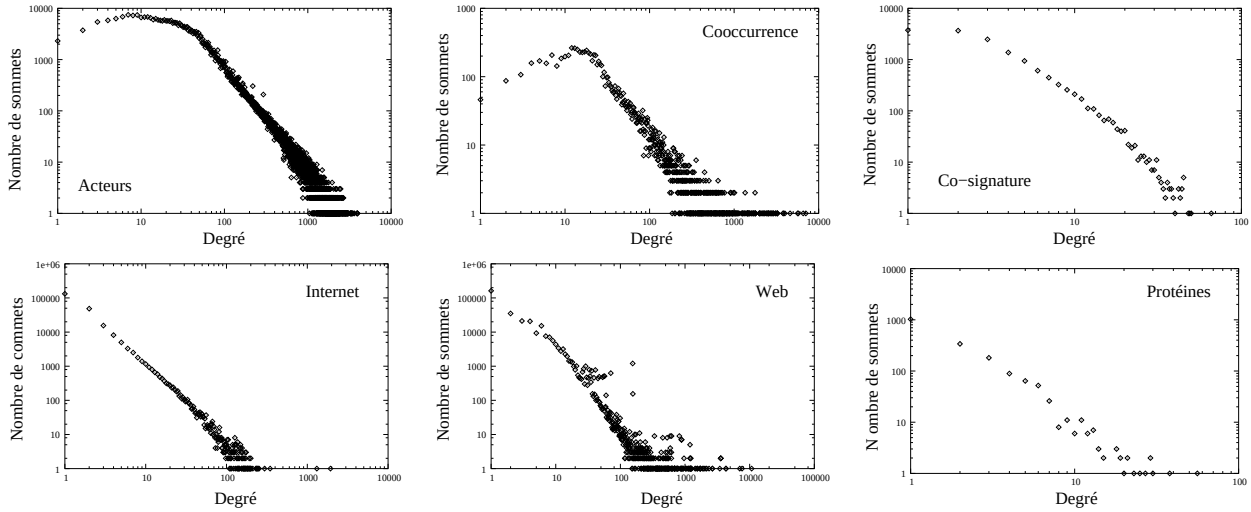

FIG. 2.3 – *Distribution des degrés pour nos six exemples.*

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
α	2.5	2.3	2.2	2.4	1.8	2.4

TAB. 2.3 – *Exposant de la loi de puissance pour nos six exemples.*

Les graphes dont la distribution des degrés suit une loi de puissance sont souvent appelés graphes sans-échelle. En effet, la valeur moyenne du degré n'est pas représentative, peu de sommets ayant ce degré et la plupart des sommets ayant des degrés très différents. En physique, on appelle souvent ce phénomène : absence d'échelle caractéristique pour les sommets, d'où le terme graphes sans-échelle.

Dans la suite de cette thèse, les différences entre loi de Poisson et loi de puissance seront souvent mises en avant car elles sont la cause de comportements radicalement différents dans de nombreux cas.

2.6 Clustering

Le *clustering*[1] est une mesure de la densité locale des sommets. Il en existe plusieurs définitions qui sont équivalentes sous certaines conditions. La plus utilisée est certainement celle introduite par Watts et Strogatz [134] définissant le clustering d'un sommet comme la densité du voisinage de ce sommet. Plus formellement, pour un sommet $u \in V$:

$$c(u) = \begin{cases} \frac{|\{(x,y)\in E, x,y\in N(u)\}|}{\binom{d(u)}{2}} & \text{si } d(u) \geq 2 \\ 0 & \text{sinon,} \end{cases}$$

1. Plusieurs traductions de ce terme ont été proposées, comme *densité locale*, *cliquicité* ou *coefficient d'interconnectivité* mais, aucune ne remportant l'unanimité, nous conservons le terme anglophone.

c'est-à-dire le rapport entre le nombre de liens dans le voisinage de ce sommet et le nombre de liens qu'il pourrait y avoir dans ce voisinage. Le clustering d'un graphe est alors défini comme étant la moyenne du clustering pour tous les sommets :

$$c(G) = \frac{1}{n} \sum_{u \in V} c(u)$$

Comme le montre le Tableau 2.4, le clustering des grands réseaux d'interactions est assez élevé, bien plus que la densité globale du graphe. Ces graphes sont donc très denses localement mais peu denses globalement. C'est ce que vise à capturer la notion de clustering.

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
c	0.171	0.466	0.785	0.638	0.822	0.153
densité	1.2e-4	2.1e-5	1.9e-4	2.2e-4	9.1e-3	9.9e-4

TAB. 2.4 – *Clustering et densité globale des grands réseaux d'interactions étudiés.*

Une autre définition du clustering, globale cette fois, est le rapport entre le nombre de triangles (trois sommets reliés les uns aux autres) et le nombre de triplets connectés (trois sommets reliés par deux liensdans le graphe) [106] :

$$c_G = \frac{3 \cdot | \triangle |}{| \wedge |}$$

Le coefficient 3 vient du fait que chaque triangle engendre 3 triplets connectés.

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
triangles	380 136	17 820 010	693 626 398	35 636	29 120 690	444
triplets	8 558 640	304 881 174	6 266 209 411	230 599	353 695 211	12 107
c_G	0.1332	0.1753	0.3321	0.4636	0.247	0.11

TAB. 2.5 – *Nombre de triangles, de triplets et clustering global des grands réseaux d'interactions étudiés.*

Ces deux définitions ne sont pas équivalentes dans la majorité des grands réseaux d'interactions étudiés (voir Tableaux 2.4 et 2.5) pour diverses raisons. Imaginons, par exemple, le cas d'un graphe dans lequel se trouve un sommet u de très fort degré d, mais de clustering nul. Dans la version locale, le clustering est moyenné entre tous les sommets : notre sommet atypique n'aura qu'une très faible influence sur le clustering du graphe. Au contraire, dans la version globale du clustering, ce sommet de très fort degré va engendrer un très grand nombre de triplets (de l'ordre de d^2), mais aucun triangle. Il va donc avoir une grande influence sur le clustering global en le faisant chuter. Or, il faut noter que cette situation n'est pas vraiment atypique puisque, dans un graphe sans-échelle, il existe toujours des sommets de très fort degré. Inversement, la présence de nombreux sommets de degré 1 peut faire chuter le clustering local sans influencer le clustering global.

Sauf mention explicite du contraire, nous utiliserons dans la suite la définition locale qui est la plus largement utilisée dans la littérature.

2.7 Centralité d'intermédiarité

La centralité d'intermédiarité (*betweenness* en anglais) exprime dans quelle mesure une personne peut servir d'intermédiaire dans la transmission d'informations entre deux autres personnes. Plus formellement, dans un graphe $G = (V, E)$, la centralité d'un sommet $u \in V$ est égale au nombre de plus courts chemins passant par u [53]. La centralité est définie de manière similaire sur les liens.

La centralité peut ainsi servir de mesure de la charge induite sur un sommet ou un lien. Dans l'Internet, par exemple, un routeur de forte centralité gère certainement de très nombreux messages. De même, un câble de forte centralité sera très souvent emprunté par des messages et devrait donc bénéficier d'un débit élevé sous peine d'être engorgé en permanence.

Dans la plupart des grands réseaux d'interactions, la centralité est extrêmement variable entre les individus [98] et, l'individu de plus grande centralité en a une bien plus élevée que le second, qui à son tour en a une bien plus élevée que le troisième, etc. La décroissance est très rapide, au moins pour les quelques individus ayant une très forte centralité.

Un autre point, déjà noté dans l'expérience de Milgram, est que la majorité des plus courts chemins passant par un sommet proviennent d'une petite proportion des voisins de ce sommet. En conséquence, un sommet de fort degré peut avoir une centralité relativement faible (par exemple, s'il n'est relié qu'à des sommets de degré 1), et réciproquement. La centralité et le degré ne sont donc pas forcément liés.

Une application de la centralité concerne la recherche de communautés. En remarquant que les liens de forte centralité font souvent le pont entre deux communautés, un algorithme de recherche de communautés a été proposé consistant à supprimer un par un les liens de plus forte centralité, jusqu'au moment ou l'on estime qu'il ne reste que des communautés disjointes [130]. Cet algorithme est malheureusement assez coûteux : le calcul de la centralité a une complexité en $\mathcal{O}(nm)$ et il faut refaire les calculs à chaque suppression de lien.

2.8 Corrélations

Jusqu'à présent, nous nous sommes contentés de décrire les graphes par des mesures globales (diamètre, densité) ou purement locales (degré ou clustering d'un sommet). Les grands réseaux d'interactions qui nous servent d'exemples types sont très similaires sur ces propriétés en ayant tous une distance moyenne faible, un clustering élevé et une distribution des degrés en loi de puissance. Ils rejoignent en cela la grande majorité des grands réseaux d'interactions.

Pour tenter de les distinguer les uns des autres sans introduire de nouveaux paramètres (comme la centralité, par exemple), une solution consiste à étudier les corrélations entre les paramètres déjà calculés. Ces corrélations peuvent de plus, comme nous le verrons dans le Chapitre 6, apporter une information fine sur le réseau considéré.

Dans cette section, nous allons présenter plusieurs types de corrélations simples mettant en jeu les degrés et qui permettent d'observer des différences comportementales importantes entre les différents grands réseaux d'interactions.

2.8.1 Corrélations entre degrés

La corrélation la plus simple met en jeu les degrés uniquement : il s'agit de savoir quel est le degré moyen des voisins d'un sommet de degré donné. Plusieurs comportements sont attendus : les sommets peuvent être connectés à des sommets de même nature qu'eux (fort degré avec fort degré) auquel cas le réseau est dit assortatif, ou à des sommets de nature opposée (fort degré avec faible degré) dans des réseaux dissortatifs. Dans le cas où aucun de ces comportements n'est observé, le réseau est dit neutre.

En pratique, il existe deux méthodes pour calculer cette corrélation. Tout d'abord, un paramètre introduit dans [99, 100], appelé coefficient d'assortativité, permet de quantifier la similarité de degrés entre voisins :

$$r = \frac{\sum_i j_i k_i - (\sum_i j_i + k_i)^2/4m}{\sum_i (j_i^2 + k_i^2)/2 - (\sum_i j_i + k_i)^2/4m},$$

où j_i et k_i sont les degrés des sommets à l'extrémité du lien i.

Ce paramètre vaut 0 si le réseau est neutre ; il est positif (resp. négatif), si le réseau est assortatif (resp. dissortatif). Le Tableau 2.6 montre les valeurs de ce paramètre sur les grands réseaux d'interactions étudiés. On peut remarquer que les réseaux sociaux (*Acteurs* et *Co-signature*) sont assortatifs, alors que tous les autres sont dissortatifs. Cet état de fait était déjà connu en sociologie pour les réseaux sociaux, mais il n'y a pas vraiment de raisons connues pour expliquer la dissortativité des autres réseaux. Des hypothèses peuvent être avancées pour *Internet* : par exemple, les machines de faible degré sont probablement des machines d'utilisateurs, connectées à des routeurs de bordure de fort degré.

Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
-0.0086	-0.0534	0.2267	0.1027	-0.1873	-0.1615

TAB. 2.6 – *Coefficient d'assortativité des grands réseaux d'interactions étudiés.*

Une autre manière d'observer les corrélations entre degrés est de tracer la courbe exprimant le degré moyen des voisins des sommets de degré donné [21, 128, 129]. Alors que le coefficient d'assortativité nous donne une mesure globale, cette courbe est beaucoup plus précise et permet de distinguer différents comportements pour les réseaux dissortatifs (voir Figure 2.4).

Mis à part pour *Internet*, les courbes confirment les valeurs du coefficient d'assortativité du Tableau 2.6. *Acteurs* et *Co-signature* ont une tendance légèrement croissante : plus les individus sont connectés, plus leurs voisins le sont aussi. Au contraire, *Web*, *Cooccurrence* et *Protéines* exhibent le comportement inverse : les sommets de fort degré sont principalement connectés à des sommets de faible degré.

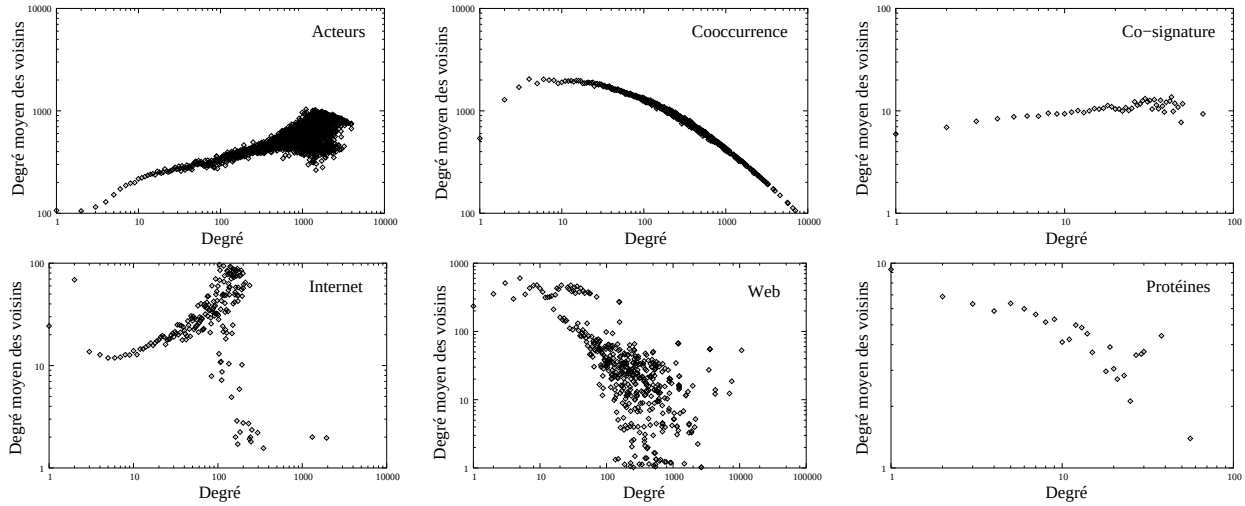

FIG. 2.4 – *Corrélation degré-degré exprimant pour chaque valeur k du degré, le degré moyen des voisins des sommets ayant degré k.*

Le comportement de *Internet* n'est pas monotone : croissant au début, puis clairement décroissant. Ceci explique que le coefficient d'assortativité soit proche de 0 : les sommets de faible et moyen degrés sont assortatifs, ceux de fort degré sont dissortatifs. La moyenne des deux donne un réseau globalement presque neutre.

2.8.2 Corrélations degré-clustering

De même que pour les corrélations degré-degré, il est possible d'observer les corrélations entre le degré et le clustering, c'est-à-dire de calculer le clustering moyen des sommets ayant un degré donné (Figure 2.5).

Toute les courbes montrent une tendance similaire : plus le degré est élevé, plus le clustering est bas. Ceci est assez naturel, pour les raisons suivantes :

- si un réseau est dissortatif, les sommets de fort degré sont reliés à des sommets de faible degré. Ce grand nombre de voisins ayant un faible degré, ils ne peuvent pas être tous reliés entre eux et donc le clustering est naturellement faible ;
- pour les réseaux assortatifs, les sommets de fort degré sont connectés à d'autres sommets de fort degré, mais si l'on regarde plus attentivement la Figure 2.4, pour *Acteurs* et *Co-signature*, on se rend compte qu'en moyenne, même les sommets de fort degré ont des voisins de degré plus faible qu'eux. L'argument précédent peut donc être réutilisé, même dans ce contexte ;

Le seul cas qui déroge à cette règle est *Internet*. Dans la partie asssortative (degrés inférieurs à 100), les sommets ont des voisins de plus fort degré qu'eux : ils peuvent donc potentiellement avoir un fort clustering, et c'est effectivement le cas.

L'allure des courbes de corrélation permet de comprendre les valeurs obtenues pour le

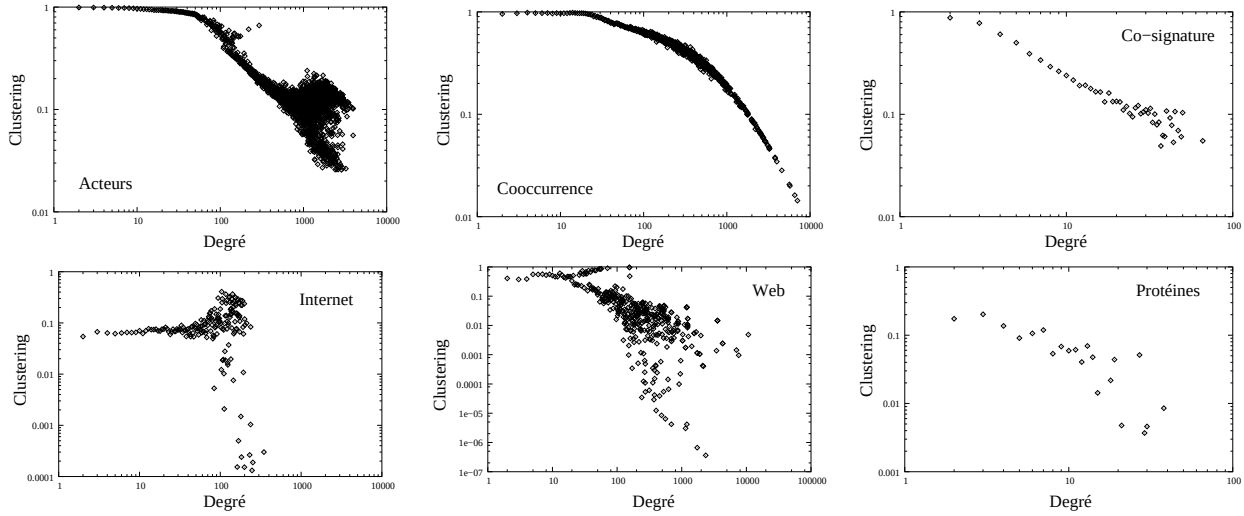

FIG. 2.5 – *Corrélation degré-clustering.*

clustering dans la section 2.6. En effet, les réseaux *Acteurs*, *Cooccurrence*, *Co-signature* et *Web* ont de très nombreux sommets de faible degré (distribution en loi de puissance) et ces sommets ont un très fort clustering. Ce sont donc eux qui sont majoritaires lorsque l'on calcule le clustering moyen. Pour les deux autres réseaux, *Internet* et *Protéines*, le comportement est un peu différent, mais même les sommets de faible degré ont un faible clustering, ce qui explique le clustering moyen plus faible de ces deux réseaux.

Conclusion

Nous avons introduit les paramètres que nous rencontrerons le plus fréquemment dans l'ensemble de la thèse. Bien que de nombreux autres paramètres aient été introduits, le plus souvent dans des cas particuliers, nous n'entrons pas plus dans leurs détails. Lorsque certaines propriétés plus spécifiques apparaîtront (notamment dans la troisième partie), nous les définirons au moment opportun.

Soulignons qu'il reste encore aujourd'hui un important travail à effectuer dans le cadre de l'analyse des grands réseaux d'interactions. Par exemple, la dynamique des grands réseaux d'interactions commence tout juste à être étudiée et permettrait à terme de décrire l'évolution de ces réseaux. De même, l'analyse des réseaux valués, hétérogènes (bipartis par exemples) et/ou hybrides (plusieurs réseaux définis sur un même ensemble d'éléments) en est aujourd'hui à ses tous débuts. Il existe aujourd'hui une très forte attente pour des outils et méthodes d'analyse raffinés, mais un cadre général et satisfaisant reste encore largement à définir. Nous proposons quelques contributions dans ce sens dans les chapitres suivants, notamment concernant les graphes bipartis (Chapitre 4) et les graphes dynamiques (Chapitre 6).

Conclusion

Nous avons présenté dans cette partie les grands réseaux d'interactions actuellement étudiés, ainsi que des propriétés caractéristiques de ces réseaux. Ils sont issus de contextes très variés, certains étant physiquement créés par l'homme, d'autres étant issus de la nature, d'autres encore reflètant des structures sociales. La grande diversité de ces réseaux apporte non seulement un ensemble très varié d'objets à étudier, mais soulève aussi un grand nombre de questions similaires dans des disciplines variées. Ces questions peuvent parfois être transposées d'un contexte à un autre, ou se généraliser pour soulever des questions plus fondamentales.

Parmi les propriétés identifiées, le clustering et la distribution des degrés montrent que les grands réseaux d'interactions ne sont généralement pas homogènes: ils ont, pour la plupart, une densité globale très faible alors qu'il sont localement denses (fort clustering), et la distribution des degrés suit une loi de puissance qui indique qu'il y a une grande disparité entre les degrés.

Le point fondamental est que tous ces réseaux sont similaires concernant ces propriétés de base, ainsi que la faible distance moyenne. Ceci permet de les considérer globalement et de les étudier dans leur ensemble.

Cette constatation a notamment donné naissance à une intense activité de modélisation des grands réseaux d'interactions en général, c'est-à-dire visant à produire des graphes artificiels ayant les trois grande propriétés citées. La deuxième partie est consacrée à cette problématique. De même, un important travail d'évaluation des conséquences de ces propriétés, notamment sur les phénomènes se déroulant sur les grands réseaux d'interactions, a été mené. Nous verrons ainsi dans le Chapitre 7 les conséquences de la distribution des degrés sur la résistance d'un réseau aux pannes et aux attaques.

Deuxième partie

Modélisation

Introduction

La modélisation des grands réseaux d'interactions vise avant tout à produire des objets artificiels *similaires* aux objets étudiés. La première étape consiste à identifier un ensemble de propriétés caractéristiques de l'objet que le modèle devrait reproduire. Nous avons vu précédemment que plusieurs propriétés de base sont assez générales : les graphes sont très peu denses, tout en ayant une forte densité locale (fort clustering), la distribution des degrés suit presque toujours une loi puissance et ces graphes ont une distance moyenne faible. Les modèles doivent donc s'attacher à capturer ces trois propriétés en premier lieu. D'autres propriétés ont été identifiées mais sont plus complexes et leurs conséquences ne sont pas forcément bien comprises. Nous nous concentrerons donc sur les trois que nous avons citées.

Une fois l'ensemble des propriétés à capturer défini, il y a principalement deux méthodes pour proposer un modèle.

Tout d'abord, on peut vouloir tirer aléatoirement un graphe dans l'ensemble de ceux ayant ces propriétés et obtenir ainsi un graphe *typique* de cette classe de graphes. Lorsque des graphes sont générés de cette manière, il est ensuite possible de décider si le jeu de propriétés retenues était suffisant : est-ce que les graphes produits par le modèle ressemblent effectivement au graphe original (au sens d'autres propriétés) ? Par exemple, si l'on dispose d'un modèle générant des graphes avec une distribution des degrés en loi de puissance [18, 22, 82], on pourra déterminer si ce modèle capture aussi le clustering et, par conséquent, savoir si le clustering est une propriété typique des graphes sans-échelle ou non.

L'autre approche consiste à définir un processus de construction *mimant* celui par lequel le graphe considéré est effectivement construit en réalité. On espère ainsi capturer les propriétés du graphe induites par sa construction. De tels processus de construction consistent généralement à partir d'un graphe initial et à le faire évoluer selon des règles plus ou moins complexes. On espère qu'à partir d'un moment le graphe aura les propriétés souhaitées. Il sera aussi possible d'étudier les autres propriétés engendrées par le modèle. Un exemple typique dont nous discuterons dans le Chapitre refchap:biparti concerne certains graphes particuliers, tel le graphe des acteurs. Dans ce graphe, à chaque fois qu'un film est tourné, il engendre une clique reliant tous les acteurs qui ont joué dedans. Nous proposerons un modèle basé sur ce principe consistant à ajouter de manière itérative des cliques à un graphe initialement vide.

En pratique, la première méthode est généralement plus adaptée si l'on souhaite analyser formellement le modèle, mais il peut être très difficile d'obtenir certaines propriétés. Le

clustering, qui est pourtant une propriété de base, reste ainsi réfractaire à cette approche : on ne sait pas générer un graphe aléatoire tiré uniformément dans l'ensemble des graphes ayant un clustering donné. Au contraire, la seconde méthode permet d'obtenir assez simplement des graphes qui ont l'avantage d'être évolutifs. Ainsi, si l'on arrive à définir un modèle générant un graphe semblable au Web par un processus de construction quelconque, alors, en laissant tourner le programme plus longtemps, on peut espérer obtenir une vision future du Web. Toutefois, cette méthode induit généralement de nombreuses propriétés non souhaitées et les analyses rigoureuses sont, en général, complexes à mener.

Les caractéristiques que l'on recherche pour un bon modèle sont donc la simplicité qui permet de faire des études formelles, mais rend le modèle plus utilisable par la communauté scientifique et le réalisme qui permet de mieux comprendre les phénomènes en jeu et justifie plus naturellement la pertinence du modèle. L'objectif est bien entendu de définir des modèles toujours plus efficaces suivant ces critères.

Cette partie dédiée à la modélisation des grands réseaux d'interactions est divisée en trois chapitres.

Le Chapitre 3 présente un état de l'art des principaux modèles introduits ces dernières années en distingant les deux classes de modèles, ceux à base de tirage aléatoire et ceux par construction itérative.

Le Chapitre 4 introduit deux nouveaux modèles qui sont en fait une version par tirage aléatoire et une version itérative du même principe : le modèle itératif est réaliste par sa construction et il est possible de prouver formellement les propriétés du modèle par tirage aléatoire.

Malgré tout, ces deux modèles, bien que constituant un progrès significatif, sont encore trop aléatoires et perdent de nombreuses corrélations. Le modèle présenté dans le Chapitre 5 améliore ces aspects, ce qui en fait l'un des modèles les plus performants disponibles actuellement.

Chapitre 3

État de l'art

Dans ce chapitre nous allons nous attacher à présenter les grands courants qui ont animé la modélisation des grands réseaux d'interactions ces dernières années. Comme nous l'avons déjà signalé, deux méthodes se détachent. La première consiste à tirer de manière aléatoire, et si possible uniforme, un graphe parmi tout ceux ayant les propriétés souhaitées. La seconde propose des processus de construction imitant les processus réels, avec l'objectif que cette construction induise les propriétés souhaitées.

Nous présentons dans la suite le premier modèle aléatoire de graphes introduit puis, divers modèles introduits pour capturer les distributions des degrés, le clustering, ainsi que des modèles plus spécifiquement conçus pour modéliser l'Internet. Finalement, nous ferons un bilan comparatif de ces modèles, d'un point de vue théorique tout d'abord, et ensuite en analysant les performances des modèles concernant les six grands réseaux d'interactions que nous avons choisis comme exemples tout au long de cette thèse.

Soulignons que de très nombreux modèles ont été proposés, que nous ne présenterons pas tous ici. Tous ont leurs avantages, mais notre objectif n'est pas ici d'être exhaustifs ; nous voulons plutôt dresser un état des lieux afin de décrire le contexte dans lequel nos contributions, présentées dans les chapitres suivants, se situent.

3.1 Le modèle aléatoire pur

Le modèle de graphe aléatoire le plus classique est celui introduit par Erdös et Rényi et étudié en profondeur par la suite [22, 46]. Dans le modèle $\mathcal{G}_{n,p}$, un graphe à n sommets est créé en considérant que chacun des $\frac{n\cdot(n-1)}{2}$ liens possibles existe avec une probabilité p fixée. De manière similaire, on peut générer un graphe aléatoire en choisissant $m = p \cdot \frac{n\cdot(n-1)}{2}$ liens au hasard (modèle $\mathcal{G}_{n,m}$).

Ce modèle a engendré un grand nombre d'études sur les propriétés des graphes obtenus en fonction de la probabilité de connexion p. Le degré moyen dans le graphe est simplement $\lambda = pn$. De nombreux phénomènes se produisent quand λ est proche de 1, c'est-à-dire quand $p = 1/n$ [22] :

– tant que $\lambda < 1$, le graphe est composé de petits arbres de taille au plus logarithmique

et d'un nombre constant de cycles. Quand λ augmente, ces petites composantes commencent à se regrouper pour en former de plus grosses ;
- pour $\lambda = 1$, la structure du graphe change brutalement, celui-ci devenant soudainement composé d'un composante géante dont la taille est de l'ordre de $n^{2/3}$;
- dès que $\lambda > 1$, cette composante géante devient encore plus grosse et contient une proportion de sommets linéaire en n. Toutes les autres composantes restent de taille négligeable et fusionnent avec la composante géante quand k augmente.

Ce phénomène à reçu le nom de *double saut* ou de *transition de phase*. De nombreuses autres études ont été menées pour déterminer la distribution des degrés du graphe, son clustering, son diamètre ou encore le nombre de sous-graphes d'un type donné (triangles, par exemple). Nous allons en présenter maintenant quelques unes.

Autres propriétés du modèle

La distribution des degrés des graphes aléatoires est connue [22]. La probabilité que le degré d'un sommet vaille d est :

$$P(degré = d) = \binom{n-1}{d} \cdot p^d (1-p)^{n-1-d},$$

c'est-à-dire une loi binômiale : le sommet a d liens choisis avec probabilité p vers d destinations parmi les $n-1$ possibles. Cette distribution peut être approximée par une loi de Poisson quand n est suffisamment grand :

$$p_k \sim e^{-\lambda} \frac{\lambda^k}{k!}$$

Une telle distribution est centrée sur la valeur moyenne $\lambda = pn$, et la déviation est très faible : la majorité des sommets a donc un degré moyen très proche de λ. Les premiers travaux sur le sujet [46] étaient aussi consacrés à l'étude du degré maximum et minimum dans un graphe aléatoire et ont montré que ces extrêmes sont de l'ordre du degré moyen.

Il est aussi très simple d'évaluer le clustering, défini comme la probabilité que deux voisins d'un même sommet soient reliés. Comme les liens sont placés au hasard, la probabilité que deux sommets soient reliés est indépendante de leurs voisins. Le clustering est donc exactement égal à la probabilité de connexion p : la densité locale est égale à la densité globale. Cela signifie notamment que le clustering d'un graphe aléatoire de taille comparable aux grands réseaux d'interactions étudiés sera extrêmement faible.

Le diamètre des graphes $\mathcal{G}_{n,p}$ est faible. Diverses études ont été menées pour mieux comprendre ses variations en fonction de p [1, 22, 30]. Il ressort de ces études que dès que le seuil, $\lambda = 1$, est dépassé, le diamètre est logarithmique en n. Le diamètre étant une borne supérieure pour la distance moyenne, cette dernière peut donc aussi être considérée comme logarithmique pour ce modèle.

Enfin, il est aussi possible de prévoir à quel moment un sous-graphe donné va apparaître dans un graphe aléatoire et, par exemple, quelle est la taille de la plus grande clique d'un

graphe aléatoire en fonction de p. Si l'on considère un sous-graphe ayant s sommets et t liens, alors ce type de graphe va apparaître dès que $p \sim n^{-s/t}$. Ainsi, les triangles vont faire leur apparition presque en même temps que la composante géante ($p \sim 1/n$), les cliques de taille 4 quand p sera de l'ordre de $n^{-2/3}$, etc. Dans des graphes $\mathcal{G}_{n,p}$ peu denses, dont la densité est du même ordre que celle des grands réseaux d'interactions étudiés, la probabilité de trouver une clique de taille supérieure à 4 ou 5 est donc extrêmement faible.

Au vu des propriétés discutées dans la partie précédente, ce modèle n'est donc pertinent que pour la distance moyenne, et ne l'est pas pour le clustering, ni pour la distribution des degrés. Il est toutefois très utile comme *modèle test*, ou sans hypothèses, puisqu'il fournit un graphe uniformément aléatoire parmi ceux ayant une taille et une densité donnée.

3.2 Modélisation des graphes sans-échelle

Nous venons de voir que les modèles aléatoires $\mathcal{G}_{n,p}$ et $\mathcal{G}_{n,m}$ génèrent des graphes dont la distribution des degrés suit une loi de Poisson. Or, les grands réseaux d'interactions étudiés ont presque tous une distribution des degrés en loi de puissance. Ceci a conduit à l'introduction de plusieurs modèles générant des graphes ayant cette propriété.

3.2.1 Distribution des degrés fixée

Des travaux antérieurs à la découverte des lois de puissance dans les grands réseaux d'interactions avaient conduit à l'étude des classes de graphes ayant une distribution des degré fixée [18, 22, 82, 92, 93, 99]. L'objectif est de pouvoir obtenir un graphe aléatoire dont la distribution des degrés soit celle souhaitée (loi de Poisson, loi de puissance, etc.). Pour atteindre cet objectif, il existe plusieurs algorithmes basés soit sur des échanges de liens, soit sur des appariements. Avec ces modèles, il est donc possible, par construction, de générer des graphes avec les distributions des degrés rencontrées en pratique.

La génération basée sur les mélanges [99] consiste à prendre un graphe ayant la distribution des degrés souhaitée et d'effectuer une suite d'échanges de liens : à chaque étape, on choisit deux liens (v_1, w_1) et (v_2, w_2) que l'on remplace par (v_1, v_2) et (w_1, w_2) si cet échange ne crée ni liens multiples ni boucles. On sait qu'après un temps infini le graphe devient complètement aléatoire. Cependant, le nombre de mélanges à effectuer[1] avant d'aboutir à un graphe suffisamment aléatoire est connu seulement dans certains cas particuliers [72]. D'après [99], effectuer $100m$ mélanges permet d'obtenir un graphe suffisamment aléatoire ayant la bonne distribution en un temps raisonnable.

Une autre méthode est basée sur l'appariement de demi-liens [18, 92, 93]. La procédure est la suivante : pour chaque sommet du graphe, on fixe son degré suivant la distribution souhaitée et ce sommet reçoit autant de demi-liens que son degré. Ensuite, il ne reste plus qu'à relier les demi-liens deux par deux de manière aléatoire (voir Figure 3.1).

1. Plus formellement, c'est le temps de mélange pour la chaîne de Markov associée au processus.

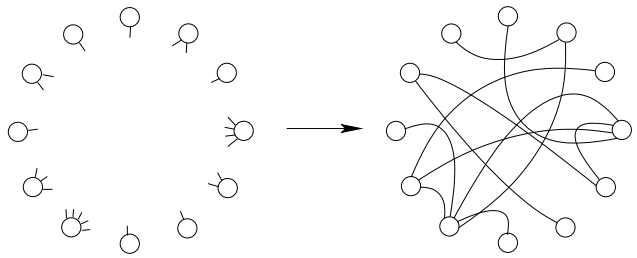

FIG. 3.1 – *Modèle par appariements : On tire le degré de chaque sommet selon la distribution souhaitée, puis on affecte à chaque sommet autant de demi-liens que son degré. Enfin, des paires de demi-liens sont choisies aléatoirement pour créer les liens du graphe.*

La somme des degrés doit être paire pour assurer que les demi-liens puissent tous être reliés. Avec cette procédure, rien n'empêche *a priori* de créer des liens multiples ou des boucles. Si l'on souhaite éviter ceux-ci, le graphe est alors rejeté et la procédure recommence. Dans le cas d'une loi de puissance, les quelques sommets de fort degré ont une forte probabilité d'être reliés entre eux plusieurs fois : cette méthode risque donc de ne jamais aboutir. En pratique, pour des graphes de grande taille, cette méthode ne crée que quelques boucles et liens multiples que l'on se contente généralement d'ignorer.

Il existe un troisième type d'algorithme utilisant la méthode par appariement avec rejets [10]. Cette méthode se nomme "go with the winners" et consiste à générer simultanément un certain nombre de graphes en ajoutant des liens progressivement, exactement comme pour la méthode par appariements. À chaque fois qu'un lien multiple ou une boucle est généré, le graphe correspondant est rejeté et, de temps en temps, tous les graphes ayant survécu sont dupliqués pour éviter que l'ensemble ne se vide. Quand tous les liens ont été placés, un graphe est choisi au hasard parmi ceux qui restent.

Cet algorithme est plus correct que les deux autres d'un point de vue théorique et pratique (pas de problème de mélange de chaîne, ni de boucles ou de liens multiples), mais est beaucoup plus lourd et lent à mettre en œuvre. Pour cette raison, il n'est généralement utilisé que pour valider les performances d'autres algorithmes.

Ces trois méthodes donnent des graphes aléatoires ayant une distribution des degrés fixée, et il est bien entendu possible de les utiliser pour générer des graphes ayant une distribution des degrés en loi de puissance. Dans ce cas particulier, les graphes obtenus ont une distance moyenne logarithmique tant que $\alpha < 3,48$ [81]. Après cette limite, le graphe n'est plus connexe. D'autres propriétés ont été étudiées ainsi [5, 33, 81, 102, 110] et, sous des hypothèses raisonnables, le clustering de ces graphes tend vers 0 quand leur taille augmente [102].

3.2.2 Modèles à base d'attachement préférentiel

Au milieu du siècle dernier, de nombreux travaux avaient déjà mis en avant la présence de lois de puissance dans de nombreux contextes et plusieurs explications y avaient été

apportées. Herbert A. Simon avait ainsi introduit le concept de *préférence* en 1955 [123] en présentant un modèle pour expliquer la distribution de la fréquence des mots dans un livre.

Ce modèle considère un livre en cours d'écriture qui comporte déjà k mots. Le $k+1^e$ mot est un mot nouveau avec une probabilité constante, ou un mot qui est déjà apparu i fois avec une probabilité proportionnelle au nombre de mots qui sont apparus i fois[2]. Cette façon de procéder qui privilégie les mots déjà fréquents amène à une distribution des fréquences en loi de puissance. C'est la première apparition formelle du concept "la popularité est attractive", ou "rich get richer".

Quelques travaux lui ont fait écho, notamment Price [40] qui étudie en 1965 le réseau des citations scientifiques et s'interroge sur les raisons qui font que certains papiers sont plus cités que d'autres. En particulier, il essaie de savoir si les papiers beaucoup cités ont une plus forte probabilité de l'être encore plus.

Plus récemment, en 1999, Albert et Barabási ont réintroduit ce concept en l'appliquant à la modélisation de graphes, initialement pour le graphe du Web [6, 8]. Dans ce graphe, quand quelqu'un crée une nouvelle page, il ajoute des hyperliens vers d'autres pages, hyperliens qui pointent généralement vers des pages connues. Or, les pages les plus connues sont celles qui sont le mieux référencées. Les pages très pointées vont donc l'être encore plus via ce phénomène.

Ce principe peut être dérivé en un modèle où les sommets arrivent un par un dans le graphe et choisissent leurs voisins en fonction du degré de ces voisins : l'attachement est dit préférentiel. La probabilité de choisir un voisin donné est généralement linéaire en le degré de ce voisin. Avec une telle règle, on peut montrer que le degré est distribué suivant une loi de puissance d'exposant 3. En effet, la variation du degré d'un sommet i, introduit au temps t_i et de degré k_i est :

$$\frac{dk_i}{dt} = \frac{k_i}{2t},$$

et, par conséquent, le nombre de liens dans le graphe est égal au temps écoulé t. Choisir un sommet de degré k revient à choisir un de ses k demi-liens parmi les $2k$ disponibles. La résolution de cette équation différentielle donne $k_i = C\sqrt{t}$. Un sommet i a degré 1 quand il est introduit dans le système au temps t_i, donc $t_i = C^{-2}$. La probabilité que $k_i < k$ est égale à la probabilité que $t_i > t/k^2$. Or, les sommets sont introduits à chaque instant donc t_i est uniforme dans $\{1; t\}$ et vaut $1/t$. La probabilité que $k_i < k$ est donc de l'ordre de $1 - 1/k^2$. Finalement :

$$P(k) \sim \frac{\partial}{\partial k}\left(1 - \frac{1}{k^2}\right) \sim k^{-3}.$$

Ce modèle a été intensément étudié, surtout pour son aspect itératif, avec l'introduction de nouveaux sommets à chaque instant. Les propriétés de ce modèle sont maintenant bien connues et, en particulier, le modèle peut être modifié pour contrôler l'exposant de la loi

2. Une autre version considère qu'un mot va apparaître avec une probabilité proportionnelle à son nombre d'apparitions précédentes.

de puissance. Si la fonction de préférence n'est plus linéaire, alors le comportement change radicalement : si elle est sous-linéaire la distribution devient exponentielle, si elle est sur-linéaire alors un seul sommet devient très vite le seul auquel tous les nouveaux se lient (*Winner takes all*).

D'autre part, la distance moyenne est logarithmique, mais le clustering tend vers 0 quand la taille du graphe augmente. Pour un survol plus complet de ces propriétés, on peut se référer avantageusement à [7].

3.3 Capturer le clustering

Le modèle purement aléatoire et celui avec distribution des degrés fixée permettent de tirer uniformément des graphes ayant des propriétés données : nombre de sommets et degré moyen pour le premier, nombre de sommets et distribution des degrés pour le second. Cette approche pourrait en principe être prolongée afin de produire un modèle ayant aussi un clustering donné mais, jusqu'à présent, il n'existe aucune méthode atteignant cet objectif et le problème semble difficile.

Plusieurs modèles basés sur un processus de construction ont donc été introduits; mais il n'y a pas aujourd'hui de consensus. Nous présentons ci-dessous les deux modèles les plus largement utilisés.

3.3.1 L'anneau de Watts et Strogatz

Le premier modèle générique pour les grands réseaux d'interactions a été introduit en 1998 par Watts et Strogatz [134] pour capturer le phénomène dit de *petit-monde* (*small-world* en anglais) qui se base, d'après les auteurs, sur trois propriétés : densité faible, distance moyenne faible et fort clustering.

Leur modèle se base sur deux constatations : il est facile de trouver des graphes ayant une distance moyenne faible, typiquement des graphes aléatoires, et on peut trouver simplement des graphes ayant un clustering élevé. C'est le cas, par exemple, pour un anneau dans lequel les sommets sont reliés à leurs k plus proches voisins (voir Figure 3.2).

Dans un tel anneau à n sommets, pour lequel chaque sommet est relié à ses $2k$ plus proches voisins (k voisins de chaque coté), il est aisé de calculer le clustering d'un sommet. Considérons un sommet u de ce graphe. Ses voisins à distance 1 sont reliés à $2k-2$ de ses autres voisins, ses voisins à distance 2 sont reliés à $2k-3$ de ses voisins, etc. Ceci n'est valable que si $k << n$ ce qui est toujours le cas car k doit être petit pour que le graphe ait une densité comparable à celle des grands réseaux d'interactions. Le clustering de u peut donc être calculé exactement par :

$$c(u) = \frac{\sum_{i=1}^{k} 2k - i - 1}{2 \cdot \binom{2k}{2}} = \frac{3k-3}{4k+2}$$

Tous les sommets ont le même clustering qui est donc aussi le clustering du graphe indépendemment de sa taille. Ce clustering est supérieur à 0.3 dès lors que chaque sommet

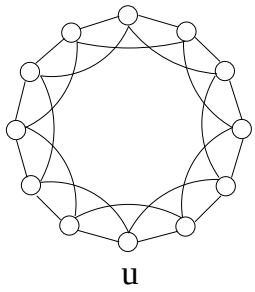

FIG. 3.2 – *Un anneau régulier à* 12 *sommets dans lequel chaque sommet est relié à ses* 4 *plus proches voisins. Dans ce graphe, les voisins de* u *à distance* i *dans l'anneau sont reliés à* $2k - i - 1$ *voisins de* u.

est relié à au moins 4 de ses voisins. Il faut malgré tout noter qu'avec cet anneau il est impossible d'obtenir un clustering supérieur à $3/4$ (si $k << n$). D'autre part, la distance moyenne dans l'anneau est très élevée (de l'ordre de n).

Mais, ni l'anneau, qui a un fort clustering et une grande distance moyenne, ni les graphes aléatoires, qui sont à l'opposé, ne capturent les deux propriétés souhaitées. L'idée de Watts et Strogatz a été de trouver un juste milieu en rendant l'anneau plus aléatoire. Le modèle qu'ils ont introduit est le suivant : initialement, on prend un anneau à n sommets en connectant chaque sommet à ses k plus proches voisins dans les deux sens (soit $2k$ voisins au total). Ensuite, on recâble chaque lien avec une probabilité p fixée en choisissant une nouvelle extrémité au hasard uniformément (voir Figure 3.3).

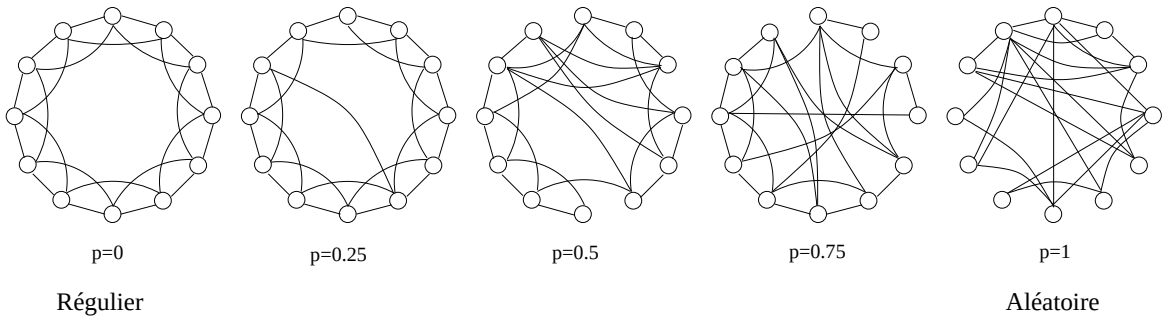

FIG. 3.3 – *Le modèle de Watts et Strogatz : introduction d'aléa dans un graphe initialement régulier.*

Les simulations de ce modèle confirment que quand p augmente, le clustering diminue, les voisinages étant moins denses, et la distance moyenne diminue car le fait de recâbler des liens crée des raccourcis dans le graphe. La distance moyenne décroît plus vite que le clustering ; il y a donc un moment où le clustering est encore élevé alors que la distance moyenne est déjà faible (voir Figure 3.4). Ceci correspond aux deux propriétés souhaitées pour les graphes petit-monde.

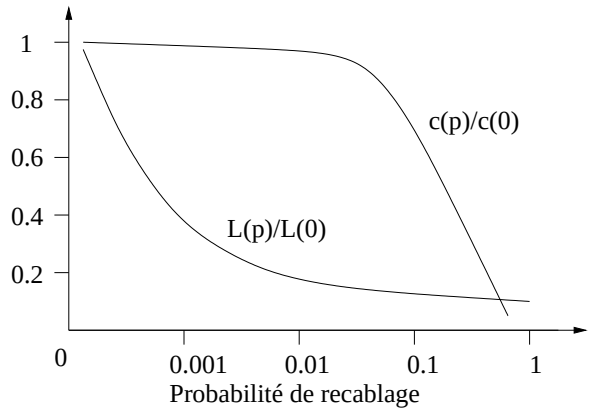

FIG. 3.4 – *Évolution du clustering, $c(p)$, et de la distance moyenne, $L(p)$ en fonction de la probabilité de recâblage p. Chaque courbe montre le rapport entre la valeur calculée et la valeur sans recâblage, $c(0)$ et $L(0)$.*

De nombreux autres modèles similaires ont été introduits. Ainsi, au lieu du recâblage, il est possible de rajouter des liens [103, 104], d'introduire un sommet central relié à certains sommets de la périphérie [41], etc. Les propriétés de ces variantes sont similaires au modèle original.

Malgré la distance moyenne faible et le fort clustering, le point faible de tous ces modèles est la distribution des degrés, qui suit une loi de Poisson.

3.3.2 Création de triangles

D'autres modèles ont été introduits pour générer des graphes à fort clustering. Certains de ces modèles se basent sur une croissance avec création de triangles à chaque étape.

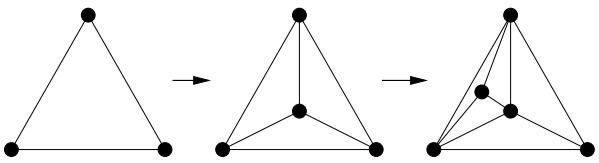

FIG. 3.5 – *Un modèle par création de triangles : initialement le graphe est un triangle et, à chaque étape, un triangle est choisi et les sommets qui le composent sont reliés à un nouveau sommet.*

Un modèle itératif a été introduit par Dorogovstev et Mendes [42] pour générer des graphes ayant une distribution des degrés en loi de puissance et un fort clustering : le graphe est initialement un triangle et, à chaque étape, un sommet est créé et relié aux trois sommets d'un triangle choisi au hasard (voir Figure 3.5). Ce modèle génère des graphes

ayant un clustering global valant 0.5 en moyenne. De plus, la distribution des degrés suit une loi de puissance [43].

Un autre modèle très similaire des mêmes auteurs est basé sur un mélange d'ajout de triangles et d'attachement préférentiel comme suit [42] : à l'instant $t = 0$, le graphe est un triangle (ou tout autre graphe) puis, à chaque étape, un sommet est créé et est relié aux deux extrémités d'un lien choisi au hasard (voir Figure 3.6). Chaque nouveau sommet génère donc un triangle, ce qui tend à augmenter le clustering. D'autre part, les sommets ayant un fort degré ont beaucoup de liens qui leur sont rattachés, et ont donc plus de chance d'être choisis, ce qui fait que la distribution des degrés de ce modèle est en loi de puissance.

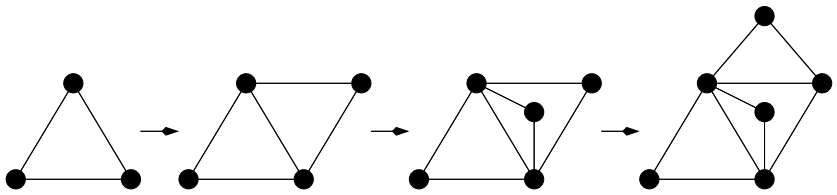

FIG. 3.6 – *Le modèle de Dorogovstev et Mendes : initialement le graphe est un triangle et, à chaque étape, un nouveau sommet est créé et relié aux deux extrémités d'un lien choisi au hasard.*

Ce modèle capture les trois propriétés principales des grands réseaux d'interactions, mais n'est pas vraiment paramétrable. Ainsi, le nombre de triangles du graphe est égal au nombre d'étapes t, le degré moyen est $2 \cdot \frac{(1+2t)}{t} \sim 4$ et, bien que ce modèle génère des graphes sans-échelle, ceux-ci n'ont aucun sommet de degré 1. De plus, le processus de construction induit des propriétés non souhaitées. Par exemple, les graphes obtenus sont planaires (ils peuvent être dessinés sur le plan sans croisement de liens).

Ce modèle manque donc fortement de réalisme mais c'est l'un des rares à capturer les trois propriétés de base des grands réseaux d'interactions : nous l'utiliserons donc souvent dans la suite à titre de comparaison avec d'autres modèles.

3.4 Autres modèles

Les modèles de Watts et Strogatz, d'Albert et Barabási, et ceux qui leur sont apparentés, ont été introduits pour capturer les propriétés principales des grands réseaux d'interactions, clustering et distribution des degrés principalement. Ils échouent tous deux à capturer simultanément ces deux propriétés et le modèle de Dorogovtsev et Mendes, qui lui les capture, est beaucoup moins réaliste.

Les principes de base qu'ils introduisent – présence de nombreux triangles pour augmenter le clustering et attachement préférentiel pour générer des graphes sans-échelle – sont souvent utilisés dans d'autres modèles comme des briques de base, comme nous le verrons à plusieurs reprises par la suite.

D'autre modèles n'utilisant pas vraiment ces principes ont aussi été introduits pour modéliser les grands réseaux d'interactions. Le plus étudié actuellement est le modèle dit de "fitness" que nous allons maintenant présenter. Nous allons aussi présenter d'autres modèles (déterministes) qui sont plus souvent utilisés à titre de comparaison avec d'autres modèles que dans des contextes pratiques.

3.4.1 Le modèle avec *fitness*

Le modèle avec attachement préférentiel génère des graphes sans-échelle mais, dans ce modèle, les sommets les plus anciens ont eu plus de temps pour collecter des liens. L'attachement préférentiel aidant, cet avantage ne cesse de s'accroître : il y a une très forte corrélation entre l'âge et le degré d'un sommet. Or, il s'avère que sur certains réseaux qui sont construits de manière itérative, des sommets très "jeunes" ont un fort degré malgré tout. Ceci est le cas, par exemple, pour *Web* dans lequel certains sites nouvellement créés sont rapidement très connus et très bien référencés [3].

Cette remarque a conduit à l'introduction d'un modèle très similaire à celui par attachement préférentiel mais dans lequel chaque sommet reçoit un poids (sa *fitness*) au moment de sa création. La *fitness* est choisie suivant une distribution fixée à l'avance et représente le potentiel d'attraction du sommet. De manière similaire au modèle par attachement préférentiel, les nouveaux sommets choisissent leurs voisins en fonction du degré et de la *fitness* de ces voisins [19] avec la règle suivante. La probabilité qu'un nouveau sommet choisisse de se connecter à un ancien sommet de degré k_i et de *fitness* η_i est :

$$p = \frac{k_i \eta_i}{\sum_{u \in V} k_u \eta_u}$$

Selon le choix de la distribution de la *fitness*, plusieurs phénomènes peuvent être observés (et prouvés). Tout d'abord, si la *fitness* est constante, on se ramène naturellement au modèle classique de l'attachement préférentiel. Dans le cas contraire, deux comportements se dégagent selon la distribution de la *fitness* :

- soit les sommets ayant la plus forte *fitness* deviennent les sommets les plus connectés et ce, même s'ils arrivent plus tard dans le système (*fit get rich*). Cela conduit à une distribution des degrés en loi de puissance dont l'exposant peut être calculé ;
- soit on assiste à un phénomène appelé condensation de Bose-Einstein, où le sommet de plus forte *fitness* va finalement obtenir une proportion linéaire des liens, quel que soit l'instant où il arrive.

Les conditions qui permettent d'aboutir à ces deux comportements sont très complexes et ne peuvent pas être résumées en quelques lignes. Nous renvoyons le lecteur intéressé à [19].

Ce modèle permet donc de généraliser le modèle à base d'attachement préférentiel et plusieurs autres. Une critique peut malgré tout être faite à ce modèle : la *fitness* est le seul paramètre qui va décider du degré d'un sommet. Ce modèle est donc presque identique au modèle avec distribution des degrés fixée dans lequel les degrés sont définis explicitement et pas de manière cachée comme avec la *fitness*. Ce modèle permet néanmoins d'unifier plusieurs modèles en un seul qui, de plus, est itératif.

3.4.2 Quelques modèles déterministes

Plusieurs modèles déterministes ont aussi été introduits, capturant les propriétés voulues, et qui sont assez simplement analysables. Ces modèles sont basés sur des processus de construction artificiels et ne peuvent pas être considérés comme réalistes.

Dans [15], un modèle est présenté pour générer simplement un graphe dont la distribution des degrés suit une loi de puissance d'exposant $1 + \frac{ln3}{ln2} \sim 2,58$. Dans ce modèle, à chaque étape, le graphe généré à l'étape précédente est reproduit en trois exemplaires qui sont connectés : la racine de l'un des trois graphes est reliée aux feuilles des deux autres (voir Figure 3.7).

