Partitionnement des Liens d'un Graphe : Critères et Mesures

Noé Gaumont[1] et François Queyroi[1]

[1] *CNRS & LIP6, Université Pierre et Marie Curie – Sorbonne Universités, Paris, France.*

La recherche de communautés chevauchantes est un enjeu important pour l'analyse des réseaux complexes. Une piste souvent envisagée est la recherche d'un partitionnement des arêtes du graphe. L'évaluation de cette décomposition tient cependant rarement compte du fait que les communautés recherchées correspondent à des groupes d'arêtes. Nous discutons dans ce papier l'utilisation de nouveaux critères pouvant répondre à ce problème. L'idée principale s'exprime simplement : un groupe d'arêtes doit contenir relativement peu de sommets tandis que le voisinage du groupe doit contenir relativement plus de sommets. La mesure dérivée de ce concept peut être optimisée par un algorithme glouton. Nous présentons des premiers résultats à travers une analyse de la mesure et des tests empiriques.

Keywords: réseaux complexes, communautés chevauchantes, mesure de qualité

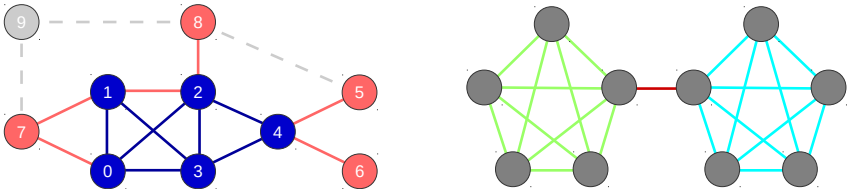
1 Introduction

La structure en communautés d'un réseau est généralement capturée par le biais d'une partition des sommets. Toutefois, la recherche d'une couverture du graphe formée par une partition de ses arêtes peut être pertinente [For10]. Ce dernier cas permet d'incorporer le fait qu'un individu peut appartenir à plusieurs groupes ce qui est un intérêt majeur dans de nombreuses applications. Les algorithmes de partitionnement cherchent généralement à maximiser une *mesure de qualité* qui capture quantitativement un critère définissant un ensemble de *communautés*. Les solutions de bonne "qualité" sont généralement des compromis entre un nombre de connexions importantes au sein d'un groupe et un nombre faible de connexions entre les groupes.

Peu de mesures ont été proposées pour la partitionnement de liens à l'exception de [ABL10, EL09]. Une problématique majeure est que la transcription directe des critères évoqués ci-dessus est, dans ce cadre, inadaptée. En effet, le chevauchement entre différentes communautés va mécaniquement accroître la connectivité des communautés entre elles. Nous suggérons d'adopter des critères différents pour évaluer la qualité d'un groupe de liens. L'approche nouvelle que nous proposons peut être illustrée ainsi : supposons que nous interceptons un ensemble de messages entre différentes personnes. On voudrait regrouper ces messages dans le cas où ils concernent un nombre relativement faible de personnes. De même, les autres messages impliquant ces personnes doivent être destinés à un nombre relativement important de personnes différentes dans le reste du réseau. Nous proposons une mesure basée sur ces critères. Elle dépend du nombre moyen de sommets obtenus en tirant aléatoirement des arêtes dans un graphe avec la même distribution de degré. Nous proposons un algorithme agglomératif pour l'optimiser et analysons les résultats obtenus sur des exemples simples.

2 Définition de la Qualité d'une Partition de Liens

Nous présentons dans cette section une formulation quantitative des idées développées en introduction de ce papier. Pour un groupe de lien donné, nous nous intéressons donc au nombre de sommets incidents à ce groupe. Une première possibilité est de comparer le nombre de sommets incidents à un groupe de liens au nombre maximum de sommets incidents à un groupe ayant le même nombre de liens. Cette stratégie est similaire à celle utilisée dans [ABL10] : la qualité d'un groupe est la densité du sous-graphe correspondant. Cette approche souffre de plusieurs défauts, notamment, de n'être définie que pour des graphes simples.



