Estimation de la densité d'arcs dans les graphes de grande taille: une alternative à la détection de clusters

Marc Boullé *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. La recherche de structures dans les graphes est un sujet étudié depuis longtemps, qui a bénéficié d'un regain d'intérêt avec la mise à disposition de graphes de grande taille sur le web, tels les réseaux sociaux. De nombreuses méthodes de recherche de clusters "naturels" dans les graphes ont été proposées, fondées notamment sur la modularité de Newman. On introduit dans cet article une nouvelle façon de résumer la structure des graphes de grande taille, en utilisant des estimateurs de densité des arcs exploitant des modèles en grille, basés sur un co-partitionnent des noeuds source et cible des arcs. Les structures identifiées par cette méthode vont au delà de la "classique" détection de clusters dans les graphes, et permettent d'estimer asymptotiquement la densité des arcs. Les expérimentations confirment le potentiel de l'approche, qui permet d'identifier des structures fortement informatives dans les graphes, sans faire l'hypothèse d'une décomposition en clusters denses.

1 Introduction

Le partitionnement de graphe est un sujet étudié depuis longtemps dans le domaine de la recherche opérationnelle. Une des plus ancienne approche est celle de la coupe minimale, dans laquelle les noeuds d'un graphe sont partitionnés en un nombre prédéterminé de clusters, habituellement de tailles approximativement égales, de telle façon que le nombre d'arcs inter-clusters soit minimisé. Ce problème combinatoire est intéressant dans de nombreuses applications, telles que le partitionnement de graphe de télécommunications, la conception de circuits VLSI (very large-scale integration) ou la distribution des traitements en calculs parallèles de façon à minimiser les échanges entre processeurs. Ce problème étant NP-dur (Garey et Johnson, 1979), de nombreuses heuristiques ont été proposées dans la littérature. Par exemple, l'algorithme de (Kernighan et Lin, 1970) est souvent utilisé pour améliorer localement des bi-partitionnements. De nombreuses méta-heuristiques ont également été utilisées pour le partitionnement de graphe, comme le recuit simulé (Johnson et al., 1989), les algorithmes génétiques ou la recherche tabou (Battiti et Bertossi, 1999). L'approche multi-niveau (Hendrickson et Leland, 1995) est particulièrement adaptée aux très grands graphes avec des contraintes fortes sur le temps de calcul. Ces familles d'heuristique représentent un large éventail d'options pour le compromis entre temps de calcul et qualité de la solution.

Avec la disponibilité de graphes de grande taille, tels le graphe du web, les réseaux sociaux, les réseaux de co-auteurs pour les publications scientifiques (Albert et Barabási, 2002), le problème de partitionnement de graphe a bénéficié d'un regain d'intérêt, spécialement pour la découverte automatique de structure de communauté dans les graphes de grande taille. Alors que la recherche classique de partition avec clusters de tailles similaires répond bien aux besoins applicatifs à l'origine de ces méthodes, on s'oriente désormais vers la recherche de de clusters "naturels" dont le nombre et la taille sont censés représenter les caractéristiques intrinsèques d'un graphe.

De nombreuse méthodes ont été proposées pour résoudre ce problème, basées sur des approches hiérarchiques, divisives, spectrales, utilisant des marches aléatoires (voir (Schaeffer, 2007) pour une étude approfondie). Le critère de modularité proposé par (Newman et Girvan, 2003) est largement utilisé dans la littérature pour l'évaluation d'un clustering de graphe, indépendamment du nombre de clusters. Ce critère est également exploité comme fonction objectif à optimiser dans plusieurs algorithmes de clustering (Clauset et al., 2004; Danon et al., 2005; Blondel et al., 2008). Ce critère vise à obtenir des clusters denses, pour lesquels la densité d'arcs intra-cluster est supérieure à la densité attendue dans le cas d'arc distribués aléatoirement, en respectant la même distribution des degrés (nombres d'arc adjacents) des noeuds.

Dans cet article, nous présentons une nouvelle façon d'analyser et de résumer la structure des graphes de grande taille, basée sur une estimation de la densité des arcs constante par morceaux. Notre approche est apparentée aux méthodes de modélisation stochastique par blocs des graphes étudiées depuis longtemps dans le domaine de la sociométrie (Holland et al., 1983; Wasserman et Anderson, 1987; Copic et al., 2009), et les étend en étant entièrement non paramétrique, le nombre de blocs étant un paramètre libre, avec des algorithmes d'optimisation performants pouvant traiter des graphes de plusieurs millions d'arcs. Notre méthode repose sur l'application des modèles en grille (Boullé, 2010) aux graphes, où chaque arc est considéré comme un individu statistique avec deux variables, les noeuds source et cible de l'arc. L'objectif est de rechercher une corrélation entre ces deux variables au moyen d'un modèle en grille bivarié non supervisé, qui dans ce cas s'interprète comme un coclustering des noeuds source et cible du graphe. Les cellules résultant du produit cartésien des deux clusterings résument la densité des arcs du graphe. Le meilleur modèle de corrélation est sélectionné au moyen de l'approche MODL (Minimum Optimized Description Length) (Boullé, 2006), et optimisé en utilisant des heuristiques combinatoires de complexité super-linéaire par rapport au nombre d'arcs.