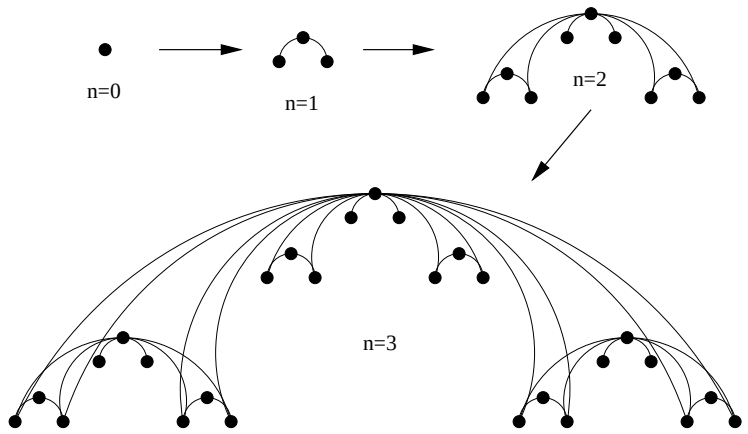

FIG. 3.7 – *Premières étapes du processus de construction déterministe. Chaque étape consiste en la duplication du graphe de l'étape précédente, la racine de l'un des graphes étant reliée à toutes les feuilles des deux autres graphes.*

Un autre modèle déterministe a été introduit pour générer des graphes sans-échelle qui ont aussi un fort clustering [36]. Ce modèle commence au temps 0 avec un triangle puis, à chaque étape, un sommet est créé pour chaque triangle du graphe et est relié aux trois sommets du triangle : c'est le modèle $K(3,t)$ (voir Figure 3.8). Ce modèle se généralise avec des cliques de taille q (modèle $K(q,t)$) : à chaque étape, un sommet est créé pour chaque clique de taille q et est relié aux q sommets de cette clique.

Selon la taille q des cliques considérées, ce modèle permet de générer des graphes sans-échelle pour lesquels, en fonction de q, le clustering varie entre 0.8 et 1 et l'exposant de la loi de puissance est compris entre 2 et $2,58$ (ces valeurs sont obtenues formellement).

3.5 Modèles spécifiques au graphe de l'Internet

Quelques grands réseaux d'interactions importants ont reçu une attention plus particulière avec la définition de modèles spécifiques. En particulier, de nombreux modèles ont été introduits pour plusieurs réseaux (*Web*, *Protéines*, *Co-signature*, etc.) mais présenter

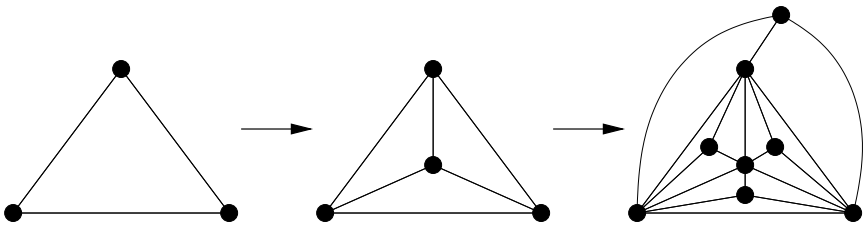

FIG. 3.8 – *Modèle déterministe engendrant des graphes avec une distribution des degrés en loi de puissance et un fort clustering. Chaque étape du processus de construction consiste à créer un nouveau sommet par clique de taille donnée (un triangle dans le cas présent) et à relier ce sommet à tous les sommets de la clique.*

tous ces modèles n'est pas l'objectif de ce chapitre. Nous allons simplement présenter ci-après succinctement quelques modèles proposés pour le graphe de l'Internet. Ils permettent d'illustrer l'approche et de montrer comment certains principes généraux présentés plus tôt sont utilisés dans un cas pratique.

La modélisation du réseau de l'Internet a des applications importantes dans de nombreux contextes tels que la simulation, la gestion du réseau ou le développement d'algorithmes spécifiques (routage avec qualité de service, communications de groupes, etc.). Des efforts de modélisation réaliste ont été menés depuis 1988 pour remplacer l'utilisation du modèle aléatoire. Pour avoir de plus amples informations, on peut se référer avantageusement à [27, 138].

Dans le modèle de Waxman [135], n sommets sont disposé de manière aléatoire dans un espace euclidien. Ensuite, les sommets sont reliés avec une probabilité $\alpha \cdot e^{-d/\beta L}$, où d est la distance euclidienne entre les sommets, L est le diamètre, et α et β sont deux paramètres du modèle : α permet de choisir le nombre de liens dans le graphe, β permet de régler le rapport entre le nombre de liens courts et de liens longs. Ce modèle est utilisé dans un modèle hiérarchique [138] : chaque sommet d'un graphe obtenu par le modèle de Waxman est remplacé par un réseau local, ce processus pouvant être répété plusieurs fois[3].

En 1999, Faloutsos et al. [48] ont mis en évidence le fait que de nombreuses lois de puissance apparaissent dans la description de l'Internet : distribution des degrés, des valeurs propres de la matrice de transition, du nombre de voisins à distance donnée, etc. La distribution des degrés en loi de puissance était complètement inattendue, même si elle est très discutée depuis [29]. Malgré tout, cela a mis en évidence le fait que les modèles utilisés jusque là étaient peu représentatifs de la réalité. De nouveaux modèles, ou des adaptations d'anciens, ont alors vu le jour pour combler cette lacune. L'attachement préférentiel est notamment beaucoup utilisé dans ce contexte, par exemple :

– BRITE [89] divise une surface en carrés et assigne à chacun un nombre de sommets suivant une distribution fixée (loi de Poisson, loi de puissance, etc.). Dans chaque

3. On peut aussi prendre un graphe issu d'un modèle quelconque.

carré les sommets sont placés de manière aléatoire et les liens sont placés en suivant des règles comme l'attachement préférentiel ou une connexion avec des sommets proches de préférence. Ce modèle est surtout destiné à modéliser l'Internet au niveau des systèmes autonomes ;
- INET [71] permet de générer un réseau de taille donnée en utilisant des équations de croissance obtenues à partir de l'évolution mesurée du niveau des systèmes autonomes de l'Internet entre 1997 et 2000. Ce modèle incorpore entre autres des règles similaires à l'attachement préférentiel ;
- GPL [27] est très similaire au modèle de Albert et Barabási mais, à chaque étape le choix est laissé entre l'ajout de sommet avec attachement préférentiel, ou l'ajout d'un lien. Ce modèle génère des graphes ayant un clustering plus élevé que le modèle original (ce résultat est purement expérimental) ;
- HOT [47] : dans ce modèle, un premier sommet privilégié, la source, est placé dans un carré unité, puis d'autres sommets sont ajoutés un par un. Chaque nouveau sommet u se relie au sommet v qui minimise une fonction linéaire de la distance euclidienne entre u et v, et de la distance en nombre de sauts à la source.

Il y encore de nos jours une intense activité autour de la modélisation de l'Internet [11, 139], pour laquelle aucune solution proposée n'est encore totalement satisfaisante.

3.6 Comparaison des différents modèles

Dans cette section, nous allons comparer les performances des modèles concernant les propriétés de base des grands réseaux d'interactions. Afin d'abréger les notations, nous allons utiliser dans toute la suite de la thèse les abréviations suivantes pour les cinq modèles que nous utiliserons fréquemment :
- ER pour le modèle aléatoire d'Erdös et Rényi dans lequel les liens existent avec une probabilité fixée [22, 46] ;
- MR pour le modèle aléatoire avec distribution des degrés donnée, introduit par Molloy et Reed [18, 92] ;
- AB pour le modèle d'Albert et Barabási avec attachement préférentiel [6, 8] ;
- WS pour le modèle de Watts et Strogatz qui consiste à recâbler les liens d'un anneau [134] ;
- DM, enfin, pour le modèle de Dorogovstev et Mendes qui consiste à relier les nouveaux sommets aux extrémités de liens pris au hasard [42].

Le Tableau 3.1 synthétise les propriétés capturées par les différents modèles. Le seul capturant les trois propriétés, celui de Dorogovstev et Mendes (DM), est malheureusement le moins réaliste des cinq.

Le Tableau 3.2 présente les performances de ces mêmes modèles dans des cas pratiques. Insistons sur le fait que ces modèles tentent de capturer les propriétés de manière *qualitative* : clustering élevé, distance moyenne courte et distribution des degrés en loi puissance.

	densité	distance moyenne	degrés	clustering
ER	OUI	OUI	NON	NON
MR	OUI	OUI	OUI	NON
AB	OUI	OUI	OUI	NON
WS	OUI	OUI	NON	OUI
DM	OUI	OUI	OUI	OUI

TAB. 3.1 – *Propriétés capturées par les principaux modèles pour les grands réseaux d'interactions.*

Leur objectif n'est pas d'obtenir des graphes ayant exactement les valeurs originales. Mais comme le Tableau 3.1 le laisse présager, les graphes obtenus sont significativement différents des graphes réels concernant au moins un de ces aspects.

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
n	75885	325729	392340	16401	9297	2113
m	357317	1090108	15038083	29552	392066	2203
α	2.5	2.3	2.2	2.4	1.8	2.4
c	0.171	0.466	0.785	0.638	0.822	0.153
c_{ER}	0.0001	0.00002	0.0002	0.0002	0.009	0.001
c_{MR}	0.0694	0.017	0.0057	0.001	0.26	0.007
c_{AB}	0.0024	0.0005	0.0015	0.003	0.028	0
c_{WS}	0.171	0.461	0.74 (*)	0.523 (*)	0.74 (*)	0.06 (*)
d	5.80	7	3.6	7.18	2.13	6.74
d_{ER}	5.25	5.47	2.97	7.57	2.06	10.4
d_{MR}	3.25	4.48	2.95	5.77	2.36	5.73
d_{AB}	4.15	5.1	2.93	5.5	2.38	8.15
d_{WS}	5.90	11.23	2559 (*)	2269 (*)	55.6 (*)	509 (*)

TAB. 3.2 – *Performance des principaux modèles pour les grands réseaux d'interactions. Pour chaque réseau, on trouve son nombre de sommets, son nombre de liens, la valeur de l'exposant de la loi de puissance, son clustering et sa distance moyenne. Nous donnons à titre de comparaison les valeurs du clustering et de la distance moyenne obtenus avec les modèles : le modèle ER (c_{ER} et d_{ER}), le modèle MR (c_{MR} et d_{MR}), le modèle AB (c_{AB} et d_{AB}) et le modèle WS (c_{WS} et d_{WS}). Dans les cas marqués avec une étoile (*), le clustering réel est trop élevé pour être obtenu avec le modèle WS. Nous avons donc utilisé dans ce cas le jeu de paramètres qui induit un clustering maximal, la distance moyenne n'est donc pas représentative.*

Conclusion

Ce survol des modèles pour les grands réseaux d'interactions montre que, malgré les récentes avancées effectuées dans ce domaine, il n'existe toujours pas de modèle satisfaisant pour produire des graphes ayant les trois principales propriétés des grands réseaux d'interactions : distance moyenne faible, fort clustering et distribution des degrés en loi de puissance.

Les méthodes à base de tirage aléatoire de graphes parmi ceux ayant un ensemble de propriétés choisies ont encore beaucoup à apporter. Elles permettent de faire des études rigoureuses et permettent de mieux analyser le comportement des grands réseaux d'interactions et l'influence des propriétés. Mais certaines propriétés sont très difficiles à capturer par ce biais. Ainsi, obtenir un graphe aléatoire avec un clustering donné est toujours un problème ouvert.

À l'opposé, les modèles basés sur l'itération d'un processus de construction sont souvent inspirés des processus de construction réels, et peuvent donc expliquer l'origine de certaines propriétés des grands réseaux d'interactions. Ces modèles ont aussi l'avantage d'être évolutifs et permettent donc de générer des graphes plus petits ou plus gros, mais avec des propriétés similaires, ou encore de faire évoluer des graphes existants. Malheureusement, il y a des inconvénients inhérents à ces processus. Ces modèles sont généralement assez complexes à analyser et certaines propriétés cachées peuvent être créées par le processus. De plus, le tirage n'est pas uniforme, il est donc assez difficile de distinguer les propriétés induites par le processus des propriétés générales de ces graphes.

Jusqu'à maintenant, aucun modèle n'a été présenté qui soit réaliste, qui capture les trois propriétés de base et qui soit suffisamment simple pour qu'on puisse les prouver.

Dans les deux chapitres qui suivent nous allons proposer plusieurs modèles qui constituent des avancées significatives dans cette direction. Ils sont issus d'un concept réaliste, sont simples et permettent de générer des graphes ayant les trois propriétés de base. Nous verrons aussi que ces modèles ne sont pas parfaits et ouvrent le voie pour de nouvelles améliorations.

Chapitre 4

Le modèle biparti

Dans ce chapitre, nous proposons un modèle aléatoire général pour tous les grands réseaux d'interactions. Il est issu de l'observation que tous les grands réseaux d'interactions ont une structure bipartie sous-jacente qui capture beaucoup de leurs propriétés. Le modèle consiste à engendrer un graphe biparti ayant la même structure que celle observée dans les grands réseaux d'interactions, puis à construire à partir de ce graphe un réseau classique.

Ce modèle permet de générer des graphes réalistes au regard des principales propriétés définies dans le Chapitre 2 et il est suffisament simple pour que l'on puisse prouver formellement ses propriétés, ce que nous ferons dans ce chapitre.

Nous commençons ici par décrire la structure bipartie sous-jacente à tous les grands réseaux d'interactions, et nous étudions ses principales propriétés. Nous proposons ensuite plusieurs variantes d'un modèle pour générer des structures biparties aléatoires similaires aux structures réelles. Nous prouvons ensuite que les réseaux obtenus à partir de ces structures biparties aléatoires ont les propriétés voulues. Nous terminons enfin en présentant les résultats expérimentaux obtenus avec ce modèle.

4.1 Graphes bipartis

Certains grands réseaux d'interactions ont, de par leur définition même, une structure bipartie sous-jacente. *Acteurs* est un graphe dont les sommets sont les acteurs de cinéma qui sont reliés s'ils ont joué dans un même film. On peut donc également le voir comme une structure reliant les acteurs aux films dans lesquels ils ont joué. À partir de cette structure bipartie, on peut reconstruire facilement *Acteurs*: deux acteurs sont reliés dans la vision classique s'ils sont reliés à un même film dans la vision bipartie. De la même manière, *Co-signature* est défini comme l'ensemble des liens entre auteurs ayant co-signé un article et peut être également vu comme un ensemble d'articles reliés aux auteurs qui y ont participé. Enfin, tout graphe de cooccurrence peut-être défini comme un ensemble de liens entre des mots et les phrases qui les contiennent.

Ces structures dans lesquelles les liens n'existent qu'entre deux catégories de sommets sont appelées graphes bipartis. Un graphe biparti non orienté est défini comme un triplet

$G = (\top, \bot, E)$ où $\top$ et $\bot$ sont deux ensembles disjoints de sommets et $E \subseteq \top \times \bot$ est l'ensemble des liens. Dans un tel graphe, les liens ne peuvent exister qu'entre des sommets de $\top$ et des sommets de $\bot$.

Ainsi, *Acteurs* peut être vu comme un graphe biparti dans lequel $\top$ est l'ensemble des films, $\bot$ est l'ensemble des acteurs et où on trouve un lien entre un film et un acteur si cet acteur a joué dans le film.

Si un graphe nous est donné sous sa forme bipartie $G = (\top, \bot, E)$, obtenir la vision classique $G' = (\bot, E')$, que l'on appelle projection sur $\bot$, est très aisé : $\{u, v\}$ appartient à E' si et seulement si u et v sont tous deux connectés à un même sommet de $\top$ dans G. La Figure 4.1 l'illustre sur un exemple. Dans la projection sur $\bot$ on peut remarquer que chaque sommet de $\top$ induit une clique (sous-graphe complet) entre tous les sommets de $\bot$ auquel il est relié. Par exemple, tous les acteurs ayant joué dans un film donné seront reliés dans la projection. La présence de cliques induit des voisinages très denses et nous verrons plus loin que ces cliques sont en grande partie responsables du fort clustering de ces grands réseaux d'interactions.

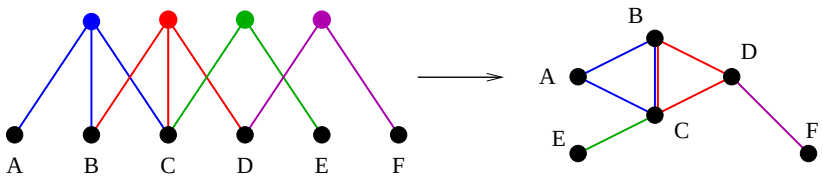

FIG. 4.1 – *Un graphe biparti et la vision classique du même graphe.*

Les graphes que l'on peut définir naturellement de manière bipartie sont effectivement obtenus sous cette forme : la base de données des films contient une liste de films avec pour chacun la liste des acteurs ayant joué dedans, donc directement les liens bipartis. Il en est de même pour les deux autres exemples. Il est donc possible de savoir le nombre de films dans lesquels un acteur a joué, ou le nombre d'acteurs d'un film. Cela définit deux distributions pour $\top$ et $\bot$:

$$\top_k = \frac{|\{t \in \top : d(t) = k\}|}{|\top|} \qquad \bot_k = \frac{|\{t \in \bot : d(t) = k\}|}{|\bot|}.$$

La Figure 4.2 présente ces distributions sur les trois grands réseaux d'interactions naturellement bipartis. Toutes ces distributions ont une propriété en commun : la présence de lois de puissance dans les distributions des degrés de $\bot$. Au contraire, les distributions des degré $\top$ sont de deux types : en loi de Poisson pour *Cooccurrence* et *Co-signature*, en loi à queue lourde pour *Acteurs*.

Effectuer le cheminement inverse, qui consiste à retrouver le graphe biparti à partir de sa projection, est en général impossible. Cela est dû au fait qu'il n'y a pas de bijection entre les graphes bipartis et leur projection : un graphe donné est en général la projection de plusieurs graphes bipartis. Donc, si l'on prend un graphe quelconque et que l'on essaye

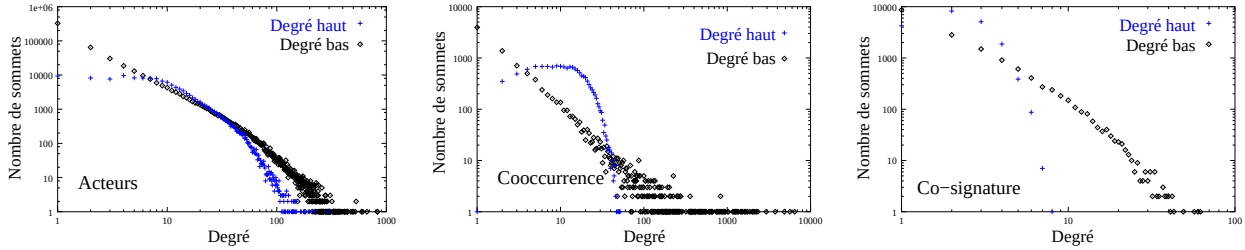

FIG. 4.2 – *Distributions des degrés de* $\top$ *et* $\bot$ *pour les graphes naturellement bipartis* Acteurs, Cooccurrence *et* Co-signature.

de construire un graphe biparti dont il soit la projection, il y a de multiples façons de procéder.

Retrouver une structure bipartie

La structure bipartie de certains graphes semble être la cause de plusieurs de leurs propriétés (telles que le clustering). Il paraît donc naturel de penser que tout graphe qui admet une vision bipartie ayant des propriétés similaires à celles que l'on observe naturellement aura les mêmes propriétés que les graphes naturellement bipartis.

Il est donc intéressant de pouvoir trouver une décomposition en graphe biparti de tout graphe, telle que la projection sur $\bot$ du biparti obtenu soit le graphe original. De plus, une telle décomposition devrait fournir, si possible, un biparti avec des propriétés similaires aux bipartis réels. On peut déjà donner quelques critères pour juger de la qualité d'une décomposition :

1. la décomposition doit couvrir le graphe, c'est-à-dire que la projection d'une décomposition doit être exactement le graphe original ;
2. la distribution de $\bot$ doit suivre une loi de puissance ;
3. la distribution de $\top$ peut suivre une loi de Poisson ou de puissance. Remarquons cependant que, dans tous les cas, on trouve des sommets dans $\top$ de degré relativement élevé, c'est-à-dire des cliques de taille non négligeable ;
4. enfin, le nombre de sommets dans $\top$ est du même ordre que le nombre de sommets de $\bot$. Cela signifie notamment que le degré moyen pour $\top$ et $\bot$ est faible.

Le problème posé est donc celui de la couverture d'un graphe par des cliques (*clique covering problem*) : en effet, les sommets de $\top$ que l'on essaie de trouver sont des cliques dans le graphe original, et tout le graphe doit être recouvert par de telles cliques. Il existe de nombreuses méthodes triviales pour couvrir un graphe par des cliques si l'on ne se fixe pas de contraintes. Par exemple, on peut considérer que chaque lien du graphe est une clique de taille 2, ou encore qu'une bonne couverture contient l'ensemble de toutes les cliques maximales du graphe. Ces deux approches ne sont pas vraiment satisfaisantes : la première ne va construire que des cliques de taille 2, alors que la seconde risque de construire un nombre de cliques exponentiel [96]. Cela contredit les critères 2 et 3 présentés plus haut.

L'approche intermédiaire que nous avons choisie consiste à ne garder qu'une clique de taille maximale pour chaque lien, au moyen d'un algorithme que nous décrirons plus loin. Si un lien est contenu dans plusieurs cliques de taille maximale, on en choisit une aléatoirement. Cette décomposition satisfait trois des critères que nous avons fixés :

- chaque lien est contenu dans au moins une clique, c'est donc bien une couverture ;
- le nombre de cliques découvertes est au plus égal au nombre de liens du graphe et est donc du même ordre que le nombre de sommets, si le degré moyen est faible ;
- enfin, la plupart des grandes cliques seront découvertes.

Cet algorithme non déterministe va donc retourner un ensemble de cliques et le graphe biparti associé sera $B = (\top, \bot, E)$, où $\top$ est l'ensemble des cliques et $\bot$ est l'ensemble des sommets du graphe initial. Une clique sera reliée à un sommet si le sommet appartient à la clique.

Si l'on applique cette procédure à l'exemple de la Figure 4.1, on obtient deux cliques de taille 2 ($\{C, E\}$ et $\{D, F\}$). Le lien $\{B, C\}$ est contenu dans $\{A, B, C\}$ et $\{B, C, D\}$. Notre algorithme choisit donc une de ces deux cliques de manière aléatoire. Remarquons cependant que, quelle que soit la clique choisie à cette étape, les deux cliques seront découvertes grâce aux liens $\{A, B\}$ et $\{B, D\}$. Ceci donnera finalement une décomposition unique qui est, dans ce cas particulier, le graphe biparti original.

Calcul effectif de la décomposition

La décomposition en cliques n'est pas, en général, un problème facile. Vouloir minimiser le nombre de cliques est NP-complet (*minimal clique covering problem*) [94, 107]. De même, calculer des cliques maximales dans un graphe, ou la clique maximale contenant un lien, est également NP-complet [2, 23]. Vouloir utiliser un tel processus de décomposition sur des graphes de grande taille serait donc irréalisable dans le cas général.

Malgré tout, il existe de nombreuses heuristiques pour de tels calculs. Dans le cas qui nous occupe, on peut utiliser les remarques suivantes : une clique maximale qui contient un lien $\{u, v\}$ dans G est aussi une clique maximale dans le sous-graphe induit par $N(u, v) \cup \{u, v\}$, où $N(u, v)$ est le voisinage des sommets u et v. On peut donc se contenter de chercher la clique maximale, $\mathcal{C}$, dans le sous-graphe induit par $N(u, v)$, et alors la clique que l'on recherche est $\mathcal{C} \cup \{u, v\}$. La Figure 4.3 illustre ce processus.

Ce processus de décomposition est NP-complet, donc se limiter au voisinage d'un lien ne permet pas, dans le cas général, de trouver les cliques maximales en temps polynômial. Heureusement, dans la plupart des grands réseaux d'interactions observés, les sous-graphes induits par $N(u, v)$ pour tous les liens $\{u, v\}$ sont en général soit denses et très petits, soit plus grands mais très peu denses (Figure 4.4). Ceci est principalement dû au fort clustering et à la distribution des degrés en loi de puissance. En pratique, on sait calculer les cliques maximales de manière très efficace dans ces cas.

Seuls deux cas posent problème : *Acteurs* et *Cooccurrence* pour lesquels certains voisinages sont à la fois denses et de grande taille. La décomposition sur ces graphes est donc beaucoup plus longue, mais reste encore calculable.

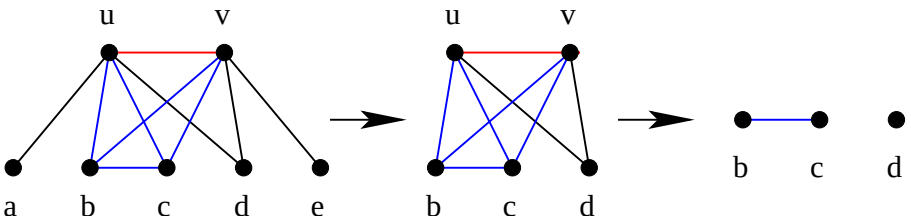

FIG. 4.3 – *Soit $G = (V, E')$ un graphe dans lequel on cherche une clique maximale contenant $\{u, v\}$. Cette clique est nécessairement contenue dans le sous-graphe induit par $N(u, v) \cup \{u, v\} = \{b, c, d, u, v\}$. Il suffit de calculer la clique maximale dans le sous-graphe induit par $N(u, v) = \{b, c, d\}$, ce qui donne $\{b, c\}$. La clique recherchée est donc $\{u, v, b, c\}$.*

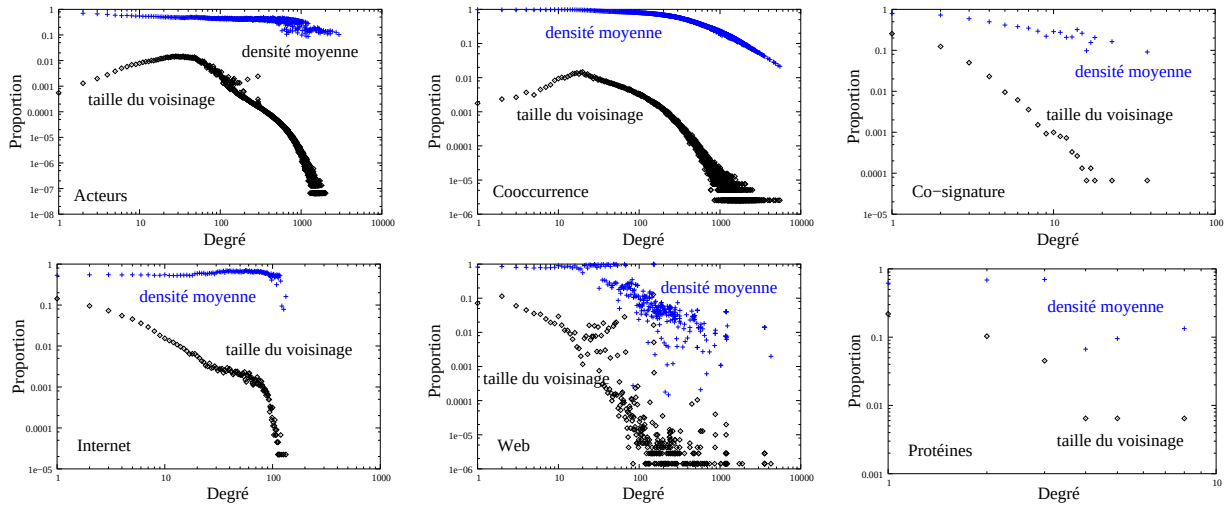

FIG. 4.4 – *Distribution de la taille des voisinages, $N(u, v)$, pour tous les liens (u, v), et densité moyenne de ces voisinages.*

Propriétés de la décomposition

Le but de l'algorithme de décomposition est de construire un réseau biparti ayant des propriétés similaires aux bipartis originaux. Une manière de l'évaluer est de prendre un graphe naturellement biparti $G = (\top, \bot, E)$, de le projeter sur $\bot$ et de décomposer cette projection. Ainsi, il est possible de comparer le biparti original et le biparti reconstruit. La Figure 4.5 présente les distributions des degrés de $\top$ et $\bot$ pour le graphe biparti initial et pour celui reconstruit par l'algorithme de décomposition.

Les distributions montrent que les degrés de $\bot$ sont quasiment identiques entre les graphes originaux et les décompositions. Pour les degrés de $\top$, il y a des différences pour les trois graphes décomposés. Tout d'abord, le processus de décomposition ne construit aucune clique de taille 1 puisqu'il cherche une clique maximale contenant chaque lien. De nombreuses cliques de taille 2 (et de taille faible en général) n'ont pas été découvertes, ce

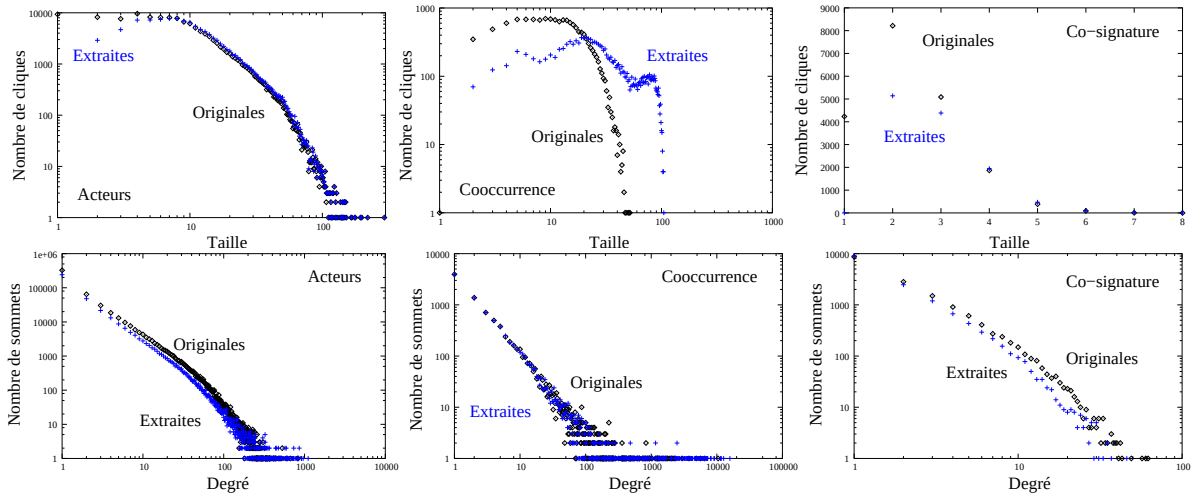

FIG. 4.5 – *Distributions des degrés de* $\top$ *(en haut) et de* $\bot$ *(en bas) originales et obtenues par l'algorithme de décomposition pour* Acteurs, Cooccurrence *et* Co-signature.

qui signifie que ces cliques ne sont en fait pas maximales dans la projection sur $\bot$. Cela arrive fréquemment pour deux raisons : tout d'abord, une clique de taille faible peut être strictement incluse dans une clique de taille plus grosse (le voisinage d'un sommet de $\top$ est inclus dans le voisinage d'un autre). Ensuite, l'union de petites cliques peut créer des cliques plus grosses. Ce sont ces cliques "artificielles" et non les plus petites qui sont alors retournées par l'algorithme. Ceci explique aussi l'apparition de sommets dans $\top$ ayant un degré supérieur au degré maximal dans le biparti original, notamment sur *Cooccurrence*.

Malgré tout, les distributions des degrés après décomposition sont très similaires aux distributions originales concernant les deux points primordiaux : certains sommets ont un fort degré et le nombre de sommets dans $\top$ est très similaire entre le biparti original et le biparti reconstruit.

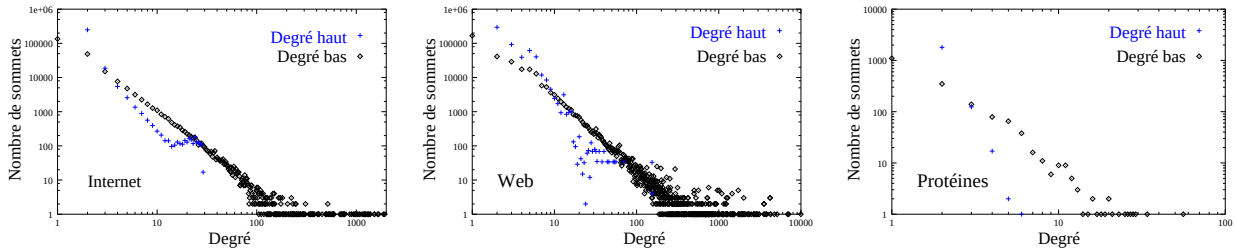

FIG. 4.6 – *Distributions des degrés de* $\top$ *et* $\bot$ *pour les visions biparties de* Internet, Web *et* Protéines *obtenues par l'algorithme de décomposition.*

Il est maintenant possible de décomposer tout graphe en un graphe biparti grâce à l'algorithme que nous avons décrit, même si ce graphe n'a *a priori* aucune structure bipartie

naturelle. La Figure 4.6 présente les distributions des degrés pour $\top$ et $\bot$ des décompositions de *Internet*, *Web* et *Protéines*.

Ces distributions ont les mêmes propriétés que les distributions vues précédemment : les distributions des degrés de $\bot$ sont en loi de puissance et les distributions des degrés de $\bot$ montrent l'existence de cliques de grande taille.

Observer les grands réseaux d'interactions sous l'angle biparti et sous l'angle classique permet de s'interroger sur les éventuelles relations existant entre ces deux visions d'un même graphe. On peut, par exemple, s'intéresser aux liens entre le degré d'un sommet dans $\bot$ et dans la projection. Il faut remarquer que le degré d'un sommet dans la projection sur $\bot$ est égal à la somme des degrés (moins 1) des sommets de $\top$ auxquels il est relié dans le graphe biparti, moins un éventuel recouvrement entre les cliques associées. Ainsi, si un sommet u est contenu dans deux cliques de taille 4, on peut espérer que son degré dans la projection soit 6, à moins que les deux cliques n'aient un autre sommet que u en commun.

La Figure 4.7 présente les corrélations entre les degrés d'un sommet dans le biparti et dans la projection. Les courbes montrent une corrélation qui laisse à penser que le degré d'un sommet est lié au nombre de cliques qui le contiennent. Nous montrerons plus loin que c'est effectivement le cas.

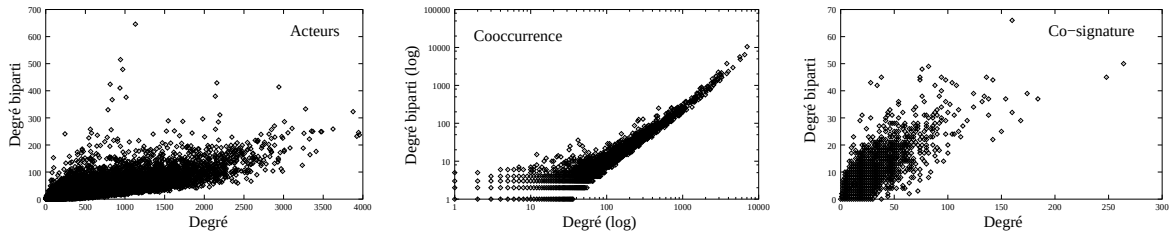

FIG. 4.7 – *Corrélations entre les degré de sommets de $\bot$ dans le biparti et dans la projection. Chaque point (x, y) correspond à un sommet de degré x dans le graphe et de degré y dans $\bot$.*

Tous les grands réseaux d'interactions ont donc une structure bipartie sous-jacente non triviale pouvant être calculée en utilisant le processus de décomposition que nous avons décrit. La question qui se pose maintenant est de savoir si les propriétés des grands réseaux d'interactions sont une conséquence de cette structure sous-jacente. Nous tentons de répondre à cette question dans les sections suivantes.

4.2 Deux modèles bipartis

Nous allons maintenant utiliser la structure bipartie sous-jacente dans tous les grands réseaux d'interactions pour proposer un modèle capturant leurs principales propriétés : clustering, distance moyenne et distribution des degrés.

Comme discuté dans le Chapitre 3, il y a principalement deux méthodes pour proposer un modèle : par tirage aléatoire ou en simulant un processus d'évolution. Notre objectif ici est de générer un graphe biparti aléatoire similaire aux vrais bipartis et dont la projection sur $\perp$ ait de bonnes propriétés. On peut soit tirer aléatoirement un graphe biparti en fixant les distributions des degrés $\top$ et $\perp$, soit utiliser un processus de construction biparti similaire au processus réel. Dans ce qui suit, nous allons présenter deux modèles sur ces principes, puis les évaluer formellement et expérimentalement.

Tirage aléatoire de graphe biparti

Générer de manière uniforme un graphe biparti avec des distributions de degrés données peut être fait de la manière suivante [32, 105, 106] :

1. créer les sommets de $\top$ et $\perp$ et assigner à chacun un degré choisi suivant la distribution ;
2. créer pour chaque sommet autant de demi-liens que son degré ;
3. relier de manière aléatoire des demi-liens de $\top$ à des demi-liens de $\perp$ pour former des liens.

La Figure 4.8 illustre ce principe.

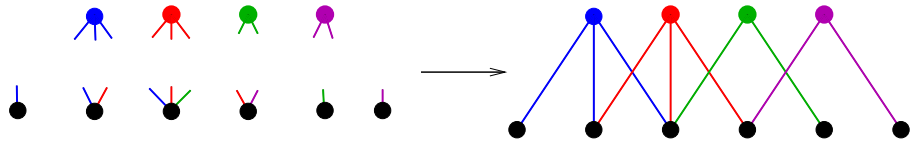

FIG. 4.8 – *Construction d'un graphe biparti aléatoire avec distributions des degrés fixées : on assigne aux sommets un nombre de demi-liens choisi suivant la distribution, puis ceux-ci sont appariés de manière aléatoire.*

Ce processus permet de générer des graphes bipartis aléatoires tirés uniformément dans l'ensemble des graphes bipartis ayant les distributions données. Ce processus ne fonctionne que si le nombre de demi-liens pour $\top$ et $\perp$ sont égaux après la deuxième étape. S'assurer que les nombres de demi-liens de $\top$ et $\perp$ sont égaux peut se faire de l'une des façons suivantes :

- si les distributions sont explicites, en prenant les distributions originales, par exemple, alors la compatibilité est assurée automatiquement ;
- si, au contraire, les distributions sont données de façon implicite sous forme d'une loi (loi de puissance, par exemple) alors, avant même le tirage, il faut s'assurer que les distributions sont compatibles : le produit du nombre de sommets par leur degré moyen doit être le même pour $\top$ et $\perp$ (en pratique on n'a pas besoin d'une égalité stricte) ;
- pendant le tirage effectif du degré des sommets, il peut arriver que le nombre de demi-liens ne soit malgré tout pas égal pour $\top$ et $\perp$. Dans ce cas, on retire le degré d'un

sommet pris au hasard de chaque côté [105, 106]. Avec des distributions compatibles, le nombre de retirages à faire est très faible avant d'obtenir un nombre de demi-liens similaire des deux cotés.

Cette méthode permet d'obtenir très simplement des graphes bipartis aléatoires avec distributions fixées [32].

Modèle incrémental avec attachement préférentiel

Pour *Acteurs*, à chaque fois qu'un film est tourné, un certain nombre d'acteurs connus est choisi (une proportion λ de tous les acteurs des films) et un certain nombre d'acteurs y jouent leur premier rôle. Un processus incrémental pour modéliser *Acteurs* pourrait donc consister à créer de nouveaux films en choisissant aléatoirement des acteurs connus et des acteurs débutants pour chacun. Dans le cas d'un modèle biparti, cela se traduit par la procédure suivante : à chaque étape, on crée un nouveau sommet dans $\top$ et on le relie à certains sommets de $\bot$ déjà existants et à certains nouveaux sommets qu'on ajoute à $\bot$.

Comme nous l'avons déjà remarqué, la distribution des degrés de $\bot$ suit une loi de puissance pour la plupart des grands réseaux d'interactions. On peut donc utiliser l'attachement préférentiel pour retrouver cette distribution en répétant la procédure suivante :

- ajouter un sommet u dans $\top$ et choisir son degré d en accord avec une distribution donnée (qui varie selon les grands réseaux d'interactions à modéliser) ;
- pour chacun des d liens du sommet u, ajouter un lien vers un sommet de $\bot$ selon l'attachement préférentiel avec probabilité λ ou, avec probabilité $1 - \lambda$, créer un nouveau sommet dans $\bot$ et le relier à u.

Le paramètre λ, que l'on appelle *taux de recouvrement*, est donc la proportion de sommets de $\bot$ préexistants à laquelle un nouveau sommet de $\top$ est relié. Il n'est pas toujours possible de connaître exactement l'ordre dans lequel les cliques sont créées dans les vrais réseaux bipartis et donc de connaître ce que vaudrait le paramètre λ pour chaque clique, mais sa valeur moyenne peut-être calculée par $\lambda = 1 - \frac{|\bot|}{\sum d_\top}$. Ceci donne des valeurs de 0.733 pour *Acteurs* (chaque film contient en moyenne un quart de nouveaux acteurs), 0.877 pour *Co-signature* et 0.949 pour *Cooccurrence*.

Avec cette définition du taux de recouvrement, on peut remarquer que $1 - \lambda$ est simplement l'inverse du degré moyen de $\bot$, étant donné que $\sum d_\top = \sum d_\bot$. Un grand recouvrement correspond donc à un grand degré moyen pour $\bot$: peu de sommets sont crées à chaque étape et donc les sommets existants accumulent un plus grand nombre de liens.

À chaque étape du processus de construction, le réseau biparti a les distributions des degrés suivantes : la distribution pour $\top$ est celle choisie par construction et celle pour $\bot$ est une loi de puissance.

Une version non bipartie

Les deux modèles introduits précédemment peuvent être définis de manière non bipartie par un processus où des cliques sont ajoutées au fur et à mesure directement dans la

projection. Dans le biparti, relier un sommet de $\top$ à plusieurs sommets de $\bot$ équivaut à ajouter une clique sur ces sommets dans la projection. Il est donc possible de définir les modèles de manière classique : initialement, le graphe contient n sommets déconnectés et, à chaque étape, d sommets sont choisis, suivant la distribution de $\top$, et sont tous reliés les uns aux autres (voir la Figure 4.9).

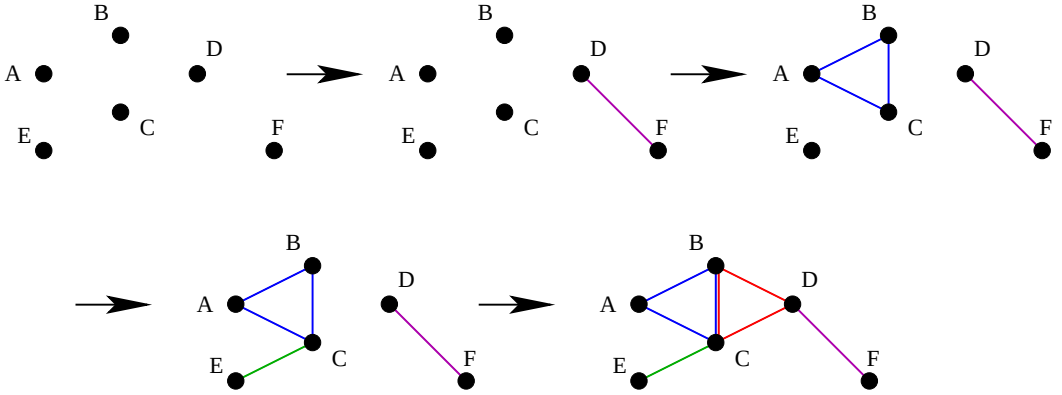

FIG. 4.9 – *Version classique du modèle : les sommets sont initialement déconnectés et, à chaque étape, une clique est ajoutée entre un ensemble de sommets choisis au hasard.*

La façon dont les d sommets sont choisis – uniformément, par attachement préférentiel ou avec toute autre loi – influence le graphe résultant. Ainsi, si les sommets sont choisis uniformément au hasard, ce modèle est équivalent à un modèle biparti avec une loi de Poisson pour $\bot$. En particulier, si d est toujours pris égal à 2, alors un seul lien est ajouté à chaque étape, ce qui revient au modèle de graphe aléatoire classique $\mathcal{G}_{n,m}$ [22, 46].

Il est possible de définir un modèle incrémental de manière similaire, modèle dans lequel de nouveaux sommets sont crées à chaque étape et des cliques ajoutées en choisissant un certain nombre d'anciens et de nouveaux sommets. Il est bien sûr possible de choisir les anciens sommets avec la règle de l'attachement préférentiel. Le modèle AB est ainsi un cas particulier de ce modèle si les cliques ajoutées sont toutes de taille 2 et si le taux de recouvrement vaut $0,5$.

Finalement, les deux modèles bipartis génèrent des graphes bipartis similaires à ceux obtenus à partir des grands réseaux d'interactions (du moins en termes de distribution de $\top$ et $\bot$). Dans les sections suivantes, Nous allons étudier formellement et expérimentalement les propriétés de la projection sur $\bot$ (distance moyenne, distribution des degrés et clustering) et nous montrerons qu'elles sont conformes à celles des grands réseaux d'interactions initiaux.

4.3 Analyse du modèle avec distributions données

Dans cette section, nous allons nous attacher à donner des preuves formelles pour les principales propriétés de la projection sur $\bot$ d'un graphe biparti aléatoire avec distributions

des degrés données.

On va par la suite se restreindre au cas où la projection sur $\perp$ est connexe. Il est montré dans [105] que, sous certaines conditions raisonnables concernant les distributions de degrés, c'est effectivement le cas.

Distribution des degrés de la projection

Nous allons tout d'abord nous intéresser à la distribution des degrés de la projection sur $\perp$ d'un graphe biparti aléatoire $G = (\top, \perp, E)$. Étant donné un sommet $u \in \perp$, on note $d(u)$ son degré dans le graphe biparti et $d_U(u)$ son degré dans la projection. On s'intéresse donc à la distribution de $d_U(u)$.

Lemme 4.3.1 *Soit un sommet $u \in \perp$. Le nombre de sommets dans $\perp$ ayant un voisin dans $\top$ en commun avec u,* i.e. *$d_U(u)$, vaut :*

$$\frac{d(u)}{|\top|} \cdot \sum_{t \neq u, t \in \perp} d(t) + \mathcal{O}\left(\frac{d(u)^2}{|\top|^2} \cdot \sum_{t \neq u} d(t)^2\right)$$

Preuve : La valeur exacte de $d_U(u)$ est donnée par :

$$d_U(u) = \sum_{t \neq u} \left(1 - \frac{\binom{|\top| - d(u)}{d(t)}}{\binom{|\top|}{d(t)}}\right)$$

car la probabilité qu'un sommet t ait un voisin en commun avec u dépend seulement du degré des deux sommets et du nombre de sommets dans $\top$. Pour simplifier cette formule, on peut approximer le quotient $\binom{|\top| - d(u)}{d(t)} / \binom{|\top|}{d(t)}$ par :

$$\begin{aligned}
\frac{\binom{|\top| - d(u)}{d(t)}}{\binom{|\top|}{d(t)}} &= \frac{(|\top| - d(u))!(|\top| - d(t))!}{|\top|!(|\top| - d(u) - d(t))!} \\
&\sim \frac{(|\top| - d(t))^{d(u)}}{|\top|^{d(t)}} \\
&\sim 1 - \frac{d(t)d(u)}{|\top|} + \mathcal{O}\left(\left(\frac{d(t)d(u)}{|\top|}\right)^2\right)
\end{aligned}$$

et donc :

$$\begin{aligned}
d_U(u) &\sim \sum_{t \neq u} \left(\frac{d(t)d(u)}{|\top|} + \mathcal{O}\left(\left(\frac{d(t)d(u)}{|\top|}\right)^2\right)\right) \\
&\sim \frac{d(u)}{|\top|} \sum_{t \neq u} d(t) + \mathcal{O}\left(\frac{d(u)^2}{|\top|^2} \sum_{t \neq u} d(t)^2\right)
\end{aligned}$$

ce qui est la formule annoncée. □

En utilisant ce lemme, il est possible de calculer la probabilité qu'un sommet u dans la projection sur $\perp$ ait un degré donné k si la distribution des degrés de $\perp$ suit une loi de puissance d'exposant β (ce qui est toujours le cas que nous considérons) :

$$P[d_U(u) = k] \sim P[d(u) = \frac{n}{\sum_{t \neq u} d(t)} \cdot k]$$
$$\sim \frac{1}{(\sum_{t \neq u} d(t)) \cdot k)^{\beta}} \sim k^{-\beta}$$

On prouve donc bien que, pour autant que la distribution des degrés de $\perp$ suive une loi de puissance, la distribution des degrés dans la projection sur $\perp$ suivra aussi une loi de puissance de même exposant.

Distance moyenne

Afin d'étudier la distance moyenne dans la projection sur $\perp$ d'un graphe obtenu avec le modèle biparti aléatoire, nous utilisons un résultat de L. Lu qui concerne le diamètre (*i.e.* la plus grande distance entre deux sommets) de certains graphes [81] :

Théorème 4.3.2 *Soit $G = (V, E)$ un graphe à n sommets dont les sommets sont pondérés avec des poids $w_1, \cdots, w_n$, tels que chaque lien $\{i, j\}$ apparaisse avec probabilité $w_i \cdot w_j \cdot p$. Si les poids sont tels que les degrés des sommets dans V suivent une loi de puissance d'exposant β strictement supérieur à 2, alors le diamètre de g vaut $\Theta(\log(n))$ avec forte probabilité.*

Ce théorème, associé au lemme concernant la distribution des degrés de la projection sur $\perp$ amène au résultat suivant :

Théorème 4.3.3 *Soit $G = (\top, \perp, E)$ un graphe biparti dont la distribution de degrés de $\perp$ suit une loi de puissance d'exposant strictement supérieur à 2, alors la distance moyenne dans la projection sur $\perp$ de G vaut $\mathcal{O}(\log(|\perp|))$ avec forte probabilité.*

Preuve : Soient u et v deux sommets de $\perp$. La probabilité qu'ils soient reliés dans la projection sur $\perp$ est égale à la probabilité qu'ils soient connectés à un même sommet de $\top$ dans G. Cette probabilité est exactement proportionnelle à $d_{\perp}(u) \cdot d_{\perp}(v)$, il est donc possible d'appliquer le Théorème 4.3.2 en considérant que le poids de chaque sommet est son degré, ce qui assure l'hypothèse concernant la probabilité de connexion. Dès lors que la distribution des degrés de $\perp$ suit une loi de puissance d'exposant strictement supérieur à 2, le diamètre de la projection sur $\perp$ est presque sûrement $\mathcal{O}(\log(|\perp|))$. Le diamètre est une borne supérieure pour la distance moyenne qui varie donc aussi logarithmiquement. □

Clustering

Nous donnons ci-dessous une borne inférieure sur le clustering d'un graphe G', projection sur $\bot$ d'un graphe biparti aléatoire. Nous montrons que sous des hypothèses raisonnables concernant les distribution des degrés de $\top$ et $\bot$, le clustering est borné inférieurement par une valeur indépendante de la taille du graphe, ce qui prouve que le modèle biparti aléatoire génère un clustering non nul.

Une approximation du clustering de tels graphes a été donnée dans [105, 106]. Ici, nous donnons une borne inférieure indépendante de la taille du graphe. Les deux résultats se complètent, le premier donnant une valeur attendue qui est effectivement très proche de la valeur réelle, alors que notre approche garantit que le clustering ne tend pas vers 0 quand la taille du graphe augmente.

Tout d'abord, on peut remarquer que la probabilité que deux sommets dans $\top$ aient plus d'un voisin dans $\bot$ en commun tend vers 0 quand la taille du graphe augmente. Par la suite, nous allons donc considérer des sommets dans la projection sur $\bot$ et supposer que leur voisinage est composé de cliques disjointes, ce qui est vrai pour les graphes de grande taille. Nous allons prouver que le clustering d'un sommet de $\bot$ dans la projection peut être borné inférieurement par une valeur qui ne dépend que de son degré.

Lemme 4.3.4 *Soient $G = (\top, \bot, E)$ et $b \in \bot$. Soit $\top_{>2}$ l'ensemble des sommets de $\top$ voisins de b de degré strictement supérieur à 2, et $\bot_{>2}$ le voisinage de $\top_{>2}$. Soient p la proportion de voisins de b appartenant à $\bot_{>2}$ et α le clustering de b restreint à $\bot_{>2}$.*

Alors le clustering de b peut être approximé par $p^2 \cdot \alpha$.

Preuve : Le voisinage de b est réparti entre les sommets de $\bot_{>2}$ et des sommets qui n'ont qu'un lien vers b. Le fait que le clustering de b restreint à $\bot_{>2}$ soit α implique que $|\triangle_{\bot_{>2}}(b)| = \alpha \cdot \binom{p \cdot d}{2}$. Si l'on considère le voisinage complet de b au lieu de $\bot_{>2}$, alors le nombre de triangles ne change pas et le nombre de triplets augmente :

$$\begin{aligned} c(b) &= \frac{\alpha \cdot \binom{p \cdot d}{2}}{\binom{d}{2}} \\ &= \alpha \cdot \frac{p \cdot d((p \cdot d - 1)}{d(d-1)} \\ &\sim p^2 \cdot \alpha \end{aligned}$$

ce qui est la formule annoncée. □

Donc tant que p est constant non nul, on peut négliger les sommets dans $\top$ de degré 2 pour le calcul du clustering d'un sommet donné : si le clustering du graphe quand on néglige ces sommets est non nul, alors le clustering du graphe sera non nul.

Lemme 4.3.5 *Si b n'est relié qu'à des sommets de $\top$ de degré au moins 3, alors :*

$$c(b) \geq \frac{1}{2 \cdot d(b) - 1}$$

Preuve : Supposons que b ne soit connecté qu'à deux sommets de $\top$, t_1 et t_2, de degré au moins 3 (le cas général sera traité plus bas). Alors le clustering de b est :

$$c(b) = \frac{\binom{d(t_1)-1}{2} + \binom{d(t_2)-1}{2}}{\binom{d(t_1)+d(t_2)-2}{2}}$$

Supposons maintenant que b est relié à t_2 et t_1' tel que $d(t_1') = d(t_1) + 1$, alors le clustering de b vaut :

$$c'(b) = \frac{\binom{d(t_1)+1-1}{2} + \binom{d(t_2)-1}{2}}{\binom{d(t_1)+d(t_2)-1}{2}}$$

et :

$$\begin{aligned} c'(b) - c(b) &= \frac{2 \cdot (d(t_2) - 1)}{(d(t_1) + d(t_2) - 2) \cdot (d(t_1) + d(t_2) - 3)} \\ &> 0. \end{aligned}$$

Ceci peut être généralisé pour montrer que le clustering croît avec le degré de t_1 et t_2. Une borne inférieure est donc obtenue quand t_1 et t_2 ont le plus petit degré admissible, à savoir 3.

