Recherche et représentation de communautés dans des grands graphes

Nathalie Villa, Taoufiq Dkaki, Sébastien Gadat, Jean-Michel Inglebert, Quoc-Dinh Truong

Recherche et représentation de communautés dans des grands graphes

Nathalie Villa (*,**), **Taoufiq Dkaki** (***), **Sébastien Gadat** (*), **Jean-Michel Inglebert** (***), **Quoc-Dinh Truong** (****)
nathalie.villa@math.univ-toulouse.fr, dkaki@irit.fr, sebastien.gadat@math.univ-toulouse.fr, inglebert@iut-blagnac.fr, truong@univ-tlse2.fr

(*) Institut de Mathématiques de Toulouse, UMR CNRS 5219, Université de Toulouse (Toulouse III), 118 route de Narbonne, F-31062 Toulouse cedex 9, France.
(**) Toulouse School of Economics, Manufacture des tabacs, 21 allées de Brienne, F-31000 Toulouse, France.
(***) Institut de Recherche en Informatique de Toulouse, Université Toulouse III, 118 Route de Narbonne, F-31062 Toulouse cedex 4, France.
(****) Université de Cantho, 1 Rue Ly Tu Trong, Cantho, Vietnam

Mots clefs :
Graphes, Représentation de graphes, Classification, Réseaux sociaux, Recherche de communautés, Fouille de données
Keywords:
Graphs, Graphs vizualisation, Clustering, Social networks, Communities, Data mining
Palabras clave :
Grafos, visualización de grafos, Clasificación, Relación social, comunidades, Minería de datos

Résumé

Le travail présenté ici concerne l'analyse, la compréhension et la représentation de grands graphes. En effet, ce type de données se retrouve de manière naturelle dans un nombre croissant de problèmes concrets : web, recherche d'information, réseaux sociaux, réseaux d'interaction biologiques... La progression des moyens de recueil et de stockage des données rend la taille de ces graphes croissante : le développement de méthodes permettant leur analyse et leur représentation est donc un domaine de recherche dynamique et important à l'interface de nombreuses disciplines (mathématiques, statistique, informatique, ...). Dans cet article, nous développons une méthode de représentation de graphes basée sur une classification préalable des sommets. En effet, dans [18], les auteurs font remarquer que « *reducing [the] level of complexity [of a network] to one that can be interpreted readily by the human eye, will be invaluable in helping us to understand the large-scale structure of these new network data* » : nous nous appuyons sur ce constat pour utiliser une classification des sommets du graphe comme étape préliminaire et simplificatrice à la représentation du graphe dans son ensemble. La phase de classification consiste en l'optimisation d'une mesure de qualité spécialement adaptée à la recherche de groupes denses dans les graphes : la modularité. Cette mesure quantifie la « distance » à un modèle nul dans lequel les arêtes sont placées indépendamment de la classification. Elle a montré sa pertinence dans la résolution de problèmes de recherche de groupes denses dans un graphe. L'optimisation de la modularité est effectuée par le biais d'un algorithme stochastique de recuit simulé. Enfin, la représentation, en tant que telle, est basée sur un algorithme de « forces » contraint décrit dans [22]. Après une courte introduction au problème (partie 1), nous détaillons, dans la partie 2, la procédure de classification des sommets et, en partie 3, l'algorithme de représentation utilisé. Un exemple issu de l'analyse de réseaux sociaux est ensuite présenté en partie 4.