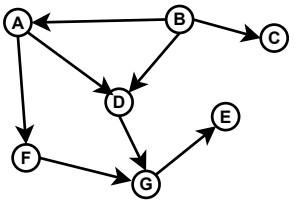
Le reste de l'article est organisé de la façon suivante. La partie 2 rapelle l'approche MODL pour les modèles en grille et l'applique à l'estimation de la densité d'arcs dans les graphes. Les apports de l'approche sont illustrés dans la partie 3, puis l'approche est évaluée sur des graphes réels dans la partie 4. Enfin, la partie 5 conclut cet article.

2 L'Approche MODL pour l'Estimation de Densité des Arcs

Les modèles en grille (Boullé, 2010) ont été introduits pour la phase de préparation du processus de fouille des données (Chapman et al., 2000), qui est une phase clé, à la fois coûteuse en temps d'analyse et critique pour la qualité des résultats. Ces modèles permettent d'estimer de façon automatique, rapide et fiable la densité conditionnelle dans le cas supervisé et la densité jointe dans le cas non supervisé. Les modèles en grille sont basés sur le partitionnement de

chaque variable numérique ou catégorielle, en intervalles ou en groupes de valeurs. Le produit cartésien de ces partitions univariées forme une partition multivariée de l'espace de représentation sous la forme d'un ensemble de cellules. Cette partition multivariée, dénommée grille, constitue un estimateur non paramétrique de densité constant par morceau. La meilleure grille est recherchée au moyen d'une approche Bayesienne de la sélection de modèles et en utilisant des heuristiques combinatoires efficaces.

Comme le montre la figure 1, un multigraphe orienté peut être représenté sous forme tabulaire avec une ligne par individu, les arcs, décrits par deux variables, noeuds source et cible. L'utilisation de modèles en grille bivariés non supervisés permet alors d'estimer la densité jointe entre ces deux variables, c'est à dire la densité des arcs dans le graphe. Nous rappelons ci dessous l'approche MODL permettant de sélectionner le meilleur modèle d'estimation de densité, en la reformulant dans le contexte des graphes.



Source	Target
A	D
A	F
B	A
B	C
B	D
D	G
F	G
G	E

FIG. 1 – *Graphe orienté et sa représentaion tabulaire.*

L'objectif est de décrire conjointement les noeuds source et cible, ce qui revient à décrire les arcs du graphe. On se propose ici de résumer la position des arcs à un niveau macroscopique en introduisant des clusters de noeuds source et des clusters de noeuds cible, et en mémorisant le nombre d'arcs entre chaque cocluster, c'est à dire par paire (source, cible) de clusters de noeuds[1]. Un tel modèle de clustering du graphe peut être considéré comme un estimateur de densité des arcs, constant par cocluster. Le modèle le plus grossier comprend un seul cluster de noeuds contenant tous les arcs, alors que le modèle le plus fin contient un noeud par cluster avec le nombre d'arcs exact par paire de noeuds. Les modèles grossiers sont plus robustes, et les modèles fins plus précis. L'enjeu est trouver un compromis entre robustesse et précision de l'estimation de densité des arcs, sur la base du niveau de grain du clustering.

Cette famille de modèles d'estimation de densité des arcs basée sur un partitionnement joint des noeuds sources et cible est formalisée dans la définition 1.

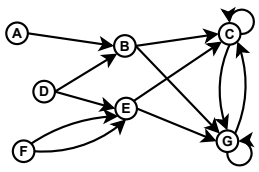
Définition 1 *Un modèle d'estimation de densité des arcs est défini par :*
- *un nombre de clusters de noeuds source et cible,*
- *la répartition des noeuds source (resp. cible) dans les clusters source (resp. cible),*
- *la distribution des arcs du graphe sur les coclusters de noeuds,*
- *pour chaque cluster de noeuds source (resp. cible), la distribution des arcs de ce cluster sur les noeuds de ce cluster.*

Notation.