Ce résultat peut être étendu dans le cas où b a plus de deux voisins :

$$\begin{aligned} c(b) &= \frac{\sum_{t_i} \binom{d(t_i)-1}{2}}{\binom{\sum_{t_i}(d(t_i)-1)}{2}} \\ &\geq \frac{\sum_{t_i} \binom{3-1}{2}}{\binom{\sum_{t_i}(3-1)}{2}} \geq \frac{1}{2 \cdot d(b) - 1} \end{aligned}$$

cette borne inférieure est celle annoncée dans le lemme. □

Le clustering de la projection peut maintenant être facilement approximé :

$$c(G') \sim \frac{1}{n} \sum_{b \in \bot} \frac{1}{2d(b) - 1}$$

Tant qu'il y a un nombre linéaire $c \cdot n$ de sommets de $\bot$ de degré 2, cette somme varie linéairement avec n :

$$\sum_{b \in \bot} \frac{1}{2 \cdot d(b) - 1} \geq \sum_{b, d(b)=2} \left(\frac{1}{2 \cdot 2 - 1} \right) = \frac{c \cdot n}{3}.$$

La borne inférieure pour le clustering est donc indépendante de n. En particulier, pour les graphes ayant une distribution des degrés en loi de puissance, le nombre de sommets de degré 2 est de l'ordre de $N \cdot 2^{-\alpha}$. Cet argument aurait été valable en considérant un degré autre que 2.

D'autre part, nous n'avons pas considéré les sommets de degré 2 dans la dernière formule, car nous pouvions l'ignorer d'après le Lemme 4.3.4. Nous devons cependant nous assurer que le nombre de tels voisins d'un sommet dans $\top$ représente au plus une fraction constante (ne tendant pas vers 1) des voisins. Ceci est effectivement le cas en pratique pour la plupart des distributions et en particulier dans tous les cas réels. Sous ces conditions, le clustering d'un graphe obtenu par projection d'un biparti aléatoire est donc supérieur à une constante non nulle, et ce, indépendemment de la taille du graphe.

4.4 Résultats expérimentaux

Les résultats formels de la section précédente donnent une intuition assez précise du comportement du modèle biparti aléatoire avec distribution des degrés donnée. Il est aussi possible de vérifier expérimentalement ces propriétés en générant des graphes avec les modèles à partir des distributions observées en pratique sur les grands réseaux d'interactions réels. C'est ce que nous allons faire dans cette section sur les six graphes habituels en utilisant le modèle par tirage aléatoire et le modèle incrémental avec attachement préférentiel.

Il peut paraître inutile de faire des expériences alors que les preuves de la section précédente garantissent ces propriétés. Or, ces preuves apportent des résultats approximés ou des bornes; il s'agit donc de vérifier qu'en pratique, les constantes négligées n'influent pas significativement sur les propriétés effectives des graphes générés.

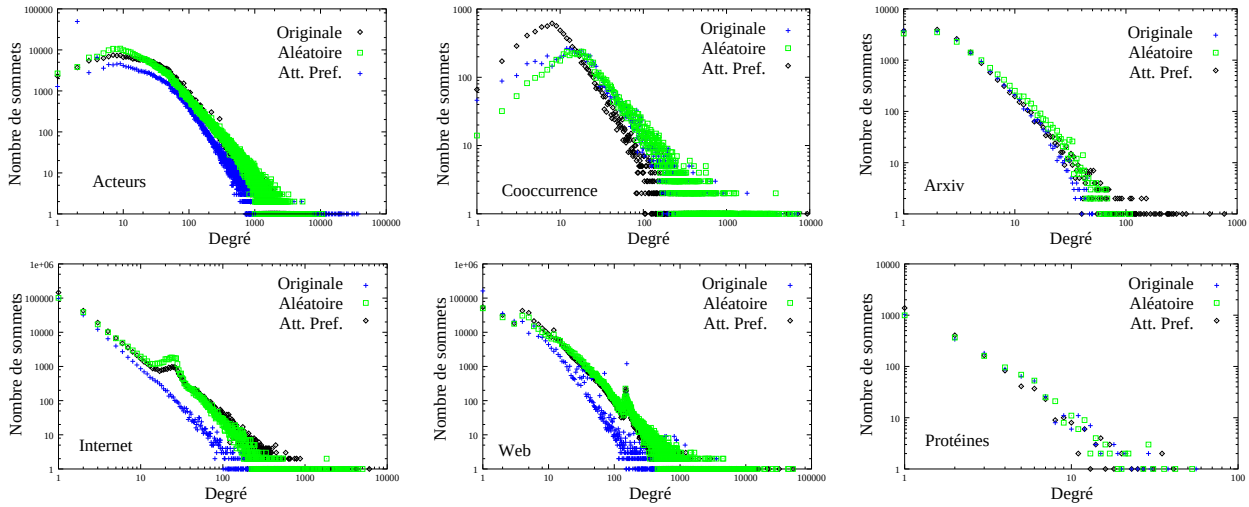

FIG. 4.10 – *Distribution des degrés originale et distributions des projections sur $\perp$ obtenues avec le modèle biparti aléatoire et le modèle biparti incrémental.*

La Figure 4.10 montre une comparaison entre les distributions des degrés originales et celles obtenues avec les deux modèles bipartis. Les Tableaux 5.4 et 4.2, quant à eux, nous donnent les valeurs obtenues pour le clustering et la distance moyenne.

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
c	0.171	0.466	0.785	0.638	0.822	0.153
c_{ER}	0.0001	0.00002	0.0002	0.0002	0.009	0.001
c_{MR}	0.0694	0.017	0.0057	0.001	0.26	0.007
c_{AB}	0.0024	0.0005	0.0015	0.003	0.028	0
c_{WS}	0.171	0.461	0.74 (*)	0.523 (*)	0.74 (*)	0.06 (*)
c_{rb}	0.32	0.663	0.767	0.542	0.831	0.187
c_{gb}	0.65	0.708	0.793	0.632	0.768	0.244
borne	0.0244	0.0404	0.024	0.0373	0.0375	0.0254

TAB. 4.1 – *Clustering obtenu avec les modèles classiques et les modèles bipartis. Pour chaque graphe, on peut trouver son clustering réel, la borne inférieure issue de la section précédente, et le clustering c_{ER} obtenu avec le modèle purement aléatoire, c_{MR} avec le modèle à distribution des degrés prescrite, c_{AB} pour le modèle AB, c_{WS} pour le modèle WS, c_{rb} pour le modèle biparti aléatoire à distributions des degrés prescrites et c_{gb} pour le modèle biparti incrémental. Dans les cas portant une étoile (*), le clustering réel est trop grand pour pouvoir être capturé avec le modèle WS : les valeurs indiquées sont donc celles obtenues avec les paramètres induisant le clustering maximal.*

Comme prévu du fait du théorème sur les distributions des degrés de la section précédente, les projections des graphes bipartis ont une distribution des degrés très similaire à la distribution originale. La seule différence notable se retrouve sur *Internet* avec une bosse qui apparaît au niveau des degrés proches de 30. La présence de cette bosse sera expliquée un peu plus loin.

	Internet	*Web*	*Acteurs*	*Co-signature*	*Cooccurrence*	*Protéines*
d	5.80	7	3.6	7.18	2.13	6.74
d_{ER}	5.25	5.47	2.97	7.57	2.06	10.4
d_{MR}	3.25	4.48	2.95	5.77	2.36	5.73
d_{AB}	4.15	5.1	2.93	5.5	2.38	8.15
d_{WS}	5.90	11.23	2559 (*)	2269 (*)	55.6 (*)	509 (*)
d_{rb}	2.97	3.2	3.06	5.07	2.06	5.8
d_{gb}	2.81	3.53	2.83	3.98	2.6	5.45

TAB. 4.2 – *Distances moyennes obtenues avec les modèles classiques et les modèles bipartis. Pour chaque graphe, on peut trouver sa distance moyenne réelle et celles obtenues avec le modèle purement aléatoire d_{ER}, le modèle avec distribution des degrés prescrite d_{MR}, le modèle AB d_{AB}, le modèle WS d_{WS}, le modèle biparti aléatoire à distributions des degrés prescrites d_{rb} et le modèle biparti incrémental d_{gb}. Dans les cas portant une étoile (*), les distances ne sont pas significatives à cause du fort clustering.*

De même, les graphes obtenus ont une faible distance moyenne et un fort clustering, très similaires à ceux des graphes originaux (sauf dans le cas de *Internet* pour le clustering). On peut aussi remarquer que la borne inférieure calculée précédemment est loin d'être

atteinte : le clustering réel est donc supérieur à cette borne. D'autre part, et par définition, les graphes obtenus ont la même distribution de tailles de cliques que les graphes originaux. Les simulations donnent des résultats qualitativement proches des valeurs réelles, ce qui prouve que la structure bipartie sous-jacente est une propriété essentielle qui permet de caractériser les grands réseaux d'interactions.

Malgré tout, on peut trouver des différences entre les valeurs réelles et celles obtenues par les modèles. Ces différences sont principalement la conséquence de l'intersection des voisinages des sommets de $\top$: si deux cliques ont un sommet en commun alors, en pratique, elles en ont très probablement d'autres. Ceci n'est pas capturé par les modèles bipartis. Ce comportement peut être vu comme du clustering biparti et il est responsable des irrégularités dans certaines distributions des degrés et des valeurs imprécises de certains paramètres. La Figure 4.11 montre la distribution de l'intersection des cliques c'est-à-dire, pour tout couple de sommets de $\top$, la taille de l'intersection de leur voisinage dans $\bot$. Ces intersections sont très petites pour les graphes aléatoires et beaucoup plus grandes pour les graphes réels.

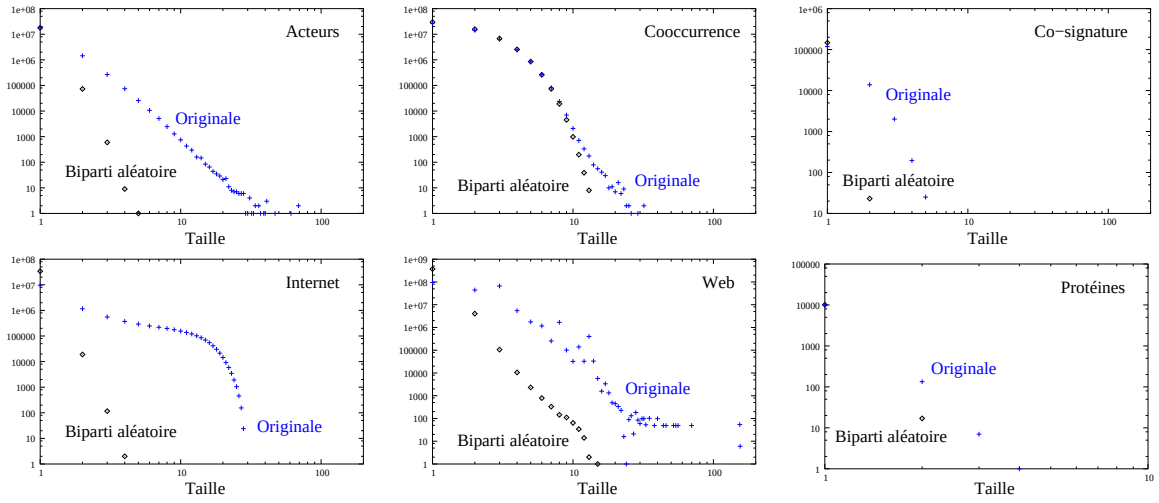

FIG. 4.11 – *Distribution de la taille des intersections de cliques pour les graphes originaux (ou leur décomposition) et les graphes obtenus avec le modèle biparti aléatoire.*

Dans le cas d'*Internet*, qui est le graphe le plus réfractaire à la modélisation bipartie, nous avons mis en évidence l'existence d'un ensemble de 94 sommets qui contient à lui seul toutes les cliques (494 pour être précis) de taille 14 et plus. Ce sous-graphe est très dense et les sommets qui le composent ont un clustering très élevé. Malgré tout, ceci n'a quasiment aucun impact sur le clustering global : comme ces sommets sont en petit nombre, le faible clustering de la majorité des autres sommets donne un clustering global plus faible. Au contraire, dans la projection sur $\bot$ d'un graphe biparti aléatoire, toutes ces grosses cliques sont disséminées dans le graphe, ce qui implique qu'un grand nombre de sommets ont un degré entre 14 et 29, d'où la bosse sur la distribution (un phénomène similaire peut

être observé sur *Web* au degré 130 environ). D'autre part, tous ces sommets se retrouvent dispersés : les 494 cliques contenaient 94 sommets à l'origine et en contiennent environ 50 000 dans le graphe aléatoire. Ces 50 000 sommets ont naturellement un fort clustering qui, cette fois, influence le clustering global du graphe.

4.5 Conclusion

Dans ce chapitre, nous avons présenté l'utilisation des graphes bipartis comme un outil général pour la modélisation des grands réseaux d'interactions. En utilisant ces graphes bipartis nous avons pu définir deux modèles de graphes aléatoires atteignant les objectifs suivants :

- les graphes obtenus par cette méthode possèdent les trois principales propriétés identifiées sur les grands réseaux d'interactions (distance moyenne faible, fort clustering et distribution des degrés en loi de puissance) ;
- les modèles sont basés sur un processus de construction *réaliste* similaire à ce qui se produit dans plusieurs grands réseaux d'interactions réels ; et,
- leur définition est suffisamment simple pour qu'on puisse à la fois donner une bonne intuition de leurs propriétés et prouver celles-ci formellement.

Alors que de nombreux modèles ont déjà été introduits, nos modèles sont les premiers qui atteignent tous ces buts. Ils constituent donc un progrès significatif dans la modélisation réaliste des grands réseaux d'interactions. De plus, il est facile et rapide d'obtenir des graphes avec ces modèles (nous fournissons un générateur [32]), ce qui les rend utilisables pour la simulation.

Malgré tout, comme nous l'avons déjà évoqué précédemment, le recouvrement entre cliques n'est pas du tout pris en compte par les modèles qui distribuent les cliques de manière aléatoire sur les sommets du graphe. Nous nous sommes même servis de cette absence d'intersection pour prouver les propriétés des modèles. Au contraire, il semble évident qu'en pratique la répartition des cliques ne se fait pas au hasard. Si l'on considère *Acteurs* par exemple, les acteurs choisis pour un film ne le sont pas au hasard mais suivant des critères assez précis, des critères de nationalité, de type de film (comédie, horreur, etc.), de préférence du réalisateur et de nombreux autres critères qui font que certains acteurs jouent très souvent ensemble (il suffit de regarder la Figure 4.11 pour s'en convaincre). Ainsi, les cliques ont une intersection généralement assez importante en pratique.

On peut faire une analogie avec le clustering dans les graphes aléatoires (les graphes ER, par exemple), dans lesquels le voisinage des sommets est très peu dense alors que le voisinage des graphes réels est au contraire très dense en général. On pourrait dire que les graphes bipartis réels sont très bi-clusterisés alors que les graphes bipartis aléatoires ne le sont pas. Dans le chapitre suivant, nous présentons un autre modèle qui résoud ce problème en prenant en compte l'intersection des cliques. Ce modèle est par conséquent encore plus performant que le modèle biparti.

Chapitre 5

Vers un modèle multiparti

Nous avons étudié dans le chapitre précédent la structure bipartie naturelle de certains grands réseaux d'interactions, notamment des graphes de collaboration (*Acteurs* et *Co-signature*), ainsi que des graphes de cooccurrence. Nous avons montré que cette structure sous-jacente explique au moins partiellement les propriétés communes de ces grands réseaux d'interactions, le fort clustering, la distribution des degrés en loi de puissance et le faible diamètre.

Nous avons aussi vu que cette structure bipartie est en fait commune à la majorité des réseaux, y compris ceux qui ne la possèdent pas naturellement. Il est ainsi possible de transformer tout réseau en graphe biparti, ce qui met en évidence la présence de cliques de grande taille et le fait que les individus appartiennent à un nombre très variable de cliques. Nous avons ainsi montré que, dans tous les cas, la structure bipartie sous-jacente peut être considérée comme responsable des principales propriétés du réseau.

Une fois cette structure non triviale identifiée, nous avons présenté plusieurs méthodes pour générer des structures aléatoires capturant certaines propriétés des graphes originaux. Malgré tout, les cliques générées aléatoirement sont généralement déconnectées, ce qui ne correspond pas à la réalité. Par exemple, si deux acteurs ont déjà joué ensemble dans un film, il y a une probabilité non négligeable qu'ils rejouent ensemble dans un autre film.

Les performances de la modélisation bipartie reposent sur l'utilisation de la remarque suivante : alors qu'on sait faire très peu de choses concernant le clustering, il est possible de faire beaucoup de choses avec les distributions de degrés, notamment construire un graphe (éventuellement biparti) aléatoire à distribution(s) de degrés prescrite(s). Nous avons ainsi capturé des informations pertinentes sur le réseau (en l'occurence la distribution de la taille de ses cliques ainsi que sa distribution de degrés) par deux distributions de degrés. Bien sûr, toutes les propriétés du graphe ne sont pas codées dans ces deux distributions, mais elles constituent un apport significatif.

Une idée naturelle pour aller plus loin dans la même direction est de tenter de capturer plus d'information en augmentant le nombre de *parties* dans le modèle. Notamment, on peut ainsi espérer coder les recouvrements de cliques, qui ne sont pas capturés dans le modèle biparti.

Dans ce chapitre nous allons présenter une proposition dans ce sens, prenant en compte ces intersections entre cliques. Dans un premier temps, ces intersections nous permettront d'obtenir une vision tripartie du graphe. Puis, nous introduirons un modèle triparti basé sur ces observations et nous montrerons expérimentalement que ce modèle affine effectivement le modèle biparti.

Nous terminerons sur les nombreuses perspectives ouvertes par cette approche, et notamment quelques améliorations possibles du modèle par l'utilisation de niveaux supplémentaires.

5.1 Graphes tripartis

Comme précédemment, nous allons considérer des graphes bipartis $G_2 = (\top, \bot, E)$ dont la projection sur $\bot$, $G_1 = (\bot, E')$, est définie par :

$$\{u, v\} \in E' \Leftrightarrow \exists t \in \top, \{u, t\} \in E \text{ et } \{v, t\} \in E$$

Nous reproduisons dans la Figure 5.1 les courbes de la Figure 4.10, qui montrent que les voisinages des sommets de $\top$ s'intersectent de façon non triviale dans les structures bipartites naturelles ou obtenues par décomposition au chapitre précédent. Nous y montrons également les intersections de voisinages des sommets de $\top$ sur les graphes bipartis aléatoires obtenues avec le modèle biparti. Ces intersections sont de natures qualitativement différentes pour les grands réseaux d'interactions et pour le modèle.

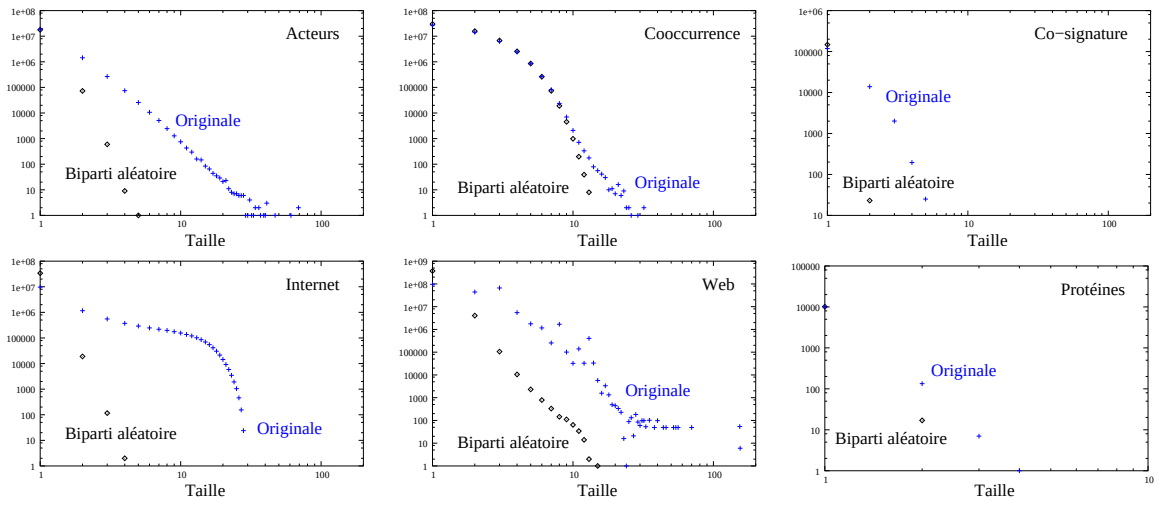

FIG. 5.1 – *Pour chaque couple de sommets de $\top$, on calcule la taille de l'intersection des voisinages de ces deux sommets, puis la distribution de ces tailles sur nos six exemples, pour les graphes bipartis originaux et les bipartis aléatoires.*

En effet, sur les graphes aléatoires les intersections entre cliques sont négligeables, le nombre de couples de cliques qui s'intersectent sur plus de deux ou trois sommets étant

très faible (sauf pour *Cooccurrence*), alors que pour les grands réseaux d'interactions de nombreux couples de cliques ont une grande intersection. En particulier, pour les graphes de grande taille et les graphes denses, *Acteurs*, *Cooccurrence* et *Web*, certaines cliques s'intersectent sur des dizaines de sommets, voire des centaines. Si on regarde l'intersection de triplets de cliques, de quadruplets, etc, la tendance s'accentue : alors que cliques s'intersectent peu dans le modèle, il y a un recouvrement non négligeable dans les graphes originaux.

Le modèle biparti aléatoire, bien que très satisfaisant pour les propriétés de base, l'est moins sur ce point. Ceci semble en particulier expliquer ses mauvaises performances dans certains cas particuliers, comme *Internet*.

Afin d'améliorer ces performances, nous allons considérer des graphes tripartis de la forme $G_3 = (\bot, \top, I, E)$. Dans un tel graphe, il n'y a de liens qu'entre des sommets appartenant à des ensembles différents : $E \subseteq (\bot \times \top) \cup (\bot \times I) \cup (\top \times I)$.

Les graphes tripartis que nous allons considérer seront en fait, par construction, très particuliers : si un sommet $i \in I$ est relié à tous les sommets de $B \subseteq \bot$ et de $T \subseteq \top$ alors $B \times T \subseteq E$. Autrement dit, tous les liens bipartis existent entre les sommets de B et ceux de T.

Décompositions en graphes tripartis

L'objectif d'une décomposition tripartie est d'associer à un graphe classique un graphe triparti qui capture des propriétés pertinentes du graphe initial par le biais des propriétés du graphe triparti, notamment ses distributions des degrés. Nous allons dans cette partie discuter différentes possibilités pour proposer une telle décomposition.

Une clique bipartie est définie comme une paire d'ensembles de sommets de $\top$ et $\bot$: $\{t, b\}$, $t \subseteq \top$ et $b \subseteq \bot$, telle que tous les sommets de t soient reliés à tous les sommets de b. Une telle clique est maximale si elle n'est incluse dans aucune autre. Dans la suite on parlera de clique bipartie k, l si $|t| = k$ et $|b| = l$.

Les cliques biparties jouent dans la modélisation tripartie le même rôle que les cliques (classiques) dans la modélisation bipartie : elles représentent une propriété importante non capturée par le modèle aléatoire (le clustering dans le cas classique et le recouvrement de cliques dans le cas biparti). La décomposition tripartie va donc consister, une fois une décomposition bipartie obtenue, en le calcul de cliques biparties dans ce graphe et leur codage par un nouvel ensemble I de sommets. La Figure 5.2 illustre un exemple de décomposition tripartie.

Avant d'aller plus loin, remarquons que nous désirons certaines propriétés pour la décomposition tripartie. Tout d'abord, il faut capturer les intersections non triviales ; on va donc se restreindre aux cliques biparties k, l telles que $k \geq 2$ et $l \geq 2$. Il faut aussi garder à l'esprit le coût relativement élevé du calcul des cliques biparties, ne serait-ce que parce que le nombre de telles cliques peut être exponentiel : l'intersection de n cliques peut générer $\binom{n}{2}$ cliques $2, x$ maximales, $\binom{n}{3}$ cliques $3, y$ maximales, etc.

Le niveau I du graphe triparti code les cliques biparties k, l. Chaque clique bipartie que

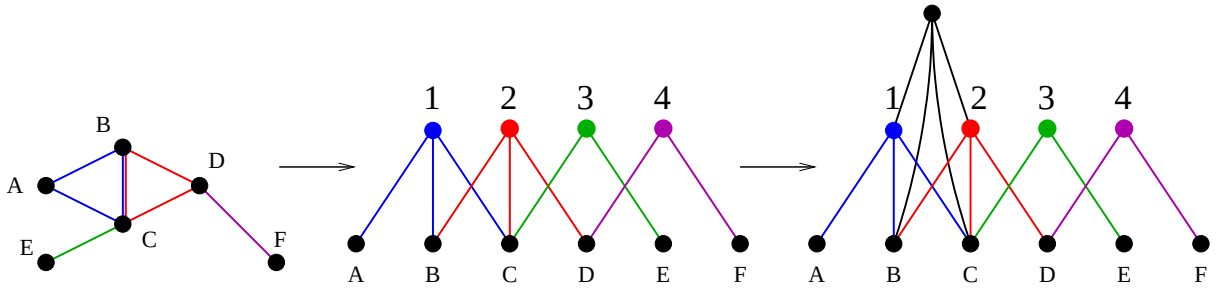

FIG. 5.2 – *Un graphe classique, une décomposition en graphe biparti, puis en triparti.*

l'on souhaite conserver va donc correspondre à un sommet au dernier niveau, sommet relié aux différents sommets de la clique bipartie.

En pratique, le nombre de cliques biparties présentes dans les grands réseaux d'interactions est effectivement très élevé : il n'est possible de calculer toutes les cliques biparties que pour de relativement petits graphes, tels *Protéines* et *Co-signature*. Le Tableau 5.2 présente le nombre de cliques biparties extraites pour ces deux grands réseaux d'interactions. Pour tous les autres graphes, le temps de calcul est trop élevé.

	Co-signature	*Protéines*
nombre de cliques biparties	5450	64

TAB. 5.1 – *Nombre de cliques biparties k, l pour $k \geq 2$ et $l \geq 2$ pour deux grands réseaux d'interactions.*

Comme pour le modèle biparti, nous ne garderons donc qu'un sous-ensemble des cliques biparties du graphe, mais nous les choisirons de taille aussi grande que possible afin de capturer un maximum d'informations sur les intersections non triviales entre cliques.

Nous allons dans la suite décrire une décomposition qui possède plusieurs des propriétés souhaitables évoquées précédemment : tous les recouvrements non triviaux sont capturés, le nombre de cliques biparties est relativement faible et des cliques biparties de grande taille sont découvertes. De plus, cette décomposition est calculable en un temps raisonnable.

5.2 Une décomposition calculable

La décomposition a donc pour but de ne garder qu'un certain nombre de cliques biparties, de préférence des cliques de grande taille. Notre approche a été de comparer plusieurs décomposition possibles sur des graphes de petites tailles afin d'évaluer les avantages et inconvénients de ces décompositions.

Une première méthode, qui n'est utilisable que sur de petits graphes, consiste à calculer toutes les cliques biparties puis à n'en garder qu'un certain nombre, suffisant pour avoir une couverture de tous les intersections de cliques non triviales. Plus précisément, on peut

calculer les cliques biparties, puis les considérer par ordre de taille décroissante, et ne garder que celles qui capturent au moins une clique $2,2$ qui ne l'était pas par un clique bipartie déjà selectionnée.

Cette méthode garde donc un ensemble suffisant de cliques en privilégiant la taille. Il reste toutefois à définir ce qu'est la taille d'une clique bipartie. Nous avons utilisé quatre méthodes pour cela qui sont, pour une clique bipartie k, l : le produit du nombre de sommets dans $\top$ et $\bot$ $(k \cdot l)$, leur somme $(k + l)$, ou plus simplement le nombre de sommets dans $\top$ (k) ou dans $\bot$ (l).

Le Tableau 5.2 présente le nombre de cliques qui seront finalement gardées selon ces quatre stratégies. Aucune des quatre stratégies n'est vraiment plus efficace que les autres pour *Co-signature*. Pour *Protéines*, toutes les cliques biparties du graphe codent une information qui n'est codée par aucune autre, donc aucune méthode ne permet de diminuer ce nombre de cliques biparties.

Il faut noter que nous avons utilisé le vrai graphe biparti de *Co-signature* et non sa décomposition. Cela ne fait aucune différence pour ce graphe en particulier, *Co-signature* et sa décomposition ayant les mêmes distributions des degrés pour $\top$ et $\bot$, comme nous l'avons vu dans la Section 4.1.

	Co-signature	*Protéines*
nombre de cliques	5450	64
$k \cdot l$	4860	64
$k + l$	4820	64
k	4705	64
l	5099	64

TAB. 5.2 – *Nombre de cliques biparties k, l gardées selon l'ordre de traitement des cliques pour* Co-signature.

Nous allons maintenant présenter la décomposition que nous avons retenue, principalement parce qu'elle est calculable en temps raisonnable et possède des propriétés similaires aux autres, notamment en termes du nombre de cliques biparties conservées.

Algorithme de décomposition

Notre décomposition va donc consister à ne garder qu'un sous-ensemble des cliques biparties, suffisamment grand pour capturer toutes les recouvrements non triviaux, mais sans les calculer toutes de manière exhaustive. Cette décomposition s'appuie sur le fait que les sommets de $\top$ ont généralement un degré assez peu élevé, et que donc les cliques de G_1 auxquelles elles correspondent ne peuvent pas s'intersecter sur un grand nombre de sommets de $\bot$. En d'autres termes, les cliques k, l ne peuvent donc pas avoir un l très élevé.

Finalement, notre algorithme de décomposition procède comme suit : pour chaque couple de sommets u et v de $\top$, on calcule l'intersection de leur voisinage $I_\bot = N(u) \cap N(v) \subseteq \bot$, puis on calcule l'ensemble $I_\top$ des sommets w de $\top$ tels que $N(w) \subseteq I_\bot$.

On obtient ainsi des cliques biparties maximales, qui sont les paires $\{I_\perp, I_\top\}$. Ce sont les cliques que nous retiendrons.

Cet algorithme génère effectivement un ensemble de cliques k, l couvrant le graphe biparti, et nous allons voir qu'il capture effectivement des propriétés non triviales.

Notons toutefois que, malgré l'effort fait pour rendre l'algorithme le plus rapide possible, les graphes de très grande taille, *Acteurs* et *Web*, n'ont pas pu être traités. Dans la suite nous allons donc nous restreindre à l'étude de la décomposition sur nos quatre autres exemples de grands réseaux d'interactions.

Le nombre de cliques biparties obtenues avec cet algorithme sur ces quatre graphes sont données dans le Tableau 5.3. On remarque que cette stratégie est un peu plus efficace sur *Co-signature* que les stratégies présentées plus tôt, en gardant environ 25% de cliques biparties en moins. Pour *Protéines*, il n'y a aucune différence, toutes les cliques étant nécessaires, comme nous l'avions déjà signalé dans le Tableau 5.2.

	Internet	*Co-signature*	*Cooccurrence*	*Protéines*
nombre de cliques biparties	145 143	3 584	869 641	64

TAB. 5.3 – *Nombre de cliques biparties obtenues avec l'algorithme pour quatre grands réseaux d'interactions.*

Propriétés des graphes tripartis

Les distributions des degrés obtenues après décomposition en triparti sont indiquées sur les Figures 5.3 et 5.4. Ces figures montrent que pour les cliques biparties k, l capturées par l'algorithme présenté précédemment, k suit une loi de puissance pour tous les graphes (Figure 5.3), alors que l a une décroissance beaucoup plus rapide (Figure 5.4).

Le fait que la décroissance de l soit rapide provient de l'absence de sommets dans $\top$ de très fort degré (mis à part pour *Acteurs* et *Web* que nous n'avons pas décomposés). Dans une clique bipartie k, l les k sommets de $\top$ doivent avoir un degré au moins l, les valeurs maximales de k sont donc égales au degré maximal dans $\top$. Nous reviendrons un peu plus loin sur la décroissance lente des k.

La Figure 5.3 présente aussi le degré moyen vers $\perp$ des sommets de I qui ont un degré vers $\top$ donné : pour k fixé, on trace la moyenne des l pour toutes les cliques biparties k, l. C'est l'équivalent biparti de la corrélation degré-degré présenté dans le Chapitre 2. Cette valeur est à peu près constante pour tous les k possibles et assez faible, entre 1 et 10 pour tous les exemples. Ceci confirme que la plupart des cliques biparties trouvées sont reliées à peu de sommets de $\perp$, et indique que cette tendance est uniforme.

La Figure 5.4 montre la moyenne réciproque : pour un degré de $\perp$ fixé l, on trace la moyenne des k pour toutes les cliques biparties k, l. Cette fois on se rend compte que les cliques biparties k, l ayant un faible l ont un k en moyenne légèrement plus élevé.

Enfin, la Figure 5.5 montre un nuage de points dans lequel chaque point (k, l) correspond à une clique bipartie k, l. Ces courbes montrent clairement la présence de cliques biparties k, l avec un grand k ou un grand l, mais pas les deux.

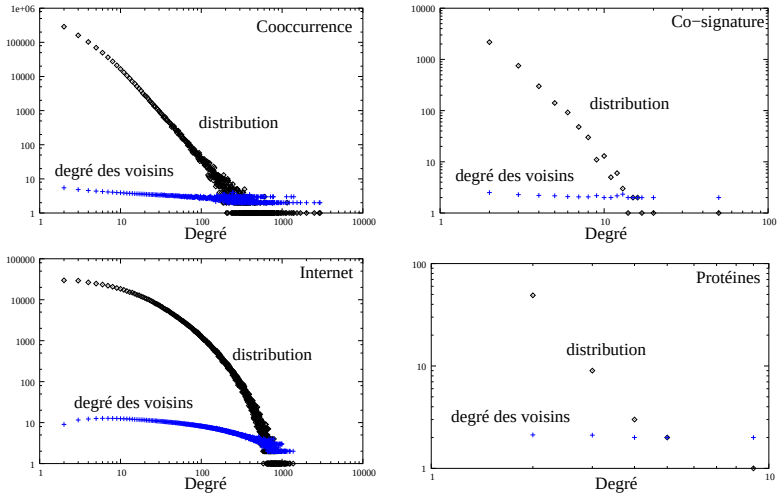

FIG. 5.3 – *Distribution des degrés de I vers $\top$ et degré moyen vers $\bot$ des sommets de I ayant un degré donné vers $\top$, pour les quatre exemples décomposés en triparti.*

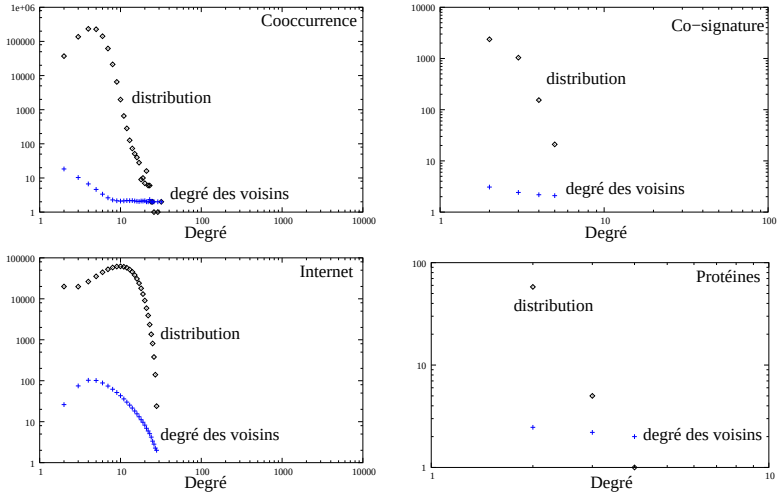

FIG. 5.4 – *Distribution des degrés de I vers $\bot$ et degré moyen vers $\top$ des sommets de I ayant un degré donné vers $\bot$, pour les quatre exemples décomposés en triparti.*

Toutes ces figures montrent que nous capturons bien le fait que de nombreuses cliques s'intersectent sur quelques sommets, plutôt que le fait que des cliques s'intersectent sur beaucoup de sommets. Cette remarque est non triviale, car notre algorithme de décomposition aurait plutôt le tendance inverse, par définition. En effet, l'algorithme calcule des cliques $2, l$ qu'il essaye d'agrandir. Rien ne garantissait *a priori* qu'elles auraient pu être agrandies en cliques k, l, avec $k >> 2$.

Cette propriété n'est donc pas induite par la décomposition, mais est bien une propriété intrinsèque des intersection de cliques rencontrées dans nos exemples : il y a de grandss

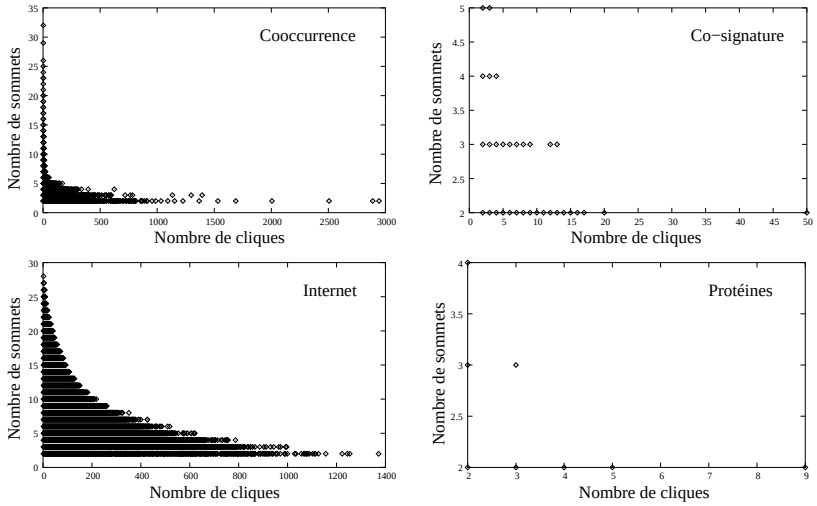

FIG. 5.5 – *Nuage de points pour les tailles des cliques biparties. Chaque point* (k, l) *correspond à une clique bipartie* k, l*, donc à un ensemble de* k *sommets de* $\top$ *(*k *cliques) tous reliés à* l *sommetsde* $\bot$*.*

ensembles de cliques qui s'intersectent toutes sur un nombre non négligeable de sommets.

5.3 Un modèle triparti

La décomposition bipartie présentée dans le chapitre précédent associée à la décomposition que nous venons de présenter permet de construire un graphe triparti associé à un graphe quelconque. Ce graphe triparti code les cliques du graphe initial au niveau $\top$ et les intersections entre ces cliques au niveau I.

Nous voulons maintenant utiliser cette structure tripartie pour générer des graphes ayant les propriétés rencontrées en pratique. Comme dans le cas du biparti, nous allons générer des graphes tripartis aléatoires ayant certaines propriétés, à savoir les distributions de degrés originales :

- les distributions de $\bot$ vers $\top$ et de $\top$ vers $\bot$, comme pour le modèle biparti ;
- la distribution couplée de I vers $\bot \times \top$, c'est-à-dire un couple (k, l) par sommet de I relié à k sommets de $\top$ et à l sommets de $\bot$.

Il s'agit maintenant, en utilisant uniquement ces informations, de générer un graphe biparti pour lequel les intersections de cliques correspondent à ces informations.

Algorithme de génération

Cet algorithme se déroule en deux phases. Une première phase consiste à placer les cliques biparties issus du niveau I dans le graphe biparti, et une deuxième phase rajoute

d'éventuels liens manquants. Plus précisément, l'algorithme est le suivant (voir Figure 5.6) :

1. Associer à chaque sommet du graphe triparti des degrés suivant les distributions
2. pour tout sommet u de I de degrés (k, l) :
   - choisir k sommets de $\top$ de degré au moins l,
   - choisir l sommets de $\bot$ de degré au moins k,
   - relier u aux k sommets de $\top$ et aux l sommets de $\bot$.
3. S'il manque des liens, les rajouter aléatoirement.

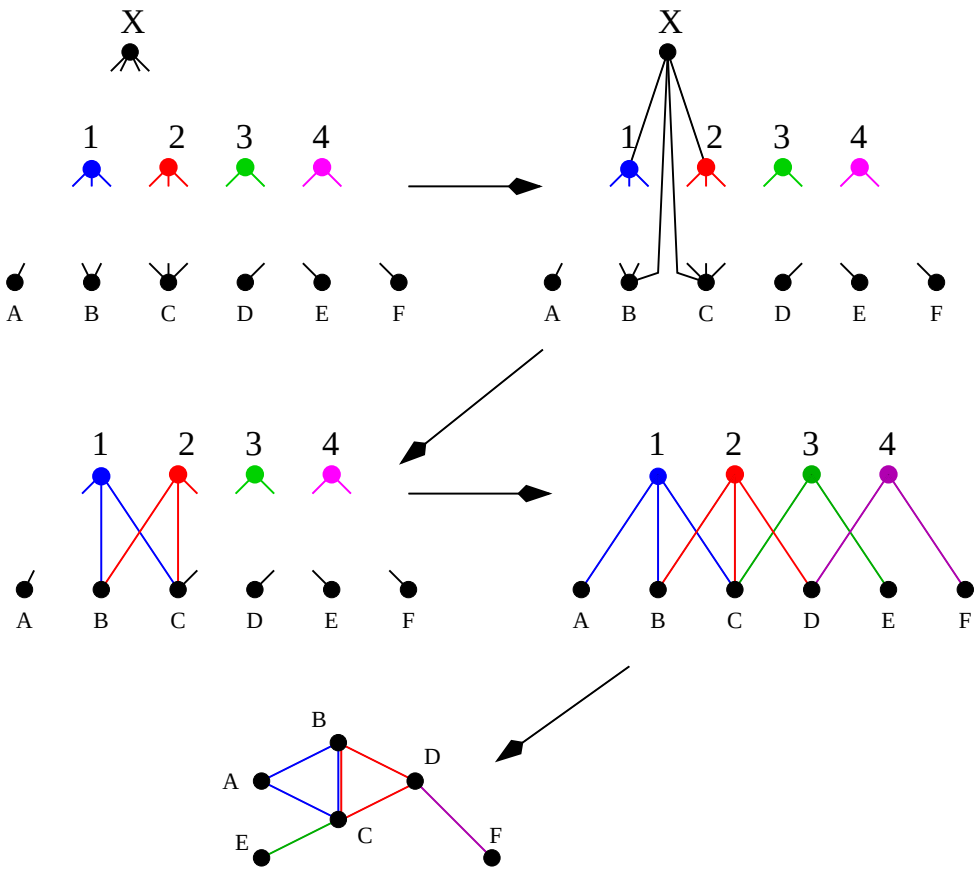

FIG. 5.6 – *Un graphe triparti, recomposé en biparti, puis projeté sur* $\bot$.

Pour chaque sommet du niveau I, qui correspond donc à une clique bipartie k, l, il faut choisir des sommets dans $\top$ et $\bot$ à relier pour former cette clique bipartie. Bien entendu, les sommets choisis, à la fois pour $\top$ et $\bot$, doivent avoir un degré qui va permettre d'effectuer les ajouts de liens. C'est ce que fait l'étape 2 de l'algorithme.

Un problème peut par ailleurs survenir. De manière intuitive, le problème est le suivant : chaque sommet (de $\top$ et $\bot$) reçoit au début de l'algorithme un degré suivant une distribution fixée *a priori*. Ensuite, on va choisir les sommets en fonction de leur degré et

leur rajouter des liens pour les cliques biparties. Il ne faut pas que le degré d'un sommet dépasse le degré qui lui avait été alloué au début de l'algorithme, sinon il ne sera plus possible d'obtenir la bonne distribution.

En pratique, cette règle n'est pas toujours vérifiée. On observe en effet de nombreux dépassements, surtout si le nombre de sommets au niveau I est élevé. Nous avons choisi d'ignorer le problème durant l'ajout de liens entre I et les niveaux plus bas, puis de supprimer autant de liens qu'il le faut à tous les sommets qui ont un degré trop élevé par rapport à ce qui leur avait été attribué. Cette solution simple s'est avérée n'avoir que peu d'influence sur nos résultats.

Performances

Nous allons donner maintenant quelques résultats expérimentaux obtenus en décomposant des grands réseaux d'interactions en graphes tripartis, puis en générant les graphes correspondant à partir des distributions des degrés suivant l'algorithme que nous venons de présenter. Nous allons nous concentrer sur les propriétés usuelles, le clustering et la distribution de degrés, mais aussi sur deux propriétés plus fines afin d'évaluer la capacité du modèle à les capturer (le modèle biparti, comme les modèles antérieurs, n'y arrivant que très mal). Il s'agira des corrélations entre degrés et des corrélations degré-clustering.

Concernant le clustering, on se rend compte qu'il est très bien capturé pour *Co-signature* et *Protéines*, mais toujours pas de manière complètement satisfaisante pour *Internet*, même si cela représente une amélioration par rapport au modèle biparti. Nous allons expliquer plus loin la raison de cette mauvaise estimation et les solutions qui peuvent être apportées.

	Internet	*Co-signature*	*Protéines*
c	0.171	0.638	0.153
c_{bip}	0.320	0.542	0.187
c_{tri}	0.247	0.601	0.173

TAB. 5.4 – *Clustering*

Concernant les distributions des degrés et les corrélations, les Figures 5.7 et 5.8 montrent que le modèle triparti est très performant pour les trois propriétés sur les graphes *Co-signature* et *Protéines*. On peut juste noter que la décroissance du clustering moyen en fonction du degré est légèrement plus rapide sur *Co-signature* pour le graphe aléatoire que pour le graphe original (Figure 5.7, deux courbes de droite).

Sur le troisième que nous avons décomposé de cette manière, *Internet*, les résultats sont plus décevants. La Figure 5.9 concerne toujours les distributions des degrés ainsi que les corrélations entre degrés et clustering. La ligne du haut montre ces propriétés pour le graphe original, et l'on avait déjà remarqué dans le Chapitre 2 que les corrélations sont plus complexes sur *Internet* que sur les autres graphes.

La deuxième ligne de la Figure 5.9 montre que le modèle triparti n'est pas complètement satisfaisant. En particulier, la distribution des degrés exhibe une bosse similaire à celle que

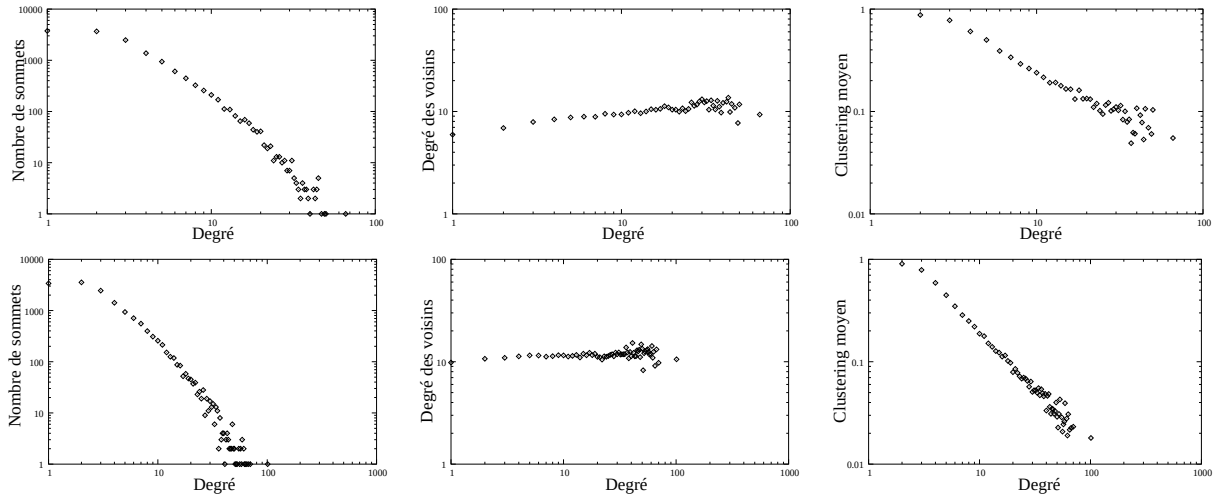

FIG. 5.7 – Co-signature*: de gauche à droite, la distribution des degrés, la corrélation degrés-degrés et la corrélation degré-clustering. En haut pour le graphe original, en bas pour un graphe obtenu avec le modèle triparti.*

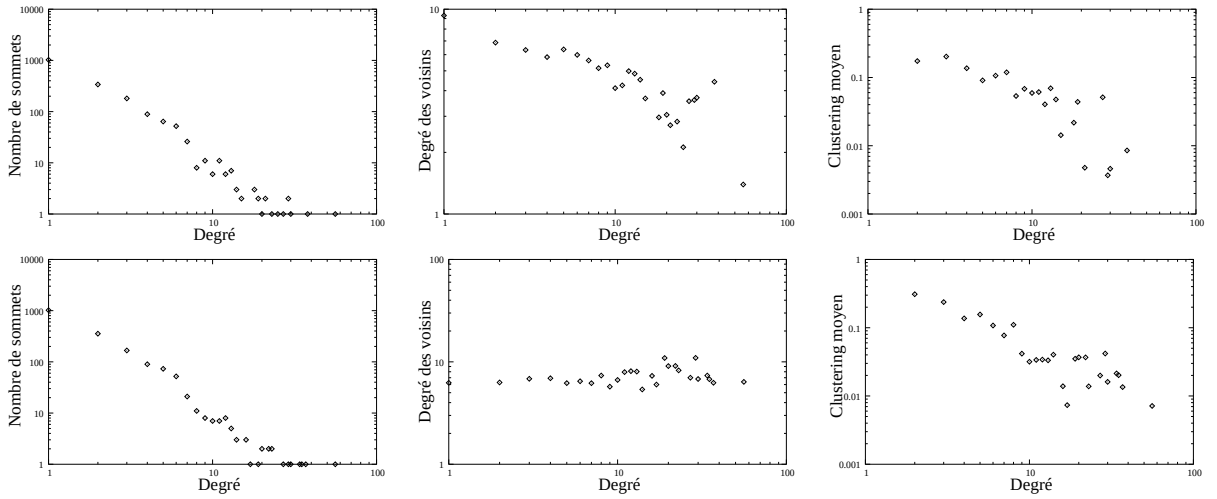

FIG. 5.8 – Protéines*: de gauche à droite, la distribution des degrés, la corrélation degrés-degrés et la corrélation degré-clustering. En haut pour le graphe original, en bas pour un graphe obtenu avec le modèle triparti.*

l'on avait observée avec le modèle biparti dans la Section 4.4. Cela provient de la perte d'information décrite précédemment à cause du trop grand nombre de cliques biparties ajoutées.

La troisième ligne permet de vérifier cette hypothèse. Elle correspond au même graphe décomposé puis recomposé de manière aléatoire avec le modèle triparti, mais en supprimant plusieurs cliques biparties. Les trois propriétés sont mieux capturées qu'avec le modèle

exact.

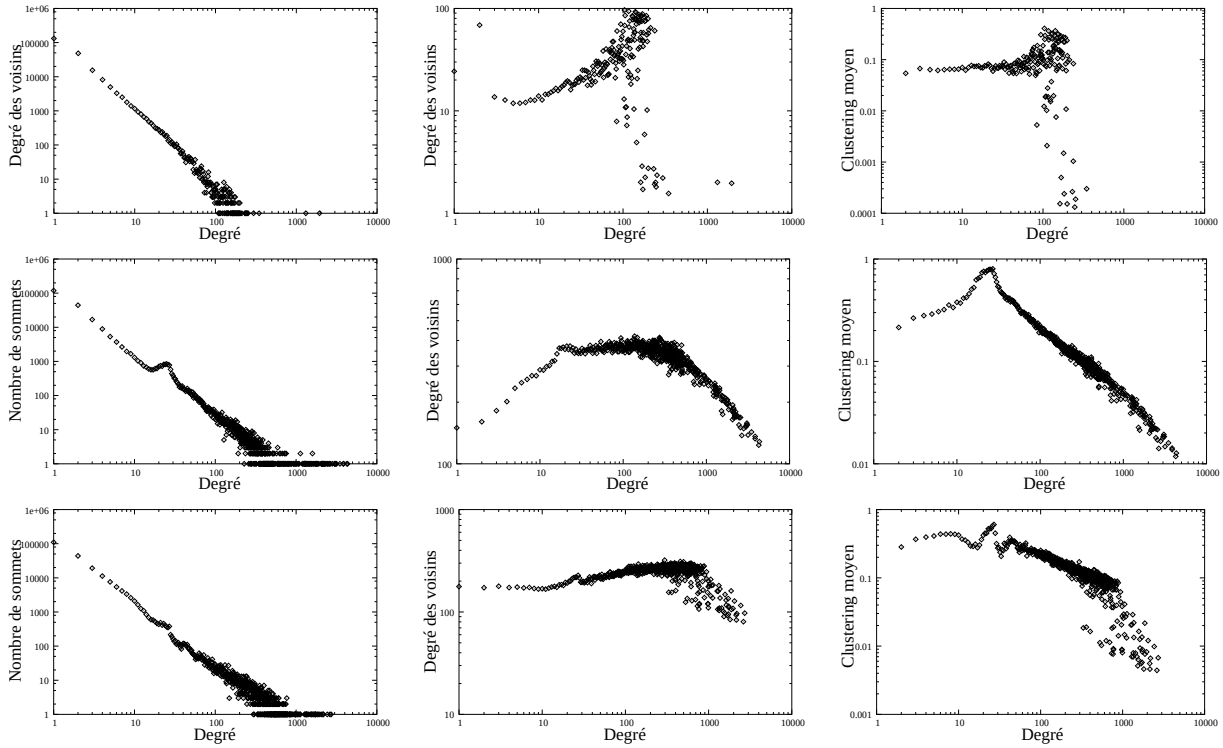

FIG. 5.9 – Internet *: de gauche à droite, la distribution des degrés, la corrélation degrés-degrés et la corrélation degré-clustering. En haut pour le graphe original, au milieu pour un graphe obtenu avec le modèle triparti, et en bas pour un graphe obtenu avec le modèle triparti en supprimant des cliques biparties.*

La qualité des graphes générés semble donc dépendante du nombre de cliques biparties qui sont extraites durant la décomposition en triparti. Le problème provient du fait que les cliques biparties ont, en pratique, une intersection assez forte. Ainsi certains ensembles de sommets de $\perp$ et de $\top$ appartiennent à plusieurs cliques biparties. Or, le modèle aléatoire distribue les cliques biparties de manière similaire à ce que fait le modèle biparti avec les cliques.