(a) (b)

FIGURE 1: (a) Exemple d'un groupe d'arêtes L (en bleu). Les arêtes rouges sont les arêtes externes L_{out} connectant les sommets internes V_{in} (en bleu) aux sommets externes V_{out} (en rouge) (b) Partition en trois groupes d'un graphe composé de deux cliques reliées par une arête.

Une deuxième approche consiste à comparer les quantités observées à celles espérées dans un modèle aléatoire nul où il n'existe pas *a priori* de structure en communautés. Le modèle le plus utilisé, notamment dans la mesure de *modularité* [NG04], est le *"configuration model"* qui correspond à un graphe aléatoire avec la même distribution des degrés. Dans ce cas, les sommets conservent le même nombre de voisins mais ces derniers sont tirés uniformément. C'est cette deuxième approche que nous allons utiliser mettre ici. La mesure de modularité repose sur la différence entre le nombre d'arêtes observé et espéré dans un groupe de sommets. Nous renversons donc ici ce paradigme en évaluant le nombre de sommets observés et espérés incidents à un groupe d'arêtes.

Les notations utilisées sont les suivantes. Soit $G = (V,E)$ un graphe non-orienté avec V l'ensemble des sommets et $E \subseteq V \times V$ l'ensemble des arêtes. Le degré d'un sommet u de G est noté $d_G(u)$. Une partition des arêtes en k groupes est notée $\mathcal{L} = (L_1, L_2, \ldots, L_k)$. Pour un groupe d'arêtes $L \in \mathcal{L}$, on pose $V_{in} = \{u \in V, \exists (u,v) \in L\}$ l'ensemble des sommets incident au groupe L, $L_{out} = \{(u,v) \in E, u \in V_{in} \wedge v \notin V_{in}\}$ l'ensemble des arêtes incidentes à un sommet de V_{in} enfin $V_{out} = \{u \in (V \setminus V_{in}), (u,v) \in L_{out} \wedge v \in V_{in}\}$ représente les sommets externes au groupe L (voir Figure 1(a)).

2.1 Tirage aléatoire d'arêtes dans le modèle nul

Nous cherchons à définir la quantité espérée de sommets différents, notée $\mu_G(m)$, pour un ensemble d'arêtes de taille m pris aléatoirement et sans remise dans un graphe aléatoire avec la même distribution de degrés. Soit B_u la variable aléatoire correspondant au nombre de fois où le sommet u est tiré. Cette variable suit une loi hypergéométrique $B_u \sim \mathcal{G}(2|E|, d_G(u), 2m)$. On a ainsi :

$$\mu_G(m) = E\left(\sum_{u \in V} \mathbb{1}_{B_u > 0}\right) = \sum_{u \in V} P(B_u > 0) = \sum_{u \in V} 1 - \frac{\binom{2|E| - d_G(u)}{2m}}{\binom{2|E|}{2m}} \tag{1}$$

Voici quelques propriétés de la fonction $\mu_G(m)$:

— La quantité $\mu_G(m)$ dépend essentiellement de la séquence de degrés $\{d_G(v)\}_{v \in V}$. On peut montrer que cette fonction est Schur-concave, ainsi plus les degrés sont uniformément répartis plus il sera surprenant d'observer un groupe d'arêtes correspondant à peu de sommets.
— Si $m = |E|$, alors le nombre de sommets attendus est bien $|V|$.
— On a $\mu_G(1) \leq 2$, en effet le modèle nul n'interdit pas la présence de boucles.

2.2 Formulation de la mesure

À partir de μ_G défini précédemment, on peut définir la qualité d'un groupe $L \subseteq E$ notée $Q(L)$ de la façon suivante :

$$Q(L) = \frac{\mu_G(m_{in}) - |V_{in}|}{\mu_G(m_{in})} - \frac{\mu_{G \setminus L}(\frac{m_{out}}{2}) - |V_{out}|}{\mu_{G \setminus L}(\frac{m_{out}}{2})} \tag{2}$$

avec $m_{in} = |L|, m_{out} = |L_{out}|$ et $G \setminus L$ correspond au graphe $G' = (V, E \setminus L)$. La partie gauche (nommée *qualité interne*) capture l'éloignement entre les quantités observée et théorique de $|V_{in}|$. La partie droite (nommée *qualité externe*) fournit le même type d'information concernant $|V_{out}|$. Toutefois seulement la moitié du

nombre d'arêtes externes est tirée dans ce cas. En effet, on ne tient pas compte d'une des extrémités de l'arête : celle correspondant à un nœud externe. Nous discutons dans la section 4 l'impact de ces deux composantes.