[1]Par commodité, on utilise le terme cluster pour désigner tout sous-ensemble de noeuds, ayant une densité d'arcs quelconque, pas nécessairement plus importante qu'en moyenne.

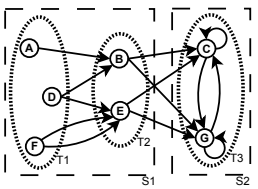
- $G = (V, E)$: graphe avec ensemble de noeuds V et d'arcs E
- S, T : ensembles des noeuds source et cible
- $n = |V|, n_S = |S|, n_T = |T|$: nombres de noeuds, de noeuds source et cible
- $m = |E|$: nombre d'arcs
- k_S, k_T : nombre de clusters de noeuds source et cible
- $k_E = k_S k_T$: nombre de coclusters
- $k_S(i), k_T(j)$: index du cluster contenant le noeud source i (resp. noeud cible j)
- n_i^S, n_j^T : nombre de noeuds du cluster source i (resp. cluster cible j)
- $m_{i.}, m_{.j}$: nombre d'arcs du noeud source i (resp. noeuds cible j), i.e. degré sortant du noeud i (resp. degré entrant du noeud j)
- $m_{i.}^S, m_{.j}^T$: nombre d'arcs sortant du cluster source i (resp. entrant dans le cluster cible j)
- m_{ij} : nombre d'arcs pour la paire de noeuds (i, j)
- m_{ij}^{ST} : nombre d'arcs pour la paire de clusters (i, j)

Ces notations sont illustrées sur la figure 2, où un multigraphe orienté est représenté avec sa matrice d'adjacence. Une version partitionnée du graphe est présentée sur la figure 3, qui équivaut à un coclustering de sa matrice d'adjacence.



	A	B	C	D	E	F	G	∑
A	0	1	0	0	0	0	0	1
B	0	0	1	0	0	0	1	2
C	0	0	1	0	0	0	1	2
D	0	1	0	0	1	0	0	2
E	0	0	1	0	0	0	1	2
F	0	0	0	0	2	0	0	2
G	0	0	1	0	0	0	1	2
∑	0	2	4	0	3	0	4	13

FIG. 2 – *Multigraphe orienté représenté avec sa matrice d'adjacence. Les nombre m_{ij} dans la matrice d'adjacence sont les nombre d'arcs par paire de noeuds (par exemple, deux arcs de F à E). Les totaux $m_{i.}$ sur la colonne de droite sont des degrés sortants des noeuds, et les totaux $m_{.j}$ sur la ligne du bas sont les degrés entrants des noeuds. Le nombre total d'arcs se trouve dans le coin bas droite de la matrice d'adjacence.*



	A	D	F	B	E	C	G	∑
A	0	0	0	1	0	0	0	1
D	0	0	0	1	1	0	0	2
F	0	0	0	0	2	0	0	2
B	0	0	0	0	0	1	1	2
E	0	0	0	0	0	1	1	2
C	0	0	0	0	0	1	1	2
G	0	0	0	0	0	1	1	2
∑	0	0	0	2	3	4	4	13

	T1	T2	T3	∑
S1	0	5	0	5
S2	0	0	8	8
∑	0	5	8	13

FIG. 3 – *Multigraphe orienté avec deux clusters source et trois clusters cible. La matrice d'adjacence du graphe (réorganisée par clusters) est représentée au centre, et celle du graphe partitionné sur la droite. Les nombres m_{ij}^{ST} de la matrice d'adjacence partitionnée sont les nombres d'arcs par cocluster (par exemple, 5 arcs de S1 à T2).*

On suppose que les nombres d'arcs m et de noeuds source et cible n_S et n_T sont connus à l'avance et on cherche à modéliser la distribution jointe des m arcs sur la base de ces deux ensembles de noeuds. Ce type de modélisation est suffisamment général pour prendre en compte les graphes orientés, les graphe bipartites (où les noeuds source et cible appartiennent à deux ensembles disjoints) et les graphes non orientés, pour lesquels on considère que chaque arc apparaît en deux exemplaires, pour chaque sens de sa paire de noeuds.