5.4 Aller plus loin

Nous avons vu qu'il existe de nombreuses façons de coder une graphe par un graphe triparti, et que ceci peut permettre de capturer finement les propriétés des grands réseaux d'interactions. Cette approche améliore significativement les performance du modèle biparti. Elle reste cependant encore largement à creuser, le choix des informations les plus pertinentes à conserver pour la génération n'étant pas encore bien compris.

Un tel modèle, qui repose sur l'utilisation exclusive de distributions de degrés, permet de générer des graphes de tailles diverses ayant par ailleurs des propriétés semblables à celle d'un graphe initial donné. Il est donc particulièrement adapté à une utilisation dans des simulations pour lesquelles la taille joue un rôle important (ce qui est presque toujours le cas). Son caractère aléatoire permet aussi d'espérer l'utiliser dans des analyses formelles, même s'il peut être délicat de manipuler plus de deux distributions de degrés en même temps.

Un autre point, qui n'a été qu'esquissé ici, semble intéressant. Il s'agit d'étudier les propriétés statistiques des versions multiparties des grands réseaux d'interactions, comme nous l'avons fait pour les bipartis. On pourrait ainsi découvrir que la façon donc les cliques s'intersectent induit (et explique) certaines propriétés observées sur les grands réseaux d'interactions.

Enfin, comme nous l'avons souligné à plusieurs reprises, les temps de calculs induits par la décomposition sont non négligeables et sont même parfois un facteur bloquant. Il semble donc essentiel de trouver une alternativee moins couteuse.

Nous avons évoqué le problème inhérent aux modèles utilisant deux ou trois niveaux : dans le modèle biparti, les cliques aléatoires sont trop dispersées dans le réseau et leur intersection est donc presque toujours négligeable. Dans le modèle triparti, il en est de même pour les cliques biparties qui ne se recouvrent pas. Or, en pratique, les cliques d'une part et les cliques biparties d'autre part se recouvrent fortement.

Le modèle triparti arrive à faire en sorte que les cliques se recouvrent. Il parait donc naturel de penser qu'ajouter un quatrième niveau permettrait d'obtenir des recouvrements entre les cliques biparties. La Figure 5.10 présente un exemple simple de ce que pourrait être une décomposition d'un graphe en quatre niveaux.

Il est malgré tout prévisible qu'un modèle à quatre niveaux ne va pas capturer les intersections de cliques triparties. Pour généraliser, il conviendrait donc d'introduire un modèle multiparti. Dans un tel modèle, chaque niveau codant des intersections non triviales dans le graphe des niveaux précédents, il serait possible de décomposer un graphe avec le nombre de niveaux souhaités. Rajouter des niveaux coderait des informations plus précises.

De même qu'un graphe contenant de grosses cliques qui ne s'intersectent pas est très bien modélisé par le modèle biparti, on peut aussi penser qu'à partir d'un certain niveau, il n'y aura plus d'intersections non triviales (on peut même définir le modèle de façon à assurer ce point), et dès lors générer un graphe avec uniquement les distributions de degré devrait donner d'excellents résultats.

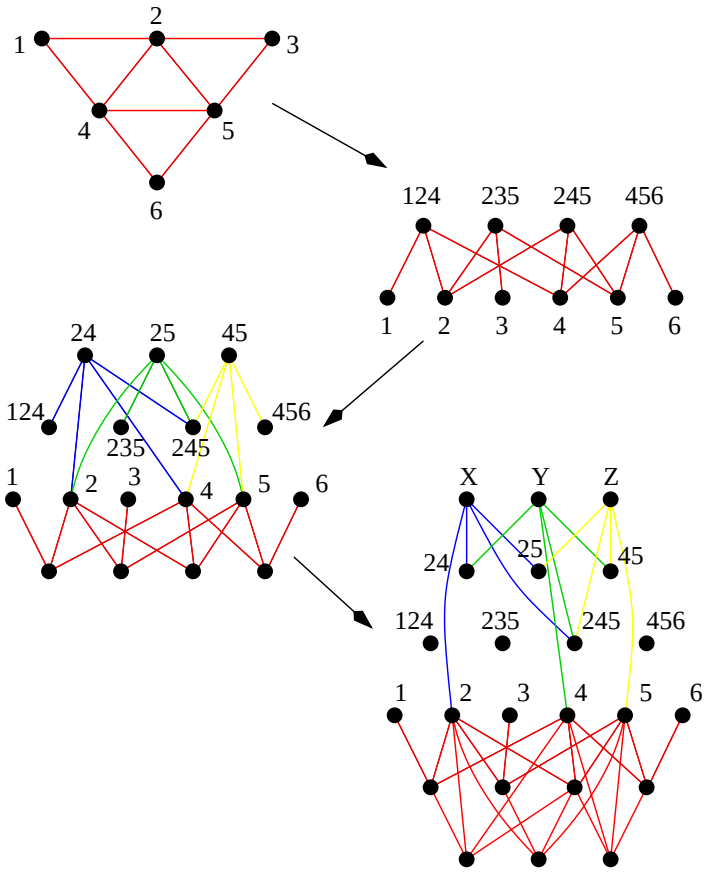

FIG. 5.10 – *Un exemple de décomposition en quatre niveaux pour coder les intersections entre cliques biparties.*

Conclusion

Comme nous l'avons vu dans la partie précédente, l'analyse des grands réseaux d'interactions permet de d'identifier des ensembles de propriétés générales de ces réseaux. À partir de ces résultats, les modèles (basés sur le tirage aléatoire de graphes ayant certaines propriétés ou la construction incrémentale suivant certaines règles) tentent de produire des graphes *ressemblant* aux grands réseaux d'interactions rencontrés en pratique. Il apportent également un éclairage important sur les propriétés, en distingant des propriétés générales des graphes aléatoires (la distance moyenne courte par exemple) et celles qui sont plus spécifiques (le clustering par exemple). Enfin, les modèles donnent aussi une intuition de l'origine de certaines propriétés observées, notamment à l'aide de principe proches de l'attaehement préférentiel.

Nous avons également vu dans cette partie que, bien que de très nombreux modèles aient été proposés dans le passé, aucun ne réussit à capturer de façon satisfaisante les trois propriétés générales des grands réseaux d'interactions : la distance moyenne courte, la distribution des degrés en loi de puissance et le fort clustering. Un important travail de modélisation reste donc à effectuer dans cette direction, et au delà pour capturer des propriétés plus fines (corrélations par exemple).

Nous avons dans cette optique adopté l'approche suivante : puisque des techniques bien maîtrisées existent pour générer et étudier les graphes aléatoires à distribution de degrés donnée, nous avons tenté de capturer les autres propriétés des grands réseaux d'interactions (et notamment le clustering) dans des informations de ce type.

Ceci a en particulier donné lieu à l'introduction du modèle biparti, qui capture mieux que toutes les propositions antérieures les trois propriétés générales des grands réseaux d'interactions, en codant la distribution de la taille des cliques par *deux* distributions de degrés (celles d'un graphe biparti). Ce modèle a de plus l'avantage de de dériver en deux versions (l'une par tirage aléatoire, l'autre par construction itérative), de reposer sur des remarques issues de considérations réalistes, d'être utilisable pour faire des simulations, et d'être accessible aux études formelles.

Continuant dans cette voie, nous avons montré que l'idée peut être poussée plus loin en codant les propriétés souhaitées par des graphes tripartis, et, au delà, multipartis. Les recherches dans cette direction restent encore largement à approfondir, mais les premiers réslutats montrent clairement que cette approche permet de surpasser le modèle biparti, et donc constitue une avancée importante du domaine. On peut ainsi espérer capturer, outre les trois propriétés générales des grands réseaux d'interactions, des propriétés plus fines

comme les corrélations entre ces propriétés, et d'autres.

Soulignons que nous nous sommes focalisés dans cette thèse sur un des principaux objectifs de la modélisation : la production d'objets *similaires* à un objet donné (un grand réseau d'interaction, ici). La modélisation a toutefois d'autres buts, comme la définition d'un cadre formel permettant ensuite d'étudier des propriétés (comme nous le verrons au Chapitre 7) ou l'explication intuitive des propriétés modélisées (comme avec l'attachement préférentiel par exemple).

Le modèle biparti s'inscrit également dans ces perspectives : il s'agit, dans le cas du biparti aléatoire, d'un modèle ouvrant la voie à de nombreuses études formelles. Nous avons ainsi pu démontrer certaines de ses propriétés, et des études formelles du type de celles que nous ferons au Chapitre 7 peuvent se transposer à ce modèle. Il permet ainsi pour la première fois de faire des études de l'impact du clustering (associé aux autre propriétés) dans un cadre formel et réaliste. De plus, comme nous l'avons vu, il est basé sur la façon dont certain grands réseaux d'interactions sont *réellement* constitués. Il permet donc également de donner une explication intuitive de certaines propriétés.

Le modele triparti, et ses extensions multiparties, en devenant de plus en plus complexes afin de capturer de plus en plus d'information, ne peuvent par contre apporter ce type de progrès que de façon bien moindre. Par exemple, effectuer des preuves formelles reposant sur un grand nombre de distributions devient très délicat, d'autant plus que ces distributions sont difficiles à caractériser. De même, on voit mal comment interpréter intuitivement les différents niveaux des multipartis, autrement que comme intersections de cliques, intersections d'intersections de cliques, etc. Les perspectives dans cette direction sont toutefois extrêmement prometteuses, et il est possible que ces carences trouvent une solution naturelle. Il est en l'étant actuel des choses trop tôt pour le dire.

Le modèle triparti, et ses extension multiparties, doivent donc en un premier temps être vus comme des modèles permettant de faire des *copies statistiques* de divers grands réseaux d'interactions. En effet, ils sont capables de produire des graphes partageant de nombreuses propriétés statistiques avec le grand réseau d'interactions initial. Ils permettent ainsi, par exemple, d'étudier l'impact de la taille d'un grand réseau d'interactions sur divers phénomènes, en offrant une méthode pour générer des graphes de diverses tailles ayant des propriétés statistiques semblables.

Troisième partie

Quelques applications

Introduction

Cette partie présente des applications des deux parties précédentes à trois cas particuliers. Ces applications ne doivent toutefois pas être vues comme des corollaires des travaux présentés avant. Au contraire, nous insisterons sur les aspects méthodologiques et notamment sur le fait que c'est souvent de l'étude de cas particuliers qu'émerge la définition de problèmes fondamentaux pertinents. Ainsi, si cette partie repose naturellement sur les notions introduites précédemment, elle est aussi l'occasion d'en évaluer la pertinence et d'identifier des problèmes clés.

Au delà, nous serons amenés à introduire des méthodes utiles dans ces cas particuliers, mais qui sont en fait beaucoup plus générales. Nous insisterons sur ces aspects faisant de cette partie un travail *autant* méthodologique qu'applicatif.

Le Chapitre 6 montre comment l'analyse des grands réseaux d'interactions permet d'obtenir des informations fines sur ceux-ci. Pour ce faire, nous étudions un système pair-à-pair dans lequel il est possible de connaître les échanges effectués entre les personnes à une grande échelle. Il est non seulement possible de connaître les volumes d'informations échangés entre les pairs ou le nombre de fichiers partagés par un pair donné, mais aussi de quantifier leur évolution au cours du temps. Ce dernier point a trait à la dynamique du graphe lui-même, point actuellement très mal compris et peu appréhendé.

L'approche de ce domaine par le biais des réseaux pair-à-pair soulève de nombreuses questions et ouvre la voie à des études sur d'autres objets naturellement dynamiques, ainsi qu'à des études plus fondamentales dans ce domaine. Cet exemple montre comment un travail sur un cas particulier peut amener à introduire des concepts et des paramètres statistiques qu'on peut ensuite réutiliser dans un cadre plus général.

Le Chapitre 7 présente une étude approfondie de l'effet de pannes et d'attaques sur un réseau. Sur un réseau comme l'Internet, cela revient à étudier de quelle manière la panne d'un routeur peut affecter le fonctionnement du réseau et, de manière similaire, l'impact qu'aurait la destruction ciblée (selon des critères à définir) de routeurs. Si un virus informatique se propage sur le réseau, on peut alors évaluer l'impact de la vaccination en certains points du réseau sur la faculté de propagation du virus.

Dans ce chapitre, nous verrons que les propriétés du graphe ont un impact fort sur les résultats et ce, de manière qualitative. Par exemple, un graphe sans-échelle est résistant aux pannes et ne s'effondrera que si la plupart des sommets sont détruits alors que ce même graphe s'effondrera très rapidement s'il est attaqué de manière ciblée. Ce chapitre illustre aussi l'utilisation de modèles (graphes aléatoires ou sans-échelle, par exemple) pour

étudier formellement des phénomènes sur les grands réseaux d'interactions.

Finalement, le Chapitre 8 est consacré à la métrologie de l'Internet, et tente de répondre à la question suivante : quel est le biais introduit par la mesure sur la topologie observée ? La question est en fait très générale et peut s'appliquer à presque tous les grands réseaux d'interactions étudiés, mais sera traitée ici dans le cas plus particulier du graphe de l'Internet. Pour ce graphe, il s'agit en pratique de récupérer les chemins suivis par les messages et ensuite de reconstruire le graphe en fusionnant tous les chemins.

À l'aide de simulations sur divers modèles de graphes, nous montrerons qu'il faut être très prudent en ce domaine, les observations étant très dépendantes à la fois du graphe et de la méthode d'observation. Au-delà, ce chapitre montre comment la gamme de modèles de grands réseaux d'interactions que nous avons présentée peut être utilisée dans des simulations afin de fournir un éclairage pertinent sur un problème précis.

Chapitre 6

Pair-à-pair

Le paradigme pair à pair (P2P), qui considère que tous les utilisateurs d'un système sont égaux, est principalement utilisé pour partager des données. Dans ces systèmes, les pairs mettent des fichiers à disposition sur leur machine et peuvent en récupérer chez les autres pairs. Ils aident aussi les pairs qui recherchent un fichier à le trouver. Idéalement, un système P2P est complètement décentralisé : il n'y a aucune autorité centrale pour contrôler ou pour aider au fonctionnement du système.

Certains réseaux P2P, dits semi-centralisés (`eDonkey`, par exemple), possèdent au contraire un certain nombre de serveurs qui aident au fonctionnement du système, alors que d'autres sont complètement distribués et donc gérés uniquement par les pairs eux-mêmes. Dans cette dernière catégorie, certains systèmes, tels que Gnutella 0.6 et `KaZaA`, distinguent deux types de clients, ordinaires ou privilégiés [69, 78]. Les clients privilégiés ont généralement de plus grosses ressources et peuvent assurer la gestion du réseau en jouant le rôle d'intermédiaires entre les clients ordinaires.

Créer de tels systèmes ainsi que les protocoles de communication entre les pairs est actuellement au centre de nombreuses recherches pour assurer notamment la robustesse du système et l'anonymat des utilisateurs [83, 108, 121]. Le simple fait de permettre aux utilisateurs d'entrer dans le système, de le quitter en le laissant opérationnel ou encore de localiser efficacement une donnée n'est pas aisé, comme le montre le nombre grandissant d'études sur ces sujets [14, 52, 86, 115, 127].

Les échanges de données entre utilisateurs d'un système P2P ne sont pas aléatoires : si un utilisateur possède une donnée qui intéresse un autre utilisateur (un morceau de musique, par exemple), alors il en a probablement d'autres (autres morceaux de musique du même artiste, par exemple). Ceci est particulièrement vrai dans les systèmes P2P où les fichiers sont découpés en plusieurs parties pour augmenter le nombre de fournisseurs potentiels d'un morceau. Dans ce cas, si un pair possède plusieurs parties d'un fichier, alors tout pair intéressé par l'une d'elle sera certainement intéressé par d'autres, voire toutes.

Bien qu'ils aient des conséquences très importantes sur l'efficacité du système, on sait aujourd'hui encore très peu de choses sur ces aspects socio-culturels [88, 119].

Plusieurs catégories d'utilisateurs ont toutefois été identifiées. La première est celle des utilisateurs qui fournissent de nouvelles données, rares ou récentes : ils accroissent ainsi

le nombre et la diversité des fichiers disponibles. Une deuxième catégorie concerne les utilisateurs qui ne font que télécharger des données et laissent l'ensemble de leurs fichiers disponibles sur leur machine : ils assurent la réplication des données et rendent ainsi le système plus stable. Enfin, la majorité des utilisateurs téléchargent des données et ne les laissent pas disponibles longtemps, ou pas du tout. Ces clients consommateurs ont *a priori* moins d'utilité pour le système.

Ces premiers résultats confirment que le comportement des utilisateurs n'est pas aléatoire. Beaucoup reste toutefois à faire dans l'analyse de ces comportements et c'est dans cette optique que ce chapitre s'inscrit.

Dans ce qui suit, nous allons étudier les échanges entre pairs dans le réseau `eDonkey` qui est semi-centralisé et repose donc sur un ensemble de serveurs dédiés gérant un certain nombre d'utilisateurs. En se plaçant sur un serveur, il est donc possible d'obtenir des données riches et précises, ce qui serait très difficile dans un système totalement distribué.

La Section 6.1 présente le contexte dans lequel s'inscrit cette étude et décrit plus précisément le fonctionnement du protocole `eDonkey` et les mesures effectuées. La Section 6.2 présente des résultats préliminaires sur ces mesures et les Sections 6.3 et 6.4 se concentrent respectivement sur l'étude du système du point de vue des pairs et des données échangées.

6.1 Préliminaires

Ces dernières années, diverses réseaux P2P ont été observés, par des mesures actives avec des crawlers [119, 121] ou des mesures passives et de l'analyse de flux [4, 12, 77, 88, 122, 125]. Notre analyse se basant sur des mesures passives, nous allons détailler plus précisément cette catégorie.

Adar et Huberman [4] ont étudié le trafic sur Gnutella 0.4 durant 24 heures en se plaçant sur un client particulier, afin de catégoriser les comportements des autres utilisateurs. Il est apparu qu'environ 70% des utilisateurs ne partagent pas de données alors que 1% des clients répondent à 50% des requêtes.

Dans [12], le trafic sur Gnutella a été observé pendant 35 heures par un client pour quantifier le trafic de signalisation[1], d'une part, et en déduire des propriétés sur les échanges, d'autre part. Les auteurs ont en particulier étudié la distance entre le client de l'étude et ceux qui le contactent. De manière similaire et toujours concernant le trafic Gnutella, [88, 125] ont étudié différentes traces de durée et de localisation variables afin de définir des stratégies d'optimisation des caches pour diminuer le trafic sur l'Internet.

Les données de différents systèmes P2P (Fastrack, Gnutella, Directconnect) ont aussi été collectées directement au niveau d'un routeur d'un fournisseur d'accès [77, 122]. L'objectif était de proposer des mécanismes pour que les opérateurs puissent réduire le trafic P2P. En particulier, [77] montre que les données sont suffisamment redondantes pour justifier des stratégies de cache.

1. Ce sont les informations nécessaires au bon fonctionnement du protocole mais inutiles pour l'utilisateur en tant que telles.

Plus récemment, [69, 78] ont utilisé des mesures passives sur KaZaA et KaZaA Lite (réseau Fastrack). Le fait que le trafic soit composé principalement de quelques fichiers populaires est confirmé dans [78] et les résultats montrent que cette tendance est encore plus élevée qu'on ne le supposait auparavant. Enfin, [69] a utilisé 3 clients KaZaA dans le réseau de NY Polytechnic, ce qui a permis de mieux comprendre le fonctionnement de KaZaA dont le protocole n'est pas ouvert. Le nombre de clients privilégiés a pu être estimé à 30 000, chacun ayant des contacts avec 40 à 60 autres clients privilégiés et 100 à 200 clients ordinaires.

Toutes ces études étaient basées sur l'utilisation de clients ou de routeurs afin de mieux appréhender les caractéristiques du trafic. La conclusion qui s'en dégage est que, indépendamment du protocole P2P sous-jacent, la majorité du trafic est concentrée sur quelques fichiers spécifiques.

Notre approche est une étape supplémentaire dans la mesure des réseaux P2P. Nous nous sommes en effet penchés sur le protocole semi-centralisé eDonkey en étudiant un serveur particulier (gérant environ 50 000 clients simultanément) et en collectant toutes les requêtes faites sur ce serveur, ainsi que les réponses du serveur. Ceci nous permet de savoir qui échange avec qui et de découvrir les différents comportements des utilisateurs.

6.1.1 Le protocole eDonkey

Le système eDonkey repose sur l'utilisation de quelques serveurs dédiés qui mettent les pairs en contact. Dans ce but, il existe un protocole permettant aux pairs de communiquer avec ces serveurs. Plusieurs logiciels (eMule, MLDonkey, etc.) implémentent ce protocole.

Celui-ci protocole fonctionne soit par TCP, soit par UDP. Le protocole TCP établit une connexion et est, en général, utilisé pour les communications qui durent dans le temps. Au contraire, UDP est un service sans connexion et est donc plus utilisé pour des messages qui peuvent être transférés en une fois. Chaque pair peut se connecter par TCP à un serveur de son choix et interroger d'autres serveurs par UDP s'il le souhaite. Les connexions TCP étant plus lourdes à mettre en œuvre et plus coûteuses pour les serveurs, les clients n'ont droit qu'à un seul serveur privilégié.

Quand un serveur reçoit une demande de connexion d'un client (toujours en TCP), il se fait passer pour un pair et tente de le contacter en retour. S'il y arrive, cela signifie que le client est joignable et ce dernier obtient un identifiant privilégié (HighId). Si, au contraire, le client n'est pas joignable à cause d'un pare-feu ou de problèmes de réseaux ponctuels, alors il obtient un LowId. Les clients LowId ne peuvent donc être contactés ni par des clients LowId, ni par des clients HighId. Seul le serveur peut communiquer avec eux en utilisant la connexion TCP établie par le client. Les clients HighId peuvent, quant à eux, être contactés par tout le monde, leur HighId contenant leur adresse IP.

L'idéal pour le système serait que tous les clients aient un HighId afin de pouvoir tous dialoguer entre eux. Les clients HighId sont donc avantagés par rapport aux LowId, le serveur acceptant de leur servir de relais s'ils souhaitent contacter un client LowId. Pour cela, le serveur contacte le LowId et lui signale l'intention du HighId. Le LowId peut ensuite

joindre le HighId pour accéder à sa requête. La Figure 6.1 illustre les échanges entre deux clients HighId et entre un HighId et un LowId.

Une fois la connexion établie, le client donne au serveur une liste de tous les fichiers qu'il souhaite mettre à disposition. Ces fichiers sont codés sous forme de métadonnées qui contiennent, entre autres, un code de hachage du fichier. Ces métadonnées sont stockées par le serveur dans une table spéciale pour répondre aux requêtes futures des clients. Une fois cette procédure terminée, le protocole se base sur quelques requêtes simples.

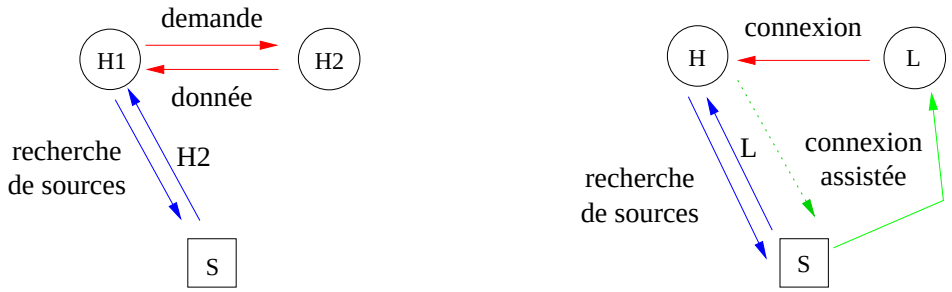

FIG. 6.1 – *À gauche : échanges entre deux clients HighId. Dans un premier temps H1 fait une recherche de sources vers le serveur S qui lui répond avec l'Id de H2. H1 contacte directement H2 qui lui envoie la donnée. À droite : échanges entre un client HighId (H) et un client LowId (L). H fait une recherche de sources et obtient l'Id de L. H demande alors à S de prévenir L qu'il veut le contacter. S contacte alors L qui envoie la donnée directement à H.*

Le premier type de requête, que nous ne discuterons pas plus avant, concerne la recherche de fichiers. De manière interactive, l'utilisateur peut demander au serveur l'ensemble des fichiers disponibles vérifiant certains critères, comme le nom, la taille, le type, etc. Le serveur répond avec les métadonnées de tous les fichiers vérifiant les critères. L'analyse de ce type de requête est un sujet de recherche prometteur, mais relève *a priori* plus du domaine de la sémantique que de celui des graphes. Il ne sera pas discuté dans la suite.

L'autre type de requête concerne la recherche de sources. Après avoir obtenu une liste de métadonnées correspondant à ses désirs, un utilisateur peut demander à un serveur quels sont les clients qui possèdent un fichier donné. Ces requêtes peuvent être effectuées par TCP si le client est connecté au serveur qu'il interroge, ou par UDP sinon. Les clients peuvent donc interroger plusieurs serveurs pour obtenir plus de réponses. En effet, les serveurs n'échangent pas leurs données. Si le client a un HighId, alors il va obtenir une liste de clients à la fois HighId et LowId, puisqu'il peut contacter les premiers directement et se faire aider par le serveur pour les seconds. Si le client a un LowId ou a interrogé le serveur par UDP, il n'aura que des HighId en réponse.

Afin de ne pas surcharger les serveurs, chaque client doit émettre ses recherches de source de manière groupée toutes les 5 minutes. Les clients ont aussi le droit de réinterroger un serveur toutes les 15 minutes pour obtenir de nouvelles sources, s'il y en a. Chaque serveur gère couramment plusieurs dizaines (ou centaines) de milliers de clients

TCP simultanément et doit donc répondre à quelques centaines de requêtes par seconde. Les clients TCP représentant environ un quart des requêtes (le reste étant en UDP), on comprend mieux la raison de ces mesures de restriction.

Les mesures

Nous allons par la suite nous intéresser principalement aux échanges entre les pairs et par conséquent nous concentrer sur les recherches de sources. Un serveur recevant ce type de requête va répondre par un message que nous enregistrons sous la forme suivante :

$$[T] \quad S \quad C \quad H \quad P_1, P_2, \ldots, P_n$$

où T est la date à laquelle la requête a eu lieu (avec une précision à la seconde), S est l'identifiant du pair qui a émis la requête, C est le protocole utilisé (UDP ou TCP), H est le code de hachage du fichier demandé par S et $(P_i)_{i=1\ldots n}$ est une liste de pairs qui ont signalé au serveur qu'ils fournissent le fichier demandé. Pour alléger ses communications, le serveur ne retourne qu'un nombre borné de pairs, nombre qui dépend de ses capacités (typiquement de l'ordre de 1 000 environ). Nous dirons dans la suite que S a fait une requête pour H et que P_i a été cité par le serveur pour H.

Nous stockons aussi toutes les requêtes de `connexion` et de `déconnexion` sous la forme :

$$[T] \quad S$$

si le pair S se connecte ou se déconnecte au temps T.

Pour pouvoir stocker les réponses du serveur, Lugdunum [131], qui est le serveur `eDonkey` le plus populaire et le plus efficace actuellement, a été modifié par son concepteur. Les mesures que nous allons présenter plus loin ont été effectuées sur un AMD Opteron 246 avec 3.2 Go de mémoire et un noyau Linux 64 bits. Cette configuration permet de gérer jusqu'à 250 000 clients simultanément. En pratique, le serveur Lugdunum le plus connu, "Razorback 2", gère couramment 600 000 clients (avec un bi-Opteron 248 et 6 Go de mémoire). Les autres serveurs Lugdunum gèrent entre 50 000 et 100 000 clients et ne sont vraiment utilisés à plein régime que lorsque "Razorback 2" est arrêté.

Nous avons mémorisé plusieurs traces de diverses longueurs, la plus longue durant 48 heures. Pendant cette période, le serveur gère environ $1,5$ million de connexions et de déconnexions, et environ 210 millions de requêtes de recherche de sources. Les données brutes représentent plusieurs giga-octets par jour et le coût de traitement est assez élevé, ce qui explique que nous n'ayons pas, à ce jour, effectué de mesure plus longue. Dans la suite, nous présenterons principalement des résultats sur cette trace de 48 heures, sauf dans certains cas particuliers où nous nous sommes restreints aux 800 premières minutes (13 heures et 20 minutes). Cette durée est la limite pour laquelle certains calculs sont encore possibles dans un temps raisonnable (quelques jours).

Il faut noter que toutes les traces ont été lancées en même temps que le serveur était mis en place, ce qui permet d'observer la phase de lancement puis la phase de stabilisation de l'activité du serveur.

D'autre part, nous considérons ces traces comme assez représentatives des échanges qui sont vraiment effectués dans un réseau P2P quelconque pour trois raisons principales :

- les requêtes effectuées par les pairs dépendent de leurs goûts et non du protocole qu'ils utilisent ;
- les phénomènes observés ne varient pas de manière significative d'une trace à l'autre et certains résultats sont confirmés par des études antérieures sur d'autres protocoles ;
- la taille de la trace nous assure que la plupart des comportements sont capturés, très peu de pairs restant connectés sur de longues périodes.

Dans ce qui suit, nous allons tout d'abord présenter quelques résultats sur la trace, puis effectuer des mesures sur la dynamique des degrés du graphe des requêtes.

6.2 Analyse des requêtes

Les résultats présentés dans cette section concernent la trace de 48 heures. La Figure 6.2 montre l'évolution du nombre de clients durant cette période, chaque ligne verticale correspondant à une période de 6 heures. La courbe met en évidence plusieurs phases : la première correspond au lancement du serveur et dure environ 6 heures. Elle permet d'observer que le remplissage du serveur se fait assez lentement, mais atteint finalement un nombre de clients qui reste relativement stable par la suite.

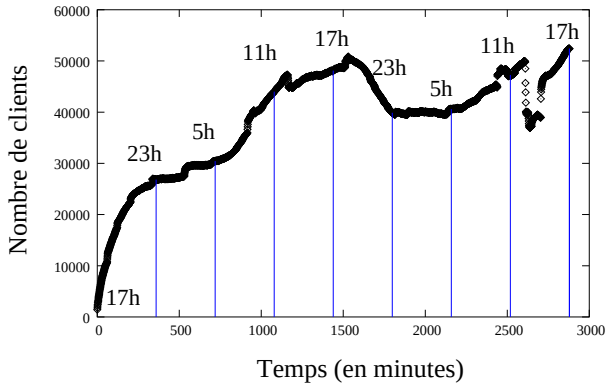

FIG. 6.2 – *Évolution du nombre de clients connectés au serveur pendant* 48 *heures. L'instant* 0 *correspond au lancement du serveur.*

Après cette phase de démarrage, on peut observer une succession de phases stables entrecoupées de phases de vidage et de remplissage correspondant respectivement aux moments des passages du jour à la nuit et de la nuit au jour. Le serveur a été démarré à 17 heures et on peut voir que la première croissance du nombre de connectés débute environ à 7 heures le lendemain. Après une légère croissance, le nombre de clients reste stable puis décroît à la fin de la journée.

Ceci peut aussi se vérifier sur la Figure 6.3 qui montre l'évolution du nombre de `connexions` et `déconnexions` par minute au cours du temps. On observe très clairement une lente variation sinusoïdale atteignant son maximum en milieu de journée et son minimum au milieu de la nuit.

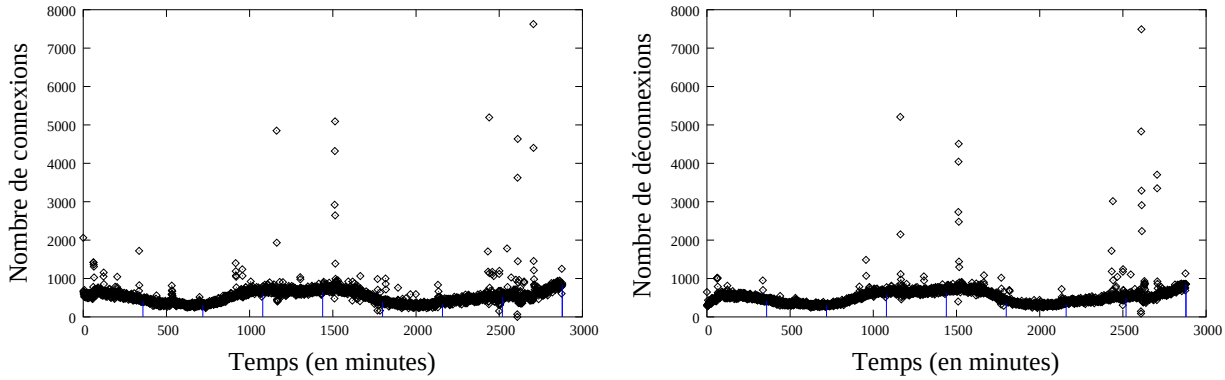

FIG. 6.3 – *Nombre de requêtes de* `connexion` *(à gauche) et de* `déconnexion` *(à droite) minute par minute, durant* 48 *heures.*

Un autre comportement observable, mais moins prévisible, est l'existence de pics dans les requêtes de `connexion` et `déconnexion` à différents instants. De tels pics sont facilement observables aux temps 1 200, 1 500, 2 500 et 2 800 minutes après le début de la trace. Ce comportement massif des utilisateurs peut s'expliquer par l'arrêt brutal d'un autre serveur : dès qu'un serveur s'arrête, tous les clients cherchent immédiatement un nouveau serveur auquel se connecter (le suivant dans une liste). Cela est possible car la majorité des serveurs fonctionnent en sous-régime et peuvent donc accueillir de nouveaux clients en cas de panne. Ces clients ne restent pas forcément connectés, ce qui explique les pics similaires pour les requêtes de `déconnexion` aux mêmes moments.

L'autre type de requête est celui de la recherche de sources. La Figure 6.4 montre le nombre de requêtes de ce type traitées par le serveur, en UDP (pour les clients non connectés au serveur), en TCP (pour les clients connectés) et au total. Le nombre de requêtes UDP est naturellement plus élevé que le nombre de requêtes TCP puisqu'elles correspondent aux clients non connectés au serveur. Les clients interrogent de nombreux serveurs auxquels ils ne sont pas connectés en utilisant UDP et font donc en moyenne plus de requêtes UDP que TCP.

Le phénomène jour-nuit est clairement visible sur cette courbe, le nombre de requêtes étant fortement corrélé avec le nombre de clients connectés. D'autre part, les pannes d'autres serveurs ont un impact direct sur le nombre de requêtes auxquelles le serveur répond, celles-ci diminuant fortement au moment de l'arrivée de nouveaux clients. Il semble donc que le serveur privilégie les clients voulant se connecter au dépens des recherches de sources.

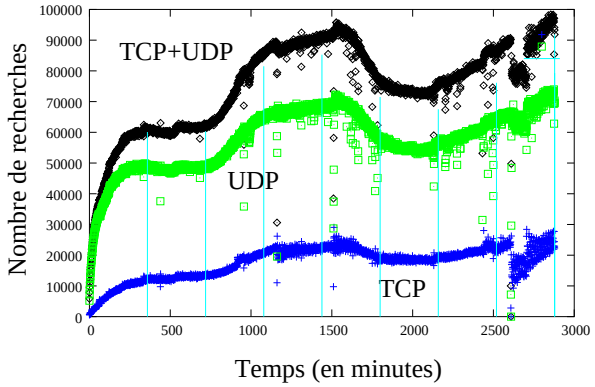

FIG. 6.4 – *Nombre de recherches de sources par minute durant* 48 *heures. De bas en haut sur la courbe :* TCP, UDP *et la somme des deux.*

6.3 Le point de vue des pairs

Il existe plusieurs méthodes pour étudier les propriétés des données récupérées comme décrit précédemment. Ici, nous allons utiliser l'analyse statistique des grands réseaux d'interactions, méthode qui permet d'obtenir une bonne compréhension de certaines propriétés fondamentales de la trace. Cette approche pourrait, et devrait, être complétée par d'autres méthodes, telles que le traitement de signal ou l'analyse sémantique des traces.

Notre méthode consiste à coder les données à l'aide d'un graphe biparti orienté et pondéré, $\mathcal{Q} = (P, D, E, w)$, que nous appelons *le graphe des requêtes*. Ce graphe est défini comme suit :

- P est l'ensemble des pairs dans le réseau, et D est l'ensemble des données (codes de hachage) ;
- $E \subseteq (P \times D) \cup (D \times P)$ est l'ensemble des arcs (orientés) reliant pairs et données : $(p, d) \in E$ si le pair p a effectué une requête concernant la donnée d, et $(d, p) \in E$ si p a été cité par le serveur comme fournissant la donnée d ;
- w est une fonction de poids sur les arcs de E telle que $w(x, y)$ est le nombre de fois où x a demandé y, ou que y a été cité comme fournissant x, pour tout $(x, y) \in E$.

Ce graphe ne stocke pas toute l'information disponible : il n'indique notamment pas les moments auxquels les requêtes sont faites. Il permet, malgré tout, d'observer certaines propriétés fondamentales très simplement. Par exemple, les données les plus demandées sont celles qui ont le degré entrant le plus élevé.

Il serait aussi possible de considérer le graphe des échanges, qui est un graphe orienté $\mathcal{G} = (P, E)$, où P est l'ensemble des pairs dans le réseau et $E \subseteq (P \times P)$ est l'ensemble des arcs tels que $(x, y) \in E$ si x envoie une donnée à y. Ces arcs peuvent être pondérés si un pair est interrogé plusieurs fois par un même autre pair pour la même donnée. Ce graphe peut être calculé à partir des réponses du serveur, mais nous ne l'étudierons pas dans la suite.

Dans la suite de cette section, nous étudions le graphe des requêtes $\mathcal{Q}$, construit à partir d'une trace de 800 minutes, vu sous l'angle des pairs. Nous allons tout particulièrement nous intéresser au degré des pairs. Rappelons que $p \in P$ a un arc vers $d \in D$ si p a fait une requête pour la donnée d, et que d a un arc vers p si ce dernier est cité par le serveur pour d. Il est donc possible d'étudier les valeurs suivantes :

- *le degré sortant* d'un pair est le nombre de données distinctes qu'il a recherchées ;
- *le degré entrant* d'un pair est le nombre de données pour lesquelles il a été cité. Les données fournies par le pair qui ne sont jamais demandées ne sont pas comptabilisées ;
- *le poids d'un lien de* $(P \times D)$*, ou poids sortant*, exprime le nombre de fois où un pair donné demande un fichier donné ;
- *le poids d'un lien de* $(D \times P)$*, ou poids entrant*, exprime le nombre de fois où un pair donné est cité par le serveur comme possédant un fichier donné ;
- *le degré sortant pondéré* d'un pair est la somme des poids des liens sortants de ce pair,. C'est donc le nombre de requêtes qu'il a faites, y compris les requêtes multiples pour une même donnée ;
- enfin, *le degré entrant pondéré* d'un pair, *i.e.* la somme des poids des liens entrants de ce pair, est le nombre de fois où il a été cité par le serveur.

Toutes ces valeurs jouent un rôle important dans la description d'un pair. Par exemple, un pair avec un fort degré sortant recherche beaucoup de données. Un pair avec un fort degré entrant a certainement beaucoup de données en partage, alors qu'un pair avec un fort degré entrant pondéré partage peut-être quelques fichiers très demandés. Si un pair a un fort degré entrant pondéré alors il va être très souvent sollicité par les autres pairs : c'est donc une mesure de sa charge. Nous utiliserons toutes ces notions pour décrire le comportement des pairs.

Considérons, tout d'abord, les distributions des degrés non pondérés (Figure 6.5, à gauche). Ces courbes montrent que les degrés sont très hétérogènes et suivent une loi de puissance : la plupart des pairs ont un faible degré, mais quelques-uns ont un degré très élevé. Il n'y a donc pas de comportement typique qui puisse, par exemple, être pris en compte pour modéliser un système P2P.

La Figure 6.5, à droite, présente les poids entrants et sortants. On peut remarquer que les poids entrants (pairs cités) suivent aussi une loi de puissance. Si l'on considère les liens (p, d), alors quelques pairs sont très souvent sollicités pour quelques fichiers, mais la majorité des pairs ne sont que rarement sollicités pour presque tous leurs fichiers.

Le poids sortant, qui mesure le nombre de fois où un pair demande un même fichier, à une limite. Ceci est la conséquence du fait que les pairs ne peuvent pas demander le même fichier plus d'une fois toutes les 15 minutes : le nombre maximal de requêtes qu'un pair peut envoyer durant une période de temps est fixe ($800/15 \sim 54$ pour la durée de la trace considérée). Les pairs qui ont un degré plus faible que cette limite sont ceux qui envoient moins de requêtes ou se sont connectés au serveur plus tard. Au contraire, les quelques points[2] au delà de cette limite correspondent à des clients qui demandent un même fichier

2. L'échelle étant doublement logarithmique, il y en fait très peu de points au-delà de cette limite, même

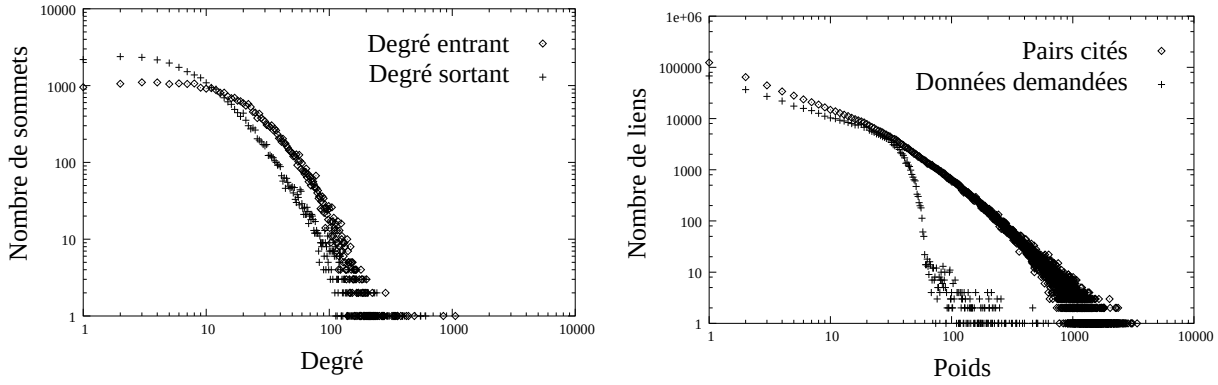

FIG. 6.5 – *À gauche : distributions des degrés entrants et sortants pour les pairs. À droite : distribution des poids entrants et sortants.*

plus souvent que ne l'autorise le protocole. Cette courbe permet donc d'évaluer le nombre de pairs déloyaux, qui peuvent nuire au système en le surchargeant.

La suite naturelle de cette analyse est d'étudier les éventuelles corrélations entre degrés entrants et sortants (Figure 6.6). La courbe de gauche présente un nuage de points, chaque point (x, y) correspondant à un pair de degré entrant x et de degré sortant y. Le nuage de points montre que les pairs avec un fort degré entrant ne sont pas ceux avec un fort degré sortant : les pairs fournissant beaucoup ne sont pas ceux qui interrogent le plus le serveur et réciproquement. Cela permet de détecter la présence de machines dédiées au partage et de pairs déloyaux qui ne fournissent rien mais récupèrent beaucoup. Cette tendance est malgré tout assez peu prononcée, comme le montre la courbe de droite de la Figure 6.6. Cette courbe montre le degré sortant moyen des pairs ayant un degré entrant donné et permet d'observer une assez forte corrélation entre le degré entrant et le degré sortant pour les petites valeurs du degré : en moyenne, les gens se comportent correctement en partageant autant qu'ils récupèrent, à un facteur multiplicatif 2 près.

Les courbes de la Figure 6.7 sur les degrés pondérés, exhibent des phénomènes un peu différents. Ainsi, les pairs qui ont le plus fort degré entrant pondéré envoient assez peu de requêtes, ce qui confirme l'hypothèse de pairs jouant le rôle de fournisseurs. La courbe montrant le degré sortant moyen en fonction du degré entrant (Figure 6.7, à droite) est très bien corrélée pour les sommets ayant un degré inférieur à 1 000. Cette corrélation est beaucoup moins nette au delà de cette valeur à cause du faible nombre de points dans cette région (voir la Figure 6.7, à droite). En moyenne, les pairs reçoivent donc 3 requêtes environ par requête émise, ce qui laisse à penser que le serveur cite en moyenne 3 pairs à chaque requête.

Étudions maintenant l'évolution des degrés au cours du temps. Tout d'abord, la Figure 6.8 présente la distribution des degrés entrants et sortants à divers instants (au tiers, aux deux tiers et à la fin de la trace), ainsi que l'évolution du degré moyen entrant et

si la courbe peut laisser croire le contraire.

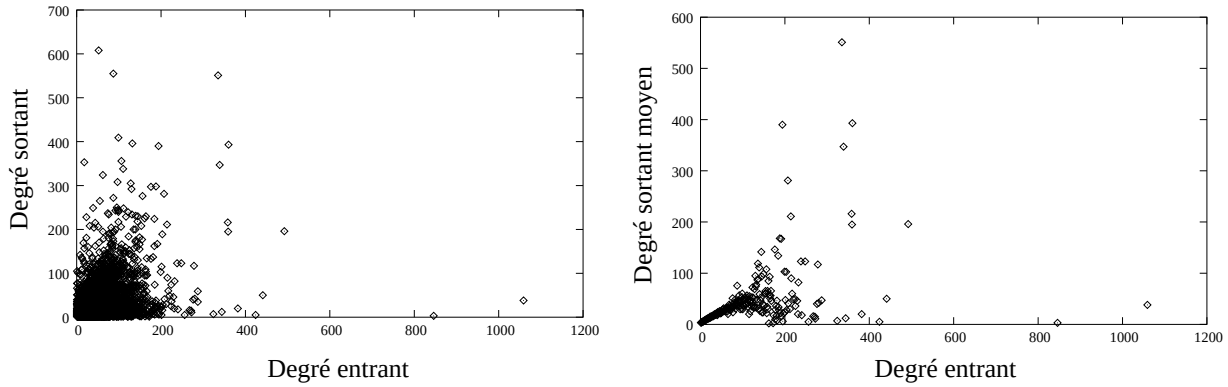

FIG. 6.6 – *À gauche : nuage de points des degrés entrants et sortants. À droite : degré moyen sortant des clients ayant un degré entrant donné.*

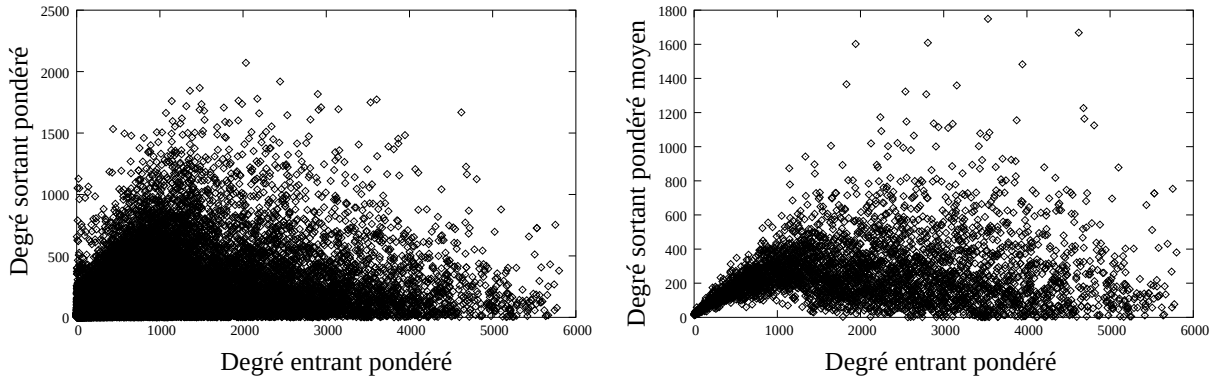

FIG. 6.7 – *À gauche : nuage de points des degrés pondérés entrants et sortants. À droite : degré moyen sortant pondéré des clients ayant un degré entrant pondéré donné.*

sortant. Les courbes de distributions des degrés sont très stables au cours du temps et celles sur les degrés moyens montrent que les pairs ont en moyenne un degré entrant plus élevé que leur degré sortant. Le rapport 2 évoqué précédemment se retrouve ici : ceci peut être expliqué par le fait que les fichiers qui sont récupérés sont rendus disponibles pendant et après leur récupération.

Malgré tout, comme nous l'avons déjà signalé, ce comportement moyen est à considérer avec précaution. Nous avons montré que les comportements sont très variables, la grande majorité des pairs n'ayant ni ce degré entrant, ni ce degré sortant, et encore moins les deux simultanément. L´évolution du degré moyen doit donc être vue comme une propriété globale du système non comme une propriété des pairs.

Pour obtenir des informations plus précises sur le comportement des pairs, nous allons maintenant étudier séparément les clients de faible et de fort degré. Pour les clients de faible degré, on peut étudier la proportion de pairs de degré i (entrant et sortant, pondéré

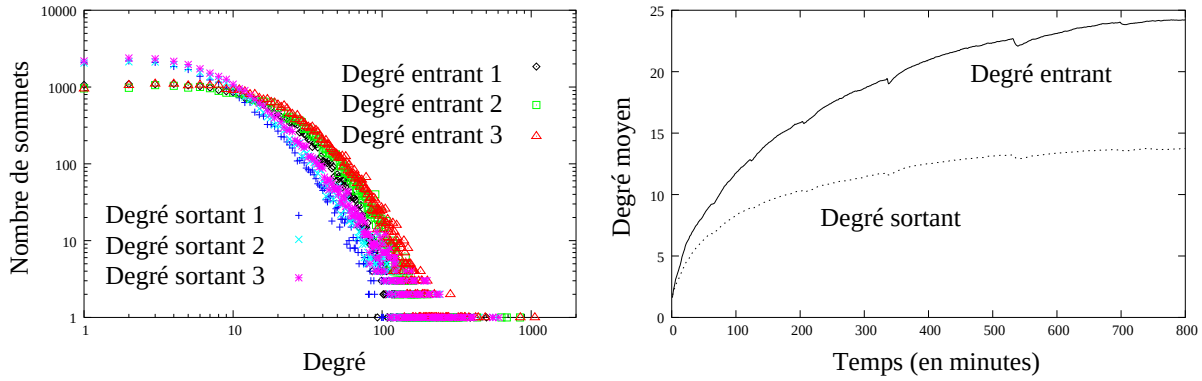

FIG. 6.8 – *À gauche : distribution des degrés entrants et sortants à trois instants. À droite : évolution du degré moyen entrant et sortant au cours du temps. Les irrégularités sont dues à des pics dans les arrivées et les départs à cause du redémarrage d'autres serveurs.*

ou non), pour de petites valeurs de i. Cette proportion est très stable comme on aurait pu le prévoir (d'après la Figure 6.8, à gauche). Il est aussi possible d'étudier l'évolution du degré des sommets de fort degré. La Figure 6.9 présente l'évolution du degré entrant et sortant des pairs ayant un degré maximal à la fin de la trace : à gauche pour les pairs qui ont un degré entrant maximal, à droite pour ceux qui ont un degré sortant maximal.

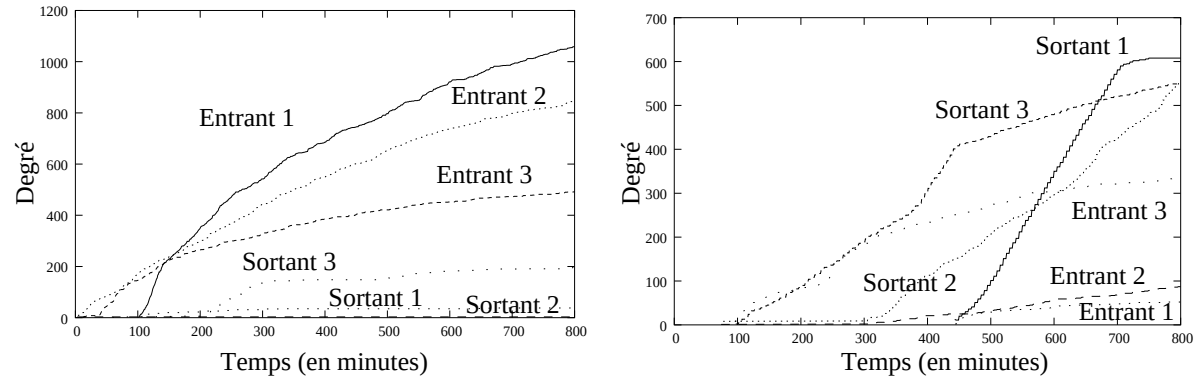

FIG. 6.9 – *Évolution des degrés entrant et sortant des trois pairs qui ont un degré maximal entrant (à gauche) et sortant (à droite).*

Ces deux courbes confirment que les sommets avec fort degré entrant ne sont pas ceux de plus fort degré sortant, et réciproquement. De plus, elles montrent que les sommets de fort degré ont un degré qui croît régulièrement. Tous ces sommets ont un comportement similaire, que l'on peut donc considerer comme typique des pairs très actifs (c'est aussi vrai pour les pairs ayant un fort degré entrant). Cela nous donne des pistes pour modéliser ces types de comportements plus finement et pour pouvoir les simuler.

Dans le cas des pairs ayant un fort degré sortant, on peut remarquer que les courbes

sont croissantes par petits paliers. Ceci est dû au fonctionnement des clients qui n'envoient des requêtes que toutes les 5 minutes, mais n'influence en rien l'allure générale de la courbe.

Étudions maintenant les degrés pondérés. L'évolution temporelle des degrés pondérés pour les pairs ayant un degré entrant maximal est tracée sur la Figure 6.10, à gauche. On peut observer que ces pairs de fort degré entrant pondéré (ceux qui sont souvent cités) ont un faible degré sortant pondéré (ils émettent peu de requêtes). De plus, l'évolution du degré entrant est très régulière et similaire pour les différents pairs, ce qui peut permettre de les modéliser simplement.

Il est possible de considérer le degré entrant pondéré d'un pair comme une mesure de sa charge. La grande hétérogénéité des degrés permet de mettre en évidence le fait que le serveur n'arrive pas à répartir la charge uniformément sur les pairs alors que c'est un des objectifs du protocole. Deux hypothèses peuvent être formulées : des pairs fournissent des données rares, le serveur est donc obligé de les citer, ou ils fournissent de nombreuses données et le serveur, même s'il ne les cite pas à chaque fois, va le faire souvent malgré tout.