La qualité globale d'une partition des arêtes $\mathcal{L}$ d'un graphe G correspond alors à la moyenne pondérée des qualités marginales :

$$Q_G(\mathcal{L}) = \frac{\sum_{L\in\mathcal{L}} |L| Q(L)}{|E|} \tag{3}$$

Nous détaillons certaines propriétés des formules (2) et (3) découlant des propriétés de μ_G :

— En s'intéressant aux sommets externes V_{out}, on pénalise la présence de sommets externes fortement connectés avec les sommets incidents à L.

— Ainsi la qualité d'une arête isolée dépend du nombre de triangles dans laquelle elle se trouve. Une arête séparant deux groupes de sommets disjoints peut avoir une qualité positive (voir Figure 1(b)).

— La qualité de $\mathcal{L} = \{E\}$ est nulle.

3 Algorithme d'Optimisation Glouton

La mesure définie en (3) peut être optimisée par une heuristique agglomérative détaillée ici. Dans cet algorithme, deux types de modifications sont envisagées, à savoir la fusion de deux groupes d'arêtes et le changement d'une arête d'un groupe à un autre. Ces opérations sont associées à un gain correspondant à la différence pondérée des nouvelles et des anciennes communautés. Le gain de fusion de deux groupes disjoints L_A et L_B est :

$$gain(L_A, L_B) = |L_A \cup L_B| Q(L_A \cup L_B) - (|L_A| Q(L_A) + |L_B| Q(L_B)) \tag{4}$$

Le gain résultant du transfert d'une arête l de L_A à L_B est :

$$gain(l, L_A, L_B) = |L_A \setminus \{l\}| Q(L_A \setminus \{l\}) + |L_B \cup \{l\}| Q(L_B \cup \{l\}) - (|L_A| Q(L_A) + |L_B| Q(L_B)) \tag{5}$$

Le déroulement de l'algorithme est le suivant : au début chaque arête correspond à un groupe, puis les meilleures opérations sont appliquées itérativement. L'algorithme se termine lorsqu'il n'existe plus aucun mouvement améliorant la qualité. Lors de la recherche du meilleur mouvement, il n'est pas nécessaire d'évaluer toutes les opérations possibles : on peut facilement montrer qu'une amélioration n'est possible que si les communautés impliquées sont voisines. Notons que l'ordre d'évaluation des mouvements est arbitraire et impacte fortement les résultats de la méthode.

4 Analyse des Résultats

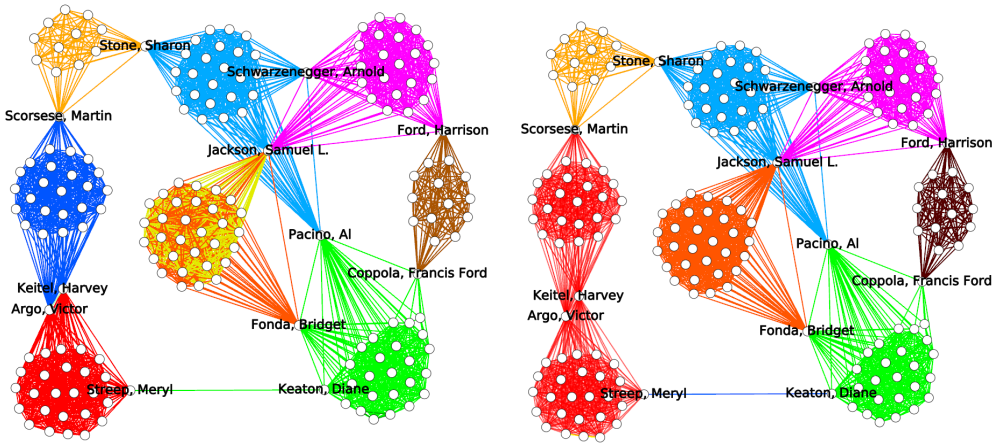
Nous discutons maintenant de résultats obtenus sur des exemples simples : dans un premier temps sur des cas théoriques et dans un second temps sur un graphe issu de la base de données IMDB [†]. En particulier, nous présentons l'influence des qualités internes et externes sur les résultats de l'algorithme présenté dans la section 3.