1 Introduction

La classification de sommets d'un grand graphe a déjà fait l'objet de nombreux travaux. Cette approche a été présentée depuis longtemps comme un moyen de simplifier la structure d'un grand graphe, dont le nombre de sommets peut être égal à plusieurs milliers, pour en faire ressortir sa structure macroscopique (voir [18]) et mettre en valeur des « communautés » (pour reprendre la terminologie couramment employée dans le domaine des réseaux sociaux). Même si aucune définition formelle de ce qu'est une communauté n'est aujourd'hui consensuelle, on désigne généralement sous ce terme un sous-ensemble des sommets du graphe, fortement connectés entre eux et faiblement connectés aux autres sommets. Aucune structure euclidienne, *a priori*, n'est définie entre les sommets d'un graphe et la classification des sommets requiert donc l'implémentation de méthodes spécifiques par rapport aux données multidimensionnelles usuelles. Une très grande variété de méthodes s'attachent à trouver des solutions satisfaisantes au problème de la partition des sommets d'un graphe : [21] présente un grand nombre de ces méthodes basées sur la définition de similarités entre sommets ou sur l'optimisation de critères de qualité. Parmi toutes ces méthodes, la classification spectrale [17] est une approche devenue très populaire mais qui est très sensible à la taille des données puisqu'elle nécessite la décomposition spectrale d'une matrice carrée dont la taille est égale au nombre de sommets du graphe.

Nous proposons, dans cet article, une méthode originale de classification des sommets d'un graphe basée sur une mesure de qualité couramment utilisée dans ce but : la modularité. L'originalité de notre approche réside dans l'utilisation d'un algorithme stochastique d'optimisation de ce critère, algorithme permettant sa mise en œuvre pour des graphes dont la taille peut être importante. Enfin, nous montrons comment cette méthode de classification peut être combinée avec un algorithme de représentation de graphes dont les sommets ont été partitionnés en classes (graphes « clusterisés ») pour aboutir à une représentation lisible de l'ensemble du graphe, des principaux groupes mis en valeur par la phase de classification et des liens existant entre sommets et entre groupes.

2 Classification de sommets dans un grand graphe

2.1 La modularité comme mesure de qualité d'une classification

Dans [21], l'auteure passe en revue les principales méthodes de classification de sommets : certains auteurs se sont concentrés sur la construction d'une mesure de similarité ou d'une distance entre sommets du graphe pour pouvoir ensuite appliquer directement les méthodes de classification existantes pour des données définies par des tableaux de similarités (voir par exemple, [2], [5] ou [12] pour des exemples de telles approches). Une approche alternative consiste à optimiser une mesure de qualité de la classification, mesure définie à partir de la structure du graphe, c'est-à-dire, de la données de ses arêtes. Classiquement, on peut minimiser les « coupes » d'arêtes entre classes : si G est un graphe non orienté, pondéré, de taille n, de sommets $V = \{1, 2, ..., n\}$ et dont l'ensemble des arêtes, E, est décrit par la matrice (symétrique, à diagonale nulle et coefficients positifs) des poids W, la coupe d'une classification $c(i)$ (pour $i \in V$) est la quantité :

$$\frac{1}{2}\Sigma_{i,j\in V, c(i)\neq c(j)} W_{ij} \ .$$

Cette mesure de qualité est fortement reliée à la définition d'un plongement du graphe dans un espace euclidien (voir par exemple, [14]) et son optimisation est fréquemment approchée par des méthodes de classification de type k-means appliquées à ce plongement. Cette méthode est connue sous le nom de classification spectrale (*spectral clustering*).

Cependant, si la classification spectrale est séduisante par sa simplicité et son interprétation très naturelle, elle peut s'avérer décevante en pratique. La raison est exposée dans [17] : la mesure de coupures considère toutes les arêtes de manière similaire. Or, Newman fait remarquer qu'être lié à un sommet de fort degré est moins exceptionnel qu'être lié à un sommet de faible degré. Pour tenir compte de cet aspect, il introduit une mesure de qualité basée sur le calcul d'une distance à un

modèle nul dans lequel la répartition des arêtes entre et à l'intérieur des classes ne dépend que du degré des sommets. Pour une classification donnée C_1, ... C_p des sommets du graphe, la *modularité* est la quantité :

$$Q = \frac{1}{2m} \Sigma_{k=1...p} \Sigma_{i,j \in C_k, i \neq j} \left(W_{ij} - P_{ij} \right)$$

où m est le nombre total d'arêtes du graphe et les P_{ij} sont les poids estimés des arêtes intra classes entre les sommets i et j dans le modèle nul où la répartition des arêtes dépend uniquement du degré et non des classes. Ainsi, une valeur de Q grande (et positive) est le signe d'une sur-représentation (par rapport au modèle nul) des arêtes à l'intérieur des classes alors qu'une valeur de Q fortement négative est le signe d'une sur-représentation des arêtes à l'extérieur des classes. La notion naturelle de communautés, exposée plus haut, correspond au premier cas.
Le choix de la valeur des P_{ij} est fait en fonction des hypothèses correspondant au modèle nul. Des contraintes raisonnables sur leurs valeurs sont, usuellement,