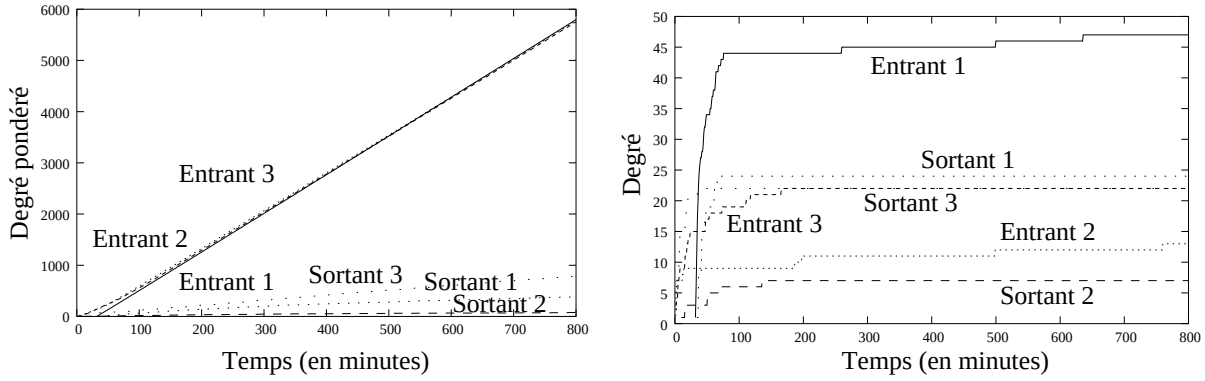

FIG. 6.10 – *À gauche : évolution du degré entrant pondéré des trois pairs ayant un degré entrant maximal à la fin de la trace. À droite : évolution du degré entrant non pondéré de ces mêmes pairs.*

Pour décider entre ces deux hypothèses, observons le degré non pondéré de ces pairs particuliers (Figure 6.10, à droite). Cette courbe montre que tous ces pairs atteignent très rapidement leur degré entrant maximal. On sait donc très vite quel est l'ensemble des données qu'ils fournissent, respectivement 47, 13 et 22 données distinctes. Ces trois pairs ont le même degré pondéré final (ou presque), mais pas le même degré non pondéré : ils ne fournissent donc pas autant de fichiers distincts et en fournissent peu. Les fichiers qu'ils proposent semblent donc assez demandés. De plus, au vu du degré pondéré très élevé des pairs, ces fichiers sont sans doute rares et le serveur est obligé de les citer à chaque fois : la charge est équilibrée entre ces pairs, mais élevée. Les données rares mais convoitées sont souvent des données nouvellement introduites sur le réseau.

De manière similaire, il est possible d'étudier l'évolution temporelle du degré sortant

pondéré pour les pairs de degré sortant pondéré maximal, c'est-à-dire les pairs qui font le plus de requêtes (Figure 6.11). Le degré sortant pondéré d'un pair peut aussi être vu comme une mesure de la charge qu'il induit sur le serveur. Il apparaît que les sommets de fort degré sortant pondéré ont aussi un fort degré entrant pondéré. Une raison pouvant être avancée pour expliquer ceci est que les pairs qui récupèrent des données les laissent disponibles sur leur machine (au moins pendant la récupération). Ce point est très positif pour le système : plus un pair est gourmand, plus il sert le système. Comme plus tôt, la présence de paliers dans l'évolution des degrés est encore dûe au protocole.

Si l'on observe le degré non pondéré de ces sommets (Figure 6.11, à droite), on observe encore une convergence qui peut être plus ou moins rapide. Ceci signifie que certains hauts degrés pondérés proviennent de clients qui envoient (trop) souvent les mêmes requêtes, alors que d'autres proviennent de clients qui font sans cesse de nouvelles recherches.

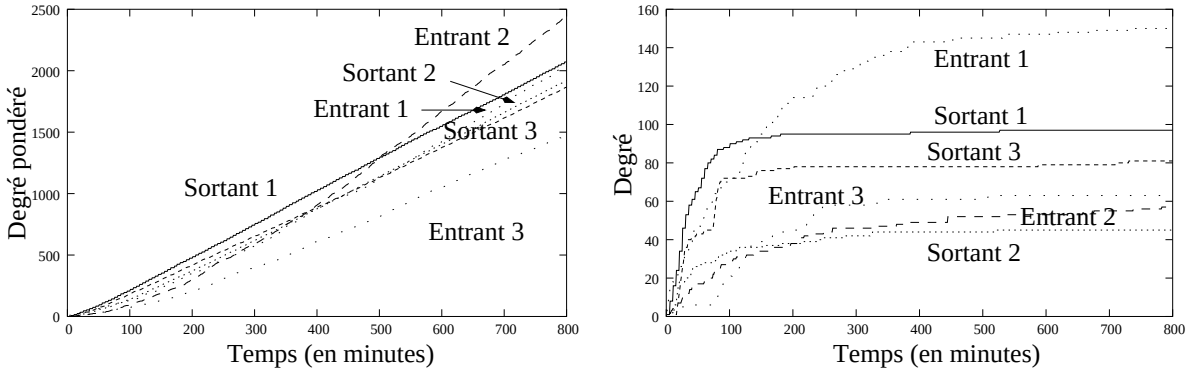

FIG. 6.11 – *À gauche : évolution du degré sortant pondéré des trois pairs ayant un degré sortant maximal à la fin de la trace. À droite : évolution du degré sortant non pondéré de ces* même *pairs.*

6.4 Le point de vue des données

Dans la section précédente, nous avons étudié le comportement des pairs en utilisant l'évolution de leurs degrés dans $\mathcal{Q}$, le graphe biparti des requêtes. Le même type d'étude peut être effectué du point de vue des données, qui constituent l'autre partie de $\mathcal{Q}$. Les résultats étant souvent similaires à ceux obtenus sur les pairs, nous allons un peu moins nous attarder sur cette partie.

Une donnée d dans D a un lien entrant venant d'un pair p si p a fait une requête concernant d. Au contraire, d a un lien sortant vers p si ce dernier a été cité pour d. Les poids sont définis comme précédemment, les degrés quant à eux sont :

- *le degré entrant* d'une donnée est le nombre de pairs (distincts) qui l'ont recherchée ;
- *le degré sortant* d'une donnée est le nombre de pairs (distincts) qui ont été cités pour elle ;

– *le degré entrant pondéré* d'une donnée est le nombre de fois où elle a été demandée ;
– *le degré sortant pondéré* d'une donnée est le nombre de pairs qui ont été cités par le serveur comme la fournissant. Un même pair cité deux fois compte double.

La Figure 6.12 présente, à gauche, la distribution des degrés entrants et sortants à différents instants (au tiers, aux deux tiers et à la fin de la trace) et, à droite, l'évolution des degré moyens. Comme pour les pairs, les degrés sont très hétérogènes et toutes les distributions suivent une loi de puissance. Ceci signifie en particulier qu'il n'y a pas de donnée typique dans le système. Le système est aussi très stable, les distributions variant très peu au cours du temps. Les courbes montrent également que les degrés moyens convergent lentement avec un degré sortant moyen supérieur au degré moyen entrant. On retrouve le facteur 2 déjà évoqué pour les pairs : le degré moyen sortant d'une donnée est deux fois plus élevé que le degré moyen entrant. Ceci est parfaitement naturel car un lien d'une donnée d vers un pair p comptant pour le degré sortant de d et pour le degré entrant de p.

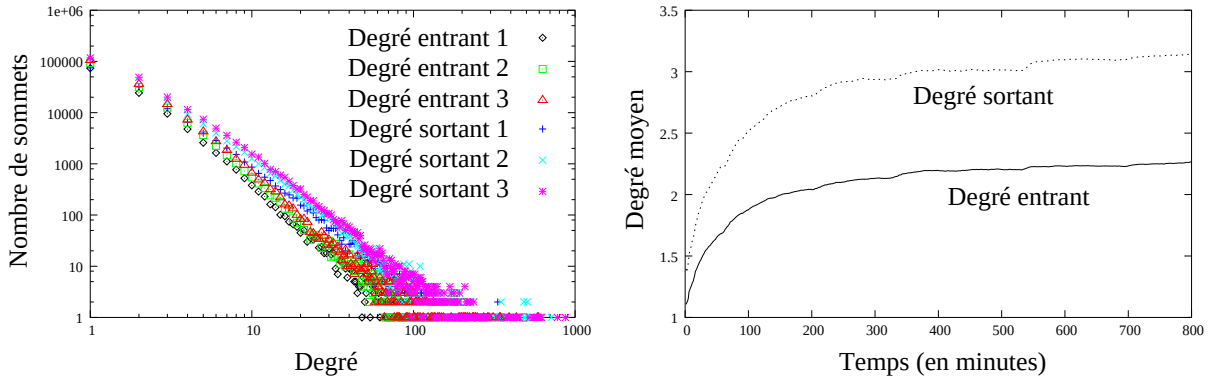

FIG. 6.12 – *À gauche : distributions des degrés entrants et sortants des données à différents instants. À droite : évolution du degré entrant et sortant moyen.*

La Figure 6.13 présente les corrélations entre les degrés entrants et sortants des données. Il apparaît très clairement que les données souvent demandées sont aussi les données très partagées : le nuage de points fait apparaître une tendance linéaire entre l'offre et la demande. Ceci est un point positif pour le protocole qui peut donc éviter de trop surcharger les pairs qui fournissent des données très demandées, ces données étant généralement très fournies aussi.

Cette courbe montre aussi que, dans la plupart des cas, les données ont un degré sortant plus élevé que leur degré entrant, la plupart des points du nuage étant au-dessus de la première bissectrice. Ce résultat était attendu, le degré sortant moyen étant supérieur au degré moyen entrant. Cette relation entre les degrés entrants et sortants semble donc assez générale.

Si l'on se penche plus précisément sur les données ayant un degré final maximal (Figure 6.14), on se rend compte que les données les plus fournies sont les plus demandées et

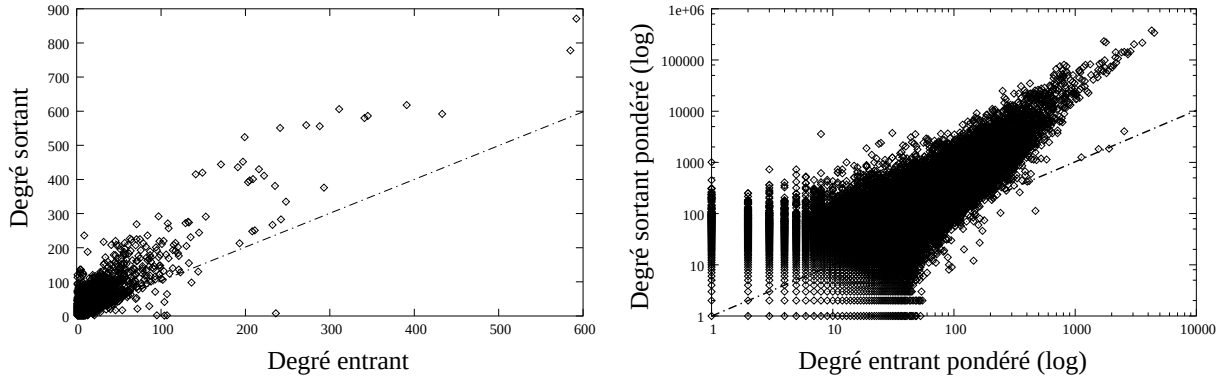

FIG. 6.13 – *À gauche : nuage de points entre les degrés entrants et sortants des données. À droite : même nuage avec les degrés pondérés.*

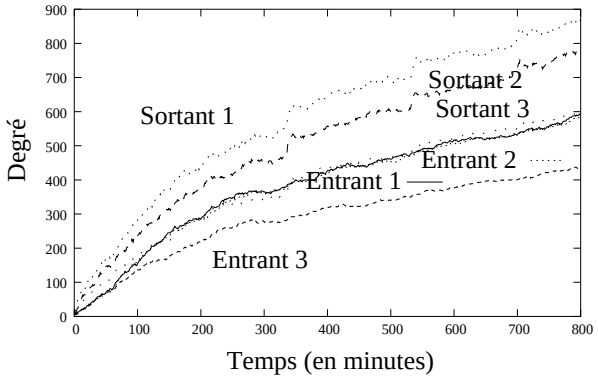

FIG. 6.14 – *Évolution temporelle des degrés entrants et sortants des* 3 *données ayant un degré maximal entrant et sortant. Ce sont les mêmes* 3 *données qui ont les degrés entrants et sortants maximaux.*

réciproquement. Ceci est plutôt positif, même si nous avons vu précédemment que certains pairs fournissent des données peu disponibles mais très demandées.

6.5 Conclusion

L'étude du système P2P `eDonkey` présentée ici à été possible grâce à la nature semi-centralisée du système, qui en fait un très bon candidat pour ce type de mesures. Nous avons utilisé une grande quantité de données provenant d'un serveur de forte capacité pour obtenir des informations sur les comportements des pairs, en termes de `connexions` et de `déconnexions`, mais aussi concernant leurs requêtes. Nous avons aussi étudié les degrés des pairs, leur évolution temporelle et les corrélations entre ces degrés. Puis, avec les mêmes techniques, nous avons étudié les propriétés des données échangées.

L'analyse de ces données met en évidence plusieurs phénomènes, certains provenant du protocole, mais la plupart étant liés au comportement des pairs.

Deux points principaux peuvent être notés :

- les pairs et les données qu'ils fournissent sont très hétérogènes, ce qui rend invalide toute notion de pair ayant un comportement moyen, ou de donnée typique ;
- les pairs peuvent être classés dans différents groupes, ceux fournissant beaucoup de données, ceux ayant des données rares, ceux faisant beaucoup de requêtes, etc. Ces groupes correspondent à des comportements bien identifiés. De même, les données peuvent être classées en diverses catégories.

Il est aussi possible d'identifier des pairs utilisant des clients modifiés. L'évaluation du nombre de tels pairs et de leur impact sur le système peut être cruciale dans la définition de nouveaux systèmes.

Ces résultats mettent en évidence un certain nombre de points utiles pour modéliser finement quelques comportements typiques. Ils doivent maintenant être approfondis pour les décrire plus précisément et proposer effectivement des modèles pouvant servir à la conception de systèmes P2P performants.

Soulignons enfin que l'étude des cas particuliers a amené, comme c'est souvent le cas, à introduire de nouvelles notions pour l'analyse des grands réseaux d'interactions. En particulier, nous avons été amenés à conduire une étude de l'évolution temporelle du graphe des requêtes, notamment en identifiant des propriétés significatives distinctes pour différents types de sommets : l'évolution de leur degré pour les sommets de fort degré, et l'évolution de leur nombre pour ceux de faible degré. Cette distinction est pertinente pour tous les graphes sans-échelle et constitue un premier pas dans la définition de propriétés pour l'analyse statistique de la *dynamique* des grands réseaux d'interactions.

Chapitre 7

Résistance aux pannes et aux attaques

Introduction.

La connexité du graphe est, pour la plupart des grands réseaux d'interactions, un point crucial. Dans le cas de l'Internet, cela signifie que chaque ordinateur peut communiquer avec tous les autres. Dans le cas des réseaux sociaux, cela conditionne la capacité d'un virus à se propager d'individu en individu, ou d'une information à se propager. Enfin, pour le graphe du Web, la connexité signifie qu'un utilisateur peut naviguer d'une page à l'autre.

On peut donc utiliser la taille de la plus grande composante connexe d'un réseau comme première mesure de la qualité du service qu'il fournit, et étudier ce paramètre quand le réseau subit des pannes ou des attaques qui suppriment une partie de ses sommets ou de ses liens.

Deux phénomènes peuvent être observés sur un réseau comme l'Internet : des pannes et des attaques. Les pannes sont accidentelles et ne touchent pas de machines en particulier. Les attaques, au contraire, visent à empêcher le réseau de fonctionner, et sont donc ciblées pour détruire ou bloquer temporairement certains routeurs. On espère que le réseau de l'Internet est résistant à ces deux phénomènes.

De même, dans un réseau social, une attaque contre un virus peut correspondre à la vaccination d'un individu, le virus ne pouvant plus se propager en passant par lui. Si un réseau social est sensible aux pannes, alors une vaccination non ciblée (consistant à vacciner des personnes aux hasard) sera efficace. À l'opposé, si le réseau est résistant aux attaques, alors aucune stratégie ne pourra empêcher la propagation du virus.

Plus généralement, dans beaucoup de contextes, étudier la résistance des réseaux aux pannes et aux attaques peut apporter un éclairage pertinent sur les méthodes pour favoriser ou gêner certains phénomènes ayant lieu sur ces réseaux.

Notre objectif dans ce chapitre est de comprendre en quoi la distribution des degrés d'un graphe influence, positivement ou négativement, la résistance de ce graphe aux pannes et aux attaques. Pour cela, nous allons étudier formellement, et pour plusieurs types de réseaux, les effets des pannes et des attaques.

Plusieurs travaux antérieurs avaient déjà abordé le problème et y avaient apporté des

réponses pertinentes. Notre approche a été d'explorer en profondeur ces résultats pour comprendre très précisément les raisons qui font que tel ou tel type de réseau est plus ou moins résistant aux pannes ou aux attaques. Nous commençons par présenter les résultats existants, puis nous introduisons deux nouvelles stratégies d'attaque, l'une sur les liens et l'autre sur les sommets.

7.1 Préliminaires.

Dans cette section, nous présentons formellement les concepts liés à la robustesse des grands réseaux d'interactions. Nous donnons ensuite quelque résultats généraux sur les graphes, qui seront utiles dans la suite du chapitre. Enfin, nous décrivons les principales techniques de preuve utilisées par la suite.

Contexte

Le concept de robustesse des réseaux face aux pannes ou aux attaques a été introduit dans [9]. Dans cet article, deux types de phénomènes pouvant endommager un réseau sont considérés : les pannes, qui sont accidentelles et n'ont pas de cible particulière, et les attaques, qui visent à déconnecter le réseau. Les pannes de sommets (resp. de liens), étant de nature aléatoire, sont modélisées par des suppressions aléatoires de sommets (resp. de liens). On peut en revanche imaginer plusieurs stratégies d'attaque. Celle qui a été considérée dans [9] consiste à supprimer les sommets de plus fort degré en priorité. Nous ferons dans la suite référence à cette stratégie sous le terme d'*attaque classique*.

Il apparaît dans cet article qu'il y a une différence qualitative entre le comportement des graphes aléatoires (modèle ER d'Erdös et Renyi) et celui des graphes sans-échelle (modèle MR de Molloy et Reed). Dans le cas des pannes, pour les graphes aléatoires la taille de la composante géante devient nulle quand une fraction strictement inférieure à 1 des sommets est supprimée, alors que cette taille décroît très lentement pour les graphes sans-échelle et n'atteint 0 que quand presque tous les sommets ont été supprimés. Les graphes sans-échelle sont plus résistants aux pannes que les graphes aléatoires. Le contraire semble également vrai pour les attaques : celles-ci sont équivalentes à des pannes dans les cas des graphes aléatoires, puisque tous les sommets ont pratiquement le même degré. Au contraire, dans le cas des graphes sans-échelle, certains sommets ont un très fort degré et jouent un rôle central : les supprimer déconnecte le réseau très rapidement. Les graphes sans-échelle s'effondrent donc beaucoup plus rapidement que les graphes aléatoires. La Figure 7.1 illustre ce phénomène. La distribution des degrés en loi de puissance pour l'Internet, qui le rendrait très résistant aux pannes, mais aussi sensible aux attaques a reçu le nom de *talon d'Achille de l'Internet* [111]. Ces résultats ont été par la suite confirmés à la fois expérimentalement et de manière formelle [9, 25, 26, 28, 35, 109].

La conclusion générale est la suivante : les graphes sans-échelle sont très résistants aux pannes mais très sensibles aux attaques. Ce dernier point peut être interprété de manière

positive en épidémiologie mais est plus inquiétant si l'on s'intéresse à la résistance du réseau de l'Internet.

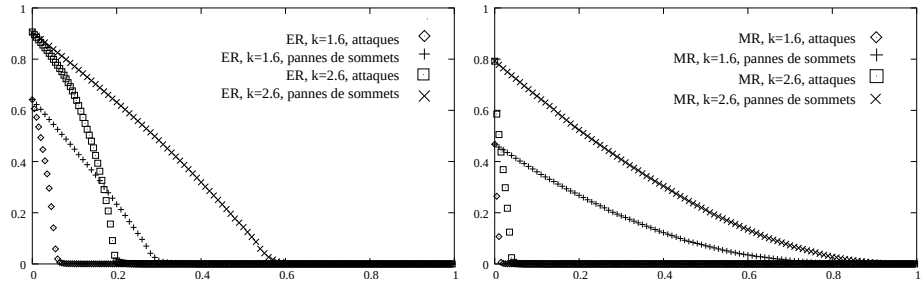

FIG. 7.1 – *Effets des pannes et des attaques sur les graphes aléatoires (à gauche) et sur les graphes sans-échelle (à droite). Les courbes présentent la taille de la plus grande composante connexe en fonction de la fraction de sommets supprimés. Différentes valeurs du degré moyen sont considérées.*

Le cas où l'on supprime des liens plutôt que des sommets n'a été que peu étudié. Cependant, les mêmes questions se posent dans ce contexte ; nous les étudierons donc également. On verra que certains résultats sont des extensions naturelles des résultats existants pour les sommets, mais que d'autres soulèvent de nouvelles interrogations.

Notations et résultats utiles

Avant d'entamer l'étude des pannes et des attaques sur des sommets ou des liens, il convient de noter quelques points particulièrement importants sur les diverses approximations que nous allons utiliser par la suite et les justifications qu'elles reçoivent.

Dans toute la suite, on note p_k la proportion de sommets de degré k, ζ la fonction de Riemann, définie si $\alpha > 1$, et qui vaut : $\zeta(\alpha) = \sum_{k=1}^{\infty} \frac{1}{k^\alpha}$. On appelle nombre harmonique, et on note $H_K^{(\alpha)}$, la quantité $\sum_{k=1}^{K} k^{-\alpha}$. Le degré moyen d'un sommet vaut $\langle k \rangle = \sum_{k=1}^{\infty} k p_k$.

Dans ce contexte, il est habituel d'observer des phénomènes de seuil (typiques de la théorie de la percolation et de la mécanique statistique en général) : il existe un seuil critique p_c tel que, tant que la fraction de sommets supprimés est inférieure à p_c, le réseau a presque sûrement une composante géante, alors que si la fraction est supérieure à p_c, le réseau est presque sûrement déconnecté [126]. Les seuils jouent un rôle central dans les phénomènes que nous considérons. Dans la suite, nous nous poserons donc la question de leur existence et de leur valeur dans divers contextes.

Nous allons maintenant nous intéresser à quelques propriétés des graphes que nous utilisons. Tout d'abord, nous aimerions insister sur le fait que, comme les graphes que nous considérons sont peu denses, les liens sont placés de façon aléatoire et leur taille tend vers l'infini, on peut considérer que ces réseaux sont localement équivalents à des arbres : quand on regarde l'ensemble des sommets à une petite distance d'un sommet donné, la partie

du graphe qu'on découvre n'a pas de cycle (rappelons que les liens sont placés de manière aléatoire).

Considérons maintenant un lien au hasard dans le graphe et suivons-le pour arriver à un sommet. Alors on peut faire les remarques suivantes sur la distribution des degrés du sommet atteint de cette manière :

- cette distribution est indépendante du sommet de départ. En effet, un lien étant formé de deux demi-liens aléatoires, le choix du second demi-lien qui est celui par lequel on arrive au sommet est indépendant du sommet de départ ;
- cette distribution est différente de la distribution des degrés des sommets, et donc différente de celle obtenue en prenant un sommet au hasard. En effet, un lien étant formé de deux demi-liens, la probabilité qu'un demi-lien donné appartienne à un sommet donné v est proportionnelle au degré de v.

Ces deux remarques permettent de conclure que la probabilité d'atteindre un sommet de degré k en suivant un lien quelconque est proportionnelle à kp_k. Pour calculer la probabilité exacte, il faut noter que la somme des probabilités doit faire 1, et donc cette probablité vaut $kp_k/\sum_{j=0}^{\infty} jp_j = kp_k/\langle k\rangle$. Soit q_k la probabilité que le sommet à l'extrémité du lien ait k voisins, autres que le sommet au bout du lien par lequel on est arrivé. Ce nombre est égal à 1 moins le degré du sommet, et donc :

$$q_k = \frac{(k+1)p_{k+1}}{\langle k\rangle} \tag{7.1}$$

De nombreux résultats concernant la robustesse des réseaux utilisent le fait qu'après des suppressions de sommets ou de liens, le réseau restant est un réseau aléatoire avec une distribution des degrés donnée p_k. Dans ce cas, il existe un critère permettant de décider si ce réseau a une composante géante, en fonction de p_k. Un tel critère peut être défini de plusieurs manières [34, 93, 101].

Théorème 7.1.1 *Un graphe dont la taille tend vers l'infini et dont la distribution des degrés est p_k a presque sûrement une composante géante si et seulement si :*

$$\langle k^2\rangle - 2\langle k\rangle = \sum k(k-2)p_k \geq 0$$

Preuve : Ce critère, prouvé formellement dans [93], peut être compris intuitivement de la manière suivante [101] : comparons le nombre moyen de voisins d'un sommet aux nombre moyen de ses voisins à distance 2 (*i.e.* les voisins de ses voisins, sans compter le sommet lui-même). Le nombre moyen de voisins d'un sommet est $\langle k\rangle$.

Pour calculer le nombre moyen de voisins à distance 2, rappelons que le nombre de voisins d'un sommet atteint en suivant un lien au hasard vaut q_k (Équation 7.1). La valeur moyenne de cette quantité est :

$$\langle q_k\rangle = \sum_{k=0}^{\infty} kq_k = \sum_{k=0}^{\infty} k(k+1)p_{k+1}/\langle k\rangle = \sum_{k=0}^{\infty}(k-1)kp_k/\langle k\rangle = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}$$

Comme le réseau est localement arborescent, deux voisins d'un même sommet n'ont aucun voisin en commun autre que le sommet initial. Donc le nombre de voisins à distance 2 d'un sommet est égal à la somme du nombre de voisins de ses voisins moins un. Le nombre moyen de voisins d'un sommet est $\langle k \rangle$, le nombre de voisins à distance 2 est donc :

$$z_2 = \langle k \rangle \langle q_k \rangle = \langle k^2 \rangle - \langle k \rangle \tag{7.2}$$

Ce raisonnement peut s'étendre aux voisins à distance n (c'est ici que l'approche n'est plus rigoureuse, puisqu'on ne regarde plus le graphe localement) :

$$z_n = \frac{z_2}{\langle k \rangle} z_{n-1} = \left(\frac{z_2}{\langle k \rangle} \right)^{n-1} z_1 \tag{7.3}$$

Donc, si $z_2/\langle k \rangle > 1$, le nombre de voisins à distance n diverge avec n et devient infini. Au contraire, si $z_2/\langle k \rangle < 1$, le nombre de voisins à distance n tend vers zéro et la composante est de taille finie.

La transition a lieu quand $z_2 = \langle k \rangle$, d'où le résultat. □

Méthodes de preuves

Deux approches principales ont été utilisées avec succès pour étudier les pannes et les attaques. Ces techniques sont peu utilisées en informatique, ainsi afin de faire une présentation claire et complète des travaux de ce domaine, nous allons expliciter les preuves de certains résultats existants, en en fournissant les références.

Newman et al. [28] utilisent le formalisme des séries génératrices. On utilisera dans la suite la série génératrice pour les distributions des degrés des sommets, G_0, définie par :

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \tag{7.4}$$

Ainsi, pour cette série, le coefficient du k-ième donne la probabilité qu'un sommet ait ce degré. On utilisera aussi G_1, la fonction génératrice pour le nombre de liens sortants d'un sommet atteint en suivant un lien au hasard. Ce nombre est distribué suivant q_k (d'après l'Équation 7.1), donc :

$$G_1(x) = \sum_{k=0}^{\infty} q_k x^k = \frac{\sum_{k=0}^{\infty} (k+1) p_{k+1} x^k}{\langle k \rangle} = \frac{\sum_{k=0}^{\infty} k p_k x^{k-1}}{\langle k \rangle} = \frac{G_0'(x)}{\langle k \rangle}, \tag{7.5}$$

où $G_0'(x)$ est la dérivée de $G_0(x)$.

Nous utiliserons dans la suite deux propriétés des séries génératrices :

- la moyenne d'une probabilité représentée par une série génératrice $F(x) = \sum_{k=0}^{\infty} p_k x^k$ vaut $\langle k \rangle = F'(1)$;
- la série génératrice pour la distribution de la somme de k tirages indépendants selon une probabilité donnée par $F(x)$ vaut $F^k(x)$.

Pour étudier les graphes sans-échelle, on utilisera deux méthodes différentes : une approche exacte et une approche utilisant une approximation continue des degrés. Pour un graphe sans-échelle, la distribution des degrés est :

$$p_k = Ck^{-\alpha}, k = m, m+1, \ldots, K \tag{7.6}$$

où m et K sont les degrés minimal et maximal dans le réseau et où C est une constante de normalisation. Nous considérerons la plupart du temps que $m = 1$. Pour les graphes dont la taille tend vers l'infini, on a $K = \infty$. Dans ce cas, $C = 1/\zeta(\alpha)$.

Si l'on utilise plutôt une approximation continue, on cherche C tel que $\int_m^K p_k = 1$. Dans ce cas :

$$\int_m^K p_k = C\left[k^{-\alpha+1}\right]_m^K/(-\alpha+1) = C(K^{-\alpha+1} - m^{-\alpha+1})/(-\alpha+1),$$

et donc $C = (-\alpha+1)/(K^{-\alpha+1} - m^{-\alpha+1})$. Quand le réseau est grand, $K \gg m$ et $K^{-\alpha+1} \ll m^{-\alpha+1}$, on peut donc approximer : $C \approx (\alpha-1)m^{\alpha-1}$.

Il est ensuite possible de déterminer le degré maximal K du réseau :

Lemme 7.1.2 *Dans un graphe sans-échelle à N sommets de degré minimal m, le degré maximal peut être approximé par $K \sim mN^{1/(\alpha-1)}$*

Preuve : K satisfait la relation suivante :

$$\int_K^\infty p_k dk = \frac{1}{N}.$$

Intuitivement, cela signifie que K est tel qu'il n'y a qu'un seul sommet de degré K ou plus, et que donc ce sommet précis a degré K (sinon $\int_{K+1}^\infty p_k dk$ vaudrait $1/N$).

Nous avons donc : $1/N = C\int_K^\infty k^{-\alpha} = C\left[\frac{k^{-\alpha+1}}{-\alpha+1}\right]_K^\infty = C\frac{K^{-\alpha+1}}{\alpha-1}$. Comme $C = (\alpha-1)m^{\alpha-1}$, on obtient le résultat. □

Le degré maximal d'un graphe sans-échelle fini peut aussi être évalué sans utiliser l'approximation continue :

Lemme 7.1.3 *Dans un graphe sans-échelle à N sommets et de degré minimal 1, le degré maximal K peut être calculé en résolvant l'équation suivante : $\sum_K^\infty p_k = 1/N$.*

La Figure 7.2 montre ces deux estimations du degré maximal, ainsi que les résultats obtenus par nos simulations. Comme on peut le voir, ces deux méthodes donnent des résultats similaires mais sous-estiment légèrement les valeurs réelles.

Lecture des courbes

Avant d'entrer dans le cœur de ce chapitre, il convient de dire quelques mots sur les courbes que nous allons présenter par la suite.

Les courbes pour les résultats expérimentaux obtenus par simulation sont toutes des moyennes sur de nombreux échantillons. Ils sont généralement représentatifs du comportement moyen, mais une simulation effectuée sur un seul échantillon peut dévier sensiblement des résultats présentés dans certains cas (en particulier pour les valeurs des seuils des graphes sans-échelle).

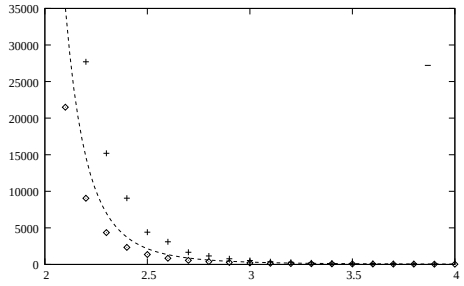

FIG. 7.2 – *Estimations du degré maximal d'un graphe sans-échelle de $N = 100\,000$ sommets, d'après le Lemme 7.1.3 (diamants), le Lemme 7.1.2 (ligne pointillée) et par simulation (croix).*

Concernant les valeurs des seuils, nous avons considéré que le réseau est complètement déconnecté quand la taille de la plus grande composante devient plus petite que 1% du total. Cette hypothèse n'a pas d'influence visible sur les résultats présentés. Toujours pour les seuils, ceux-ci sont calculés en fonction du degré moyen du graphe dans le cas des graphes aléatoires, et de l'exposant pour les graphes sans-échelle. Ce sont en effet les paramètres significatifs dans les deux contextes.

Certaines courbes utilisées pour comparer l'effet d'une stratégie de suppression de sommets ou de liens sur des graphes aléatoires et des graphes sans-échelle utiliseront des graphes avec le même degré moyen. Nous avons choisi les valeurs $1,6$ et $2,6$ qui, pour des graphes sans-échelle, correspondent à des exposants $2,5$ et 3 respectivement. Ces valeurs sont représentatives des exposants des graphes réels.

À plusieurs endroits nous présenterons des évaluations numériques pour des formules d'approximation. Ces formules analytiques ne sont pas toujours solvables exactement, car le degré des sommets d'un graphe ne peut prendre que des valeurs entières. Ces courbes ne sont donc que des approximations.

Enfin, tous les graphes sans-échelle utilisés dans ce chapitre sont des graphes MR, présentés plus tôt dans le Chapitre 3 [92]. Nous avons généré les graphes sans-échelle de N sommets et d'exposant α en tirant N degrés parmi une loi de puissance d'exposant α, avec comme valeur minimale 1 et comme valeur maximale N. Nous avons ensuite engendré des graphes aléatoires ayant cette distribution des degrés avec le modèle MR.

Le fait que ces graphes soient engendrés avec le modèle MR est très important pour toutes les preuves de ce chapitre : nous utilisons le fait que les liens soient formés de paires de demi-liens tirées au hasard.

7.2 Résistance aux pannes

Dans cette section, nous allons étudier la résistance des graphes aléatoires et des graphes sans-échelle aux pannes.

Nous allons tout d'abord présenter un résultat général, indépendant de la distribution

des degrés du graphe considéré. Nous allons détailler les deux preuves existantes pour ce résultat [28, 34]. Nous appliquerons ensuite ce résultat aux deux cas qui nous occupent, les graphes aléatoires et les graphes sans-échelle.

Nous étudierons ensuite l'impact des pannes de sommets du point de vue des liens et, enfin, nous étudierons les pannes de liens, qui se traitent de manière très similaire.

7.2.1 Résultats généraux sur les pannes de sommets

Soit un réseau dont la distribution des degrée est p_k, telle que le réseau ait une composante géante. Notre objectif est de prouver le résultat suivant :

Théorème 7.2.1 *[28, 34] Le seuil p_c pour les pannes de sommets, pour un réseau dont la taille tend vers l'infini, est donné par :*

$$p_c = 1 - \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$$

Ce résultat est cohérent avec le Théorème 7.1.1 : p_c est positif si et seulement si la distribution des degrés est telle que le réseau soit presque sûrement connexe :

$$1 - \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \geq 0 \iff \langle k^2 \rangle - 2\langle k \rangle \geq 0.$$

Ce résultat peut être obtenu de deux manières différentes. La première, proposée dans [34], consiste à calculer la nouvelle distribution des degrés après les suppressions pour pouvoir utiliser le critère du Théorème 7.1.1, et savoir si le réseau est toujours connexe ou pas. La seconde, proposée dans [28], s'intéresse à la taille moyenne des composantes connexes finies après les suppressions. Le seuil est l'instant où cette taille moyenne diverge : en effet, c'est le moment où la composante infinie devient finie. À cet instant, elle se transforme donc en une composante finie dont la taille tend vers l'infini. La taille moyenne des composantes finies diverge donc elle aussi en ce point.

Nous allons maintenant donner ces deux preuves, en entrant dans les détails des raisonnements utilisés.

Supposons qu'une fraction p des sommets soit supprimée lors de pannes et notons $p_k(p)$ la distribution des degrés du réseau après la suppression. On a alors :

Lemme 7.2.2 *[34] La distribution $p_k(p)$ après suppression aléatoire d'une fraction p des sommets est donnée par :*

$$p_k(p) = \sum_{k_0=k}^{\infty} p_{k_0} \binom{k_0}{k} (1-p)^k p^{k_0-k}$$

Preuve : Si un sommet a degré k_0 avant la suppression, alors la probabilité qu'il ait degré $k' \leq k_0$ après la suppression est $\binom{k_0}{k'}(1-p)^{k'}p^{k_0-k'}$. En effet, $k_0 - k'$ de ses voisins ont été supprimés avec probabilité $p^{k_0-k'}$, et k' n'ont pas été supprimés avec probabilité $(1-p)^{k'}$. □

Le point important à noter est que les liens du réseau original sont construits en reliant aléatoirement des paires de demi-liens. Par conséquent, les liens restants ont été construits de la même manière. Le réseau après les suppressions est donc équivalent à un réseau aléatoire avec la nouvelle distribution des degrés.

Il est donc possible d'appliquer le Théorème 7.1.1. Pour cela il faut calculer les deux premiers moments de la nouvelle distribution des degrés :

Proposition 7.2.3 *[34] Les deux premiers moments de la distribution des degrés $p_k(p)$ sont :*

$$\langle k(p)\rangle = (1-p)\langle k\rangle \qquad et \qquad \langle k^2(p)\rangle = (1-p)^2\langle k^2\rangle + p(1-p)\langle k\rangle$$

Pour prouver cette proposition, on a besoin du lemme suivant :

Lemme 7.2.4 *Pour tous k et k_0 entiers, et pour tout p réel, on a :*

$$\sum_{k=0}^{k_0} k^2 \binom{k_0}{k}(1-p)^k p^{k_0-k} = (1-p)^2 k_0^2 + p(1-p)k_0$$

et :

$$\sum_{k=0}^{k_0} k \binom{k_0}{k}(1-p)^k p^{k_0-k} = (1-p)k_0$$

Preuve : Remarquons tout d'abord que :

$$\begin{array}{rcl} (x+y)^{k_0} & = & \sum_{k=0}^{k_0}\binom{k_0}{k}x^k y^{k_0-k} \\ \frac{d}{dx}\left((x+y)^{k_0}\right) & = & k_0(x+y)^{k_0-1} \\ & = & \sum_{k=0}^{k_0}\binom{k_0}{k}kx^{k-1}y^{k_0-k} \end{array}$$

En multipliant les deux cotés de cette dernière égalité par x, on obtient :

$$xk_0(x+y)^{k_0-1} = \sum_{k=0}^{k_0}\binom{k_0}{k}kx^k y^{k_0-k}.$$

et en fixant $x = 1-p$ et $y = p$ dans cette équation on obtient la première affirmation. De la même équation, on obtient aussi :

$$\begin{array}{rcl} \frac{d}{dx}(xk_0(x+y)^{k_0-1}) & = & k_0(x+y)^{k_0-1} + xk_0(k_0-1)(x+y)^{k_0-2} \\ & = & \sum_{k=0}^{k_0} k^2\binom{k_0}{k}x^{k-1}y^{k_0-k} \\ x(k_0(x+y)^{k_0-1} + xk_0(k_0-1)(x+y)^{k_0-2}) & = & \sum_{k=0}^{k_0} k^2\binom{k_0}{k}x^k y^{k_0-k} \end{array}$$

En fixant $x = 1 - p$ et $y = p$, on obtient alors :

$$\begin{aligned} \sum_{k=0}^{k_0} k^2 \binom{k_0}{k} (1-p)^k p^{k_0-k} &= (1-p)k_0 + (1-p)^2 k_0(k_0-1) \\ &= (1-p)^2 k_0^2 + p(1-p)k_0 \end{aligned}$$

ce qui est la deuxième affirmation. □

Il est maintenant possible de prouver la Proposition 7.2.3 :
Preuve : Le résultat énoncé est obtenu à partir de cette suite d'équations :

$$\begin{aligned} \langle k(p) \rangle &= \sum_{k=0}^{\infty} k p_k(p) \\ &= \sum_{k=0}^{\infty} k \sum_{k_0=k}^{\infty} p_{k_0} \binom{k_0}{k} (1-p)^k p^{k_0-k} \\ &= \sum_{k_0=0}^{\infty} \sum_{k=0}^{k_0} k p_{k_0} \binom{k_0}{k} (1-p)^k p^{k_0-k} \\ &= \sum_{k_0=0}^{\infty} p_{k_0} \sum_{k=0}^{k_0} k \binom{k_0}{k} (1-p)^k p^{k_0-k} \\ &= \sum_{k_0=0}^{\infty} p_{k_0} (1-p) k_0 \\ &= (1-p)\langle k \rangle \end{aligned}$$

$$\begin{aligned} \langle k^2(p) \rangle &= \sum_{k=0}^{\infty} k^2 p_k(p) \\ &= \sum_{k=0}^{\infty} k^2 \sum_{k_0=k}^{\infty} p_{k_0} \binom{k_0}{k} (1-p)^k p^{k_0-k} \\ &= \sum_{k_0=0}^{\infty} \sum_{k=0}^{k_0} k^2 p_{k_0} \binom{k_0}{k} (1-p)^k p^{k_0-k} \\ &= \sum_{k_0=0}^{\infty} p_{k_0} \sum_{k=0}^{k_0} k^2 \binom{k_0}{k} (1-p)^k p^{k_0-k} \\ &= \sum_{k_0=0}^{\infty} p_{k_0} [(1-p)^2 k_0^2 + p(1-p)k_0] \\ &= (1-p)^2 \langle k^2 \rangle + p(1-p)\langle k \rangle \end{aligned}$$

□

Finalement, par application du Théorème 7.1.1 on prouve le résultat souhaité, comme énoncé dans le Théorème 7.2.1:
Preuve : D'après le Théorème 7.1.1, le seuil p_c est tel que $\langle k^2(p_c) \rangle - 2\langle k(p_c) \rangle = 0$. D'après la Proposition 7.2.3, cela est équivalent à $(1-p_c)[(1-p_c)\langle k^2 \rangle - (2-p_c)\langle k \rangle] = 0$, ce qui donne le résultat. □

Nous allons maintenant décrire une autre méthode pour obtenir ce résultat [28, 101]. On suppose qu'une fraction p des sommets ont été supprimés au hasard. De manière équivalente, on peut dire que chaque sommet est *absent* avec probabilité p et est *présent* avec probabilité $1 - p$.

Dans la suite, nous allons utiliser plusieurs séries génératrices. Comme nous l'avons définie dans la section précédente, G_0 correspond à la distribution des degrés et G_1 à la distribution des degrés des sommets obtenus en suivant un lien aléatoire. De même, on introduit F_0, la fonction génératrice pour la distribution des degrés des sommets *présents*. La probabilité qu'un sommet ait degré k et soit présent étant $(1-p)p_k$, $F_0(x)$ vaut :

$$F_0(x) = \sum_{k=0}^{\infty} (1-p) p_k x^k = (1-p) G_0(x). \tag{7.7}$$

On introduit également H_1, la série génératrice pour les tailles des *composantes finies de sommets présents* obtenues en suivant un lien au hasard. C'est-à-dire le nombre de sommets atteignables en suivant un lien.

Le seuil $p = p_c$ pour les pannes de sommets est exactement la fraction telle que la taille moyenne des composantes finies diverge. Nous allons donc étudier cette taille moyenne, donnée par $H_1'(1)$.

Lemme 7.2.5 *La fonction génératrice des tailles des* composantes finies de sommets présents *atteintes en suivant un lien aléatoire, $H_1(x)$, satisfait :*

$$H_1(x) = p + (1-p)xG_1(H_1(x)). \tag{7.8}$$

Preuve : La composante est de taille nulle si le lien mène vers un sommet absent, ce qui arrive avec probabilité p. Donc $H_1(0) = p$. Si le sommet à l'extrémité du lien est présent, il a k liens sortants, k étant distribué selon $G_1(x)$ (équation 7.5).

Chacun de ces k liens sortants mène vers une composante, dont les tailles sont distribuées suivant $H_1(x)$. La taille totale de la composante auquel le sommet appartient est donc 1 plus la somme des tailles de k composantes indépendantes. La distribution de cette somme donnée par est $H_1^k(x)$. Donc, si le sommet est présent, la fonction génératrice des tailles des composantes est :

$$x\sum_{k=0}^{\infty} q_k H_1^k(x) = xG_1(H_1(x)).$$

Le sommet étant présent avec probabilité $1-p$, cela donne le résultat. □

Grâce à ce résultat, on peut donner une autre preuve du Théorème 7.2.1 :

Preuve : La taille moyenne d'une composante à l'extrémité d'un lien aléatoire est donnée par :

$$\begin{aligned}\langle s\rangle &= H_1'(1)\\ &= (1-p)G_1(H_1(1)) + (1-p)G_1'(H_1(1))H_1'(1)\\ &= (1-p) + (1-p)G_1'(1)H_1'(1)\end{aligned}$$

(en utilisant le fait que $G_1(1) = H_1(1) = 1$).
On a donc :

$$H_1'(1) = \frac{1-p}{1-(1-p)G_1'(1)}. \tag{7.9}$$

Cette quantité diverge quand p vérifie $1-(1-p)G_1'(1) = 0$, *i.e.* quand p atteint la valeur critique

$$p_c = 1 - \frac{1}{G_1'(1)}.$$

On rappelle que $G_1(x) = \sum_{k=1}^{\infty} kp_k x^{k-1}/\langle k\rangle$. Donc $G_1'(x) = \sum_{k=1}^{\infty} k(k-1)x^{k-2}/\langle k\rangle$, et $G_1'(1) = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}$.
Finalement,

$$p_c = 1 - \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle},$$

ce qui est bien le résultat annoncé. □

Les cas des graphes aléatoires et sans-échelle

Les résultats du Théorème 7.2.1 sont valables pour tout réseau, quelle que soit sa distribution des degrés. Nous allons maintenant l'appliquer à deux cas particuliers, les graphes aléatoires et les graphes sans-échelle.

Corollaire 7.2.6 *Pour tout graphe aléatoire dont la taille tend vers l'infini, on a :*

$$p_c = 1 - \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} = 1 - \frac{1}{\langle k \rangle}.$$

La Figure 7.3 montre cette fraction critique en fonction du degré moyen du réseau.

Pour appliquer le Théorème 7.2.1 aux graphes sans échelle, il faut calculer les deux premiers moments de la distribution des degrés. Deux approches principales ont été introduites pour ce faire [28, 34].

D'après [101], pour les réseaux dont la taille tend vers l'infini, on a :

$$\langle k \rangle = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)} \text{ et } \langle k^2 \rangle = \frac{\zeta(\alpha - 2)}{\zeta(\alpha)}.$$

Ce qui entraîne le résultat suivant :

Corollaire 7.2.7 *[28] Le seuil p_c pour les pannes de sommets pour les graphe sans-échelle dont la taille tend vers l'infini vaut :*

$$p_c = 1 - \frac{\zeta(\alpha - 1)}{\zeta(\alpha - 2) - \zeta(\alpha - 1)}$$

Preuve : Application directe du Théorème 7.2.1. □

On peut obtenir une approximation de cette valeur dans le cas de graphes sans-échelle finis de la façon suivante : on obtient une évaluation du degré maximal K en fonction du nombre de sommets par le Lemme 7.1.3. On a alors le résultat suivant :

Corollaire 7.2.8 *Pour un graphe sans-échelle de degré maximal K, le seuil pour les pannes de sommets vaut :*

$$p_c = 1 - \frac{\sum_{k=1}^{K} k^{-\alpha}}{\sum_{k=1}^{K} k^{-\alpha+2} - \sum_{k=1}^{K} k^{-\alpha+1}}.$$

Preuve : La distribution des degrés du graphe est alors donnée par :

$$p_k = \frac{\sum_{k=1}^{K} k^{-\alpha}}{\sum_{j=1}^{K} j^{-\alpha}}.$$

Les premiers et seconds moments de la distribution valent :

$$\langle k \rangle = \frac{1}{\sum_{j=1}^{K} p_j} \cdot \sum_{k=1}^{K} k^{-\alpha+1}.$$

et

$$\langle k^2 \rangle = \frac{1}{\sum_{j=1}^{K} p_j} \cdot \sum_{k=1}^{K} k^{-\alpha+2}.$$

D'où le résultat. □

On peut aussi obtenir une évaluation numérique du résultat du Théorème 7.2.1 en utilisant une approximation continue :

Corollaire 7.2.9 *[34], Le seuil p_c pour les pannes de sommets dans les graphes sans-échelle à N sommets vaut :*

$$p_c = \begin{cases} 1 - \left[\frac{2-\alpha}{3-\alpha} m - 1\right]^{-1} & si\ \alpha > 3 \\ 1 - \left[\frac{2-\alpha}{\alpha-3} m^{\alpha-1} N^{\frac{3-\alpha}{\alpha-1}} - 1\right]^{-1} & si\ 2 < \alpha < 3 \\ 1 - \left[\frac{2-\alpha}{3-\alpha} m N^{\frac{1}{\alpha-1}} - 1\right]^{-1} & si\ 1 < \alpha < 2 \end{cases}$$

Preuve : Soit $\kappa_0 = \langle k^2 \rangle / \langle k \rangle$. On a alors

$$\kappa_0 = \frac{C \int_m^K k^{-\alpha+2}}{C \int_m^K k^{-\alpha+1}} = \left(\frac{2-\alpha}{3-\alpha}\right) \frac{K^{3-\alpha} - m^{3-\alpha}}{K^{2-\alpha} - m^{2-\alpha}} \tag{7.10}$$

on rappelle que $C \approx (\alpha - 1) m^{\alpha-1}$ est une constante de normalisation.

Quand $K \gg m$, on a $K^\beta \gg m^\beta$ si $\beta > 1$, et $K^\beta \ll m^\beta$ si $\beta < 1$. Les approximations suivantes peuvent alors être faites :

- si $\alpha > 3$, alors $\begin{array}{rcl} K^{3-\alpha} - m^{3-\alpha} & \approx & m^{3-\alpha} \\ K^{2-\alpha} - m^{2-\alpha} & \approx & m^{2-\alpha} \end{array}$;
- si $2 < \alpha < 3$, alors $\begin{array}{rcl} K^{3-\alpha} - m^{3-\alpha} & \approx & K^{3-\alpha} \\ K^{2-\alpha} - m^{2-\alpha} & \approx & m^{2-\alpha} \end{array}$;
- si $1 < \alpha < 2$, alors $\begin{array}{rcl} K^{3-\alpha} - m^{3-\alpha} & \approx & K^{3-\alpha} \\ K^{2-\alpha} - m^{2-\alpha} & \approx & K^{2-\alpha} \end{array}$.

D'après le Théorème 7.2.1 on sait que $p_c = 1/(\kappa_0 - 1)$, on peut donc faire l'approximation :

$$1 - p_c \longrightarrow \begin{cases} \left[\frac{2-\alpha}{3-\alpha} m - 1\right]^{-1} & \alpha > 3 \\ \left[\frac{2-\alpha}{\alpha-3} m^{\alpha-2} K^{3-\alpha} - 1\right]^{-1} & 2 < \alpha < 3 \\ \left[\frac{2-\alpha}{3-\alpha} K - 1\right]^{-1} & 1 < \alpha < 2 \end{cases} \tag{7.11}$$

Avec l'évaluation précédente de K (lemme 7.1.2), on obtient le résultat. □

Ce résultat peut facilement être étendu au cas où la taille du réseau tend vers l'infini :

Corollaire 7.2.10 *[34], Le seuil p_c pour les pannes de sommets dans des graphes sans-échelle dont la taille tend vers l'infini vaut :*

$$p_c = \begin{cases} 1 - \left[\frac{2-\alpha}{3-\alpha} m - 1\right]^{-1} & si\ \alpha > 3 \\ 1 & si\ 1 < \alpha < 3 \end{cases}$$

Pour les deux Corollaires 7.2.7 et 7.2.9, si $\alpha < 3$, alors $p_c = 1$, ce qui signifie qu'il faut supprimer tous les sommets pour détruire le réseau.

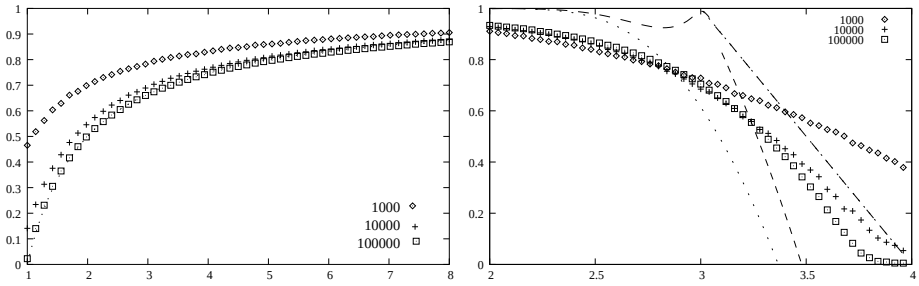

FIG. 7.3 – *Seuil pour les pannes de sommets pour des réseaux aléatoires (à gauche) et sans-échelle (à droite), pour des réseaux de taille* 1 000, 10 000 *et* 10^5. *Les lignes continues représentent les évaluations numériques. Pour les graphes sans échelle, ces lignes représentent les résultats obtenus à partir : du corollaire 7.2.8 pour* 10^5 *sommets (pointillés), du corollaire 7.2.7 (tirets courts), du corollaire 7.2.9 pour* 10^5 *sommets (tirets longs), et du corollaire 7.2.10 (pointillés confondus avec la courbe à tirets longs pour* $\alpha > 3$*).*

Pannes de sommets du point de vue des liens

Les graphes sans-échelle ont un grand nombre de sommets de faible degré. Quand on supprime un sommet aléatoirement, il a donc de grandes chances d'être de faible degré et d'entraîner la suppression de peu de liens. Il est donc naturel de se demander si la différence de comportement entre les graphes sans-échelle et les graphes aléatoires, pour les pannes, n'est pas dûe au fait que ces derniers perdent leurs liens plus rapidement. Pour répondre à cette question, nous étudions la fraction de liens supprimés quand une fraction p des sommets est supprimée lors de pannes.

Proposition 7.2.11 *Quand une fraction* p *de sommets sont supprimés aléatoirement, la fraction de liens correspondante vaut* $m(p) = 2p - p^2$.