La famille de modèles introduite en définition 1 est entièrement définie par les paramètres de spécification de la partition des noeuds en clusters

$$k_S, k_T, \{k_S(i)\}_{1 \leq i \leq n_S}, \{k_T(j)\}_{1 \leq j \leq n_T},$$

par les paramètres de la distribution multinomiale des arcs sur les coclusters

$$\{m_{ij}^{ST}\}_{1 \leq i \leq k_S, 1 \leq j \leq k_T},$$

et par les paramètres de la distribution multinomiale des noeuds sortant de chaque cluster source (resp. entrant dans chaque cluster cible) sur les noeuds du cluster

$$\{m_{i.}\}_{1 \leq i \leq n_S}, \{m_{.j}\}_{1 \leq j \leq n_T}.$$

Les nombres de noeuds par cluster n_i^S and n_j^T sont déduits de la spécification des partitions des noeuds en clusters : ils ne font pas partie des paramètres de modélisation. De façon similaire, les nombres d'arcs sortant et entrant de chaque cluster peuvent se calculer en additionnant les effectifs de chaque cocluster, selon $m_{i.}^S = \sum_{j=1}^{k_T} m_{ij}^{ST}$ et $m_{.j}^T = \sum_{i=1}^{k_S} m_{ij}^{ST}$.

Pour sélectionner le meilleur modèle, on applique une approche Bayésienne, en utilisant la distribution a priori des paramètres de modélisation de la définition 2.

Définition 2 *La distribution a priori des paramètres d'un modèle d'estimation de densité des arcs est choisie en exploitant la hiérarchie du paramétrage, de façon uniforme à chaque niveau de la hiérarchie :*

- *les nombres de clusters k_S et k_T sont indépendants entre eux, et uniformément distribués entre 1 et n_S pour les noeuds source, entre 1 et n_T pour les noeuds cible,*
- *pour un nombre k_S donné de clusters source, toutes les partitions des n_S noeuds en k_S clusters sont équiprobables,*
- *pour un nombre k_T donné de clusters cible, toutes les partitions des n_T noeuds en k_T clusters sont équiprobables,*
- *pour un modèle de taille (k_S, k_T), toutes les distributions des m arcs sur les $k_E = k_S k_T$ coclusters sont équiprobables,*
- *pour un cluster de noeuds source (resp. cible) donné, toutes les distributions des arcs sortant (resp. entrant) du cluster sur les noeuds du cluster sont équiprobables.*

En prenant le log négatif des probabilités, on obtient la probabilité a posteriori d'un modèle, qui fournit le critère d'évaluation du théorème 1 (Boullé, 2010).

Théorème 1 *Un modèle M d'estimation de densité des arcs distribué selon l'a priori hiérarchique uniforme est optimal au sens de Bayes si sa valeur pour le critère suivant est minimale*

$$\begin{aligned} c(M) = & \log n_S + \log n_T + \log B(n_S, k_S) + \log B(n_T, k_T) \\ & + \log \binom{m + k_E - 1}{k_E - 1} + \sum_{i=1}^{k_S} \log \binom{m_{i.}^S + n_i^S - 1}{n_i^S - 1} + \sum_{j=1}^{k_T} \log \binom{m_{.j}^T + n_j^T - 1}{n_j^T - 1} \\ & + \log m! - \sum_{i=1}^{k_S} \sum_{j=1}^{k_T} \log m_{ij}^{ST}! \\ & + \sum_{i=1}^{k_S} \log m_{i.}^S! - \sum_{i=1}^{n_S} \log m_{i.}! + \sum_{j=1}^{k_T} \log m_{.j}^T! - \sum_{j=1}^{n_T} \log m_{.j}! \end{aligned} \tag{1}$$

$B(n, k)$ est le nombre de répartitions de n éléments en k sous-ensembles (éventuellement vides). Pour $n = k$, $B(n, k)$ correspond au nombre de Bell. Dans le cas général, $B(n, k)$ peut s'écrire comme une somme de nombre de Stirling de deuxième espèce (nombre de partitions de n valeur en k ensembles non vides) : $B(n, k) = \sum_{i=1}^{k} S(n, i)$.

La première ligne de la formule 1 correspond à la distribution a priori des nombres de clusters k_S et k_T et à la spécification de la partition des noeuds source (resp. cible) en clusters. La second ligne représente la spécification des paramètres de la distribution multinomiale des m arcs sur k_E coclusters, suivi de la spécification des paramètres de la distribution multinomiale des arcs sortant (resp. entrant) de chaque cluster sur les noeuds du cluster. La troisième ligne correspond à la vraisemblance de la distribution des arcs sur les coclusters au moyen d'un terme de multinome, et la dernière ligne à la vraisemblance de la distribution des arcs sortant (resp. entrant) de chaque cluster sur les noeuds du cluster.

