Habilitation à Diriger des Recherches
Résumé des travaux

Structures of Complex Networks and of their Dynamics

Christophe Crespelle

Ce manuscrit présente une synthèse de mes travaux de recherche dans le domaine des réseaux complexes. Ces activités ont commencé en 2008, un an après l'obtention de mon doctorat en informatique. Ces travaux touchent à des problématiques diverses du domaine : la métrologie de l'Internet, l'analyse des réseaux dynamiques, la modélisation des réseaux statiques et les codages efficaces pour les graphes. Chacune de ces thématiques donne lieu à un chapitre. Une caractéristique essentielle de ce mémoire est qu'il mêle l'utilisation de méthodes empiriques (mesure et simulation) et de méthodes formelles (preuve). C'est une particularité de mon activité de recherche et ce manuscrit est écrit avec l'intention d'illustrer comment ces deux ensembles de méthodes, souvent séparés, peuvent s'enrichir mutuellement.

Le premier chapitre traite de métrologie, qui est la science de la mesure. Du fait qu'ils proviennent de contextes concrets, les réseaux complexes ne peuvent être connus que par une opération de mesure. Dès lors, se pose la question de savoir si le résultat de la mesure est fidèle au réseau réel ou s'il est induit par le procédé de mesure lui-même, auquel cas on parle de *biais*. Une grande controverse a éclaté à ce propos concernant la distribution des degrés de l'Internet. Toutes les mesures effectuées jusqu'à présent ont confirmé que cette distribution est en loi de puissance. Cependant, il a été montré, à la fois empiriquement et formellement, que le résultat de ces mesures ainsi que le procédé utilisé pour les obtenir sont biaisés. Il en ressort que nous n'avons actuellement à disposition aucune estimation fiable de cette distribution, qui est pourtant d'importance primordiale pour la gestion du réseau. Ce premier chapitre présente une méthode de mesure pour estimer rigoureusement la distribution des degrés du coeur de l'Internet. Deux implémentations de la méthode sont exposées : une dédiée à la topologie logique, au niveau IP, et une dédiée à la topologie physique.

Le deuxième chapitre concerne l'analyse des réseaux dynamiques. Il contient une étude de cas et deux travaux méthodologiques. L'étude de cas porte sur le réseau dynamique des contacts entre individus dans un hôpital, enregistrés avec une précision de 30s pendant une durée de six mois sur toute la population de l'hôpital, patients et personnels. Cette mesure a été effectuée dans le but de mieux comprendre la diffusion des souches de staphylocoque en milieu hospitalier et l'apparition de souches résistantes aux antibiotiques. Nous présentons une analyse de la structure des contacts, dans sa dimension topologique et sa dimension temporelle, selon un découpage prédéfini de l'hôpital en services et en catégories socio-professionnelles. Le premier des deux travaux méthodologiques concerne la structure des changements de la topologie d'un réseau dynamique au cours du temps. Le réseau est décrit comme une série de graphes et on s'intéresse aux différences entre deux graphes consécutifs de la série. La question est de savoir si les changements de topologie d'un instant à l'autre sont répartis dans l'ensemble du réseau ou s'ils sont au contraire concentrés autour d'une partie restreinte des noeuds. Le deuxième travail méthodologique concerne la description des réseaux dynamiques par une série de graphes. Nombre de réseaux dynamiques sont naturellement des *flots de liens*, c'est-à-dire un ensemble de triplets (u, v, t) signifiant qu'il y a un lien entre les noeuds u et v au temps t. La plupart des travaux sur ces réseaux commencent par les transformer en séries de graphes sur lesquelles sont menées toutes les analyses. Le procédé utilisé pour ce faire est l'*agrégation*. Il consiste à former le graphe des liens ayant existé dans une fenêtre de temps choisie. Malheureusement, cette transformation induit une perte d'information qui altère le flot

de liens original. Cette altération est d'autant plus grande que la largeur de la fenêtre utilisée est grande. Nous proposons une méthode pour déterminer quelle elle est la largeur maximale de la fenêtre d'agrégation qui garantisse que la série de graphes formée conserve, pour l'essentiel, les propriétés du flot de liens original.