- $\Sigma_{i,j \in V} P_{ij} = 2m$ (ie : le nombre d'arêtes du modèle nul est identique au nombre d'arêtes du graphe considéré) ;
- $\Sigma_{j \in V} P_{ij} = d_i$ où, pour tout $i \in V$, d_i désigne le degré du sommet i : $d_i = \Sigma_{j \in V} W_{ij}$. Cette hypothèse est simplement équivalente au fait que le degré de chaque sommet du modèle nul est le même que le degré des sommets dans le graphe considéré.

Enfin, par hypothèse, on suppose que P_{ij} est de la forme $f(d_i)f(d_j)$ qui traduit, d'une part, le fait que les poids attendus du modèle nul ne dépendent que des degrés des sommets considérés et non de leur répartition entre les différentes classes et, d'autre part, le fait que les poids entre deux arêtes données sont indépendants. [17] montre qu'alors la valeur des P_{ij} est donnée par $P_{ij} = \frac{d_i d_j}{2m}$.
L'optimisation de Q sur l'ensemble des partitions des sommets du graphe ne peut être atteinte par une recherche exhaustive sur l'ensemble des partitions possibles des sommets en p classes (voir [17]) et diverses approches d'approximation de ce critère ont été proposées récemment : l'optimisation hiérarchique [18] qui peut, dans certains cas, conduire à des solutions assez inadéquates (voir, par exemple, l'exemple donné dans [17]), l'optimisation basée sur un critère de décomposition spectrale [17], l'optimisation par recuit déterministe [13], ...
Nous présentons ici une approche par recuit simulé de l'optimisation de la modularité. Cette approche, basée sur une démarche stochastique, présente l'avantage de ne nécessiter que des opérations de très faible complexité : elle est donc particulièrement bien adaptée à des données de grande taille.

2.2 Optimisation par recuit simulé

Le recuit simulé est un algorithme stochastique [11] destiné à optimiser une fonction de coût lorsque d'autres algorithmes plus simples (de type descentes de gradient) sont inadaptés. Il est particulièrement utile lorsque l'espace des possibles est discret ou bien lorsque la fonction possède de nombreux maxima locaux. Ici, l'optimisation de la Q-modularité sur l'espace des partitions possibles du graphe est typiquement le cadre d'application d'un algorithme stochastique tel que le recuit simulé (ou encore d'un algorithme génétique). On présente ici brièvement l'algorithme de recuit simulé et, pour plus de détails, on consultera [11].
De manière similaire à l'algorithme de Métropolis-Hastings [15], le principe de l'algorithme de recuit simulé est de simuler, par une chaîne de Markov, une variable prenant ses valeurs dans l'ensemble des classifications possibles des sommets du graphe. Cette chaîne de Markov a pour loi de probabilité invariante :

$$\mu_T(C) = \frac{e^{Q(C,G)/T}}{Z}$$

où $Q(C,G)$ désigne la modularité de la classification C du graphe G, T est un paramètre de température positif et Z une constante de normalisation qu'il est inutile de calculer. L'intérêt de cette loi apparaît lorsque l'on fait tendre le paramètre de température, T, vers 0 : la chaîne de Markov a alors tendance à se figer sur les états de grande probabilité, c'est à dire, sur les maxima de Q. Un choix généralement admis pour la décroissance du paramètre de température est $T_l = \frac{\gamma}{\log l}$ où l désigne l'état l de la chaîne de Markov (ie, le nombre d'itérations de l'algorithme) et γ est une constante strictement positive.