Dans cet article, nous avons utilisé les heuristiques détaillées dans (Boullé, 2010), qui présentent l'avantage de la tenue de charge avec une complexité algorithmique de O(m) en occupation mémoire et O($m\sqrt{m}\log m$) en temps de calcul [2]. L'heuristique principale est une heuristique gloutonne ascendante, qui partant d'une solution initiale ayant autant de clusters que de noeuds, considère toutes les fusions de clusters et effectue celle qui améliore le plus le critère d'évaluation. Cette heuristique est complétée par des étapes de post-optimisation (déplacement de noeuds entre les clusters), et intégrée au sein d'une méta-heuristique consistant essentiellement à relancer l'algorithme en partant de plusieurs solutions initiales aléatoires et à retenir la meilleure solution rencontrée. Les algorithmes d'optimisation dont le principe est résumé ci-dessus ont été évalués extensivement dans (Boullé, 2010), sur la base d'une large variété de jeux de données artificiels, où la vraie densité est connue. La méthode est à la fois robuste et précise, en étant capable d'approximer n'importe quelle distribution, pourvu qu'il y ait assez d'instances dans l'échantillon d'apprentissage.

3 Illustration

Après avoir rappelé le critère de modularité (Newman et Girvan, 2003), nous illustrons les apports de notre approche de façon comparative en utilisant des jeux de données artificiels et

[2] L'outil utilisé dans les expérimentations est disponible en shareware sur www.khiops.com

montrons que les motifs découverts par notre approche étendent significativement ceux basés sur le critère de modularité.

3.1 Critère de Modularité pour le Clustering de Graphes

L'objectif de la détection de communauté est de partitionner un graphe en clusters ayant une forte densité d'arcs, avec une faible densité d'arcs entre les clusters différents. Le critère de modularité Q (Newman et Girvan, 2003) est largement utilisée dans les méthodes récentes de détection de communauté (Clauset et al., 2004; Danon et al., 2005; Blondel et al., 2008). La modularité évalue la densité des arcs dans les clusters de façon relative à la densité attendue en cas d'indépendance entre les extrémités des arcs.

Soit $G = (V, E)$ un graphe comportant $|V| = n$ noeuds et $|E| = m$ arcs, et m_{ij} un élément de la matrice d'adjacence du graphe. Dans le cas des graphes simples non orientés considérés usuellement, $m_{ij} = 1$ si les noeuds i et j sont connectés par un arc, $m_{ij} = 0$ sinon. Le degré d'un noeud i est le nombre d'arcs incidents à ce noeud. Dans le cas d'un graphe non orienté, les degrés entrant et sortant d'un noeud sont égaux. En utilisant les notations de la partie 2, on a $m_{i.} = m_{.i} = \sum_j m_{ij} = \sum_j m_{ji}$.

Un graphe non orienté ayant m_U arcs correspond à un graphe orienté symétrique ayant $m = 2m_U$ arcs. En supposant que les degrés des noeuds sont respectés et en cas d'indépendance des extrémités des arcs, la probabilité d'observer un arc entre deux noeuds i et j est $m_{i.}m_{.j}/m$. La modularité Q est définie par

$$Q = \frac{1}{m} \sum_{ij} \Big(m_{ij} - \frac{m_{i.}m_{.j}}{m} \Big) \delta(k_S(i), k_T(j)), \tag{2}$$

où $k_S(i) = k_T(i)$ est l'index du cluster contenant le noeud i, la fonction $\delta(x, y)$ vaut 1 si $x = y$ et 0 sinon, et $m = \sum_{ij} m_{ij}$ est égal à deux fois le nombre d'arcs (non orientés). La modularité prend ses valeurs entre -1 et 1, et à des valeurs positives quand les clusters ont plus d'arcs observés que dans le cas d'indépendance des extrémités des arcs. Ce critère vaut 0 dans les deux cas extrêmes d'un seul cluster et d'autant de clusters que de noeuds. La modularité possède deux propriétés intéressantes : elle est fondée théoriquement pour la découverte de clusters plus dense qu'en cas d'indépendence des extrémités des arcs, et elle ne nécessite aucun paramètre, comme par exemple le nombre de clusters.

3.2 Famille de Graphes Artificiels

Nous utilisons une famille de graphes artificiels consistant en quatre clusters de dix noeuds, nommés A, B, C, D. Les graphes considérés sont des graphes simples non orientés, et la proportion des arcs potentiels est contrôlée pour chaque cocluster, c'est à dire pour chaque paire de clusters de noeuds. Par exemple, choisir une proportion de $p = 20\%$ pour les arcs du cocluster (A, B) signifie que 20% des arcs potentiels avec une extrémité dans A et l'autre dans B (parmi $100 = 10 * 10$ arcs potentiels) sont dans le graphe. Un graphe aléatoire est produit en générant une valeur aléatoire $v \in [0, 1]$ pour chaque arc du graphe complet et en gardant l'arc dans le cocluster correspondant si $v \leq p$. Par exemple, la figure 4 illustre le cas d'un graphe artificiel où la proportion d'arcs est de 20% dans chaque cocluster, et présente un exemple de

graphe aléatoire généré selon cette distribution, en représentant les noeuds de chaque cluster sur un cercle pour une meilleure lisibilité.