Le troisième chapitre est dédié à la modélisation des réseaux statiques, c'est-à-dire la génération de topologies synthétiques réalistes. Le but est de générer aléatoirement des graphes qui ont les propriétés connues des réseaux issus du monde réel, en particulier : une faible densité globale, des distances courtes, une distribution des degrés hétérogène et une densité locale élevée (appréciée par le coefficient de *clustering*). La méthode connue sous le nom de *modèle de configuration* permet de générer des graphes présentant les trois premières propriétés. Depuis, le domaine bute sur la difficulté à générer aléatoirement des graphes ayant une forte densité locale. Nous explorons deux nouvelles voies pour lever cette difficulté. La première consiste à générer des graphes non par leurs arêtes mais par leurs cliques, en respectant la structure de chevauchement de ces dernières. Cela soulève le problème de la terminaison d'un procédé itératif de factorisation de bicliques dans un graphe multiparti, pour lequel nous élaborons deux solutions. La deuxième voie de modélisation que nous proposons consiste à approximer les réseaux réels par des graphes fortement structurés, c'est-à-dire définis par une propriété mathématique. Il s'agit de représenter un réseau complexe par une paire formée d'un graphe fortement structuré et de l'ensemble des différences entre ce graphe et le réseau original, ces deux parties de la topologie étant ensuite générées indépendamment. Dans le but d'obtenir de telles représentations pour des réseaux issus du monde réel, nous développons ou améliorons plusieurs algorithmes d'édition et de complétion minimale de graphes, notamment pour les classes des graphes d'intervalles, des graphes de permutation et des cographes. L'approche de modélisation proposée est testée en utilisant les résultats fournis par l'algorithme d'édition minimale développé pour les cographes.

Le but du quatrième et dernier chapitre est de développer des codages efficaces pour les graphes, qui soient à la fois compacts en espace et qui ne pénalisent pas le temps d'exécution des requêtes faites par les algorithmes. Cela est primordial en pratique pour stocker en mémoire limitée les immenses jeux de données constitués par les réseaux complexes sans allonger les temps de traitement de ces données. Nous nous intéressons à garantir un temps d'exécution optimal pour la requête de voisinage (lister les voisins d'un sommet donné), qui est probablement la requête la plus utilisée par les algorithmes et qui est également utilisée pour l'exploration et la visualisation. Deux paramètres de codage, la *contiguïté* et la *linéarité*, sont étudiés. Ils sont basés sur un ou plusieurs ordres linéaires des sommets du graphe considéré, dont le but est de grouper autant que possible les voisinages des sommets. La contiguïté utilise un seul ordre dans lequel les voisinages des sommets peuvent être segmentés en plusieurs intervalles. La linéarité, que nous introduisons, utilise plusieurs ordres dans lequel chaque sommet retient un unique intervalle formé par certains de ses voisins. Il découle de leurs définitions que la linéarité est toujours au plus égale à la contiguïté. Nous montrons qu'il existe des familles de graphes pour lesquelles la linéarité est asymptotiquement négligeable devant la contiguïté, ce qui implique que le codage par linéarité est strictement plus puissant que celui par contiguïté. Au passage, nous fournissons des bornes supérieures et inférieures atteintes sur la contiguïté et la linéarité dans le pire des cas des cographes à n sommets.

Le manuscrit se termine en décrivant deux directions de recherche ouvertes que je crois particulièrement importantes pour le domaine des réseaux complexes dans les prochaines années. La première est le développement d'une théorie des *flots de liens*, comme un nouvel objet mathématique, pour étudier les réseaux dynamiques sans passer par les graphes, ce qui est actuellement à l'origine de nombreux blocages. La deuxième direction concerne l'approximation des réseaux complexes par des graphes fortement structurés et l'avènement d'une théorie algorithmique des *graphes presque structurés*.