Le cas où G est constitué d'une unique clique illustre l'intérêt d'utiliser la qualité externe (voir Eq. (2)). En effet, la partition constituée d'un seul groupe a une qualité interne de 0. Toutefois il est possible d'obtenir une partition ayant une qualité interne supérieur à 0. En utilisant les parties internes et externes de la fonction de qualité, la partition regroupant toute la clique obtient toujours une qualité nulle. Cependant il n'est plus possible de trouver un autre partitionnement ayant une meilleure qualité. Pour l'exemple visible sur la figure 1(b), la partition maximisant la mesure est obtenue en séparant les deux cliques et le lien les reliant. Cela illustre le fait qu'une arête seule peut avoir une contribution positive si elle sépare deux groupes disjoints.

Nous nous intéressons maintenant aux résultats obtenus sur un réseau composé de cliques se chevauchant peu (voir la Fig. 2) issu de IMDB. Cet exemple n'est pas trivial pour autant. En effet, les algorithmes de [ABL10] et de [EL09] [‡] forment des partitions peu significatives (*i.e.* qui sont éloignés d'un découpage en

†. `http://www.imdb.com/`

‡. Sur les trois méthodes proposés par les auteurs, deux retournent des situations peu significatives



(a) Avec qualité interne (b) Avec qualité interne + externe bornée

FIGURE 2: Projection d'un sous-graphe biparti personnes-films issu de la base de données IMDB.

cliques). En particulier, certains groupes détectés par [ABL10] correspondent à des étoiles dont la racine est un sommet présent dans plusieurs cliques.

La Fig. 2(a) présente la partition obtenue avec la fonction de qualité interne. On remarque que les groupes denses sont bien retrouvés avec également les liens les reliant. Toutefois certaines cliques sont séparées en deux groupes (voir les groupes centraux jaunes et oranges). L'utilisation de la mesure en Eq. (2) mène dans ce cas à une agrégation excessive des groupes (on obtient alors trop peu de groupes au vu du nombre de cliques présentes dans le graphe). Il apparaît qu'une solution à ce problème consiste à borner la qualité externe à 0, ce qui revient à ne pénaliser que les situations où le nombre de sommets externes est relativement faible et ne pas différencier les cas où les sommets externes sont relativement nombreux. La Fig. 2(b) donne la partition obtenue en utilisant la qualité interne et cette qualité externe *bornée*. Le groupe qui n'avait pas été capturé est maintenant bien reconnu. Cependant, d'autres fusions inappropriées apparaissent, notamment dans les cas où l'intersection entre les cliques comprend plus d'un sommet. Dans tous les cas, les différentes qualités permettent d'isoler l'arête (Meryl Streep,Diane Keaton) connectant deux cliques, cela n'apparaît pas en utilisant les autres méthodes de la littérature.

5 Conclusion et Perspectives

Dans ce papier, nous introduisons de nouveaux critères pour l'évaluation des partitions de liens tenant compte du chevauchement des sommets dans la connectivité externe d'un groupe. À partir de ces critères, nous définissons une mesure de qualité basée sur le nombre de sommets induits attendus par un ensemble de liens. Nous proposons un algorithme agglomératif d'optimisation glouton reposant sur deux types de modifications. Les premiers résultats obtenus sont prometteurs mais les analyses montrent que certaines situations sont encore mal évaluées. Une perspective intéressante est, à terme, d'étendre les concepts présentés ici à l'analyse de graphes dynamiques correspondants à des flots de liens.

Références

[ABL10] Y.-Y. Ahn, J. P. Bagrow, and S Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307) :761–764, 2010.

[EL09] T.S. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1) :016105, 2009.

[For10] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5) :75–174, 2010.

[NG04] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physics Reviews E*, 69(026113), 2004.