Dans cet article, nous utilisons l'heuristique de (Blondel et al., 2008), une des plus performantes de l'état de l'art, qui construit très rapidement des partitions de graphe de grande qualité (selon le critère de modularité).

Pour les partitions basées sur l'approche MODL, nous utilisons l'outil Khiops[3]. Khiops est un outil général de préparation des données et de modélisation, qui implémente la méthode décrite dans la partie 2. Pour le problème de clustering de graphe, l'outil est appliqué à la représentation tabulaire du graphe, avec deux variables *Source* et *Cible*, et un individu par arc (deux dans le cas non orienté), pour la tâche d'analyse non supervisée bivariée.

3.3 Graphe Aléatoire

On étudie en premier lieu le cas d'un graphe aléatoire (Erdős et Rényi, 1976), pour lequel la densité d'arc est uniforme ; chaque arc potentiel a une probabilité de 20% d'être présent dans le graphe, ce qui équivaut à une probabilité de présence de 20% localement à chaque cocluster. La figure 4 montre les paramètres de cette distribution sur la gauche, puis un exemple de graphe généré selon cette distribution, et sur la droite les clusters découverts selon l'approche MODL et selon le critère de modularité, avec une couleur différente par cluster. En utilisant l'approche MODL décrite dans la partie 2, notre méthode construit un seul cluster de noeuds (Q=0), alors que la méthode basée sur la modularité construit 6 clusters (Q=0.217). Ce comportement de sur-apprentissage est reconnu par (Clauset et al., 2004) :

> "Non zero values represent deviation from randomness, and in practice it is found that a value above about 0.3 is a good indicator of significant community structure in a network."

A l'opposé, notre critère est fondé sur la théorie de l'information pour ne construire au plus qu'un seul cluster en cas de données aléatoires (Boullé, 2010), ce qui confirmé par les résultats présentés sur la figure 4.

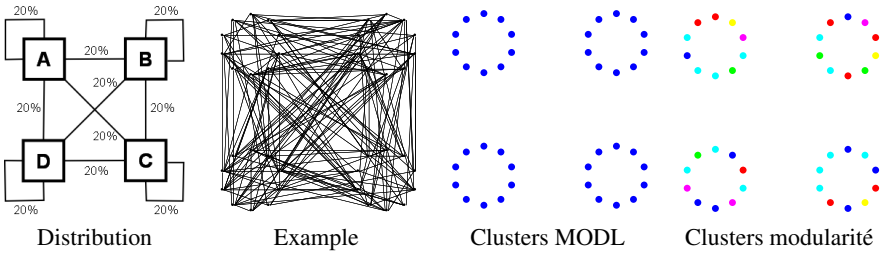


FIG. 4 – *Graphe artificiel : arcs aléatoires.*

3.4 Quasi-cliques

La figure 5 présente un motif classique consistant en quatre clusters denses, avec une densité intra cluster de 80% et une densité inter-cluster de 10%. Les méthodes basées sur l'ap-

[3]Khiops : disponible en shareware sur http ://www.khiops.com

proche MODL et sur le critère de modularité aboutissent au même résultat, en identifiant correctement les quatre clusters associés à A, B, C, D (avec Q=0.409). Dans le cas d'un graphe vérifiant l'hypothèse de décomposabilité en clusters denses, notre méthode se comporte ici comme les méthodes de clustering classiques.

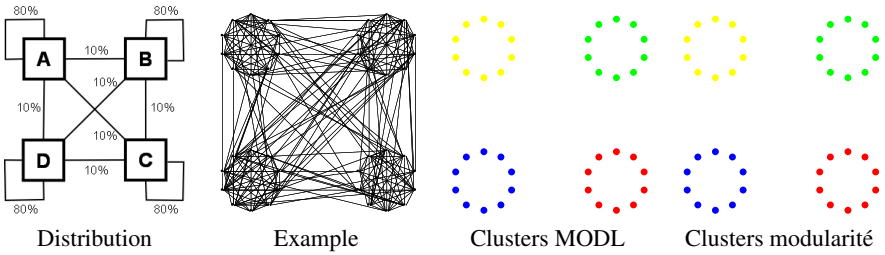


FIG. 5 – *Graphe artificiel : quasi-cliques.*

3.5 Stables

La figure 6 présente un motif inhabituel avec quatre clusters qui sont des stables, c'est à dire des sous-graphes vides d'arc, et une densité inter-cluster de 50%. Dans cet exemple, la densité intra-cluster est nettement inférieure à la densité moyenne. De fait, tous les graphes n'ont pas une structure en clusters. Cependant, tout algorithme de clustering produit une partition en clusters quelque soit le graphe en entrée, et l'algorithme basé sur la modularité crée 4 clusters sans intérêt (avec une modularité Q=0.132 en dessous du seuil de 0.3).