Preuve : Soit un réseau dans lequel une fraction p des sommets ont été supprimés. Cela correspond à une suppression d'une fraction p de demi-liens. La fraction correspondante de liens supprimés est composée de liens reliant deux demi-liens supprimés et de liens entre des sommets restants et des sommets supprimés. Chaque demi-lien a une probabilité p d'être relié à un demi-lien supprimé, et une probabilité $1 - p$ d'être relié à un demi-lien conservé, étant donné que les demi-liens sont reliés au hasard.

La fraction p de demi-liens supprimés est donc composée de p^2 demi-liens reliés entre eux et de $p(1-p)$ demi-liens reliés à des demi-liens restant (mais qui vont être supprimés). La fraction totale de demi-liens supprimés est donc $p^2 + 2p(1-p) = 2p - p^2$. La fraction de demi-liens supprimés est égale à la fraction de liens supprimés, ce qui donne le résultat. □

Grâce à ce résultat, on obtient directement, en utilisant le Théorème 7.2.1 :

Corollaire 7.2.12 *La fraction de liens supprimés au seuil* p_c *pour les pannes de sommets vaut :*

$$m(p_c) = 1 - \left(\frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \right)^2$$

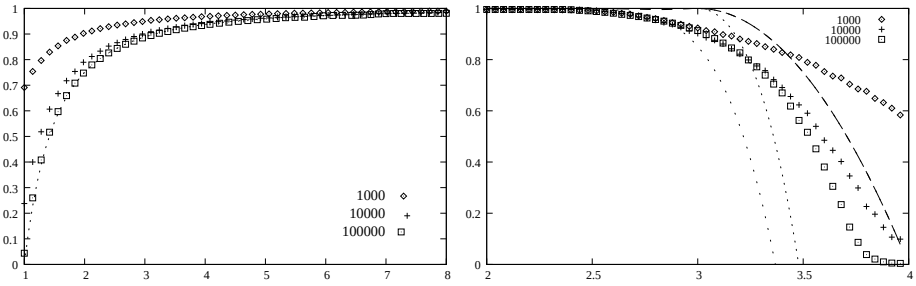

FIG. 7.4 – *Fraction de liens supprimés au seuil pour les pannes de sommets, pour des graphes aléatoires (à gauche) et des graphes sans-échelle (à droite), ainsi que les valeurs théoriques de cette fraction. Pour les graphes sans échelle, ces courbes représentent les valeurs obtenues à partir : du corollaire 7.2.15 pour* 10^5 *sommets (ligne pointillée de gauche), du corollaire 7.2.14 (ligne pointillée de droite), du corollaire 7.2.17 (lignes à tirets) (les valeurs obtenues dans le cas fini, corollaire 7.2.16, sont les mêmes).*

Preuve : Le résultat découle immédiatement du Théorème 7.2.1 qui stipule que $p_c = 1 - \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$, et de la Proposition 7.2.11. □

Ces résultats sont eux aussi valides pour tout réseau, quelle que soit sa distribution des degrés. Nous les appliquons maintenant aux deux cas qui nous intéressent.

Corollaire 7.2.13 *Pour les réseaux aléatoires dont la taille tend vers l'infini, la fraction* $m(p_c)$ *de liens supprimés au seuil* p_c *des pannes de sommets vaut :*

$$m(p_c) = 1 - \frac{1}{\langle k \rangle^2} \tag{7.12}$$

Pour les graphes sans-échelle, il y a encore deux méthodes possibles : un calcul approché qui devient exact à la limite infinie [28], ou l'approximation continue [34]. Dans le premier cas, d'après le Corollaire 7.2.7 on obtient :

Corollaire 7.2.14 *La fraction de liens supprimés lorsque le seuil des pannes de sommets est atteint, pour des graphes sans-échelle donc la taille tend vers l'infini, vaut :*

$$m(p_c) = 1 - \left(\frac{\zeta(\alpha - 1)}{\zeta(\alpha - 2) - \zeta(\alpha - 1)} \right)^2$$

Corollaire 7.2.15 *La fraction de liens supprimés lorsque le seuil* p_c *pour les pannes de sommets dans des graphes sans-échelle de degré maximal* K *vaut :*

$$m(p_c) = 1 - \left(\frac{\sum_{k=1}^{K} k^{-\alpha}}{\sum_{k=1}^{K} k^{-\alpha+2} - \sum_{k=1}^{K} k^{-\alpha+1}} \right)^2 .$$

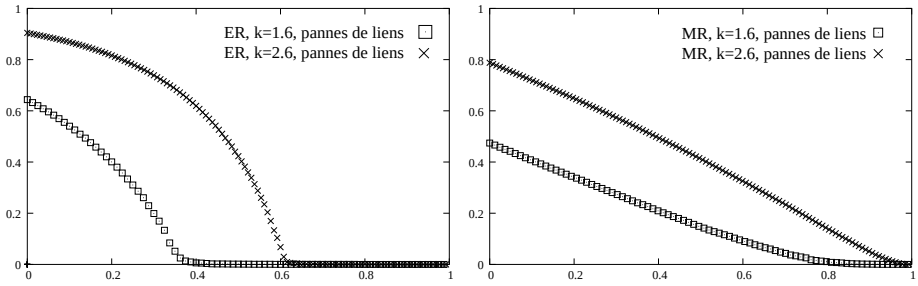

FIG. 7.5 – *Effet des pannes de liens sur des graphes aléatoires (à gauche) et des graphes sans-échelle (à droite).*

D'après le Corollaire 7.2.9, on obtient :

Corollaire 7.2.16 *Pour les graphes sans-échelle à N sommets, la fraction de liens supprimés lorsque le seuil pour les pannes de sommets, p_c, est atteint, vaut :*

$$m(p_c) = \begin{cases} 1 - \left[\frac{2-\alpha}{3-\alpha}m - 1\right]^{-2} & \alpha > 3 \\ 1 - \left[\frac{2-\alpha}{\alpha-3}m^{\alpha-1}N^{\frac{3-\alpha}{\alpha-1}} - 1\right]^{-2} & 2 < \alpha < 3 \\ 1 - \left[\frac{2-\alpha}{3-\alpha}mN^{\frac{1}{\alpha-1}} - 1\right]^{-2} & 1 < \alpha < 2 \end{cases}$$

Dans la limite infinie, on obtient :

Corollaire 7.2.17 *Pour des graphes sans-échelle dont la taille tend vers l'infini, la fraction de liens supprimés lorsque le seuil pour les pannes de sommets, p_c, est atteint, vaut :*

$$m(p_c) = \begin{cases} 1 - \left[\frac{2-\alpha}{3-\alpha}m - 1\right]^{-2} & \alpha > 3 \\ 1 & 1 < \alpha < 3 \end{cases}$$

7.2.2 Pannes de liens

Résultats généraux pour les pannes de liens

Supposons que les liens soient supprimés au hasard, avec probabilité p. De même que pour les pannes de sommets, il y a deux approches : considérer la nouvelle distribution des degrés, ou utiliser des séries génératrices.

Considérons l'effet de la suppression d'une fraction p des liens sur la distribution des degrés. Soit $p_k(p)$ la probabilité qu'un sommet quelconque ait un degré k après la suppression des liens. $p_k(p)$ vaut alors :

Lemme 7.2.18

$$p_k(p) = \sum_{k_0=k}^{\infty} p_{k_0} \binom{k_0}{k} (1-p)^k p^{k_0-k}$$

Preuve : Ceci correspond à la suppression aléatoire d'une fraction p de demi-liens. Si un sommet avait degré k_0 avant les suppressions, alors la probabilité qu'il ait degré $k' \leq k_0$ après est : $\binom{k_0}{k'}(1-p)^{k'}p^{k_0-k'}$. En effet, $k_0 - k'$ de ses voisins ont été supprimés, ce qui arrive avec probabilité $p^{k_0-k'}$, et k' de ses voisins restent, avec probabilité $(1-p)^{k'}$. D'où le résultat. □

Remarquons que cette distribution des degrés est la même qu'après la suppression d'une fraction p des sommets (voir Lemme 7.2.2). Il faut aussi noter, une fois de plus, que le réseau obtenu est équivalent à un réseau aléatoire avec la nouvelle distribution des degrés.

Il est donc possible de calculer les deux premiers moments de cette distribution pour appliquer le Théorème 7.1.1 et obtenir le même résultat que celui du Théorème 7.2.1:

Théorème 7.2.19 *Le seuil m_c pour les pannes de liens dans un réseau dont la taille tend vers l'infini, vaut :*

$$m_c = 1 - \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$$

Dans [28], une autre méthode pour retrouver ce résultat a été utilisée. Elle consiste à dire que chaque lien est marqué *présent* avec probabilité $1-p$, et *absent* avec probabilité p.

De même que précédemment, on définit $H_1(x)$ comme étant la fonction génératrice pour la probabilité que l'extrémité d'un lien choisi au hasard mène à une composante d'un certain nombre de sommets.

Lemme 7.2.20 *$H_1(x)$ satisfait une condition de la forme :*

$$H_1(x) = p + (1-p)xG_1(H_1(x)).$$

Preuve : Cette composante est de taille 0 avec probabilité p (si le lien est absent). Avec probabilité $1-p$, le lien mène à un sommet de degré $k+1$, où k est distribué suivant $G_1(x)$. Chacun des k liens sortants mène à une composante, dont les tailles sont distribuées selon $H_1(x)$. La taille totale de la composante est donc 1 plus la taille des k autres, d'où le résultat. □

Ce lemme est identique au Lemme 7.2.5 obtenu pour les suppressions aléatoires de sommets. On peut donc retrouver le Théorème 7.2.19 avec le même raisonnement que pour le Théorème 7.2.1.

Pannes de liens dans des graphes aléatoires et sans-échelle

Comme les résultats théoriques sont identiques pour les pannes de liens et de sommets, nous allons maintenant simplement donner des résultats expérimentaux. Ils sont présentés dans la Figure 7.6.

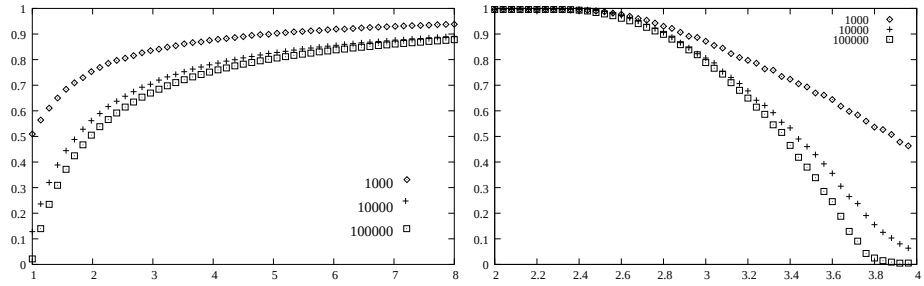

FIG. 7.6 – *Seuil expérimental pour les pannes de liens dans un graphe aléatoire (à gauche) et un graphe sans-échelle (à droite).*

7.3 Attaques classiques

Dans cette section, nous allons étudier la résistance des réseaux aux attaques classiques, comme elles sont définies dans [9], c'est-à-dire la suppression d'une fraction p des sommets de plus fort degré. Cette étude se fait avec les mêmes techniques que celles que nous avons utilisées dans la section précédente.

7.3.1 Résultats généraux

Dans [35], les deux conséquences d'une attaque sur un réseau sont présentées. Tout d'abord, le degré maximum dans le réseau diminue et devient $K(p)$, p et $K(p)$ étant liés par la relation suivante :

$$p = \sum_{k=K(p)+1}^{K} p_k \approx \sum_{k=K(p)+1}^{\infty} p_k \tag{7.13}$$

Ensuite, la distribution des degrés de ce réseau est modifiée de la manière suivante : une fraction $s(p)$ des demi-liens du réseau initial étaient reliés aux sommets supprimés par l'attaque. Cette fraction vaut :

$$s(p) = \frac{1}{\langle k \rangle} \sum_{K(p)+1}^{K} k p_k \approx \frac{1}{\langle k \rangle} \sum_{K(p)+1}^{\infty} k p_k \tag{7.14}$$

Considérons maintenant les demi-liens restants. Comme les paires de demi-liens sont reliées au hasard, chaque demi-lien est relié à un sommet supprimé avec probabilité $s(p)$. Les demi-liens reliés à un sommet supprimé sont aussi supprimés, chaque sommet restant perd donc ses demi-liens avec probabilité $s(p)$. Il faut comprendre que $s(p)$ ne représente pas la fraction globale de *liens* supprimés dans l'attaque[1], mais seulement la fraction de liens supprimés qui étaient reliés à des sommets survivants.

1. Cette fraction vaut $s(p)^2$ (fraction de demi-liens appartenant à des sommets supprimés reliés entre eux), plus $2s(p)(1-s(p))$ (fraction de demi-liens reliés à des sommets restant), et vaut donc $m(p) = s(p)(2-s(p))$, ce qui est supérieur à $s(p)$.

Le réseau après attaque est donc un graphe sans-échelle aléatoire de degré maximum $K(p)$, pour lequel une fraction $s(p)$ de liens a été supprimée aléatoirement. Il est donc possible d'appliquer les résultats pour les suppressions aléatoires de liens, à savoir le Théorème 7.2.19. Pour cela, il faut calculer les deux premiers moments de la distribution des degrés d'un réseau de degré maximal $K(p)$:

$$\langle k(p)\rangle = \sum_{k=1}^{K(p)} kp_k, \langle k^2(p)\rangle = \sum_{k=1}^{K(p)} k^2 p_k.$$

Ce qui amène le résultat suivant :

Théorème 7.3.1 *Le seuil p_c pour les attaques est la solution de l'équation :*

$$1 - s(p) = \frac{\langle k(p)\rangle}{\langle k^2(p)\rangle - \langle k(p)\rangle},$$

où p, $s(p)$ et $k(p)$ sont reliés par les équations 7.14 et 7.13, et où $\langle k(p)\rangle$ et $\langle k^2(p)\rangle$ sont donnés par la relation ci-dessus.

Le même résultat peut être obtenu avec un autre argument [101]. Considérons que, au lieu de supprimer des sommets, on les marque comme *absents*. On peut alors définir

$$F_0(x) = \sum_{k=1}^{K(p)} p_k x^k,$$

qui est la probabilité de trouver un sommet *présent* de degré k. Définissons aussi

$$F_1(x) = \frac{1}{\langle k\rangle} \sum_{k=1}^{K(p)} kp_k x^{k-1},$$

qui est la fonction génératrice pour la probabilité de trouver un sommet avec $k-1$ liens sortants au bout d'un lien choisi au hasard. Notons que $F_1(1) = (1-p)\langle k(p)\rangle$. On peut alors définir, de manière similaire à ce que l'on a fait dans la Section 7.2, la fonction génératrice pour les tailles des composantes atteignables par un lien quelconque, $H_1(x)$.

On a alors :

Lemme 7.3.2 *[28] $H_1(x)$ satisfait une relation de la forme :*

$$H_1(x) = 1 - F_1(1) + xF_1(H_1(x))$$

Preuve : La composante au bout du lien choisi au hasard est de taille 0 si le sommet est absent, ce qui arrive avec probabilité $1 - F_1(1)$. Si le sommet est présent, alors il a un nombre de liens sortants distribués selon $F_1(x)$. Chacun de ces liens mène à une composante dont la taille est distribuée selon $H_1(x)$. Si le lien est présent, la taille de la compasante est donc distribuée selon $xF_1(H_1(x))$. □

Théorème 7.3.3 *[28] Le seuil p_c pour les attaques est donné par :*

$$\frac{\sum_{k=1}^{K(p_c)} k(k-1)p_k}{\langle k \rangle} = 1,$$

où p_c et $K(p_c)$ sont reliés par l'équation 7.13.

Preuve :

p_c est atteint quand la taille moyenne des composantes diverge. On cherche donc p_c tel que $H'(1) = \infty$.

$$\begin{aligned} H_1'(x) &= F_1(H_1(x)) + xH_1'(x)F_1'(H_1(x)) \\ H_1'(1) &= F_1(1) + H_1'(1)F_1'(1), \end{aligned}$$

où l'on a utilisé le fait que $H_1(1) = 1$. Donc, $H_1'(1) = F_1(1)/(1 - F_1'(1))$, et $H_1'(1)$ diverge au point $F_1'(1) = 1$.

En écrivant la dérivée de F_1 on obtient le résultat. □

Remarquons que les Théorèmes 7.3.1 et 7.3.3 sont strictement équivalents et correspondent à deux écritures différentes de la même condition.

7.3.2 Application aux graphes sans-échelle

Pour appliquer ces résultats aux graphes sans-échelle, il y a deux méthodes [35, 101]. Nous présentons ces deux méthodes pas à pas en parallèle.

Lemme 7.3.4 *[35] Après une attaque supprimant une fraction p des sommets dans un graphe sans-échelle d'exposant α, le degré maximal $K(p)$ vaut :*

$$K(p) = mp^{1/(1-\alpha)}$$

Preuve : Rappelons que le degré maximal du réseau avant l'attaque peut être évalué par $\int_K^\infty p_k dk = 1/N$ (voir Lemme 7.1.2).

Le nouveau degré maximal peut être estimé de manière similaire :

$$\int_{K(p)}^{K} p_k dk = \int_{K(p)}^{\infty} p_k dk - \frac{1}{N} = p \tag{7.15}$$

Si la taille du réseau est très élevée, $1/N$ est négligeable devant p, et on peut ignorer le degré maximal original, et donc $\int_{K(p)}^\infty p_k dk = p$.

Rappelons que $p_k = Ck^{-\alpha}$, avec $C \approx (\alpha - 1)m^{\alpha-1}$. Il en découle : $p = C\left[\frac{k^{-\alpha+1}}{-\alpha+1}\right]_K^\infty = CK^{-\alpha+1}/(\alpha - 1) = m^{\alpha-1}K^{-\alpha+1}$. D'où le résultat □

On peut aussi lier p et $K(p)$ en appliquant directement l'Équation 7.13 :

Lemme 7.3.5 *[28]*

$$p = 1 - \frac{H_{K(p)}^{(\alpha)}}{\zeta(\alpha)}$$

Preuve :

$$p = \sum_{K(p)+1}^{\infty} p_k = \frac{1}{\zeta(\alpha)} \sum_{K(p)+1}^{\infty} k^{-\alpha} = \frac{1}{\zeta(\alpha)}(\zeta(\alpha) - H^{(\alpha)}_{K(p)}) = 1 - \frac{H^{(\alpha)}_{K(p)}}{\zeta(\alpha)}$$

□

Pour appliquer le Théorème 7.3.1, il faut calculer le nombre de demi-liens supprimés pendant l'attaque. On peut l'approximer par :

Lemme 7.3.6 *[35]*

$$s(p) = \left(\frac{K(p)}{m}\right)^{2-\alpha} = p^{(2-\alpha)/(1-\alpha)}$$

Preuve : $s(p) = \frac{1}{\langle k \rangle} \int_K^{\infty} k p_k$.

Rappelons que $p_k = Ck^{-\alpha}$, avec $C \approx (\alpha - 1)m^{\alpha-1}$. Le degré moyen $\langle k \rangle$ dans le réseau est donné par :

$$\begin{aligned} \langle k \rangle &= C \int_m^{\infty} k^{-\alpha+1} \\ &= C \left[\frac{k^{-\alpha+2}}{-\alpha+2}\right]_m^{\infty} \\ &= \frac{\alpha-1}{\alpha-2} m \end{aligned}.$$

Donc, $s(p) = \frac{1}{\langle k \rangle} C \left[\frac{k^{-\alpha+2}}{-\alpha+2}\right]_K^{\infty} = m^{\alpha-2} K^{-\alpha+2}$.

D'où le résultat. □

On peut également lier $s(p)$ et $K(p)$ de la façon suivante :

Lemme 7.3.7 *[28]*

$$s(p) = \frac{\zeta(\alpha)(\zeta(\alpha-1) - H^{(\alpha-1)}_{K(p)})}{\zeta(\alpha)\zeta(\alpha-1)} = 1 - \frac{H^{(\alpha-1)}_{K(p)}}{\zeta(\alpha-1)},$$

Preuve : Comme on l'a déjà vu dans l'Équation 7.14, la fraction de demi-liens reliés à ces sommets est donnée par : $s(p) = \sum_{k=K(p)}^{K} \frac{k p_k}{\langle k \rangle}$, où $\langle k \rangle$ est le degré moyen original. □

On obtient alors le résultat suivant :

Corollaire 7.3.8 *[35]*

$$\left(\frac{K(p_c)}{m}\right)^{2-\alpha} - 2 = \left(\frac{2-\alpha}{3-\alpha}\right) m \left(\left(\frac{K(p_c)}{m}\right)^{3-\alpha} - 1\right)$$

Preuve : Par le Théorème 7.3.1 on obtient :

$$1 - s(p) = \frac{1}{\kappa_0 - 1}, \tag{7.16}$$

où κ_0 est donné par l'Équation 7.10, où $K(p)$ remplace K.

Cette relation est équivalente à :

$$1 - s(p) = \left(\left(\frac{2-\alpha}{3-\alpha}\right) \frac{K(p)^{3-\alpha} - m^{3-\alpha}}{K(p)^{2-\alpha} - m^{2-\alpha}} - 1\right)^{-1}$$

On cherche donc la fraction critique p_c de sommets supprimés telle que l'égalité ci-dessus soit vérifiée. $s(p)$ et $K(p)$ sont reliés par le Lemme 7.3.6.

Il est donc possible de transformer l'Équation 7.3.2:

$$\begin{aligned}
1 &= \left(1-\left(\frac{K(p)}{m}\right)^{2-\alpha}\right)\left(\left(\frac{2-\alpha}{3-\alpha}\right)\frac{K(p)^{3-\alpha}-m^{3-\alpha}}{K(p)^{-\alpha}-m^{2-\alpha}}-1\right)\\
1 &= \left(1-\left(\frac{K(p)}{m}\right)^{2-\alpha}\right)\left(\frac{2-\alpha}{3-\alpha}\right)\frac{K(p)^{3-\alpha}-m^{3-\alpha}}{K(p)^{-\alpha}-m^{2-\alpha}}-1+\left(\frac{K(p)}{m}\right)^{2-\alpha}\\
\left(\frac{K(p)}{m}\right)^{2-\alpha}-2 &= \left(\frac{2-\alpha}{3-\alpha}\right)\frac{K^{3-\alpha}-m^{3-\alpha}}{K^{2-\alpha}-m^{2-\alpha}}\left(\left(\frac{K(p)}{m}\right)^{2-\alpha}-1\right)\\
&= \left(\left(\frac{K(p)}{m}\right)^{2-\alpha}-1\right)\left(\frac{2-\alpha}{3-\alpha}\right)m\frac{(K(p)/m)^{3-\alpha}-1}{(K(p)/m)^{2-\alpha}-1}
\end{aligned}$$

□

Une fois que cette équation est résolue numériquement, on obtient la valeur critique de $K(p_c)$, et la valeur p_c correspondante par le Lemme 7.3.4.

En appliquant le Théorème 7.3.3, on obtient :

Corollaire 7.3.9 *[28]*

$$H^{(\alpha-2)}_{K(p)} - H^{(\alpha-1)}_{K(p)} = \zeta(\alpha-1)$$

Une fois le degré maximum évalué, il est possible de calculer le nombre de sommets à supprimer par le Lemme 7.3.5. La Figure 7.7, à droite, présente les valeurs du seuil pour les attaques, obtenues par les corollaires 7.3.8 et 7.3.9, ainsi que des valeurs expérimentales pour ce seuil.

Les courbes théoriques ont été obtenues au moyen d'un logiciel de calcul formel de la façon suivante. Pour le Corollaire 7.3.8, il suffit de résoudre numériquement l'équation. Remarquons toutefois que le terme de droite de l'équation n'est pas défini au point $\alpha = 3$. La valeur en ce point de notre courbe est donc obtenue par continuité de la courbe des deux côtés de ce point.

Pour résoudre le Corollaire 7.3.9, on éxécute l'algorithme suivant :

$K(p) = 1$, $somme = 0$, $total = 0$
Tant que $somme < \zeta(\alpha-1)$ **Faire**
  $somme = somme + K(p)^{-\alpha+2} - K(p)^{-\alpha+1}$
  $total = total + K^{-\alpha}/\zeta(\alpha)$
  $K(p) = K(p) + 1$
Fin Tant que
$\rightarrow total$

L'algorithme s'achève quand $K(p)$ est effectivement le degré maximal du réseau après attaque. La quantité *total* vaut alors $H^{(\alpha)}_{K(p)}/\zeta(\alpha)$, et p_c vaut $1 - total$.

7.3.3 Application aux graphes aléatoires

Le Théorème 7.3.3 peut aussi être appliqué aux graphes aléatoires :

Corollaire 7.3.10

$$z = (e^{-z} \sum_{k=1}^{K(p)} \frac{k^2 z^k}{k!}) - (e^{-z} \sum_{k=1}^{K(p)} \frac{k z^k}{k!})$$

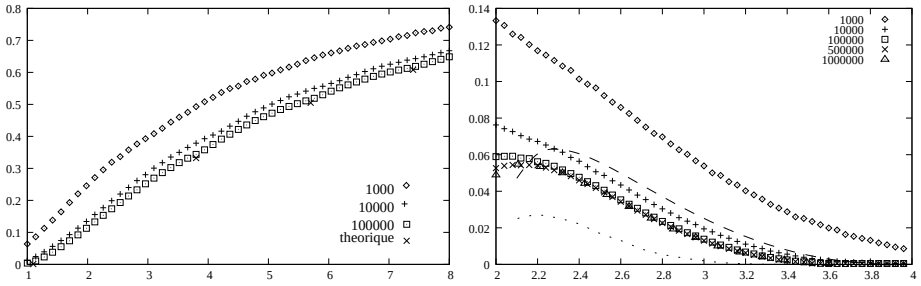

FIG. 7.7 – *Seuil critique pour les attaques classiques pour des graphes aléatoires (à gauche) et des graphes sans-échelle (à droite). Pour les graphes sans échelle, on a représenté les valeurs théoriques du seuil obtenues à partir du corollaire 7.3.9 (ligne pointillée) et du corollaire 7.3.8 (ligne à tirets).*

La Figure 7.7 (à gauche) présente les valeurs théoriques du seuil pour les attaques sur les graphes aléatoires, d'après le Corollaire 7.3.10. On obtient la valeur théorique du seuil par un algorithme du même type que celui présenté plus haut pour les graphes sans-échelle.

Cependant, dans les réseaux aléatoires, les degrés des sommets sont tous très proches de la valeur moyenne. À chaque étape de l'algorithme, la somme augmente donc fortement et peut donc dépasser de beaucoup la valeur visée. La valeur de $K(p)$ obtenue ne satisfait donc pas toujours exactement l'équation. Nous avons par conséquent choisi les points obtenus pour les valeurs de z qui génèrent l'erreur minimale. Les autres valeurs ne permettent pas de calculer correctement le seuil théorique. Il est cependant intéressant de constater que les valeurs expérimentales du seuil suivent la courbe intuitivement suggérée par ces quelques points.

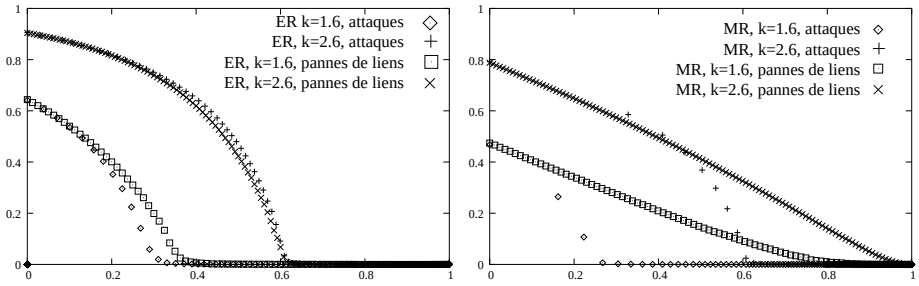

FIG. 7.8 – *Effets de l'attaque classique sur les sommets du point de vue des liens, pour des graphes aléatoires (à gauche) et des graphes sans-échelle (à droite), comparés à l'effet des pannes de liens.*

7.3.4 Attaques vues sous l'angle des liens

Résultats généraux

Nous allons maintenant évaluer précisément la fraction m_c de liens supprimés pendant une attaque classique. Comme nous l'avons déjà vu, cette fraction est donnée par :

Théorème 7.3.11 *La fraction de liens supprimés au seuil p_c pour les attaques vaut :*

$$m(p_c) = s(p_c)^2 + 2s(p_c)(1 - s(p_c)),$$

où $s(p_c)$ est la fraction de demi-liens reliés aux sommets supprimés, et est liée à p_c par l'Équation 7.14.

Application aux graphes aléatoires et sans-échelle

Pour calculer le nombre de liens supprimés dans une attaque sur les sommets, il faut donc calculer le nombre de demi-liens supprimés.

Pour les graphes sans-échelle, cela peut être fait par deux méthodes différentes, soit en appliquant le Corollaire 7.3.8 et le Lemme 7.3.6, ou alors en appliquant le Corollaire 7.3.9 pour obtenir $K(p_c)$, et obtenir $s(p_c)$ par l'Équation 7.14.

Pour les graphes aléatoires, il est possible d'appliquer le Corollaire 7.3.10 pour obtenir $K(p_c)$, et d'utiliser l'Équation 7.14 pour calculer $s(p_c)$.
Ces valeurs sont tracées sur la Figure 7.9, avec des valeurs expérimentales.

Discussions sur l'efficacité de l'attaque

Il est maintenant possible de conclure précisément sur l'efficacité de cette stratégie d'attaque. Tout d'abord, bien que le nombre de liens supprimés soit très élevé, cela ne suffit pas à expliquer l'effondrement du réseau : supprimer autant de liens au hasard ne détruit pas le réseau. D'autre part, le nombre de liens supprimés durant une attaque classique sur un graphe aléatoire et un graphe sans-échelle sont très similaires pour des valeurs du degré

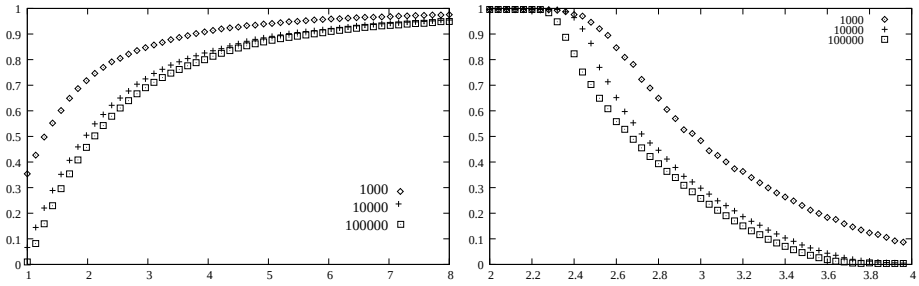

FIG. 7.9 – *Fraction critique m_c de liens à supprimer dans une attaque classique pour détruire le réseau. Les lignes représentent les valeurs théoriques et les points les valeurs expérimentales. Pour les graphes sans-échelle, les valeurs obtenues par les Corollaires 7.3.9 (ligne pointillé) et 7.3.8 (ligne en tirets) sont aussi tracées.*

moyen représentatives des graphes rencontrés en pratique. Ceci modère la conclusion selon laquelle les graphes sans-échelle sont très sensibles à cette attaque : en termes de liens, ils sont aussi résistants que les graphes aléatoires.

7.4 Nouvelles stratégies d'attaque

Nous avons vu précédemment qu'un réseau dont la distribution des degrés est p_k a presque sûrement une composante de taille linéaire en le nombre de sommets si :

$$\langle k^2 \rangle - 2\langle k \rangle > 0 \iff p_1 < \sum_{k=3}^{\infty} k(k-2)p_k$$

Le point principal de cette équation est la fraction de sommets de degré 1. Il semble donc que toute stratégie qui vise à augmenter cette fraction devrait détruire rapidement le réseau. En utilisant cette remarque, nous allons proposer deux nouvelles stratégies d'attaques, la première sur les sommets et la seconde sur les liens. Ces stratégies permettent de mieux comprendre l'efficacité des attaques classiques.

7.4.1 Attaques proches des pannes

Notre première stratégie consiste simplement à supprimer au hasard des sommets de degré au moins 2. Ceci diminue le nombre de sommets de degré strictement supérieur à 1 et augmente le nombre de sommets de degré 0 ou 1. Les effets de cette stratégie sont visibles sur la Figure 7.10.

Cette stratégie d'attaque est à peine différente des pannes et pourtant elle est beaucoup plus efficace. Elle est qualitativement différente puisqu'elle présente un seuil pour les graphes sans-échelle, ce qui n'est pas le cas des pannes.

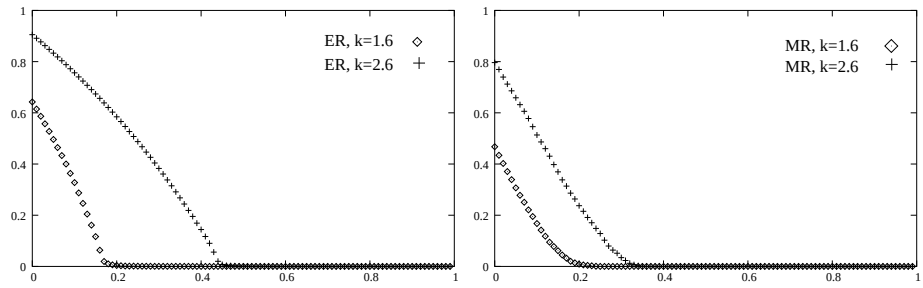

FIG. 7.10 – *Effet de la nouvelle stratégie d'attaque sur les sommets pour des graphes aléatoires (à gauche) et pour des graphes sans-échelle (à droite).*

Il est facile de montrer l'existence de ce seuil en en donnant une borne supérieure :

Théorème 7.4.1 *Pour tout réseau ayant une fraction non nulle de sommets de degré* 1*, la nouvelle attaque sur les sommets a un seuil.*

Preuve : Quand tous les sommets qui avaient initialement un degré strictement supérieur à 1 ont été supprimés, alors le réseau est complètement déconnecté car tous les sommets restant ont degré 0 ou 1. La composante géante est donc sûrement détruite quand une fraction $1 - p_1 - p_0$ des sommets a été supprimée. □

On peut calculer la valeur de cette borne supérieure pour les graphes aléatoires et les graphes sans-échelle.

Corollaire 7.4.2 *Pour les graphes sans-échelle d'exposant* α*, une borne supérieure pour le seuil pour la nouvelle attaque sur les sommets est* $1 - 1/\zeta(\alpha)$*.*

Corollaire 7.4.3 *Pour les graphes aléatoires de degré moyen* z*, une borne supérieure pour le seuil pour la nouvelle attaque sur les sommets est* $1 - e^{-z}(z + 1)$*.*

Les courbes pour ces bornes supérieures sont montrées sur la Figure 7.11, avec les valeurs expérimentales du seuil.

Remarquons pour le cas des graphes sans échelle que les valeurs expérimentales obtenues sont au-dessus de la borne supérieure théorique de $1 - p_1 = 1 - 1/\zeta(\alpha)$. Ceci est dû au fait que nos graphes sont finis, et à la façon dont nous les engendrons : la fraction de sommets de degré 1 dans nos graphes est différente de la valeur théorique attendue quand la taille du graphe tend vers l'infini. Nous avons donc représenté sur la même courbe la valeur empirique de la borne supérieure théorique, qui vaut $1 - p_1$.

Remarquons que les valeurs du seuil sont assez élevées (il faut supprimer de nombreux sommets pour détruire le réseau). Notre but n'est pas de proposer une attaque efficace, mais plutôt de montrer que la différence qualitative entre les pannes et les attaques classiques vient de la présence des sommets de degré 1 qui ne sont jamais supprimés dans une attaque.

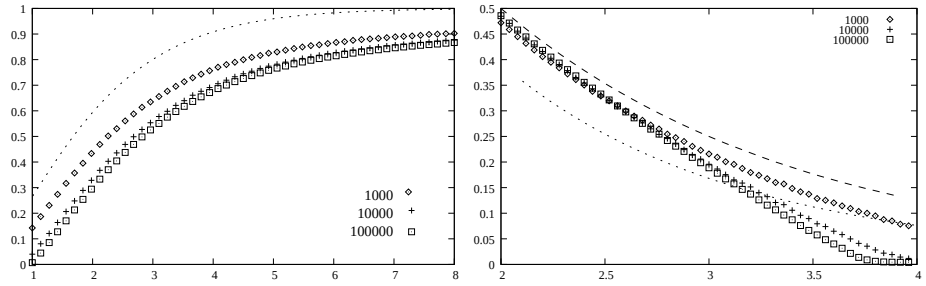

FIG. 7.11 – *Courbes pour la borne supérieure pour la nouvelle attaque sur les sommets, et pour les valeurs expérimentales du seuil pour, des réseaux de taille* 10^3, 10^4 *et* 10^5, *pour des graphes aléatoires (à gauche) et des graphes sans-échelle (à droite). Les lignes continues représentent les bornes théoriques. Pour les graphes sans-échelle, la courbe en pointillés représente la borne supérieure théorique pour des graphes dont la taille tend vers l'infini, et la courbe en tirets représente la borne supérieure mesurée en pratique sur les graphes que nous avons engendrés.*

7.4.2 Attaque sur les liens

Dans la Section 7.3.4 nous avons vu que, bien que l'attaque classique présente un seuil du point de vue des liens, elle est très peu efficace (le seuil est très élevé). Nous proposons maintenant une attaque sur les liens, toujours basée sur l'idée d'augmenter le nombre de sommets de degré 1. Elle consiste à supprimer de manière aléatoire des liens entre des sommets de degré au moins 2. Les effets de cette stratégie sont visibles sur la Figure 7.12. On peut montrer que cette attaque a un seuil en en donnant une borne supérieure :

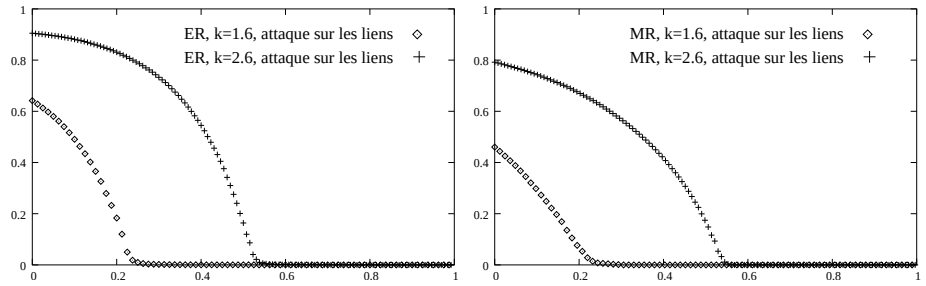

FIG. 7.12 – *Effet de la nouvelle attaque sur les liens, sur les graphes aléatoires (à gauche), et les graphes sans-échelle (à droite).*

Théorème 7.4.4 *La nouvelle stratégie d'attaque sur les liens a un seuil* m_c *pour les graphes aléatoires et les graphes sans-échelle.*

Preuve : On obtient une borne supérieure de la manière suivante : quand tous les liens entre les sommets de degré au moins 2 sont supprimés, le réseau est alors uniquement

composé d'un ensemble d'étoiles disjointes (chaque sommet central est relié à des sommets de degré 1).

Pour les graphes sans-échelle d'exposant supérieur à 2 et les graphes aléatoires, le degré maximal est sous-linéaire en N (pour les graphes sans-échelle finis à N sommets on peut l'évaluer comme $N^{\frac{1}{\alpha-1}}$ [34]). La taille de la plus grande composante connexe (*i.e.* la plus grande étoile) est donc également sous-linéaire en N.

Une borne supérieure pour la fraction m_c de liens à supprimer pour déconnecter le graphe est donc donnée par la fraction de liens entre des sommets de degré au moins 2. Cette quantité est égale à 1 moins la fraction de liens incidents à au moins un sommet de degré 1. Le nombre de tels liens est donné par le nombre de sommets de degré 1, moins le nombre de liens entre deux sommets de degré 1.

Ce dernier nombre peut être calculé comme suit[2] : il y a Np_1 sommets de degré 1, chacun ayant une probabilité $Np_1/2|E|$ d'être connecté à un autre sommet de degré 1 ($|E| = N\langle k\rangle/2$ est le nombre de liens du réseau). Le nombre de sommets de degré 1 reliés à un sommet de degré 1 est donc $N^2p_1^2/2|E| = Np_1^2/\langle k\rangle$ en moyenne. Finalement, le nombre de liens entre deux tels sommets est $Np_1^2/2\langle k\rangle$.

On peut donc calculer le nombre de liens adjacents à au moins un sommet de degré 1 : $Np_1 - Np_1^2/2\langle k\rangle$, et le nombre de liens non adjacents à un sommet de degré 1 est $|E| - Np_1 + Np_1^2/2\langle k\rangle$. La fraction de tels liens est donc :

$$1 - \frac{2p_1}{\langle k\rangle} + \frac{p_1^2}{\langle k\rangle^2}.$$

ce qui est une borne supérieure pour la nouvelle stratégie d'attaque sur les liens. □

On peut calculer la valeur de cette borne supérieure pour les graphes aléatoires et les graphes sans-échelle.

Corollaire 7.4.5 *Pour un graphe sans-échelle d'exposant α, une borne supérieure pour le seuil de la nouvelle attaque sur les liens est donnée par :*

$$1 - \frac{2}{\zeta(\alpha-1)} + \frac{1}{\zeta^2(\alpha-1)} = 1 - \frac{2\zeta(\alpha-1)-1}{\zeta^2(\alpha-1)}.$$

Corollaire 7.4.6 *Pour un graphe aléatoire de degré moyen z, une borne supérieure pour le seuil de la nouvelle attaque sur les liens est donnée par :*

$$1 - 2e^{-z} + e^{-2z}.$$

Ces bornes peuvent être évaluées numériquement. La Figure 7.13 les présente, ainsi que les valeurs expérimentales du seuil.

Pour les graphes sans-échelle, la borne supérieure théorique est en dessous des valeurs expérimentales du seuil. Ceci est un effet de la taille finie des graphes utilisés dans nos expérimentations, alors que la borne est valable pour des tailles de graphes tendant vers

2. Ceci est valable pour N suffisamment grand.

l'infini. Cet effet disparaît quand la taille du graphe tend vers l'infini et s'explique comme suit : quand tous les liens adjacents à deux sommets de degré au moins 2 ont été supprimés, le graphe est composé d'étoiles : les sommets de degré 1 sont reliés aux sommets centraux des étoiles. La taille de la plus grande composante est donc égale au plus fort degré (plus un) de ces sommets centraux.

Quand la taille du graphe tend vers l'infini, le plus fort degré du graphe est sous-linéaire en la taille du graphe, donc la composante géante du graphe est bien détruite.

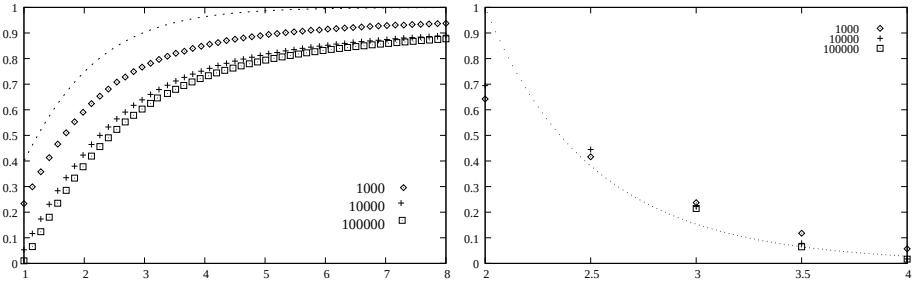

FIG. 7.13 – *Bornes supérieures pour le seuil de la nouvelle attaque, avec les valeurs expérimentales pour le seuil pour des réseaux de taille* 10^3, 10^4 *and* 10^5. *Pour des graphes aléatoires (à gauche) et des graphes sans-échelle (à droite).*

Dans le cas de graphes finis, nous avons considéré que le graphe était déconnecté quand la taille de la plus grande composante était inférieure ou égale à 1% de la taille du graphe. Or, si l'on observe ce qui se passe pour des graphes de taille 10^5, on se rend compte que le degré maximal du graphe est supérieur à 1 000 (soit 1% de la taille du graphe), tant que $\alpha < 2,9$ environ (voir Figure 7.2). Cette valeur correspond au moment où la courbe du seuil expérimental passe au-dessus de la borne supérieure empirique.

Concernant la borne supérieure théorique, on retrouve le même problème que pour le cas de la nouvelle stratégie d'attaque sur les sommets : les valeurs théoriques sont au dessous des valeurs obtenues pour des graphes finis en mesurant la fraction de liens incidents à au moins un sommet de degré 1.

Si l'on compare ces résultats avec ceux obtenus dans la Section 7.3.4, on peut observer que la nouvelle attaque est plus efficace que la stratégie classique vue sous l'angle des liens. Ceci n'est pas surprenant car l'attaque classique supprime beaucoup de liens adjacents à des sommets de degré 1, liens qui n'aident pas à maintenir la cohésion du réseau. Notre stratégie, au contraire, vise à détruire uniquement les liens utiles.

Conclusion

Dans ce chapitre, nous avons fourni une comparaison détaillée des pannes et des attaques classiques sur des graphes aléatoires et les graphes sans-échelle. Notre objectif était d'apporter un éclairage plus complet sur l'efficacité réelle des attaques sur les graphes sans-échelle.

Pour cela, nous avons étudié l'affirmation répandue selon laquelle l'efficacité des attaques sur les graphes sans-échelle provient du grand nombre de liens supprimés durant cette attaque. Nous avons montré qu'en supprimant le même nombre de liens au hasard, cela avait un impact beaucoup plus réduit, ce qui contredit cette affirmation. Cependant, si l'on considère le nombre de liens supprimés lors d'une attaque classique, alors les graphes sans-échelle ne sont pas plus fragiles que les graphes aléatoires.

Finalement nous avons utilisé un critère de connexité classique pour proposer deux nouvelles stratégies d'attaque. La première est très similaire à une série de pannes mais se comporte comme les attaques classiques (il y a un seuil pour les graphes sans-échelle), ce qui tend à montrer que la présence d'un seuil pour l'attaque classique n'est pas tant due à son efficacité qu'au grand nombre de sommets de degré 1 présents dans ces graphes. La seconde stratégie que nous avons proposée, basée sur la suppression de liens, montre qu'il est possible de définir des stratégies plus efficaces que la stratégie classique, si l'on s'intéresse à la fraction de liens supprimés.

Ces résultats amènent à la conclusion que les pannes et les attaques classiques ont un comportement différent et que la nature aléatoire ou sans-échelle influence aussi ce processus, mais qu'il faut être prudent sur les conclusions que l'on peut en tirer. La sensibilité des graphes sans-échelle aux attaques vient de la présence de nombreux sommets de faible degré. Leur résistance aux pannes vient du fait que choisir un sommet au hasard revient à choisir un sommet de faible degré la plupart du temps. De plus, le fait que les attaques classiques sur les graphes sans-échelle suppriment de nombreux liens est une raison, mais pas la seule, du fait qu'elles déconnectent rapidement le réseau.

Ces travaux peuvent être poursuivis dans de nombreuses directions. Tout d'abord, la précision des différents seuils devrait être améliorée. De même, l'impact de la finitude des graphes réels n'est pas encore très bien compris et devrait être approfondi. D'autres propriétés des grands réseaux d'interactions, telles que le clustering ou les corrélations entre degrés, ont aussi certainement un impact sur la robustesse des réseaux.

Plus généralement, l'impact des pannes et des attaques sur les grands réseaux d'interactions, tels que l'Internet, le Web, les réseaux P2P ou encore les réseaux biologiques ou sociaux, devrait être approfondi. Il est très probable que ces réseaux ont des propriétés inconnues qui affecte leur résistance aux pannes, mais qui permettent sans doute de concevoir certains types d'attaques auxquelles ils sont très sensibles.

Chapitre 8

Exploration du graphe de l'Internet

Du fait de sa structure et de son administration complètement distribuée, il est très difficile d'obtenir une carte de l'Internet. Pourtant, une telle carte du réseau serait extrêmement utile dans de nombreux problèmes où une connaissance globale est souhaitable. Citons, par exemple, et parmi beaucoup d'autres, les problèmes liés à la robustesse du réseau [9, 34, 35] ou à la simulation de protocoles et des usages à venir [85]. Explorer la topologie de l'Internet est un problème de recherche en soi [51, 56, 84, 124, 132]. En effet, de nombreuses difficultés apparaissent quand on veut le cartographier, que ce soit au niveau des routeurs ou au niveau des systèmes autonomes.

Les cartes actuelles reposent majoritairement sur l'utilisation intensive de `traceroute` : avec cet outil, un ensemble de routes est récupéré depuis quelques sources vers de nombreuses destinations. Les routes sont ensuite fusionnées (avec quelques précautions) pour obtenir une carte du réseau, comme décrit dans la Section 1.1.

Ainsi que nous l'avons évoqué dans le Chapitre 1, cette approche fournit des cartes partielles et biaisées [31, 62, 67, 76, 112, 118]. Afin d'améliorer ces mesures, plusieurs groupes de recherche s'orientent désormais vers une approche massivement distribuée de la mesure [50, 113, 114]. L'idée de base de cette approche est d'augmenter significativement le nombre de sources (typiquement limité à quelques dizaines dans les mesures les plus massives actuellement disponibles) pour améliorer la qualité des cartes.

L'objectif central de ce chapitre est d'évaluer la pertinence d'une telle approche en la comparant avec les approches classiques. Pour cela, nous avons conduit un grand nombre d'expériences de simulation. Nous avons considéré un graphe G représentant le réseau à explorer, puis nous avons simulé un processus d'exploration afin d'obtenir une vision G', *a priori* incomplète et biaisée, de celui-ci. Enfin, nous avons comparé G et G' pour tenter de quantifier le biais introduit en fonction du nombre de sources et de destinations utilisées.

La suite de ce chapitre repose fortement sur les modèles de graphes présentés dans la seconde partie de cette thèse. Tout d'abord, nous discutons la méthodologie employée, avant d'analyser les résultats de simulation pour divers modèles et propriétés (Sections 8.2–8.6). Ensuite, nous montrons (Section 8.7) que cette approche permet de définir des stratégies d'exploration plus efficaces en choisissant la localisation des sources et des destinations.

Enfin, nous comparons nos résultats avec des données réelles afin d'identifier les simulations les plus significatives (Section 8.8).

8.1 Préliminaires

Modéliser `traceroute` et le processus d'exploration

Dans la suite, nous allons supposer qu'une route obtenue par `traceroute` est en fait un plus court chemin entre la source et la destination. Cette hypothèse, bien qu'assez classique [37, 62, 76], n'est par vraiment satisfaisante [54, 67], mais nous nous en contenterons car la modélisation réaliste des routes sur l'Internet est aujourd'hui encore un problème ouvert. Nous justifions ce choix par le fait qu'il n'existe pas de modèle plus réaliste, mais aussi par l'utilisation intensive que nous allons faire de routes. Il faut donc être capable de les calculer très efficacement et les plus courts chemins ont beaucoup d'avantages d'un point de vue algorithmique.

En pratique, on peut utiliser l'outil `traceroute` une ou plusieurs fois entre une même source et une même destination dans l'espoir d'obtenir plusieurs routes différentes. Cela se traduit par deux modèles d'exploration, le premier modélisant une exploration poussée du réseau alors que le second représente une exploration plus rapide et moins complète.

Dans le premier cas, celui du modèle ASP (pour *All Shortest Paths*), nous allons mesurer, pour chaque couple source-destination, tous les plus courts chemins les reliant. Le modèle ASP n'est pas réaliste : en effet, on ne peut pas espérer mesurer tous les chemins entre une source et une destination, même en effectuant beaucoup de mesures. Malgré tout, ce modèle peut-être considéré comme un *meilleur cas*, ou une borne supérieure sur l'information que l'on peut obtenir en faisant des explorations à base de plus courts chemins.

L'autre cas, celui du modèle USP (pour *Unique Shortest Path*), consiste en une exploration rapide et nous allons donc mesurer un unique plus court chemin pour chaque couple source-destination. Étant donné qu'il peut y avoir de nombreux plus courts chemins entre deux sommets, nous avons fait le choix de considérer que les messages se déplacent sur un arbre enraciné sur la source. En conséquence, les chemins suivis seront toujours les mêmes d'une source donnée vers une destination donnée.

Nous avons aussi effectué d'autres expériences en utilisant des modèles proches à base de plus courts chemins aléatoires, ou en mesurant plusieurs plus courts chemins (mais pas tous). Les résultats n'étant pas significativement différents, nous nous concentrerons sur les deux modèles extrêmes, ASP et USP.

Enfin, nous choisirons les sources et les destinations de façon purement aléatoire. Nous allons donc calculer des plus courts chemins, avec USP ou ASP, entre chacune de ces sources et chacune de ces destinations, afin d'obtenir une mesure du réseau. Un tel modèle a déjà été introduit dans [16, 76] et a reçu le nom d'étude (k, m)-`traceroute`, k étant le nombre de sources et m le nombre de destinations.

Méthodologie

Notre approche globale a été la suivante :

1. générer un graphe G en utilisant un des modèles présentés dans le Chapitre 3. Les modèles que nous avons utilisé sont le modèle purement aléatoire ER, le modèle avec distribution des degrés fixés MR, le modèle avec attachement préférentiel AB, le modèle clusterisé sans échelle DM et enfin le modèle biparti aléatoire GL ;
2. construire une vision G' de G en utilisant un modèle d'exploration donné, USP ou ASP, avec un certain ensemble de sources et de destinations ;
3. comparer les propriétés statistiques de G' à celles de G. Dans ce chapitre, nous nous sommes concentré sur les propriétés de base : proportion découverte, degré moyen, distribution des degrés, distance moyenne et clustering.

Soulignons qu'avec cette approche, nous visons des résultats qualitatifs seulement : nous voulons connaître l'influence des propriétés du réseau sur la vision que l'on en a, et savoir dans quelle mesure les cartes sont valables du point de vue de leurs propriétés statistiques. Notre approche, qui utilise beaucoup de modélisation, ne permet pas vraiment d'apporter de résultats quantitatifs sur la qualité des cartes actuelles, mais met en évidence l'influence fondamentale de certaines propriétés sur la qualité de la mesure, et l'absence d'influence d'autres propriétés.

8.2 Explorations avec peu de sources

8.2.1 Une seule source

Nous allons tout d'abord nous placer dans le cas extrême d'une exploration avec une seule source. De nombreuses cartes actuelles de l'Internet ont été construites de la sorte (parfois en utilisant le source-routing) ; ce modèle capture donc une certaine réalité.