La mise en œuvre pratique de la simulation de la chaîne de Markov est faite par le biais de la méthode d'« acceptation / rejet » : étant donnée une classification C_I, des sommets du graphe G, on choisit, de manière aléatoire, une classification « voisine », C_F, selon un processus réversible. La classification C_F est choisie à partir de C_I avec une probabilité $P(C_I,C_F)$ (dans notre cas, on prend, pour simplifier, $P(C_I,C_F) = P(C_F,C_I)$). Enfin, on accepte la classification C_F (c'est à dire qu'on la conserve comme nouvel état de la chaîne de Markov) avec la probabilité $\min\left(1, e^{\frac{\Delta Q(C_I,C_F,G)}{T}}\right)$ où $\Delta Q(C_I, C_F, Q)$ désigne la différence de modularité entre la classification C_I et la classification C_F (plus précisément, $\Delta Q(C_I, C_F, Q) = Q(C_F, G) - Q(C_I, G)$). L'algorithme d'optimisation de la modularité par recuit simulé est décrit dans l'Algorithme 1 .

L'étape 3.d de l'Algorithme 1 ne nécessite pas le calcul intégral de la modularité $Q^{l-1}\left(C_1^{l-1},...,C_p^{l-1},G\right)$. En effet, [13] montre que la modularité peut être mise sous la forme matricielle

$$Q\left(C_1,...,C_p,G\right) = \frac{1}{2m}\mathrm{Tr}\left(S^T BS\right)$$

où S désigne la matrice de taille $p \times n$ dont les éléments sont, pour tout x dans V et tout $k = 1, \ldots, n$, $S_{kx} = \begin{cases} 1 & \text{si } x \in C_k \\ 0 & \text{sinon} \end{cases}$ (en d'autres termes, S est la matrice d'appartenance des sommets aux classes), B la matrice $(W - P)$ et $\mathrm{Tr}(S^{\mathrm{T}}BS)$ la trace de la matrice $S^{\mathrm{T}}BS$, c'est-à-dire, la somme de ses éléments diagonaux. On montre alors facilement que l'étape 3.d se réduit à :

$$\Delta Q = \frac{1}{m}\sum_{y \neq x}\left(S_{jy}^{l-1}B_{xy} - S_{iy}^{l-1}B_{xy}\right) = \frac{1}{m}\left(\sum_{y \in C_j^{l-1}} B_{xy} - \sum_{y \in C_i^{l-1} - \{x\}} B_{xy}\right)$$

ce qui nécessite moins de $(n - 1)$ additions, c'est-à-dire, bien moins que le calcul de la modularité d'une classification donnée qui nécessite de l'ordre de n^2 opérations élémentaires.

1. Choisir, de manière aléatoire, une partition initiale de V en p classes C_1^0, …, C_p^0.
2. Calculer la modularité associée à cette partition, $Q^0\left(C_1^0,...,C_p^0,G\right)$.
3. Pour l = 1 … L, répéter :
   a. Tirer deux classes, C_i^{l-1} et C_j^{l-1}, au hasard, parmi $\left(C_i^{l-1}\right)_{i=1...p}$
   b. Si $C_i^{l-1} \cup C_j^{l-1} \neq \varnothing$, tirer un sommet, x, au hasard dans $C_i^{l-1} \cup C_j^{l-1}$; sans perte de généralité, on peut supposer, dans la suite, que $x \in C_i^{l-1}$.
   c. Définir une nouvelle classification telle que $C_k = C_k^{l-1}$ pour tout $k \neq i, j$, $C_i = C_i^{l-1} - \{x\}$ et $C_j = C_j^{l-1} \cup \{x\}$ (changement de classe de x).
   d. Calculer $\Delta Q = Q\left(C_1,...,C_p,G\right) - Q^{l-1}\left(C_1^{l-1},...,C_p^{l-1},G\right)$.
   e. Selon que :
      - $\Delta Q > 0$, on accepte la nouvelle classification : $C_k = C_k^{l-1}$ pour tout k et on a donc $Q^l\left(C_1^l,...,C_p^l,G\right) = \Delta Q + Q^{l-1}\left(C_1^{l-1},...,C_p^{l-1},G\right)$.
      - $\Delta Q < 0$, on accepte la nouvelle classification avec probabilité $e^{\gamma \log(l) \Delta Q}$ (pour un $\gamma > 0$ fixé) ; on en déduit $Q^l\left(C_1^l,...,C_p^l,G\right)$.