Les quatre clusters sont correctement identifiés par la méthode MODL (avec Q=-0.251). Alors que les approches classiques de clustering de graphe sont essentiellement paramétriques et visent à identifier des clusters denses dans les graphes, notre approche se comporte comme un estimateur non paramétrique de la densité des arcs dont l'objectif est de produire un résumé de la densité au moyen d'un approximateur constant par morceaux.

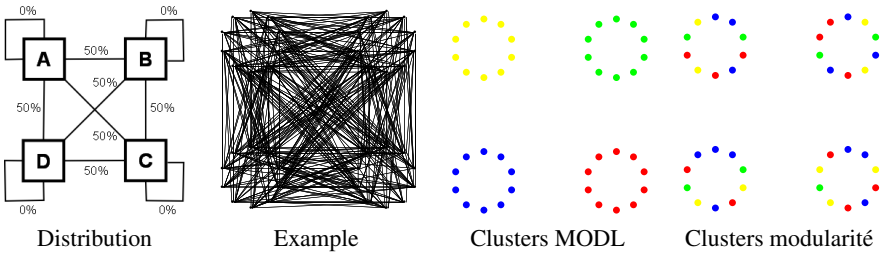


FIG. 6 – *Graphe artificiel : stables.*

4 Experimentation

Dans une première expérimentation, nous utilisons les graphes aléatoires introduits par (Johnson et al., 1989). Leurs caractéristiques sont résumées dans le tableau 1, ainsi que les résultats en termes de nombre de clusters et de critère de modularité obtenu par l'algorithme de (Blondel et al., 2008) et par notre méthode. L'algorithme d'optimisation de la modularité produit de nombreux clusters sans intérêt, avec de façon surprenante une modularité souvent supérieure au seuil de 0.3 recommandé par (Clauset et al., 2004), notamment dans le cas des graphes peu denses. En revanche, notre méthode produit un seul cluster pour tous les graphes aléatoires, confirmant ainsi sa robustesse.

			Modularité		MODL	
Graphe	Noeuds	Arcs	Clust.	Mod.	Clust.	Mod.
g500.005	451	625	23	0.676	1	0
g500.01	493	1223	15	0.446	1	0
g500.02	500	2355	12	0.289	1	0
g500.04	500	5120	12	0.194	1	0
g1000.0025	933	1272	34	0.708	1	0
g1000.005	994	2496	19	0.446	1	0
g1000.01	1000	5064	15	0.279	1	0
g1000.02	1000	10107	9	0.202	1	0

TAB. 1 – *Graphes aléatoires : notre méthode produit systématiquement un seul cluster.*

Dans une seconde expérience, nous utilisons dix graphes provenant de maillage en éléments finis (Walshaw, 2000) et de réseaux de co-auteurs (Newman, 2001) (les quatre derniers graphes). Ces graphes ainsi que les résultats de clustering sont présentés dans le tableau 2, avec jusqu'à 570 clusters selon notre approche pour le plus grand graphe contenant environ 150.000 noeuds et un millions d'arcs. Une inspection détaillée des résultats montre que notre approche produit un résumé nettement plus détaillé dans le cas des grands graphes fortement structurés, de type maillage en éléments finis, alors qu'elle fabrique un clustering plus synthétique pour les graphes de coauteurs particulièrement peu denses. Pour le graphe netscience par exemple, la méthode basée sur la modularité partitionne les 1461 noeuds en 278 clusters, dont 100 ne contiennent qu'un seul arc, alors que les clusters découverts par notre méthode contiennent entre 75 et 200 noeuds. Pour le graphe wave, notre approche construit 570 clusters de taille équilibrée (de 850 to 2250 arcs), environ 15 plus denses en arcs que les 31 clusters déséquilibrés (de 8500 à 80000 arcs) produits par la méthode basée sur la modularité.