Soit $G_u(x)$ le graphe obtenu en faisant une exploration de G d'une source donnée vers x destinations prises au hasard en utilisant le modèle USP. Soit $n_u(x)$ le nombre de sommets, et $m_u(x)$ le nombre de liens de $G_u(x)$. De manière similaire, on notera $G_a(x)$, $n_a(x)$ et $m_a(x)$ les résultats obtenus avec le modèle ASP. Pour commencer, nous allons étudier des graphes aléatoires ER. Rappelons que, pour ce modèle, chaque lien existe avec une probabilité p fixée.

La Figure 8.1 montre la proportion de sommets et de liens découverte en fonction du nombre de destinations. Plusieurs remarques naturelles peuvent être faites sur cette courbe. Tout d'abord, si l'on ne considère que quelques destinations, seule une petite partie du réseau est découverte. Cette proportion augmente avec le nombre de destinations et, évidemment, tous les sommets sont découverts si l'on prend tous les sommets comme destinations.

Plusieurs points plus subtils sont à remarquer. À la fois $n_u(x)$ et $n_a(x)$ croissent rapidement avant d'atteindre un point critique où la croissance se poursuit de manière linéaire, la croissance initiale étant beaucoup plus rapide pour $n_a(x)$ que pour $n_u(x)$. Au contraire,

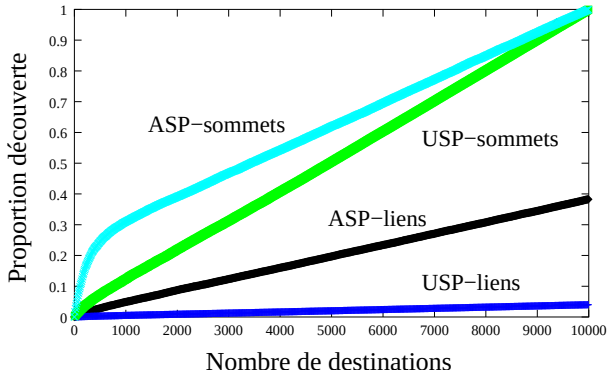

FIG. 8.1 – *Proportion du nombre total de sommets et de liens découverts pendant une exploration, en fonction du nombre de destinations. Graphe aléatoire ER avec* $n = 10\,000$ *et* $p = 0.005$.

la croissance de $m_u(x)$ et $m_a(x)$ est linéaire dès le début et atteint une valeur maximale qui est étonnament faible : même en considérant toutes les destinations possibles, un grand nombre de liens n'est pas découvert. L'exploration ASP qui considère tous les plus courts chemins, donne de meilleurs résultats, mais pas vraiment satisfaisants.

Ce comportement peut aussi être observé pour différentes valeurs de n et p, les courbes ayant toujours la même allure. Malgré tout, la valeur maximale atteinte par $m_u(x)$ et $m_a(x)$, *i.e.* la proportion maximale de liens découverts, varie avec p, la probabilité d'existence des liens. Pour mieux comprendre l'influence de p sur cette proportion, étudions la Figure 8.2 qui présente la proportion de liens découverts en choisissant toutes les destinations possibles, en fonction de p.

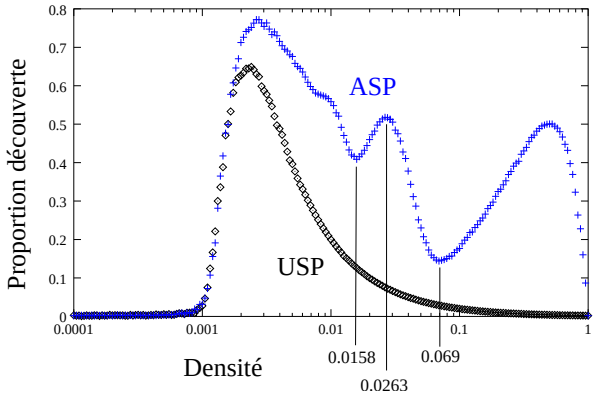

FIG. 8.2 – *Proportion de liens découverts (une source, toutes les destinations) en fonction de* p *pour un graphe ER avec* $n = 1\,000$ *sommets. La courbe obtenue pour* ASP *a reçu le nom de "courbe du chameau" à cause de sa forme bosselée.*

Les deux courbes ont plusieurs propriétés en commun qui s'expliquent simplement.

Tout d'abord, rappelons qu'en dessous d'une certaine valeur de p (tant que le degré moyen est inférieur à 1), le réseau n'est pas connecté. En dessous de ce seuil, toute exploration n'utilisant qu'une source ne peut découvrir que la composante arborescente dans laquelle cette source est située et qui ne représente qu'une très petite partie du graphe. Quand le réseau devient connecté, il est très arborescent, et il n'y a en général que très peu de chemins entre la source et les autres sommets. Les deux méthodes d'exploration permettent donc de découvrir presque tous les liens, ce qui correspond aux valeurs maximales atteintes sur la Figure 8.2. À l'opposé, quand p est très proche de 1, le graphe est presque complet, et donc presque tous les sommets sont à distance 1 de la source. Le graphe obtenu par une exploration à partir d'une source, USP ou ASP, est presque une étoile. Ce graphe contient environ $n-1$ liens, ce qui est négligeable comparé au nombre total de liens, proche de $\frac{n\cdot(n-1)}{2}$.

La courbe pour le modèle USP est très simple à comprendre étant donné que ce modèle produit un arbre à partir d'une source. Il permet donc de découvrir $n-1$ liens (si le graphe est connexe). Le nombre de liens total du graphe étant $m = p \cdot \frac{n\cdot(n-1)}{2}$, la proportion découverte est donc $\frac{n-1}{m} = \frac{2}{p\cdot n}$. Quand p est faible, il faut considérer le nombre de sommets de la composante géante au lieu de n, sinon cette proportion devient plus grande que 1. En particulier, près du seuil critique $1/n$, l'exploration donne une bonne vision du graphe original. Avec l'augmentation de p, cette proportion diminue de façon inversement linéaire, ce qui est confirmé par les simulations.

Au contraire, la forme très irrégulière obtenue avec le modèle ASP est à première vue surprenante : elle est composée de bosses et de creux de grande amplitude qui ne peuvent pas être interprétés de façon immédiate. L'explication de cette forme peut être donnée, mais en regardant un peu plus précisement les propriétés du modèle ASP.

La courbe du chameau

Intéressons-nous tout d'abord aux liens qui ne sont pas découverts lors d'une exploration. Tout lien se trouvant sur un plus court chemin entre la source et n'importe quel autre sommet sera découvert, par définition du modèle ASP. De manière réciproque, tout lien découvert est sur un plus court chemin entre la source et une destination. Les liens manqués sont, par conséquent, ceux ne se trouvant pas sur un plus court chemin. On se convainc facilement que ces liens sont exactement ceux reliant deux sommets à égale distance de la source. La courbe de la Figure 8.2 représente donc, en fait, le nombre total de liens du graphe, moins ceux entre des sommets équidistants de la source.

Pour calculer le nombre de tels liens, considérons la distribution des distances des sommets à la source. La Figure 8.3 montre que cette distribution est centrée autour d'une valeur moyenne qui décroît quand p augmente. Il faut noter que la forme générale de la courbe semble indépendante de p. La distribution à considérer est la distribution discrète de la Figure 8.3 car les distances sont entières. Or, sur cette distribution, deux cas peuvent se produire : la valeur moyenne (qui est approximativement celle pour laquelle la courbe continue est maximale) peut être entière ou centrée entre deux entiers. Dans le premier cas, la majorité des sommets est à même distance de la source alors que, dans le second

cas, une moitié se trouve à une certaine distance de la source et une autre moitié à cette distance plus un (quelques sommets sont plus proches ou plus éloignés). Ces deux cas sont illustrés sur la Figure 8.3 (premier cas pour $p = 0.0158$ et $p = 0.069$, second cas pour $p = 0.0263$).

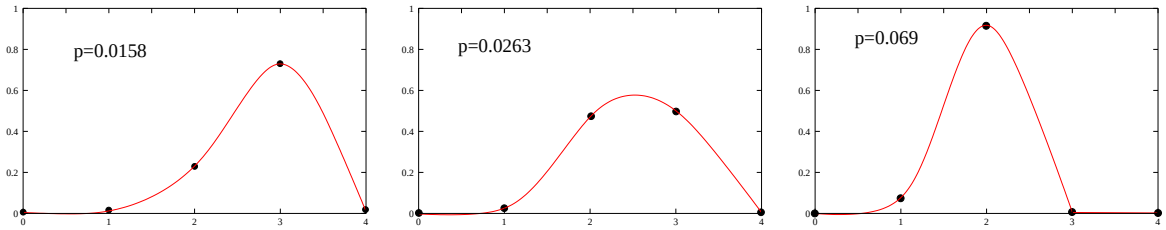

FIG. 8.3 – *Distribution des distances à la source pour des graphes aléatoires de différentes densités. Les distributions sont centrées autour d'une valeur moyenne qui décroît avec p.*

Les liens manqués sont ceux situés entre des sommets à même distance de la source. Donc, si la majorité des sommets est à même distance de la source, alors la majorité des liens sera entre ces sommetset sera donc manquée (d'où un creux sur la courbe). Au contraire, s'il y a autant de sommets à distance k qu'à distance $k+1$, alors les liens seront mieux répartis : un quart des liens seront entre des sommets à distance k, un quart entre des sommets à distance $k+1$ et le reste entre des sommets à des distances différentes. Donc, la moitié seulement des liens se trouvera entre des sommets à même distance de la source, ce qui est parfaitement confirmé par les résultats expérimentaux qui montrent des bosses dont le maximum atteint environ 0.5. On peut observer une bosses supplémentaire à environ $p = 0,008$, quand la distance moyenne des sommets à la source vaut $3,5$. Les autres bosses ne sont pas visibles car la distribution est trop étalée et le raisonnement ci-dessus n'est donc plus vraiment valide.

Ces résultats montrent clairement que des paramètres très simples comme la proportion de liens découverts ne peuvent pas être déduits simplement d'une vue partielle, même obtenue de façon très simple. L'efficacité de l'exploration dépend de la densité de liens et de faibles variations peuvent avoir un fort impact sur les résultats obtenus.

8.2.2 Quelques sources de plus

En pratique, les explorations réelles de l'Internet sont lancées depuis plusieurs sources, ou utilisent des techniques plus sophistiquées pour obtenir la vision la plus complète possible. Regardons dans quelle mesure le fait de prendre quelques sources au lieu d'une seule peut influencer les observations, en terme de proportion de liens découverts.

La Figure 8.4 montre la proportion de liens découverts en fonction du nombre de sources, toujours sur des graphes ER. Si on utilise toutes les destinations, la qualité de la vision du graphe augmente très rapidement avec le nombre de sources, ce qui est un point positif car, en pratique, il est très difficile d'utiliser beaucoup de sources.

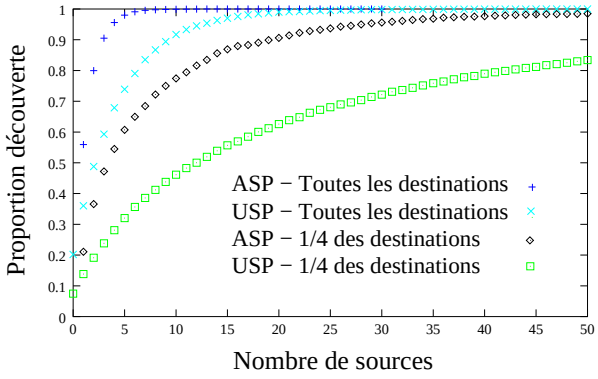

FIG. 8.4 – *Variation de la proportion de liens découverts en fonction du nombre de sources utilisées, soit en prenant tous les sommets comme destinations, soit en choisissant un quart des sommets comme destinations. Graphe aléatoire ER avec* $p = 0.005$, $n = 2000$.

Malgré tout, l'hypothèse selon laquelle on peut obtenir des chemins vers toutes les destinations est peu réaliste. Il est difficile de savoir quelle proportion des sommets servent de destination dans une exploration réelle. Aussi avons-nous choisi de faire une exploration de quelques sources vers un quart des destinations pour mesurer la qualité de la vision. Ces résultats apparaissent aussi sur la Figure 8.4. Dans ce cas, il faut beaucoup plus de sources pour obtenir une vision correcte du réseau, surtout avec le modèle USP.

Afin d'explorer plus précisement ces comportements, nous allons maintenant étudier le cas général en autorisant n'importe quel nombre de sources et de destinations, pour quantifier la proportion découverte et d'autres propriétés.

8.3 Proportion découverte

Avec des courbes telles que celles présentées ci-dessus, il n'est pas possible d'observer l'effet d'un grand nombre de paramètres, en particulier la proportion de sources et de destinations. Pour y remédier, nous allons maintenant faire usage de courbes en niveaux de gris définies comme suit : pour un graphe G à n sommets, nous allons considérer un carré de taille $n \times n$. Chaque courbe représente les variations d'un paramètre (clustering, distance moyenne, etc.) en fonction du nombre de sources et de destinations. Ainsi, le point (x, y) du carré correspond à une vision G' de G en utilisant x sources et y destinations, avec un modèle d'exploration donné. Le point $(0, 0)$ correspond à une exploration sans source ni destination : rien ne sera donc observé. Au contraire, le point (n, n) correspond à une exploration avec toutes les sources et toutes les destinations : le graphe sera donc entièrement découvert. La Figure 8.5 donne un exemple de courbe en niveau de gris, le paramètre étudié étant ici le degré moyen.

La couleur du point dans les courbes représente la valeur du paramètre p considéré allant du noir, pour $p = 0$, au blanc, pour la valeur maximale de p. Il faut noter que cette valeur maximale n'est pas forcément obtenue au point (n, n), certains paramètres pouvant

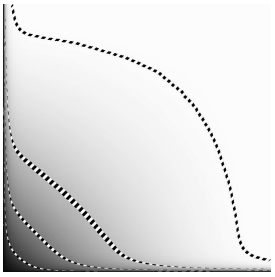

FIG. 8.5 – *Un exemple de courbe en niveau de gris représentant l'évaluation du degré moyen en fonction du nombre de sources (en abscisses) et de destinations (en ordonnées) pour un graphe ER.*

être surévalués durant l'exploration. Dans ces courbes, le point $(0, 0)$ sera donc toujours noir et le point (n, n) aura la couleur du paramètre p pour le graphe original G. Tout point plus sombre que celui-ci correspond à une zone où le paramètre est sous-estimé, tout point plus clair à une zone où le paramètre est surestimé. Enfin, la variation de gris est linéaire ; un gris médian correspond donc à un point où la valeur du paramètre vaut 50% de la valeur maximale.

Finalement, pour améliorer la visibilité des courbes, nous avons ajouté les courbes de niveau 25%, 50%, 75% et 99%. La courbe de niveau $x\%$ est l'ensemble des points où le paramètre vaut $x\%$ de la valeur maximale, à 1% près. Ces lignes de niveau sont d'une aide précieuse pour interpréter les courbes en niveau de gris.

Rappelons que chaque point correspond à une exploration G' d'un graphe G, ce qui a un coût de calcul assez élevé. Les calculs doivent donc être optimisés, et il faut les effectuer sur des graphes de taille raisonnable. Nous avons effectué des essais sur des graphes de taille 10^3, 10^4 et 10^5. Sur les petits graphes (10^3 sommets), on observe encore des effets liés à la taille, qui disparaissent pour des tailles de 10^4 et plus. Les calculs sur des graphes de taille 10^5 étant à la limite de ce qu'il est actuellement raisonnable de calculer, nous avons utilisé des graphes à 10^4 sommets qui offrent un bon compromis entre la taille et le temps de calcul.

Dans cette section, nous allons nous concentrer sur les propriétés de base : la proportion de sommets et de liens découverts, ainsi que le degré moyen, en faisant varier simultanément le nombre de sources et de destinations. Nous allons présenter les résultats les plus significatifs en utilisant les modèles de graphes ER, AB, MR et DM, pour tenter de comprendre les paramètres qui ont une influence forte sur la qualité de l'exploration.

Graphes aléatoires ER

Les remarques qui ont été faites dans la section précédente en utilisant peu de sources sont confirmées par l'utilisation de courbes en niveaux de gris (Figures 8.6 et 8.7).

Tant que le degré moyen du graphe exploré est faible, il n'y a pas de différence qualitative

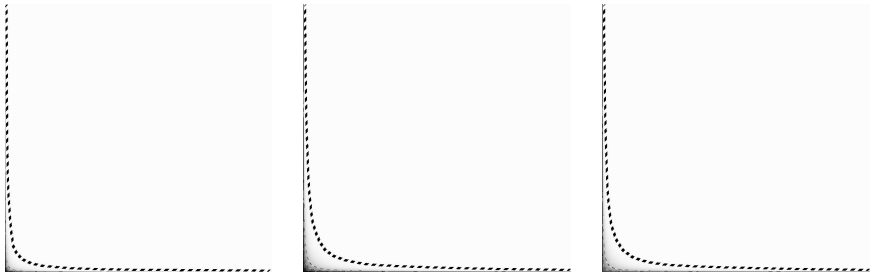

FIG. 8.6 – *Graphe ER : nombre de sommets, lienset degré moyen.* $p = 0.001$ *(degré moyen* 10*),* $n = 10^4$*,* USP*. Les courbes pour* ASP *sont très similaires.*

entre les explorations USP et ASP étant donné que le nombre de plus courts chemins entre deux sommets est faible. La qualité de l'exploration est satisfaisante même avec peu de sources et de destinations. Au contraire, si le degré moyen est plus élevé, le nombre de plus courts chemins entre deux sommets augmente ce qui entraîne une différence significative entre USP et ASP. Ceci peut être observé sur la Figure 8.7.

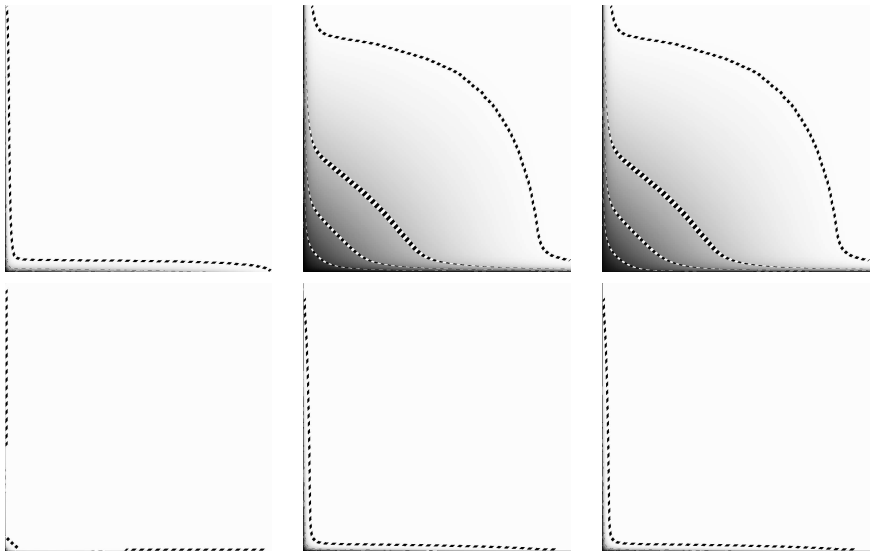

FIG. 8.7 – *Graphe ER dense : nombre de sommets, lienset degré moyen.* $p = 0.01$ *(degré moyen* 100*),* $N = 10^4$*,* USP *(première ligne) et* ASP *(seconde ligne).*

Remarquons que le degré moyen est obtenu en divisant deux quantités, le nombre de sommets et le nombre de liens découverts. Si l'une des deux propriétés est très mal estimée et pas l'autre, le degré moyen sera alors mal estimé aussi. Le fait de quotienter ces deux mesures agit comme un filtre du *pire cas*. Par exemple, la Figure 8.7 montre ce phénomène dans le cas USP : le nombre de liens est très mal estimé ce qui se répercute sur le degré moyen.

Toujours pour les graphes ER, comme nous l'avons déjà signalé, il n'y a pas de différence qualitative quand la taille du graphe observé varie. Les courbes précédentes (Figure 8.6) peuvent ainsi être comparées à celles obtenues sur des graphes ER plus petits (Figure 8.8) et sont très similaires. La proportion de sources et de destinations nécessaires pour obtenir une bonne estimation diminue légèrement quand la taille du graphe augmente, mais le nombre de sources et de destinations à considérer augmente (la proportion diminue moins vite que la taille n'augmente).

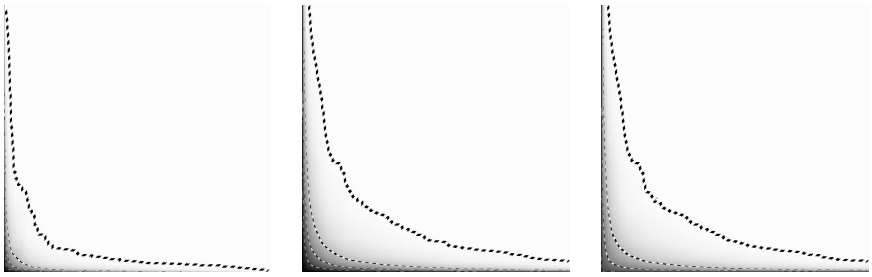

FIG. 8.8 – *Petit graphe ER : nombre de sommets, lienset degré moyen.* $p = 0.01$ *(degré moyen* 10*),* $N = 10^3$*,* USP.

Le premier paramètre qui influence la qualité de l'observation est donc la densité du graphe. À taille égale, plus un graphe est dense, plus il est difficile à explorer. Ceci est dû au grand nombre de courts chemins qui sont difficiles à capturer.

Graphes sans-échelle AB et MR

Penchons-nous maintenant sur les graphes sans-échelle, tout d'abord en considérant des graphes AB. La Figure 8.9 montre que la qualité de l'exploration est qualitativement similaire à celle obtenue pour les graphes ER (les courbes restent similaires en changeant les paramètres), bien que plus mauvaise : pour obtenir une carte très précise, il faut utiliser plus de sources et de destinations que sur un graphe ER de même taille. On observe aussi une différence assez marquée entre USP et ASP, ce qui tend à montrer qu'il existe en général de nombreux plus courts chemins, bien que le degré moyen soit faible.

La même expérience avec des graphes MR, qui sont aussi sans-échelle, ne donne pas des résultats similaires aux graphes AB comme on aurait pourtant pu s'y attendre. La Figure 8.10 montre des résultats surprenants : la qualité est beaucoup plus mauvaise pour les graphes MR que pour les graphes AB. Même en utilisant ASP, qui donne d'ordinaire de bien meilleurs résultats, il faut prendre environ la moitié des sources et des destinations pour voir 75% du graphe (en termes de sommets et de liens). Pour les graphes MR, nous avons utilisé une loi de puissance avec un exposant 2.5 qui est une valeur réaliste.

Au contraire, et cela peut paraître surprenant, le degré moyen est très bien estimé en étant à peine surestimé. Comme le nombre de sommets et le nombre de liens découverts sont mal estimés, mais avec un biais similaire, le quotient peut faire disparaître le biais.

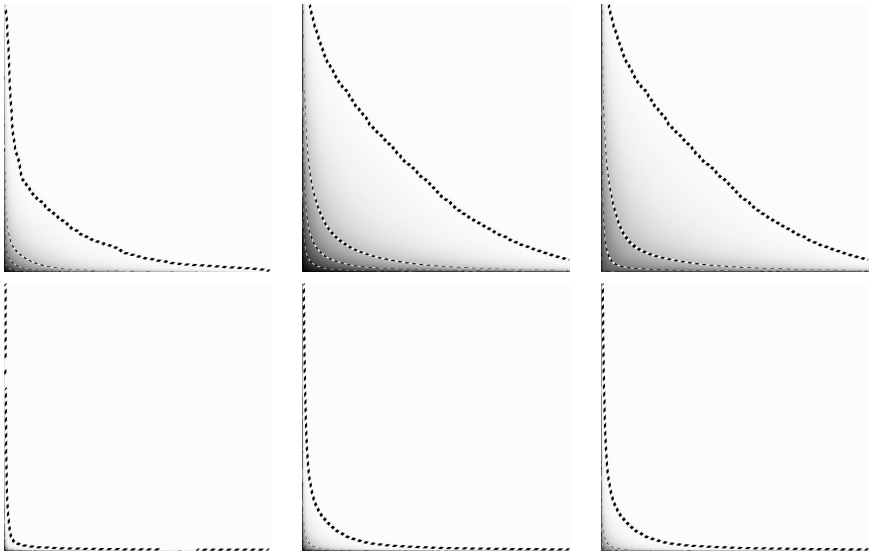

FIG. 8.9 – *Graphe AB : nombre de sommets, lienset degré moyen.* $k = 10$, $n = 10^4$, USP *(première ligne) et* ASP *(seconde ligne).*

Le degré moyen est surestimé car les sommets de fort degré et les liens qui y sont attachés sont découverts très rapidement, alors que les sommets de faible degré ne seront découverts que plus tard. C'est seulement à ce moment-là que le degré moyen baisse.

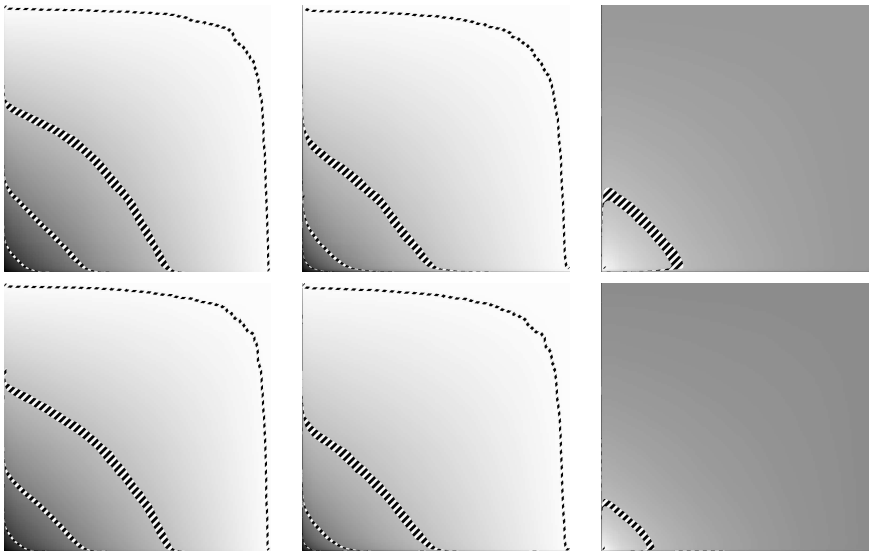

FIG. 8.10 – *Graphe MR : nombre de sommets, lienset degré moyen. Loi de puissance avec exposant* $\alpha = 2.5$, $n = 10^4$, USP *(première ligne) et* ASP *(seconde ligne).*

Un argument simple permet d'expliquer pourquoi les graphes MR sont plus difficiles à

explorer que les graphes AB. Dans un graphe AB de degré moyen k, le degré minimal est par construction $k/2$, étant donné que l'on ajoute $\frac{k}{2}$ liens à chaque étape. La distribution des degrés suit donc une loi de puissance mais uniquement pour les sommets de degré supérieur à $\frac{k}{2}$. Au contraire, et toujours par construction, les graphes MR contiennent un grand nombre de sommets de faible degré, y compris de degré 1. Pendant une exploration, ces sommets de degré 1 sont extrêmement difficiles à découvrir puisqu'il faut obligatoirement les choisir comme source ou destination pour pouvoir les observer. Les sommets de faible degré, 2 ou plus, sont un peu moins dur à capturer mais ne sont pas forcément découverts très rapidement.

Une manière de vérifier cet argument est de considérer uniquement le cœur des graphes MR. Le cœur d'un graphe est obtenu en supprimant tous les sommets de degré 1 et en itérant jusqu'à ce qu'il n'y en ait plus (la suppression de sommets de degré 1 peut créer de nouveaux sommets de degré 1 qu'il va falloir supprimer aussi). Un graphe MR est donc composé d'un cœur auquel sont accrochées des structures arborescentes. La Figure 8.11 présente les résultats obtenus sur le cœur. Pour USP, les résultats sont beaucoup plus proches de ceux obtenus sur les graphes AB. La différence qui subsiste vient de la difficulté à découvrir les autres sommets de très faible degré qui n'existent pas dans les graphes AB (aucun sommet de degré inférieur à $\frac{k}{2}$). La différence entre USP et ASP est moins prononcée sur le cœur de MR que sur les graphes AB, ce qui sous-entend qu'il y a plus de chemins multiples dans ces derniers.

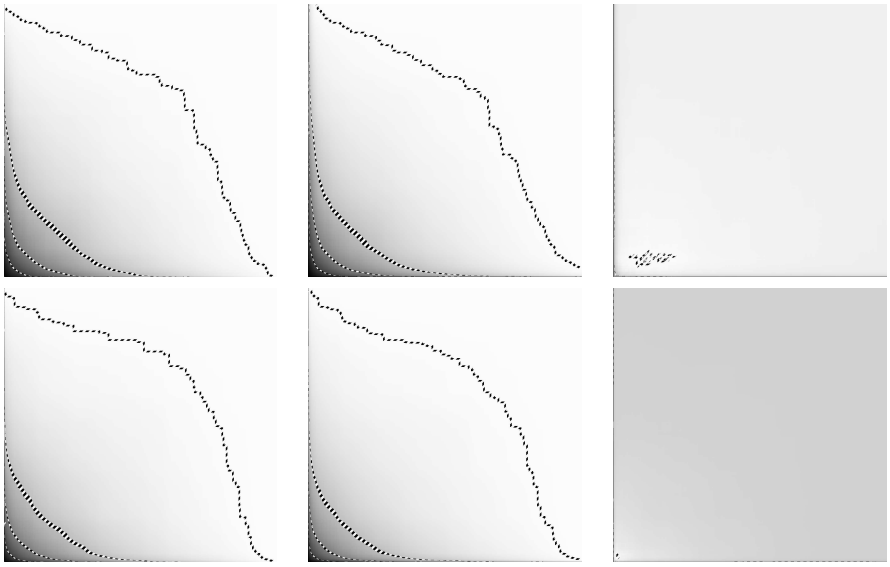

FIG. 8.11 – *Cœur d'un graphe MR : nombre de sommets, lienset degré moyen. Loi de puissance avec exposant* $\alpha = 2.5$, $n = 1311$, USP *(première ligne) et* ASP *(seconde ligne).*

Nous avons donc identifié un second paramètre qui influence la qualité de l'exploration : la présence de sommets de faible degré. De tels sommets appartiennent à peu de courts

chemins et sont donc difficiles à découvrir. Au contraire, le cœur du graphe, et notamment les sommets de fort degré, sont rapidement détectés.

Graphes clusterisés DM

Finalement, nous allons étudier l'impact du clustering en utilisant les graphes DM qui sont sans-échelle et contiennent de nombreux triangles. Comme les graphes AB, ils n'ont pas de sommets de degré 1 et sont très peu denses. Les effets identifiés précédemment sur les graphes MR et ER ne devraient donc pas se produire.

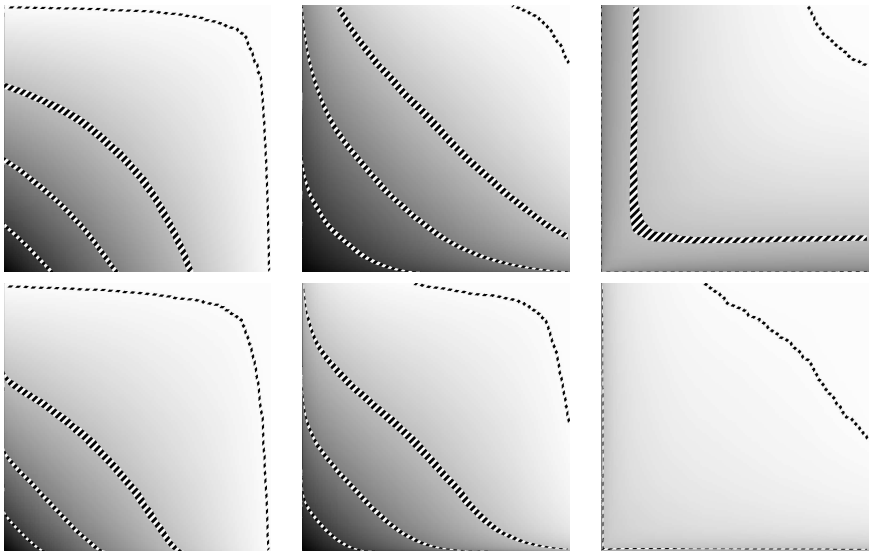

FIG. 8.12 – *Graphe DM : nombre de sommets, liens, et degré moyen.* $n = 10^4$, USP *(première ligne) et* ASP *(seconde ligne).*

Malgré tout, on peut voir sur la Figure 8.12 que la qualité de la vision est toujours mauvaise. Le fait que les courbes pour USP et ASP soient très similaires semble indiquer qu'il y a peu de courts chemins (faible densité).

Ceci, et la faible qualité obtenue, peut s'expliquer comme suit : pour explorer un graphe dense (une clique, par exemple), il faut utiliser un grand nombre de sources et de destinations. Ainsi, dans un simple triangle formé de trois sommets, il n'est pas possible de découvrir plus d'un lien par `traceroute`, que ce soit en USP ou en ASP. Le plus court chemin est toujours de longueur 1, découvrir un triangle requiert donc au minimum trois chemins et, de manière plus générale, découvrir une clique de taille k en requiert $k(k-1)/2$.

Le fort clustering des graphes DM, provenant de la présence de nombreux sous-graphes denses, explique la difficulté à explorer de tels graphes.

Cette fois encore, le degré moyen est très mal estimé. Des comportements visuellement très similaires pour la proportion de sommets et de liens découverts, comme sur les Figures 8.10 et 8.12, peuvent donner des estimations du degré moyen très différentes.

Nous avons finalement identifié deux phénomènes qui rendent les graphes difficiles à explorer correctement. La densité élevée du graphe, qu'elle soit globale ou locale (clustering), rend les liens difficiles à détecter et la distribution des degrés en loi de puissance engendre de nombreux sommets de faible degré, très difficiles à découvrir. Ces deux propriétés sont complémentaires et agissent sur différentes parties du graphe : le cœur pour le clustering et le bord (structure arborescentes) du graphe pour les sommets de faible degré.

8.4 Distributions des degrés

La distribution des degrés sur le graphe de l'Internet a reçu une forte attention récemment et c'est en particulier la propriété qui a fait l'objet du plus grand nombre d'études sur le biais introduit par la mesure [31, 62, 67, 76, 112, 118]. Nous allons approfondir ces études en considérant différents modèles de graphes et d'explorations, et en faisant varier le nombre de sources et de destinations. Il n'est pas possible d'utiliser de courbes en niveaux de gris dans ce contexte puisque l'on va s'intéresser à la vitesse de convergence vers la distribution réelle. Ceci ne peut que difficilement s'exprimer par une valeur réelle qui serait nécessaire pour une courbe en niveaux de gris. Par conséquent, nous allons présenter les distributions des degrés pour des valeurs de paramètres représentatives.

Graphes aléatoires

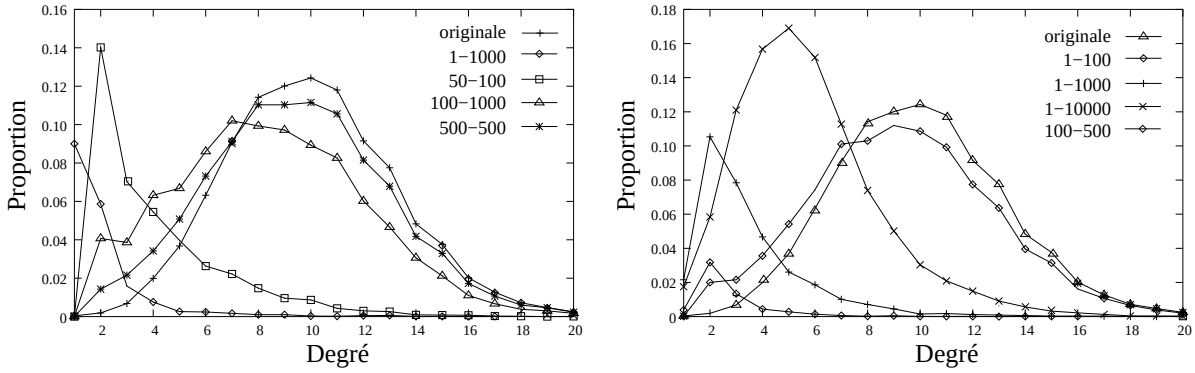

FIG. 8.13 – *Graphe ER : distribution des degrés en fonction du nombre de sources et de destinations.* $k = 10$, $n = 10^4$. USP *(à gauche) et* ASP *(à droite). Les courbes montrent aussi la distribution originale du graphe.*

Si l'on considère tout d'abord les graphes ER avec un faible degré moyen et donc peu denses, la Figure 8.13 montre qu'avec peu de sources la distribution est très éloignée de la distribution réelle. En particulier, avec une exploration USP, la distribution converge très lentement : elle est toujours notablement différente de la distribution réelle, même avec 1% de sources et 10% de destinations. Avec ASP, les résultats sont meilleurs et une distribution presque parfaite est obtenue avec seulement 0.5% des sources et 20% des destinations.

Le cas des graphes ER plus denses (Figure 8.14) est plus intéressant : la présence de sommets de fort degré fait qu'il est possible d'assimiler la distribution des degrés obtenue par exploration avec USP à une loi de puissance. Cela a déjà été étudié dans des travaux précédents [76, 112] pour montrer à quel point l'exploration peut engendrer des phénomènes qualitativement différents. Ce biais de mesure se produit quand on n'utilise que très peu de sources et beaucoup de destinations avec USP (Figure 8.14, à gauche). Il disparaît dès que le nombre de sources dépasse environ 0.5% des sommets (Figure 8.14, à droite), ou que l'on considère une exploration avec ASP, même avec peu de sources et de destinations (Figure 8.15).

Si l'on se place à la frontière entre les zones très biaisées et les zones non biaisées, on voit apparaître des phénomènes transitoires plus surprenants. Ainsi, sur la Figure 8.14 à droite, la courbe utilisant 500 sources et 5 000 destinations est clairement composée de deux bosses. Ceci est dû au fait que les liens proches des sources sont tous découverts ou presque, alors que ceux proches des destinations ne le sont pas. La distribution des degrés est donc un cumul entre deux distributions : celle des voisins des sources et celle des voisins des destinations [76].

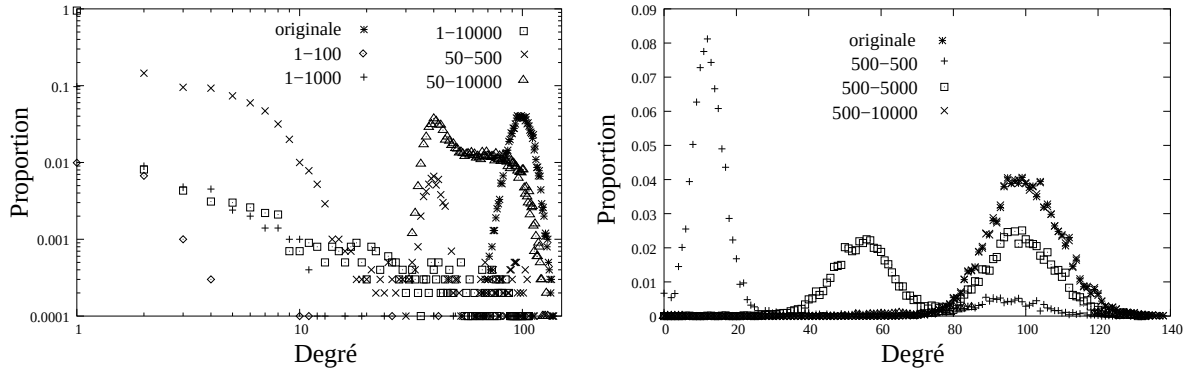

FIG. 8.14 – *Graphe ER : distributions des degrés. $k = 100$, $n = 10^4$, avec* USP. *Pour un petit nombre de sources (à gauche) et un grand nombre de sources (à droite).*

Ces résultats concernant les graphes ER, qui ont une distribution de degrés en loi de Poisson, montrent qu'il est assez difficile d'obtenir une bonne estimation de leur distribution des degrés. Celle-ci est grandement améliorée en utilisant de nombreuses sources et destinations. L'utilisation de peu de sources, comme c'est le cas sur les explorations réelles, peut conduire à des distributions très différentes des distributions réelles.

Graphes sans-échelle

Sur les graphes sans-échelle, les résultats sont complètement différents, comme le montrent les Figures 8.16 et 8.17 pour les graphes MR et DM respectivement. Les explorations USP

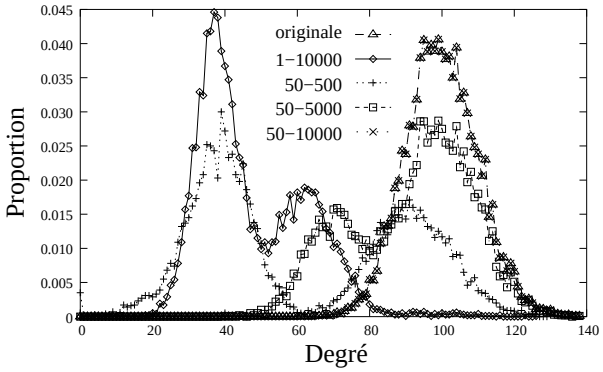

FIG. 8.15 – *Graphe ER : distribution des degrés.* $k = 100$, $n = 10^4$, ASP.

et ASP donnent toutes deux une vision correcte de la distribution réelle [1], même en utilisant peu de sources et de destinations. Dans le cas des graphes MR, les résultats sont excellents (nous omettons ceux pour les graphes AB qui sont très similaires). Dans le cas des graphes DM, l'exposant de la droite est sous-estimé au début mais converge très vite quand le nombre de sources augmente.

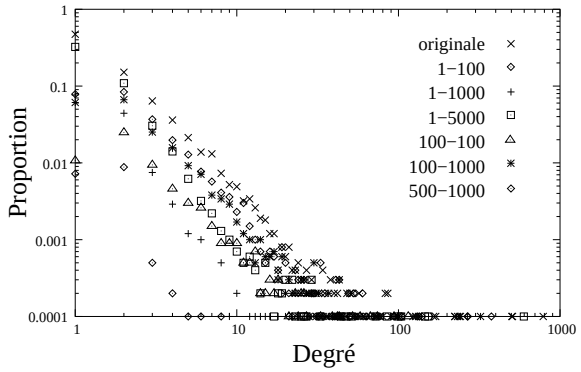

FIG. 8.16 – *Graphe MR : distribution des degrés. Loi de puissance d'exposant* $\alpha = 2.5$, $n = 10^4$, USP.

Le comportement des graphes ER et sans-échelle sont complètement différents pour ce qui est de la distribution des degrés. Alors qu'il est assez difficile d'obtenir une bonne vision des graphes ER, l'exposant (et donc la loi) des graphes sans-échelle est très bien estimé sur tous les types de graphes (MR, AB et DM), même avec peu de sources. Les cas où la distribution peut être mal évaluée ne se produisent que très rarement (en utilisant très peu de sources). Dans tous les autres cas, la convergence est en moyenne très rapide.

1. La caractéristique principale d'une loi puissance est son exposant, *i.e.* la pente de la droite en échelle log-log. Ici, nous divisons le nombre de sommets de degré donné par le nombre total de sommets, incluant les sommets non découverts. La pente est malgré tout la même la plupart du temps.

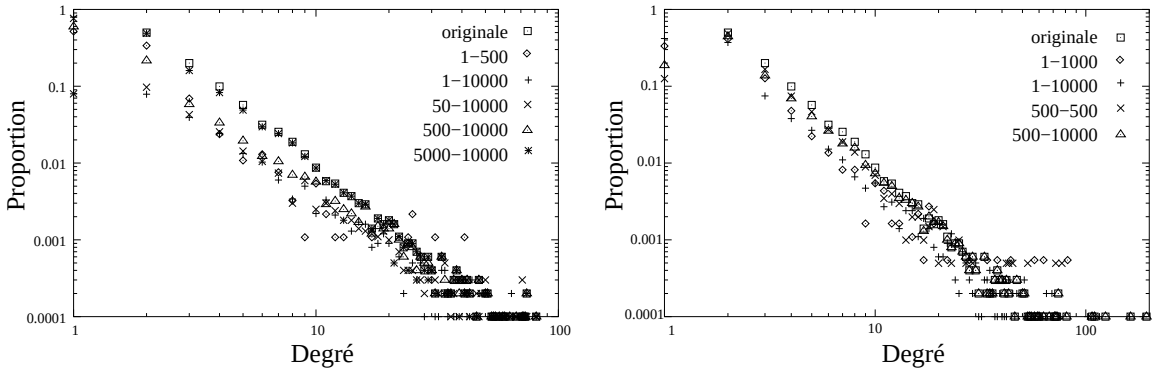

FIG. 8.17 – *Graphe DM : distribution des degrés.* $n = 10^4$, USP *(à gauche) et* ASP *(à droite).*

8.5 Distance moyenne

Le calcul de la distance moyenne est difficile en pratique sur des graphes de grande taille et, dans notre cas, la quantité de graphes sur lesquels il faut l'effectuer complexifie encore les calculs. Il est heureusement possible d'évaluer précisément cette distance en la calculant pour un petit nombre de couples source-destination. Les résultats de cette section seront donc tous basés sur cette approximation.

Remarquons qu'une fois que tous les sommets ont été découverts, ajouter de nouvelles sources ou destinations ne peut que faire diminuer la distance moyenne en créant des raccourcis. La distance sera donc surestimée au début, puis convergera vers la valeur réelle. Les explorations avec ASP sont aussi naturellement plus efficaces que celles avec USP puisqu'elles permettent de découvrir plus de liens.

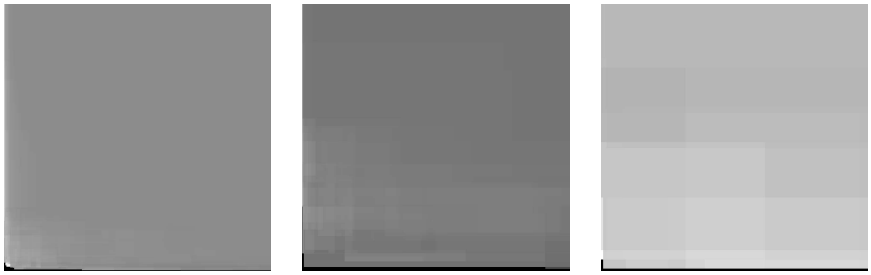

FIG. 8.18 – *Distance moyenne pour (de gauche à droite) : graphe ER* ($k = 10$, $n = 10^4$)*, graphe AB* ($k = 10$, $n = 10^4$) *et graphe DM* ($N = 10^4$)*. Explorations avec* USP.

Comme on peut le vérifier sur la Figure 8.18, l'évaluation de la distance moyenne devient très vite excellente dans tous les cas. Les courbes en niveaux de gris sont presque uniformes, ce qui signifie qu'avec très peu de courts chemins on obtient une très bonne évaluation de la distance moyenne du graphe entier. Toutes les courbes pour les graphes ER de différentes

densités sont similaires, de même que les courbes pour MR et AB. Nous ne présentons donc pas ces courbes.

Il faut juste noter que l'évaluation est un peu moins bonne pour les graphes DM (Figure 8.18 à droite). Comme la découverte des liens est très mauvaise sur ces graphes (voir Figure 8.12), la valeur de la distance moyenne met plus longtemps à converger.

8.6 Clustering

Le clustering global d'un graphe peut être calculé comme étant le rapport entre le nombre de triangles et le nombre de triplets connectés dans le graphe (voir le Chapitre 1). De manière similaire au degré moyen, l'évaluation du clustering dépend de la vitesse à laquelle les triangles sont découverts, rapportée à la vitesse de découverte des triplets : l'évaluation du clustering est correcte si triangles et triplets sont découverts au même rythme. C'est ce que nous allons étudier maintenant.

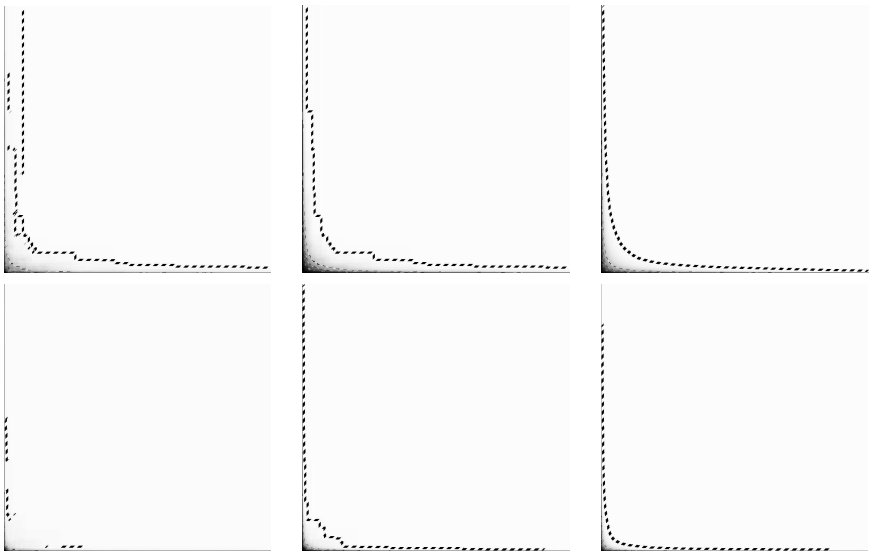

FIG. 8.19 – *Graphe ER : clustering, nombre de triangles et de triplets découverts.* $k = 10$, $n = 10^4$. USP *(première ligne) et* ASP *(deuxième ligne).*

Considérons tout d'abord le cas des graphes ER. Quand le degré moyen est faible, il n'y a presque pas de triangles, le clustering est donc presque nul. Quand le degré moyen augmente, le nombre de triangles et donc le clustering augmentent de pair. Les Figures 8.19 et 8.20 ne sont pas surprenantes : en augmentant le nombre de sources et de destinations, on découvre plus de liens, l'évaluation du clustering est donc améliorée.

Considérons maintenant les graphes AB et MR (de comportements très similaires) qui ont un faible clustering. Le clustering est surestimé dans le cas USP en prenant peu de sources et de destinations (Figure 8.21). Ceci vient du fait que l'on découvre plus de triangles que de triplets au tout début de l'exploration. Mais cette tendance s'inverse

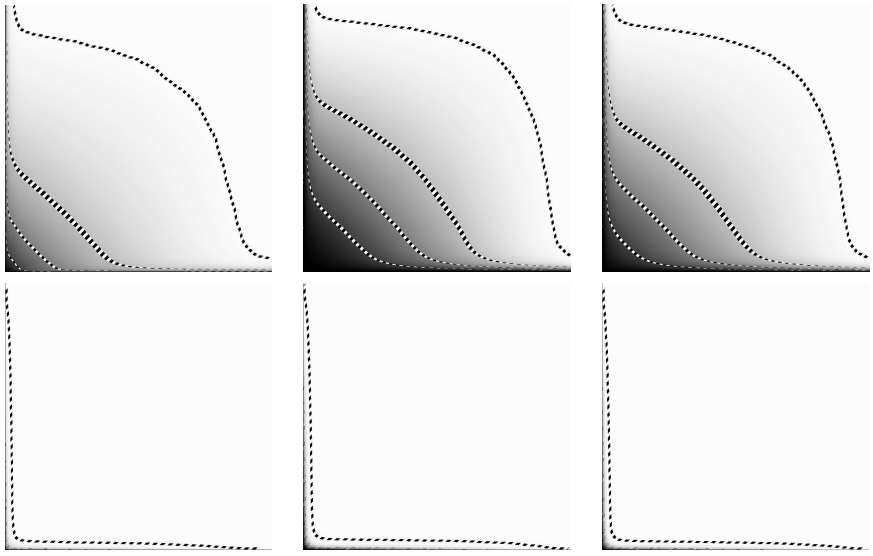

FIG. 8.20 – *Graphe ER dense : clustering, nombre de triangles et de triplets découverts.* $k = 100$, $n = 10^4$. USP *(première ligne) et* ASP *(deuxième ligne).*

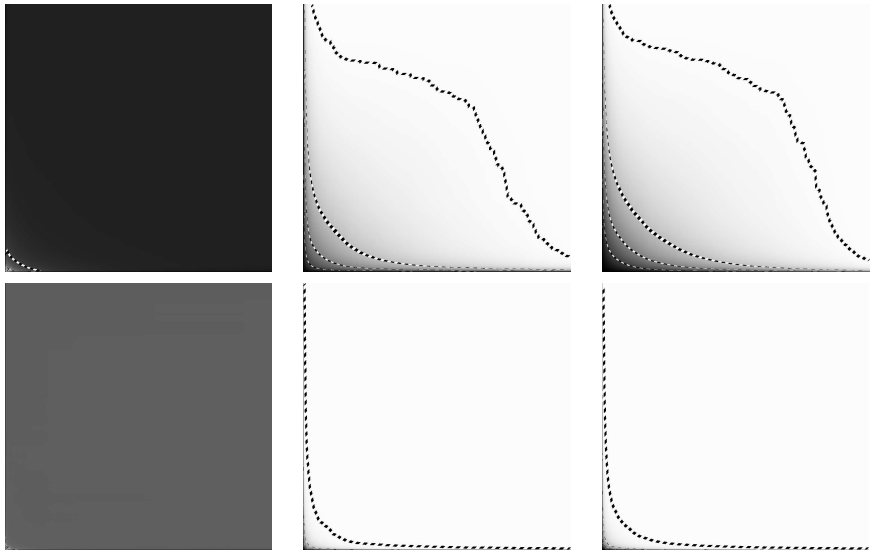

FIG. 8.21 – *Graphe AB : clustering, nombre de triangles et de triplets découverts.* $k = 10$, $n = 10^4$. USP *(première ligne) et* ASP *(deuxième ligne).*

rapidement et le clustering est très vite correctement estimé et converge vers la valeur réelle (les courbes sont presque uniformément grises), bien que les deux paramètres dont il découle soient, eux, mal estimés. Les explorations avec ASP donnent des résultats plus satisfaisants, tous les paramètres étant très vite bien estimés.