Algorithme 1 : Recuit simulé pour optimisation de la modularité

2.3 Un exemple simulé

Nous nous proposons d'analyser la qualité de la classification obtenue par l'approche décrite dans l'Algorithme 1 . Pour ce faire, nous utilisons un modèle de graphe aléatoire (non orienté et non pondéré) plausible du point de vue de l'étude des réseaux sociaux : il s'agit du « planted *p*-partition model » décrit, par exemple, dans [4]. Un graphe aléatoire généré par ce modèle dépend de 4 paramètres, notés *p* (nombre de classes, appelées dans la suite *classes naturelles*), *k* (nombre de sommets dans chaque classe), *r* (probabilité d'une arête intra-classe) et *q* (probabilité d'une arête inter-classe). Nous noterons, dans la suite, *G*(*p*,*k*,*r*,*q*) ce modèle. Une réalisation du modèle pour 3 classes de 4 sommets est donnée dans la Figure 1.
Pour évaluer si l'Algorithme 1 était capable de retrouver les classes naturelles du modèle « planted *p*-partition », nous avons généré pour des valeurs fixées des paramètres du modèle, 50 graphes aléatoires. Deux ensembles de paramètres ont été testés :

- *G*(5,100,0.7,0.02) : sur les 50 graphes aléatoires générés selon ce modèle, la modularité moyenne de la partition naturelle est de l'ordre de 0.70. Cette valeur est compatible avec les valeurs trouvées dans les réseaux sociaux réels.

- $G(50,10,0.7,0.2)$: sur les 50 graphes aléatoires générés selon ce modèle, la modularité moyenne de la partition naturelle est de l'ordre de 0.37. Cette valeur est bien inférieure aux valeurs généralement observées sur des réseaux sociaux réels. La recherche des classes naturelles dans ce modèle pouvant donc être considérée comme un problème relativement compliqué.

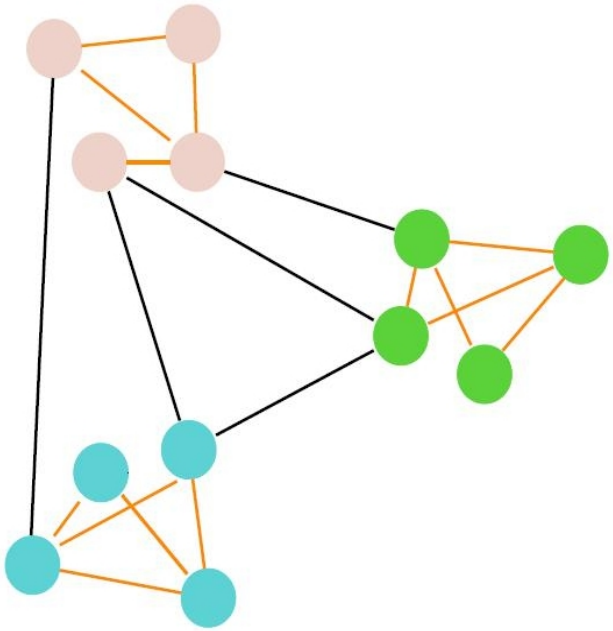

Figure 1 : Exemple de « planted p-partition graph » : le graphe est composé de 3 classes de 4 sommets avec une proportion d'arêtes intra-classes de l'ordre de 58% et une proportion d'arêtes inter-classes de l'ordre de 10 %

Par ailleurs, l'algorithme a été testé avec le nombre correct de classes à retrouver ($p_{algo} = p$) ou bien avec un nombre supérieur ($p_{algo} > p$) pour permettre de voir si l'algorithme est sensible à la donnée d'un nombre adéquat de classes ou non.
Enfin, la qualité de reconstruction des classes a été mesurée de deux manières différentes :

- par le calcul du taux de mauvais classement des sommets dans leur classe naturelle (pourcentage de sommets qui ne sont pas classés dans la classe la plus fréquemment affectée aux sommets de la classe naturelle à laquelle ils appartiennent) ;
- par calcul de la différence entre la modularité de la classification naturelle et la modularité de la classification fournie par l'algorithme (cette mesure de qualité sera appelée *défaut de modularité*).