5 Conclusion

Dans ce papier, nous avons présenté une nouvelle méthode de découverte de structures dans les graphes, en considérant les graphes comme des modèles générateurs dont les unités statistiques sont les arcs, avec une densité jointe inconnue des noeuds source et cible. L'approche MODL introduite pour les modèles en grille (Boullé, 2010) est exploitée en partitionnant conjointement les noeuds source et cible, de façon à estimer de façon non paramétrique

			Modularité		MODL	
Graphe	Noeuds	Arcs	Clust.	Mod.	Clust.	Mod.
bcsstk15	3942	56934	7	0.719	145	0.508
airfoil1	4253	12289	19	0.859	45	0.881
nasa4704	4704	50026	14	0.772	134	0.622
4elt	15606	45878	34	0.923	91	0.911
brack2	62631	366559	31	0.905	320	0.824
wave	156317	1059331	31	0.876	570	0.794
netscience	1461	2742	278	0.960	22	0.898
hep-th	7610	15751	631	0.849	48	0.777
cond-mat	16264	47594	796	0.844	127	0.785
astro-ph	16046	121251	420	0.727	233	0.586

TAB. 2 – *Graphes réels : résultats de clustering pour la méthode basée sur l'optimisation de la modularité et pour notre méthode.*

la densité des arcs. L'approche est comparée expérimentalement avec une méthode de l'état de l'art optimisant le critère de modularité. Les résultats montrent que notre méthode est à la fois plus robuste et plus informative, en étant capable d'identifier des structures nouvelles inaccessibles aux méthodes basées sur l'optimisation de la modularité. Les heuristiques exploitées dans notre méthode ont une complexité algorithmique en $O(m\sqrt{m}\log m)$, où m est le nombre d'arcs. Bien que cela soit acceptable dans nombreuses applications, il serait pertinent dans des travaux futurs d'améliorer la performance des algorithmes pour exploiter les bénéfices de l'approche sur des graphes de très grande taille, possédant potentiellement des millions de noeuds et des milliards d'arcs.

Références

Albert, R. et A.-L. Barabási (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys. 74*, 47–97.

Battiti, R. et A. Bertossi (1999). Greedy, prohibition, and reactive heuristics for graph partitioning. *IEEE Transactions on Computers 48*(4), 361–385.

Blondel, V., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment 2008*(10), P10008.

Boullé, M. (2006). MODL : a Bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Boullé, M. (2010). Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hands on pattern recognition*. Microtome. in press.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, et R. Wirth (2000). *CRISP-DM 1.0 : step-by-step data mining guide*.

Clauset, A., M. Newman, et C. Moore (2004). Finding community structure in very large networks. *Physical Review E 70*(6). 066111.

Copic, J., M. O. Jackson, et A. Kirman (2009). Identifying community structures from network data via maximum likelihood methods. *The B.E. Journal of Theoretical Economics 9*(1).

Danon, L., A. Díaz-Guilera, J. Duch, et A. Arenas (2005). Comparing community structure identification. *Journal of Statistical Mechanics : Theory and Experiment*, P09008.

Erdős, P. et A. Rényi (1976). On random graphs I. *Selected Papers of Alfréd Rényi 2*, 308–315. First publication in Publ. Math. Debrecen 1959.

Garey, M. et D. Johnson (1979). *Computers and Intractability : A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Hendrickson, B. et R. Leland (1995). A multilevel algorithm for partitioning graphs. In *Conference on High Performance Networking and Computing, Proceedings of the 1995 ACM/IEEE conference on Supercomputing*. Article No. : 28.

Holland, P., K. Laskey, et S. Leinhardt (1983). Stochastic blockmodels : First steps. *Social Networks 5*(2), 109–137.

Johnson, D., C. Aragon, L. McGeoch, et C. Schevon (1989). Optimization by simulated annealing : An experimental evaluation, part 1, graph partitioning. *Operations Research 37*, 865–892.

Kernighan, B. et S. Lin (1970). An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal 49*, 291–317.

Newman, M. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science of the USA 98*, 404–409.

Newman, M. et M. Girvan (2003). Finding and evaluating community structure in networks. *Physical Review E 69*. 026113.

Schaeffer, S. (2007). Graph clustering. *Computer Science Review 1*(1), 27–64.

Walshaw, C. (2000). The graph partitioning archive. University of Greenwich, UK.

Wasserman, S. et C. Anderson (1987). Stochastic a posteriori blockmodels : Construction and assessment. *Social Networks 9*(1), 1–36.

Summary

The discovery and analysis of structures in graphs has been long studied in the past. With the recent availability of many network data on the web, such as social networks, there is a renewed interest for these research topics, especially for the automatic discovery of community structures in large networks, exploiting for example the modularity criterion of Newman. In this paper, we present a novel way to summarize the structure of a large graph, based on the estimation of edge density. The graph is modeled using a clustering of the vertices, with a piecewise constant estimation of the density of the edges across the clusters. We evaluate our approach on numerous artificial and real graphs. The results show the validity of the approach, which automatically provides an insightful summary of large graphs.