Enfin, pour les graphes clusterisés obtenus avec le modèle DM, la Figure 8.22 montre

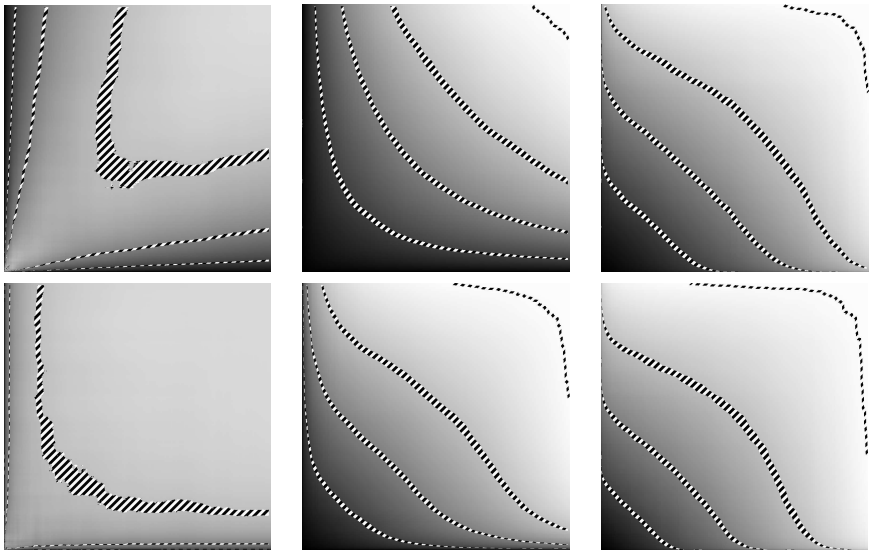

FIG. 8.22 – *Graphe DM : clustering, nombre de triangles et de triplets découverts.* $n = 10^4$. USP *(première ligne) et* ASP *(deuxième ligne).*

que le clustering est bien évalué dans tous les cas, sauf lorsque l'on utilise beaucoup de sources et peu de destinations, ou l'inverse (ce qui est le cas pour les explorations réelles). En effet, dans ces deux cas, les vitesses de découverte des triangles et des triplets sont très différentes. Au contraire, si le nombre de sources et de destinations est similaire, les deux mesures sont mal estimées mais du même ordre : le clustering se trouve donc mieux estimé (mais toujours relativement mal).

Le fait que la vision que l'on a du graphe soit construite par fusion de structures arborescentes, ce qui rend les triangles difficiles à découvrir, rend le clustering difficilement évaluable. Cette évaluation est possible uniquement si le clustering réel est faible, auquel cas peu de sources et de destinations suffisent. Dans ce cas précis, la marge d'erreur étant plus faible, les risques sont aussi plus faibles. Au contraire, si le clustering est très élevé, il faut beaucoup de sources et de destinations avant de découvrir un nombre satisfaisant de triangles.

8.7 Placement des sources et des destinations

Sur l'Internet, tous les routeurs ne jouent pas le même rôle et n'ont pas les mêmes propriétés. Du point de vue des degrés, notamment, certains ont beaucoup de connexions alors que d'autres en ont très peu. La question de savoir si le choix des sources et des destinations a une influence sur la qualité de l'exploration se pose donc.

La différence la plus simple entre les sommets de l'Internet vu comme un graphe est leur degré. Nous allons donc considérer des stratégies dans lesquelles les sources et les destinations sont choisies en fonction de leur degré, au lieu de les choisir dans un ordre

aléatoire comme nous l'avons fait jusqu'à maintenant. Les trois stratégies de base sont les suivantes :

- sources et destinations sont choisies par ordre de degrés croissants ;
- sources et destinations sont choisies par ordre de degrés décroissants ;
- les sources sont choisies par ordre de degrés croissants et les destinations par ordre de degrés décroissants. La stratégie décroissante pour les sources et croissante pour les destinations est symétrique.

Les résultats obtenus avec ces stratégies doivent être comparés avec la stratégie aléatoire étudiée dans les sections précédentes.

Notons que d'autres stratégies basées sur d'autres propriétés, telles que la centralité, seraient envisageables. Nous étudions ici les cas les plus simples qui donnent déjà quelques intuitions sur l'influence du placement des sources et des destinations.

Comme on aurait pu s'y attendre, ces différentes stratégies donnent des résultats très similaires sur les graphes ER. En effet, les sommets dans les graphes ER ont presque tous le même degré, et la qualité de l'exploration est bonne même avec très peu de sources et de destinations (Figure 8.6). Il était donc peu probable de pouvoir vraiment améliorer ces résultats avec une autre stratégie d'exploration. De même, l'exploration des graphes AB est déjà très satisfaisante avec un nombre raisonnable de sources et de destinations : il n'y a pas d'amélioration visible. Nous allons donc nous restreindre aux graphes MR et DM.

Dans le cas des graphes MR (Figure 8.23), les résultats montrent que le placement est intéressant : les quatre stratégies donnent des résultats différents. De plus, la meilleure stratégie semble être la stratégie croissante-croissante. La raison en est que les sommets de faible degré sont difficiles à découvrir dans les graphes sans-échelle (voir Section 8.3) et que les choisir comme sources ou destinations nous assure de les découvrir rapidement. Au contraire, et par les mêmes arguments, la stratégie décroissante-décroissante est beaucoup moins efficace.

Sur les graphes DM, la Figure 8.24 nous donne les résultats des différentes stratégies pour des explorations USP (les explorations ASP donnent des résultats très similaires). Dans ce cas, la stratégie à choisir dépend de l'objectif à atteindre. Pour la découverte des sommets, comme pour les graphes MR, il faut choisir la stratégie croissante-croissante. Par contre, la découverte des liens est très mauvaise avec cette stratégie, et le degré moyen est par conséquent très mal estimé. Pour bien estimer ces propriétés, la stratégie croissante-décroissante (ou la symétrique) est plus efficace, surtout si le nombre de sources et de destinations est du même ordre : la zone blanche qui correspond aux points au-dessus de la courbe de niveau 99, donc très bien estimés, est beaucoup plus grande dans ce cas.

Dans tous les cas, on se rend compte que la stratégie décroissante-décroissante est la plus mauvaise. Les sommets de fort degré sont découverts très vite, il ne sert donc à rien de les considérer en priorité.

Sans aller plus loin dans les stratégies de placement, on se rend compte sur ces quelques exemples de l'efficacité de celles-ci. Malgré tout, choisir la bonne stratégie n'est pas trivial car cela dépend à la fois des propriétés du graphe sous-jacent et des propriétés que l'on

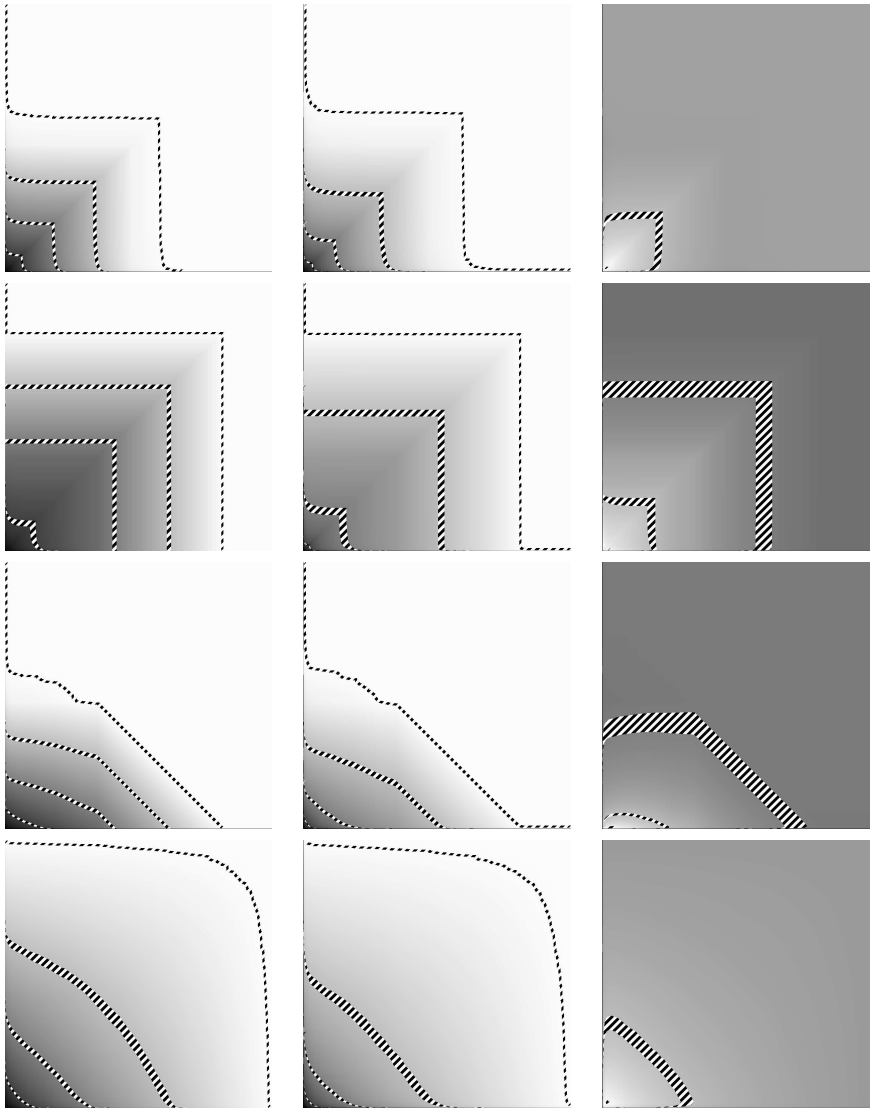

FIG. 8.23 – *Graphe MR: nombre de sommets, de liens et degré moyen. Loi de puissance d'exposant* $\alpha = 2.5$, $n = 10^4$. USP *avec les quatre principales stratégies pour les sources-destinations. De haut en bas: croissant-croissant, décroissant-décroissant, croissant-décroissant et aléatoire. Il n'y a pas de différences notable sur les courbes* ASP.

veut estimer. Or, en pratique, même si l'on connaît à peu près les propriétés du graphe, il n'est pas forcément évident de connaître les propriétés des sommets afin de savoir lesquels utiliser comme sources ou destinations.

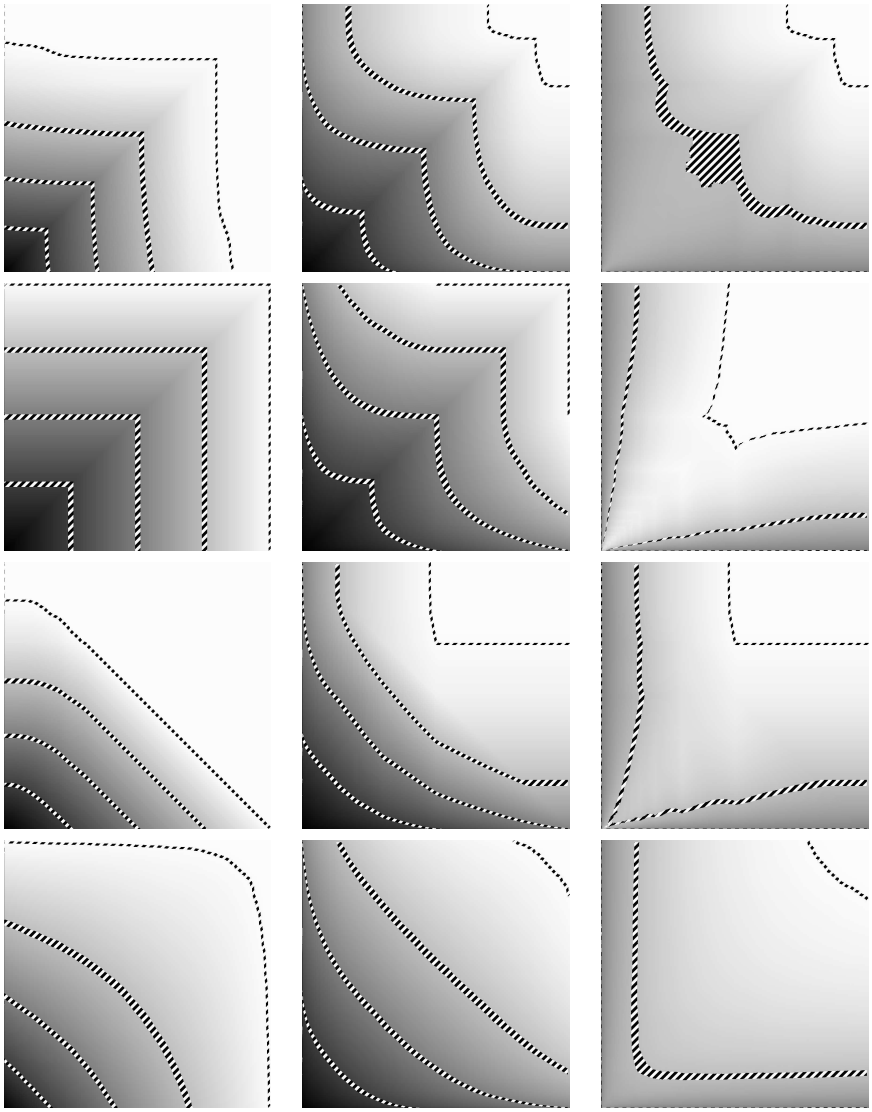

FIG. 8.24 – *Graphe DM: nombre de sommets, liens et degré moyen.* $n = 10^4$. USP *avec les quatre principales stratégies pour les sources-destinations. De haut en bas : croissant-croissant, décroissant-décroissant, croissant-décroissant et aléatoire.*

8.8 Expérience sur des données réelles

Jusque là, nous avons présenté des simulations sur des réseaux générés par des modèles en utilisant des approximations sur le fonctionnement de `traceroute` et le processus d'exploration. Nous allons maintenant répéter ces expériences sur des graphes réels afin d'évaluer la pertinence d'une telle approche.

Pour cela, nous allons utiliser le cœur de la carte de l'Internet *Mercator* [55, 56], *i.e.* le sous-réseau obtenu en supprimant itérativement tous les sommets de degré 1 de la carte

originale. Cette carte a toutes les propriétés que nous avons mentionnées plus tôt : fort clustering, distribution de degrés en loi de puissance et distance moyenne faible. Nous nous restreignons à l'étude du cœur car nous avons déjà remarqué que les structures arborescentes sont difficiles à explorer. Notre but est maintenant d'identifier d'autres propriétés.

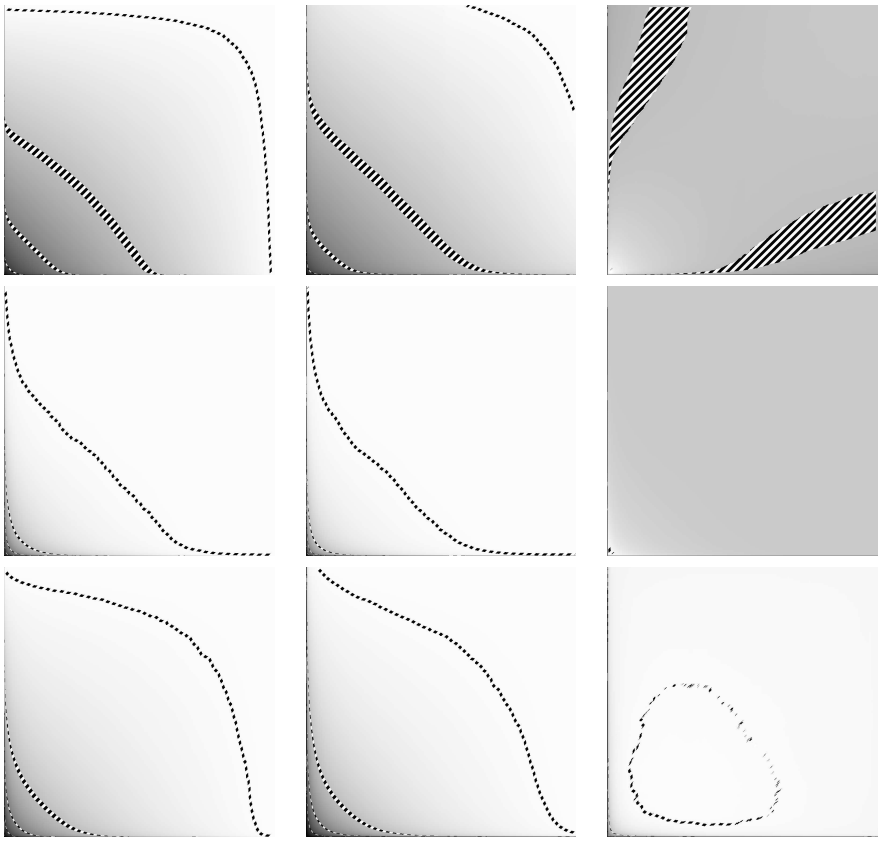

FIG. 8.25 – *Nombre de sommets, liens et degré moyen pour (de haut en bas) : le cœur de la carte* Mercator*, un graphe MR avec la même distribution des degrés et un graphe GL avec la même distribution de tailles de cliques. Exploration* USP.

À titre de comparaison, nous avons effectué les mêmes expériences sur des graphes ayant les mêmes propriétés, à savoir un graphe MR avec exactement la même distribution des degré et un graphe GL avec la même distribution de taille de cliques. Le premier a donc la même distribution des degré mais pas de clustering alors que le second a à la fois la même distribution des degrés et un clustering non trivial. La Figure 8.25 présente les résultats pour les propriétés de base et la Figure 8.26 présente le clustering [2]. Les résultats concernant

2. Les sauts dans les courbes en niveau de gris pour le clustering sont dus aux mêmes sauts observés dans le nombre de triplets découverts. Ces derniers viennent de l'existence d'un sommet de fort degré qui, lorsqu'il est choisi comme une source, produit un très grand nombre de triplets ($d(d-1)$ où d est son degré).

la distance moyenne et les distributions des degrés sont similaires à ceux observés sur les modèles.

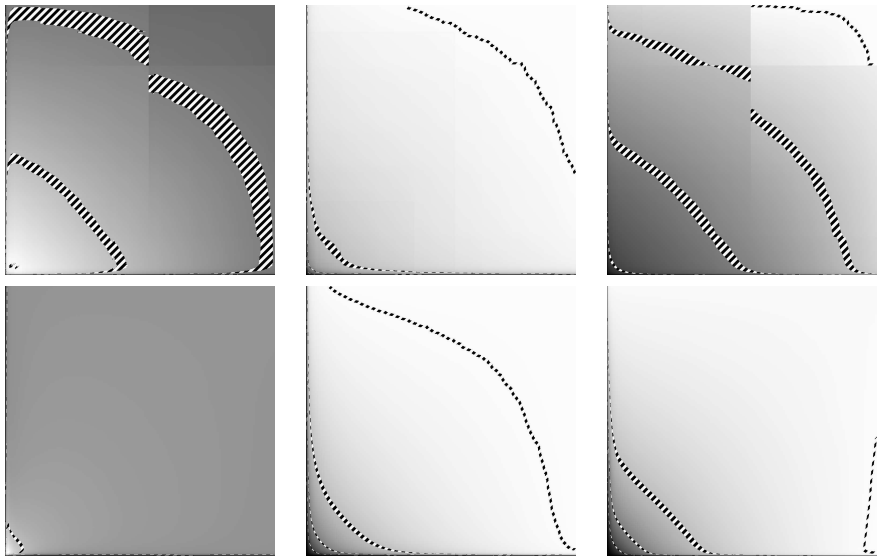

FIG. 8.26 – *Clustering, nombre de triangles et nombre de triplets pour le graphe* Mercator *original (première ligne) et un graphe GL avec la même distribution de taille de cliques (seconde ligne). Explorations* USP.

La Figure 8.25 associée aux remarques formulées précédemment nous permet de conclure sur certains points :

- la mauvaise qualité de l'exploration de la carte *Mercator* n'est pas uniquement due à la présence de structures arborescentes, celles-ci ayant été supprimées ;
- la carte *Mercator* ne peut pas être vue comme un graphe MR, l'exploration de son cœur donnant des résultats à la fois différents de l'exploration du cœur d'un graphe MR (Figure 8.11) et de l'exploration d'un graphe MR avec la même distribution des degrés (deuxième ligne de la Figure 8.25) ;
- le clustering est sans doute la propriété principale qui influence la qualité de l'exploration, les résultats sur la carte de *Mercator* étant assez similaires à ceux obtenus sur les graphes DM (première ligne de la Figure 8.12) et sur les graphes GL (troisième ligne de la Figure 8.25) ;
- les différences observées entre *Mercator*, les graphes DM et les graphes GL laissent à penser que, si la distribution des degrés et le clustering influencent le processus d'exploration, il doit exister d'autres propriétés, non capturées par ces deux modèles, qui ont aussi une influence.

La position exacte des destinations et les routes utilisées pour produire la carte de *Mercator* ne sont pas disponibles. Il n'est donc pas possible de produire des courbes en niveau de gris qui soient celles de l'exploration réelle en considérant les vraies routes au lieu

de plus courts chemins. Une telle expérience apporterait toutefois un éclairage intéressant sur le problème et serait à conduire.

8.9 Conclusion

À partir des simulations présentées ici, nous pouvons émettre quelques conclusions sur la qualité des cartes de l'Internet et sur des améliorations possibles :

- deux propriétés statistiques des grands réseaux d'interactions ont une grande influence sur la qualité des explorations obtenues par plus courts chemins : la présence de nombreux sommets de faible degré et le fort clustering. Ces deux propriétés agissent de manière indépendante et leur effets sont combinés dans le cas de l'Internet, ce qui rend le graphe difficile à explorer ;
- utiliser une exploration à partir d'un grand nombre de sources semble pertinent dans ce contexte, surtout pour les graphes sans-échelle clusterisés ;
- évaluer la distribution des degrés de ces graphes, ainsi que leur distance moyenne, se fait très bien même avec des explorations de tailles très réduites ;
- les détails du processus d'exploration n'ont pas une très grande importance quand le nombre de sources et de destinations augmente. Dans le cas de l'Internet, augmenter le nombre de points de vue pourrait donc donner des mesures plus indépendantes du processus ;
- bien que le clustering et le caractère sans-échelle jouent un rôle fondamental dans la qualité des explorations, il semble que d'autres propriétés non identifiées aient aussi un rôle important ;
- le placement des sources et des destinations peut grandement améliorer les explorations, à condition de pouvoir identifier les endroits stratégiques où placer les sources dans le réseau sans connaître celui-ci.

Enfin, soulignons que la métrologie de graphes (dont celle de l'Internet) n'en est qu'à ses débuts. Beaucoup de travail reste à faire, tant sur les aspects formels que les simulations ou les expériences.

Au delà, ce chapitre a montré comment un ensemble de modèles de graphes et un ensemble de propriétés significatives permettent, dans un contexte de simulation, d'étudier l'impact de diverses propriétés sur un phénomène complexe. Cette approche est très générale et peut être utilisée dans des contextes très divers, en informatique et ailleurs (sociologie, biologie, etc.). Notons également que les données réelles sont essentielles dans ce contexte pour valider les simulations. De même, il est crucial de disposer de modèles capturant les diverses propriétés qu'on cherche à observer, ce qui n'est pas parfaitement le cas dans le contexte des grands réseaux d'interactions comme nous l'avons vu dans la deuxième partie.

Conclusion

Les résultats présentés dans cette partie montrent que les applications fournissent non seulement des réponses à certains problèmes spécifiques, mais définissent aussi des problématiques plus larges. Les différentes approches illustrées sont basées sur des mesures et de l'analyse (Chapitre 6), des approches plus formelles (Chapitre 7), ou à base de simulation (Chapitre 8). Chacune de ces approches permet d'obtenir des résultats pertinents et d'ouvrir sur de nombreuses perspectives.

Ainsi, le chapitre sur l'analyse de réseaux pair-à-pair introduit un nouveau type de mesure et permet de mieux appréhender la nature des échanges. Dans de nombreux contextes, il est possible d'utiliser des mesures variées et le pair-à-pair en est un exemple typique : observation du trafic sur un routeur de l'Internet, comportement au niveau des utilisateurs, ou encore, comme c'est le cas dans le Chapitre 6, le comportement vu au niveau du serveur. Ces différentes approches permettent d'obtenir plusieurs visions d'un même réseau et soulèvent de nouvelles problématiques, comme l'observation de la dynamique des échanges.

Le chapitre sur l'étude de la robustesse des réseaux a permis, par une approche couplant simulation et preuves formelles, de mieux comprendre les raisons qui font que certains réseaux sont moins résistants aux pannes ou aux attaques que d'autres. Les études sur les graphes finis apportent aussi un éclairage nouveau et plus proche de la réalité des phénomènes œuvrant sur ces réseaux. Cette étude se base essentiellement sur les degrés des sommets, mais d'autres propriétés (le clustering, par exemple) ont certainement une influence forte, positive ou négative, sur la robustesse. Il reste donc un travail important à effectuer sur cette problématique.

Enfin, le Chapitre 8 montre l'apport de la simulation et de la modélisation (ici, pour estimer le biais dans les mesures) dans les contextes liés aux grands réseaux d'interactions. Les résultats montrent que les mesures sont non seulement imprécises en général mais également dépendantes des paramètres que l'on veut estimer. Cette approche à base de simulation sur des modèles permet donc d'étudier l'impact des propriétés du graphe sur l'opération de mesure, ou tout autre phénomène, et d'obtenir de nombreux résultats expérimentaux. Ceux-ci sont une première étape pour aborder le problème plus général de la métrologie de graphes.

Ces trois exemples d'applications illustrent bien la diversité des utilisations qui peuvent être faites des études générales sur les grands réseaux d'interactions. En retour, elles font émerger des problématiques nouvelles, comme la métrologie de graphes ou l'analyse et la modélisation des dynamiques de graphes, sur lesquelles existe une attente extrêmement

forte. Elles sont en effet d'une importance cruciale dans des domaines très divers, au sein même de l'informatique, mais aussi en sciences sociales, en sciences du vivant, en linguistique et dans de nombreux autres domaines. Nous ne sommes aujourd'hui qu'aux prémisses de cette mise au jour de nombreuses problématiques, tant théoriques qu'applicatives, issues des grands réseaux d'interactions.

Conclusion et perspectives

Nous avons abordé dans cette thèse deux des grandes problématiques de l'étude des grands réseaux d'interactions : l'analyse et la modélisation, avec un accent plus fort sur cette dernière.

Nous avons ainsi décrit dans une première partie un éventail relativement large de grands réseaux d'interactions rencontrés et étudiés en pratique, ainsi que les principales notions utilisées pour les décrire. Nous avons discuté leur pertinence, et les avons observées sur plusieurs cas pratiques qui nous ont servi d'illustration tout au long de la thèse. Les propriétés statistiques observées ont elle-mêmes été utilisées dans l'ensemble de la thèse. Il est ressorti de cette partie que, sauf cas exceptionnels, les grands réseaux d'interactions rencontrés en pratique ont une distance moyenne faible, une distribution des degrés assimilable à une loi de puissance et une densité globalement faible, mais localement forte (clustering). Ces propriétés, et particulièrement les deux dernières, sont non triviales : elles ne sont pas rencontrées dans les graphes en général. Elles définissent donc une classe de graphes particulière, dont la plupart des grands réseaux d'interactions font partie.

C'est cette classe de graphes que vise à étudier la modélisation, que nous avons abordée en seconde partie de cette thèse et qui a constitué l'essentiel de notre apport. En particulier, il s'agit de produire des graphes artificiels ayant les propriétés des grands réseaux d'interactions rencontrés en pratique. Nous avons dans cette partie exposé l'état de la recherche dans ce domaine qui, bien qu'intensive ces dernières années, ne répond toujours pas de façon satisfaisante à la demande. Nous avons proposé une approche consistant fondamentalement à coder les propriétés des grands réseaux d'interactions par des ensembles de distributions de degrés. Ceci a conduit à l'introduction du modèle biparti aléatoire (et de sa version incrémentale), qui constitue un progrès significatif dans le domaine : il capture les trois principales propriétés des grands réseaux d'interactions, il est suffisamment simple pour se prêter à l'analyse formelle, et il repose sur des observations issues de cas réels. Au delà, nous avons montré que la méthode peut être poussée plus loin avec profit, en introduisant des modèles tripartis et plus généralement multipartis, qui semblent pouvoir capturer très finement de nombreuses propriétés.

Dans la troisième partie, qui peut à la fois être vue comme applicative et méthodologique, nous avons présenté une étude de cas (graphe d'échanges P2P), l'étude formelle du comportement des grands réseaux d'interactions dans un cas précis (résistance aux attaques et aux pannes) et enfin une utilisation des modèles dans un contexte de simulation pour étudier un problème pratique (l'exploration de l'Internet). Ces trois études ont

apporté des résultats en elles-mêmes, mais elles ont surtout permis de mettre en lumière plusieurs points : comment l'analyse de cas particuliers amène à introduire des notions en fait générales (comme la description de la dynamique d'un graphe), comment les modèles peuvent être utilisés pour effectuer des études formelles et comment ils peuvent être utilisés dans un contexte de simulation.

Au delà de ces résultats et de leur apport au domaine des grands réseaux d'interactions, la constatation générale qui ressort de notre travail est que non seulement énormément de travail reste à effectuer pour en saisir tous les aspects, mais qu'en plus ce domaine nouveau ouvre de nombreuses perspectives plus larges. Il nous semble donc essentiel, au terme de ce travail et pour le conclure, de proposer une synthèse des différents aspects qui nous sont apparus comme étant les plus prometteurs. D'autres nous ont probablement échappé, et certains ne sont pas aujourd'hui à portée de vue, mais nous arrivons à un stade de développement du domaine où nous pouvons être confiants dans les perspectives que nous présentons ci-dessous.

Métrologie de graphes

Comme nous l'avons déjà souligné, la plupart des grands réseaux d'interactions qui nous intéressent ne sont pas accessibles directement, mais au contraire ne sont connus qu'au travers d'une opération de mesure. Cette mesure elle-même est complexe et un important travail reste à faire dans cette direction. Au delà des cas particuliers, l'*interprétation* de la mesure est un problème en soi : la vision obtenue est-elle représentative ? Quelles sont les propriétés observées les plus fiables ? Quel est le biais introduit par la mesure ? Comment le corriger ? Peut-on mettre au point des méthodes de mesure dédiées à la mesure d'une caractéristique précise ? etc.

Toutes ces questions, et celles qui s'y rapportent, fondent une nouvelle problématique de recherche sur les grands réseaux d'interactions, celle de la métrologie.

Quelques études ont déjà été menées sur cette problématique [62, 67, 76], dont celle présentée au chapitre 8. À notre connaissance, elles sont toutes focalisées sur le cas de l'Internet exploré avec `traceroute`. Pourtant, les questions soulevées se retrouvent dans des contextes très divers, et y jouent souvent un rôle essentiel. Un important travail de formalisation et d'étude est donc aujourd'hui nécessaire sur cette question.

Analyse

Bien que l'analyse des grands réseaux d'interactions soit la plus ancienne activité du domaine, et que de très nombreuses études de cas aient été menées, beaucoup de travail reste à faire dans cette direction. En particulier, il s'agit bien sûr de déterminer les propriétés communes à la majorité des grands réseaux d'interactions et celles qui les séparent en plusieurs classes.

Mais cette continuation du travail d'analyse déjà effectué, qui peut s'avérer complexe et fastidieuse, n'est pas la seule perspective ; il existe en effet aujourd'hui un très fort

besoin de notions pour l'analyse des réseaux valués (un poids est associé à chaque lien, comme dans les cas de l'Internet avec la bande passante ou des réseaux sociaux avec fréquence des contacts, par exemple), des réseaux hybrides (plusieurs réseaux définis sur un même ensemble de sommets, comme les interactions protéiques ou les réseaux d'amis et de relations professionnelles, par exemple), des réseaux hétérogènes (réseaux bipartis ou multipartis, comme par exemple les réseaux de collaborations ou de cooccurrence), ou les réseaux dynamiques (comme, parmi d'innombrables exemples, le graphe du Web, les graphes P2P, les réseaux sociaux, etc).

Certaines de ces classes de grands réseaux d'interactions sont en fait tellement générales qu'elles sont peut être des façons plus naturelles d'étudier ces objets. Par exemple, la quasi-totalité des grands réseaux d'interactions étudiés évoluent au cours du temps. L'étude de leur dynamique, qui reste aujourd'hui très largement embryonnaire, pourrait s'avérer plus pertinente que leur étude statique. De même, beaucoup de grands réseaux d'interactions peuvent avec bénéfice être considérés comme bipartis ; des paramètres statistiques adaptés à ce contexte seraient donc d'une grande utilité.

Remarquons que nous avons dans cette thèse contribué à ces deux problématiques (analyse statistique de la dynamique des grands réseaux d'interactions et des graphes bipartis) : dans le chapitre 6 nous avons été amenés à introduire des méthodes d'analyse de la dynamique, et dans le chapitre 4 nous avons introduits des notions capturant certaines propriétés des graphes bipartis. Malgré tout, l'essentiel reste à faire dans ces domaines.

Modélisation

Nous l'avons vu, l'état actuel de la modélisation des grands réseaux d'interactions est largement insatisfaisant. Aucun modèle réaliste, hormis le modèle biparti, ne capture à la fois les trois propriétés générales des grands réseaux d'interactions. Toutes les propositions, y compris la nôtre, ont leurs défauts et sont inaptes à capturer finement les propriétés voulues. De plus, un important travail de modélisation reste à faire concernant les *phénomènes* en réseau, comme la diffusion d'informations dans un réseau social, la propagation de virus, l'évolution temporelle du réseau, etc.

Nous sommes toutefois convaincus, au terme de notre étude, que l'approche que nous avons proposée peut être étendue et que divers modèles multipartis, capturant diverses propriétés, peuvent en être dérivés. Cette perspective peut s'avérer très fructueuse, du moins dans l'optique de la modélisation consistant à produire des graphes artificiels *similaires* à des grands réseaux d'interactions donnés. Elle reste toutefois à creuser.

De même, l'utilisation effective du modèle biparti est encore marginale. Il est pourtant clair que les méthodes formelles, comme celles du chapitre 7, et les simulations, comme celles du chapitre 8, s'adaptent directement à ce modèle. Il permettrait donc d'étudier l'impact du clustering dans divers contextes, comme celui de la robustesse des réseaux par exemple.

Algorithmes pour les grands réseaux d'interactions

Nous en avons très peu parlé dans cette thèse, mais la manipulation informatique des grands réseaux d'interactions pose de nombreux problèmes, du fait de la grande taille des graphes en cause (typiquement des millions de sommets). Ceci en interdit la manipulation par les algorithmes classiques. Dans ce contexte, même le calcul des plus courts chemins pose problème.

Toutefois, nous avons vu que les graphes en question ne sont pas quelconques, mais au contraire ont des propriétés statistiques non triviales en commun. Il semble donc naturel de tenir compte de ces propriétés pour la conception d'algorithmes en tirant parti. Quelques travaux ont commencé à explorer cette direction, mais ils restent très embryonnaires [**?**, **?**].

Soulignons que l'analyse et la modélisation jouent dans cette perspective un rôle essentiel : l'analyse permet de connaître les propriétés dont on espère tirer parti, et la modélisation permet d'analyser formellement les algorithmes proposés.

L'algorithmique dédiée aux grands réseaux d'interactions s'inscrit ainsi dans le contexte classique de l'algorithmique de graphes, dans laquelle les graphes purement aléatoires sont souvent utilisés pour évaluer les performances. Les considérations ci-dessus pourraient donc constituer un renouveau significatif pour cette discipline.

Réseaux sociaux de l'Internet

Jusqu'à récemment, et hormis quelques cas très particuliers, l'étude des réseaux sociaux passait par des enquêtes de terrain (questionnaires, démarchage téléphoniques) fastidieuses et fournissant une information largement incomplète et biaisée sur le réseau social. Aujourd'hui toutefois, de nombreux réseaux sociaux se développent sur l'Internet. Ainsi, le graphe du Web ou les échanges P2P peuvent être vus sous cet angle. Mais on peut également citer les graphes d'échanges de courriers électroniques, les réseaux induits par les discussions (en direct ou en différé, comme dans les forums), les réseaux de sites Web commerciaux, et de nombreux autres.

La nature *en-ligne* et *numérique* de ces réseaux, même si ses effets doivent être considérés avec la plus grande prudence, ouvre des perspectives sans précédent pour l'étude des réseaux sociaux. Cet état de fait, associé à l'arrivée à une certaine maturité des méthodes d'étude des grands réseaux d'interactions, permet d'espérer des avancées significatives dans ce domaines dans les prochaines années.

De nombreuses autres perspectives pourraient être énumérées en conclusion de cette thèse. Nous citerons simplement le cas des réseaux biologiques, et notamment ceux des interactions protéiques, qui sont aujourd'hui encore très mal compris malgré leur rôle crucial dans la compréhension du vivant et ses applications en médecine.

Finalement, un constat se dégage : l'étude des grands réseaux d'interactions n'en est qu'à ses débuts et devrait dans les années à venir, de par la quantité de perspectives, la définition de plus en plus claire de problématiques pertinentes, et l'arrivée à maturité de certaines méthodes, se développer en un domaine de recherche très actif. Ce domaine est

à la frontière de plusieurs disciplines : l'informatique, les mathématiques, la physique, les statistiques, les sciences humaines et sociales, les sciences du vivant, et d'autres, de par les objets d'étude, les problèmes posés, et les méthodes proposées pour les résoudre. Il permettra probablement, et commence déjà à s'engager dans cette voie, de jeter des ponts solides entre disciplines et de développer des collaborations fructueuses.

Bibliographie

[1] *Graph Theory and Combinatorics*, chapter The evolution of sparse graphs, pages 35–57. Academic Press, 1984.

[2] J. Abello, P.M. Pardalos, and M.G.C. Resende. On maximum clique problems in very large graphs. In AMS-DIMACS Series on Discrete Mathematics and Theoretical Computer Science, editors, *External Memory Algorithms*, volume 50, 1999.

[3] L. Adamic and B. Huberman. Power-law distribution of the world wide web. *Science*, 287, 2000.

[4] E. Adar and B.A. Huberman. Free riding on gnutella. *First Monday*, September 2000.

[5] W. Aiello, F.R.K. Chung, and L. Lu. A random graph model for massive graphs. In *ACM Symposium on Theory of Computing (STOC)*, pages 171–180, 2000.

[6] R. Albert and A.-L. Barabási. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[7] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics 74, 47*, 2002.

[8] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world wide web. *Nature*, 401:130–131, 1999.

[9] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance in complex networks. *Nature*, 406:378–382, 2000.

[10] D. Aldous and U. Vazirani. Go with the winners algorithms. In *Proc. 35th Symp. Foundations of Computer Sci.*, pages 492–501, 1994.

[11] J.I. Alvarez-Hamelin and N.s Schabanel. An internet graph model based on trade-off optimization. *European Physical Journal B, special issue on "Applications of networks"*, 38(2):231–237, 2004.

[12] K. Anderson. Analysis of the traffic on the gnutella network. 2001.

[13] arXiv.org e Print archive. http://arxiv.org/.

[14] L. Viennot A.T. Gai. Broose : A loose distributed hashtablebased on the de-brujin topology. Technical report, 2004.

[15] A.-L. Barabási, E. Ravasz, and T. Vicsek. Deterministic scale-free networks. *Physica A 299, (3-4)*, pages 559–564, 2001.

[16] Paul Barford, Azer Bestavros, John Byers, and Mark Crovella. On the marginal utility of network topology measurements. In *ACM SIGCOMM Internet Measurement Workshop 2001*, San Francisco, CA, November 2001. ACM SIGCOMM.

[17] L. Barrière, P. Fraigniaud, E. Kranakis, and D. Krizanc. Efficient routing in networks with long range contacts. In *15th International Symposium on Distributed Computing (DISC '01)*, pages 270–284, 2001.

[18] E. Bender and E. Caneld. The asymptotic number of labelled graphs with given degree sequences. *J. Combin. Theory*, Ser. A 24:296–307, 1978.

[19] G. Bianconi and A.L. Barabási. Bose-einstein condensation in complex networks. *Physical review letters*, 86(24):5632–5635, 2001.

[20] V.D. Blondel and P.P. Senellart. Automatic extraction of synonyms in a dictionary. In *SIAM Workshop on Text Mining*, 2002.

[21] M. Boguna, R. Pastor-Satorras, and A. Vespignani. Epidemic spreading in complex networks with degree correlations. In al J.M. Rubi et, editor, *XVIII Sitges Conference "Statistical Mechanics of Complex Networks"*. Springer Verlag, 2003.

[22] B. Bollobás. *Random Graphs.* Academic Press, 1985.

[23] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo. The maximum clique problem. In D.-Z. Du and P. M. Pardalos, editors, *Handbook of Combinatorial Optimization*, volume 4. Kluwer Academic Publishers, Boston, MA, 1999.

[24] S. Brin, R. Motwani, L. Page, and T. Winograd. The pagerank citation ranking: Bringing order to the web.

[25] A.Z. Broder, S.R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *WWW9 / Computer Networks*, 33(1-6):309–320, 2000.

[26] A. Broido and K. Claffy. Topological resilience in ip and as graphs. 2002. `http://www.caida.org/analysis/topology/resilience/`

[27] T. Bu and D. Towsley. On distinguishing between internet power law topology generators. In *INFOCOM*, 2002.

[28] D.S. Callaway, M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85:5468–5471, 2000.

[29] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. The origin of power laws in internet topologies revisited. In *INFOCOM*, 2002.

[30] F. Chung and L. Lu. The diameter of random sparse graphs.

[31] Aaron Clauset and Cristopher Moore. Traceroute sampling makes random graphs appear to have power law degree distributions. cond-mat/0312674.

[32] Source code for the random bipartite graph generator. http://www.liafa.jussieu.fr/~guillaume/programs/.

[33] R. Cohen, D. ben Avraham, and S. Havlin. *Handbook of graphs and networks*, chapter 4: Structural properties of scale free networks. Wiley-VCH, 2002.

[34] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Resilience of the internet to random breakdown. *Phys. Rev. Lett.*, 85:4626–4628, 2000.

[35] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685, 2001.

[36] F. Comellas, G. Fertin, and A. Raspaud. Vertex labeling and routing in recursive clique-trees, a new family of small-world scale-free graphs. In *Sirocco 2003 - The 10th Int. Colloquium on Structural Information and Communication Complexity*, pages 73–87.

[37] L. Dall'Asta, J.I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignan. A statistical approach to the traceroute-like exploration of networks: theory and simulations. cond-mat/0406404.

[38] Self-Organized Networks Database. http://www.nd.edu/~networks/database/index.html.

[39] The Internet Movie Database. http://www.imdb.com/.

[40] D.J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[41] S.N. Dorogovtsev and J.F.F. Mendes. Exactly solvable small-world network. *Euro. phys. Lett.*, 50 (1):1–7, 2000.

[42] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks. *Adv. Phys. 51, 1079-1187*, 2002.

[43] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Generic scale of the "scale-free" growing networks. *Phys. Rev. E*, 63, 2001.

[44] J.A. Dunne, R.J. Williams, and N.D. Martinez. Network structure and robustness of marine food webs. *review at Marine Ecology Progress Series*, 2003.

[45] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66, 2002.

[46] P. Erdös and A. Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[47] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *ICALP*, 2002.

[48] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[49] R. Ferrer and R.V. Solé. The small-world of human language. In *Proceedings of the Royal Society of London*, volume B268, pages 2261–2265, 2001.

[50] Cooperative Association for Internet Data Analysis. http://www.caida.org/.

[51] Cooperative Association for Internet Data Analysis Skitter tool. http://www.caida.org/tools/measurement/skitter/.

[52] P. Fraigniaud and P. Gauron. An overview of the content-addressable network d2b. Brief Announcement at 22nd ACM Symp. on Principles of Distributed Computing (PODC), July 2003.

[53] L.C. Freeman. A set of measures of centrality based upon betweeness. *Sociometry*, 40:35–41, 1977.

[54] T. Friedman, M. Latapy, J. Leguay, and K. Salamatian. What is a route on the internet. preprint.

[55] Internet Maps from Mercator. http://www.isi.edu/div7/scan/mercator/maps.html.

[56] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *IEEE INFOCOM 2000*, pages 1371–1380, Tel Aviv, Israel, March 2000. IEEE.

[57] J.-L. Guillaume and S. Le Blond. Statistical properties of exchanges in p2p systems. In *The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications*, 2004.

[58] J.-L. Guillaume and M. Latapy. The web graph: an overview. In *ALGOTEL'02 (Quatrièmes Rencontres Francophones sur les aspects Algorithmiques des Télécommunications)*, 2002.

[59] J.-L. Guillaume and M. Latapy. Modèles pour les topologies réalistes. In *ALGOTEL'03 (Cinquièmes Rencontres Francophones sur les aspects Algorithmiques des Télécommunications)*, 2003.

[60] J.-L. Guillaume and M. Latapy. Bipartite graphs as models of complex networks. In *Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN)*, 2004.

[61] J.-L. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90(5):215–221, 2004.

[62] J.-L. Guillaume and M. Latapy. Complex network metrology, 2004. preprint - http://www.liafa.jussieu.fr/~latapy/Publis/.

[63] J.-L. Guillaume and M. Latapy. *Mesures de l'Internet*, chapter Topologie d'Internet et du Web : mesure et modélisation, pages 213–226. les canadiens en Europe, 2004.

[64] J.-L. Guillaume and M. Latapy. Relevance of massively distributed explorations of the internet topology: Simulation results (short version). In *ALGOTEL'04 (Sixièmes Rencontres Francophones sur les aspects Algorithmiques des Télécommunications)*, 2004.

[65] J.-L. Guillaume, M. Latapy, and S. Le Blond. Statistical analysis of a p2p query graph based on degrees and their time-evolution. In *6th International Workshop on Distributed Computing (IWDC 2004)*, 2004.

[66] J.-L. Guillaume, M. Latapy, and C. Magnien. Comparison of failures and attacks on random and scale-free networks. In *ALGOTEL'04 (Sixièmes Rencontres Francophones sur les aspects Algorithmiques des Télécommunications)*, 2004.

[67] Y. Hyun, A. Broido, and K. Claffy. Traceroute and BGP AS path incongruities. http://www.caida.org/outreach/papers/2003/ASP/.

[68] Information Sciences Institute. Internet protocol, September 1981. RFC 791.

[69] K. W. Ross J. Liang, R. Kumar. Understanding kazaa. 2004.

[70] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature, 407, 651*, 2000.

[71] C. Jin, Q. Chen, and S. Jamin. Inet: Internet topology generator. Technical Report CSE-TR-443-00, Department of EECS, University of Michigan, 2000.

[72] R. Kannan, P. Tetali, and S. Vempala. Simple markov-chain algorithms for generating bipartite graphs and tournaments. *Random Struct. Alg.*, 14:293–308, 1999.

[73] J.M. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC)*, 2000.

[74] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A.S. Tomkins. The Web as a graph: Measurements, models, and methods. In T. Asano, H. Imai, D. T. Lee, S. Nakano, and T. Tokuyama, editors, *Proc. 5th Annual Int. Conf. Computing and Combinatorics, COCOON*, number 1627. Springer-Verlag, 1999.

[75] J. Kleinfeld. History of the small-world problem (notes), october 2000.

[76] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *IEEE INFOCOM*, 2003.

[77] N. Leibowitz, A. Bergman, R. Ben-Shaul, and A. Shavit. Are file swapping cacheable? characterizing p2p traffic. In *7th International Worshop on Web Content Caching and Distribution (WCW'03)*, 2002.

[78] N. Leibowitz, M. Ripeanu, and A. Wierzbicki. Deconstructing the kazaa network. In *3rd IEEE Workshop on Internet Applications (WIAPP'03)*, 2003.

[79] F. Liljeros, C. Edling, and L.A.N. Amaral. Sexual networks implications for the transmission of sexually transmitted infections. *Microbes and infection*, 5:189–196, 2003.

[80] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature*, (411):907–908, 2001.

[81] L. Lu. The diameter of random massive graphs. In ACM-SIAM, editor, *12th Ann. Symp. on Discrete Algorithms (SODA)*, pages 912–921, 2001.

[82] T. Luczak. Sparse random graphs with a given degree sequence, in Random Graphs, vol. 2. A.M. Frieze, T. £uczak eds. Wiley, New York, 1992. pages. 165-182.

[83] B.M. Maggs, K. Sripanidkulchai, and H. Zhang. Efficient content location using interest-based locality in peer-to-peer systems. In *INFOCOM*, 2003.

[84] D. Magoni and J.-J. Pansiot. Analysis of the autonomous system network topology. *ACM SIGCOMM Computer Communication Review*, 31(3):26 – 37, July 2001.

[85] D. Magoni and J.-J. Pansiot. Influence of network topology on protocol simulation. In *ICN'01 - 1st IEEE International Conference on Networking*, volume Lecture Notes in Computer Science, pages 762–770, July 9-13, 2001.

[86] D. Malkhi, M. Naor, and D. Ratajczak. Viceroy: A scalable and dynamic emulation of the butterfly. In *Proceedings of the 21st ACM Symposium on Principles of Distributed Computing*, 2002.

[87] G.S. Malkin. Traceroute using an ip option. Xylogics, Inc., January 1993. RFC 1393.

[88] E.P. Markatos. Tracing a large-scale peer to peer system : an hour in the life of gnutella. Technical Report 298, 2001.

[89] A. Medina, I. Matta, and J. Byers. On the origin of power laws in internet topologies. In *ACM Computer Communication Review, 30(2)*, april, 2000.

[90] S. Milgram. The small world problem. *Psychology today*, 1:61–67, 1967.

[91] S. Milgram. The small world problem, 1992.

[92] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, pages 161–179, 1995.

[93] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.*, pages 295–305, 1998.

[94] S.D. Monson, N.J. Pullman, and R. Rees. A survey of clique and biclique coverings and factorizations of (0,1)-matrices. *Bull. Inst. Combin. Appl.*, 14:17–86, 1995.

[95] J.M. Montoya and R.V. Sole. Small world patterns in food webs. *Journal of Theoretical Biology*, 2000.

[96] J.W. Moon and L. Moser. On cliques in graphs. *Israel J. Math. 3*, 3:23–28, 1965.

[97] M.E.J. Newman. Scientific collaboration networks: I. Network construction and fundamental results. *Phys. Rev. E*, 64, 2001.

[98] M.E.J. Newman. Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64, 2001.

[99] M.E.J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89, 2002.

[100] M.E.J. Newman. mixing patterns in networks. *Phy. Rev. E*, 67, 2003. cond-mat/0209450.

[101] M.E.J. Newman. Random graphs as models of networks. In Stefan Bornholdt and Heinz Georg Schuster, editors, *Hankbook of Graphs and Networks: From the Genome to the Internet*. Wiley-vch, 2003.

[102] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[103] M.E.J. Newman and D.J. Watts. Renormalization group analysis of the small-world network model. *Phys. Lett. A*, 263:341–346, 1999.

[104] M.E.J. Newman and D.J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60:7332– 7342, 1999. cond-mat/9904419.

[105] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 2001.

[106] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99 (Suppl. 1):2566–2572, 2002.

[107] J. Orlin. Contentment in graph theory: Covering graphs with cliques. *Indigationes Mathematicae*, 80:406–424, 1977.

[108] G. Pandurangan, P. Raghavan, and E. Upfal. Building low-diameter p2p networks. In *IEEE Symposium on Foundations of Computer Science*, pages 492–499, 2001.

[109] S.-T. Park, A. Khrabrov, D.M. Pennock, S. Lawrence, C. Lee Giles, and L.H. Ungar. Static and dynamic analysis of the internet's susceptibility to faults and attacks. In *IEEE Infocom 2003*, San Francisco, CA, April 1–3 2003.

[110] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001.

[111] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2004.

[112] T. Petermann and P. De Los Rios. Exploration of scale-free networks. *To appear in Eur. Phys. J. B*, 2004.

[113] DIMES@home Project. http://www.cs.huji.ac.il/~eproj/available/dimes.

[114] Traceroute@Home project. University of Paris 6, coordinator: Timur friedman.

[115] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content addressable network. In *Proceedings of ACM SIGCOMM 2001*, 2001.

[116] S. Redner. How popular is your paper? an empirical study of the citation distribution. *Eur.Phys.J.*, B(4):131–134, 1998.

[117] M.G.C. Resende. Detecting dense subgraphs in massive graphs. Talk in XVII International Symposium on Mathematical Programming, August 2000.

[118] P. De Los Rios. Exploration bias of complex networks. In *Proceedings of the 7th Conference on Statistical and Computational Physics Granada*, 2002.

[119] M. Ripeanu, I. Foster, and A. Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6(1), 2002.

[120] S. Saroiu, P. Krishna Gummadi, and S.D. Gribble. A measurement study of peer-to-peer file sharing systems. Technical report, Univ. of Washington, Dep. of Comp. Sci., 2002.

[121] S. Saroiu, P. Krishna Gummadi, and S.D. Gribble. A measurement study of peer-to-peer file sharing systems. In *Proceedings of Multimedia Computing and Networking 2002 (MMCN '02)*, San Jose, CA, USA, January 2002.

[122] S. Sen and J. Wang. Analysing peer to peer traffic accross large networks. In *Internet Measurement Workshop (IMW 2002), Marseille, France*, 2002.

[123] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[124] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with rocketfuel. In *Proceedings of ACM/SIGCOMM '02*, August 2002.

[125] K. Sripanidkulchai. The popularity of gnutella queries and its implications on scaling. 2001.

[126] D. Stauffer and A. Aharony. *Introduction to Percolation Theory*. Taylor & Francis, London, 2nd edition, 1994.

[127] I. Stoica, R. Morris, D. Karger, M. Kaashock, and H. Balakrishman. Chord: A scalable peer-to-peer lookup protocol for internet applications. In *ACM SIGCOMM*, pages 149–160, 2001.

[128] Lakshminarayanan Subramanian, Sharad Agarwal, Jennifer Rexford, and Randy H. Katz. Characterizing the internet hierarchy from multiple vantage points. In *Proc. of IEEE INFOCOM 2002, New York, NY*, Jun 2002.

[129] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. On characterizing network hierarchy. Technical Report 03-782, Computer Science Department, University of Southern California, 2001. submitted.

[130] J.R. Tyler, D.M. Wilkinson, and B.A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *Int Conf on Communities and Technologies*, 2003.

[131] Lugdunum url:. http://lugdunum2k.free.fr/kiten.html.

[132] A. Vazquez, R. Pastor-Satorras, and A. Vespignani. Internet topology at the router and autonomous system level. [cond-mat/0206084].

[133] Bible Today New International Version. http://www.tniv.info/bible/.

[134] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[135] B.M. Waxman. Routing of multipoint connections. *IEEE Journal of Selected Areas in Communications*, pages 1617–1622, 1988.

[136] R.J. Williams and N.D. Martinez. Simple rules yield complex food webs. *Nature*, (404):180–183, 2000.

[137] F. Wu, B.A. Huberman, L.A. Adamic, and J. Tyler. Information flow in social groups. cond-mat/0305305, 2003.

[138] E.W. Zegura, K.L. Calvert, and M.J. Donahoo. A quantitative comparison of graph-based models for Internet topology. *IEEE/ACM Transactions on Networking*, 5(6):770–783, 1997.

[139] S. Zhou and R.J. Mondragon. Accurately modeling the internet topology. cs.NI/0402011, 2004.