Les résultats obtenus sont résumés dans le Tableau 1. Le taux de bonne classification des sommets dans leurs classes naturelles est tout à fait satisfaisant, avec moins de 10% de sommets mal classés même dans la situation la plus difficile (celle décrite par la dernière colonne du tableau). De même, la modularité est correctement optimisée puisque le défaut de modularité représente environ 10% de la modularité attendue pour le problème le plus difficile (dernière colonne) contre 5% environ pour le problème le plus facile (première colonne). Enfin, même dans la situation où le nombre de classes recherché par l'algorithme est très supérieur au nombre de

classes naturelles (deuxième colonne), l'algorithme se comporte correctement et tend à diminuer automatiquement le nombre de classes proposées par défaut à l'utilisateur.

$r \times q$ $p \times k$ p_{algo}	0.7×0.02 5×100 5	0.7×0.02 5×100 10	0.7×0.2 50×10 50
Taux d'erreur (classification)	1,36 %	2,34 %	6,32 %
Défaut de modularité	0,036	0,040	0,036

Tableau 1 : Performances moyennes de l'algorithme d'optimisation de la modularité par recuit simulé sur le modèle « p-planted partition graph ».

Les résultats de cette section sont le signe que l'algorithme proposé est une approche de classification des sommets d'un graphe qui donne des performances tout à fait satisfaisantes. Dans la suite, nous expliquons comment utiliser une classification pertinente obtenue par cette approche pour représenter l'intégralité du graphe de manière à en faire ressortir les principales structures.

3 Représentation des sommets d'un graphe clusterisé

Il est admis que la visualisation contribue fortement à l'efficacité des processus d'analyse et d'exploitation des données. C'est dans cette optique que plusieurs travaux se sont intéressés au dessin de graphes. Construire un dessin de graphe consiste à positionner les sommets et les arcs d'un graphe dans un espace d'affichage selon des critères et des contraintes dont l'objectif est d'offrir les visualisations les plus facilement interprétables. Les travaux les plus connus se fondent sur une approche dirigée par les forces. Dans cette approche, que l'on peut attribuer à Eades (voir, par exemple, [6]), les sommets sont assimilés à des anneaux et les arcs à des ressorts. Les sommets sont placés dans une disposition initiale souvent aléatoire et laissés libres de sorte que les forces des ressorts les déplacent vers un état d'équilibre où l'énergie du système – anneaux et ressorts – est minimale. Une autre variante de cette méthode, appelée « force et repousse », consiste à modifier la manière dont les forces sont appliquées : des forces répulsives sont activées pour toutes les paires de sommets tandis que des forces attractives sont instaurées uniquement pour les sommets connectés.

Ces algorithmes de forces sont utilisés avec succès pour le dessin de graphes possédant relativement peu de sommets. La difficulté de visualiser un graphe dépend, en effet, de sa taille qui est l'obstacle majeur pour la réalisation de l'objectif de visualisation. Plus le nombre de sommets et d'arcs est grand, plus les critères de visibilité et de qualité de l'interprétation sont difficiles à atteindre. Pour prendre en charge des graphes contenant plus que quelques centaines de sommets, on a recourt soit à des algorithmes plus efficaces (voir, par exemple, [9]), soit à une classification préalable ayant pour but de réduire la complexité du graphe. Cette réduction de complexité consiste souvent à considérer le graphe des classes plutôt que le graphe initial et s'accompagne indéniablement d'une perte d'information. C'est cette perte d'information que nous cherchons à minimiser par l'utilisation d'algorithmes de représentation de graphes « clusterisés ». Ce type d'algorithme fournit une visualisation qui permet d'apprécier non seulement la place de chaque sommet dans le graphe (relativement à ses voisins) mais aussi la structure en communautés du graphe. Elle permet ainsi de montrer les relations entre communautés (ou classes) et de juger l'influence de l'appartenance à une classe sur le positionnement de chaque sommet. Elle bénéficie de la simplification de structure apportée par l'étape de classification tout en conservant la représentation du graphe dans son ensemble.

En réalité, très peu de travaux ont été dédiés à la visualisation des graphes clustérisés. Parmi ces travaux, nous pouvons mentionner ceux de Noack [19], de Chuang *et al.* [3] et de Eades et Huang [7] comme les plus représentatifs. Ces travaux se basent sur le modèle de forces traditionnel de Kamada et Kawai [10], ajoutent un sommet virtuel au centre de chaque classe et relient chacun de ces nouveaux sommets à tous les sommets de la classe qui lui est associée. Ces modèles donnent de

bons résultats pour des graphes dont les classes sont beaucoup plus denses que le graphe global. Ainsi, outre l'introduction d'un sommet virtuel dont le sens n'est pas réellement défini de manière concrète, cette approche s'avère inutilisable lorsque le rapport de la densité dans les classes à la densité globale du graphe n'est pas faible. Cela peut être le cas dans de nombreux exemples concrets, notamment lorsque le nombre de classes de la classification est faible devant la taille du graphe.
Dans cet article, nous proposons d'utiliser une approche alternative de visualisation capable de gérer la visualisation des graphes clustérisés quel que soit le rapport des densités intra et inter classes. Cette approche est présentée dans [22]. Ce modèle est basé sur le modèle de forces de Frunchterman [8] et permet d'associer à chaque classe une zone de visualisation prédéfinie dans laquelle chacun de ses sommets sera représenté. L'algorithme proposé présente deux différences par rapport à celui de Frunchterman. Tout d'abord, les forces mises en œuvre sont de deux types : celles intervenant entre les sommets d'une même classe d'une part et celles intervenant entre les nœuds de classes différentes d'autre part. Ces forces sont caractérisées par des intensités différentes dans les deux cas. Ensuite, les sommets sont contraints à l'intérieur des zones associées à leur classe d'appartenance par l'ajout de forces répulsives entre les sommets et la frontière de leur zone. Pour ce faire, les sommets et les frontières des zones associées aux différentes classes sont chargés de forces électriques de même nature. Ceci assure que chaque sommet reste toujours à l'intérieur de la zone qui lui a été initialement associée.

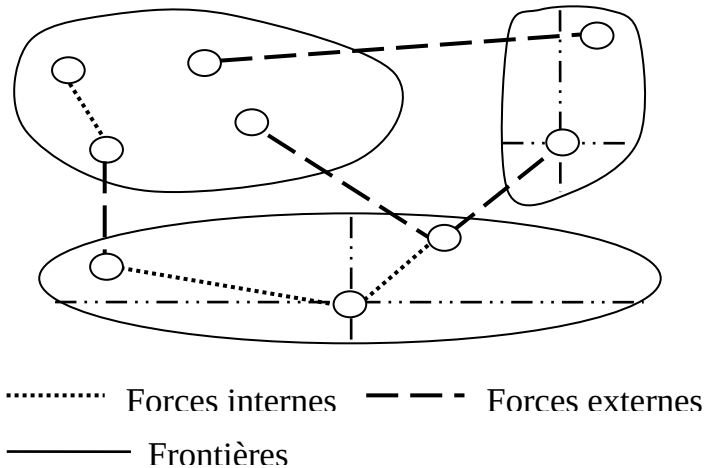

Figure 2 Modèle de forces pour la représentation d'un graphe clusterisé

La Figure 2 résume l'intégralité du modèle de forces mis en œuvre lors de la représentation d'un graphe clusterisé. C'est cette approche que nous combinons avec la méthode de classification de sommets décrite plus haut pour obtenir une représentation globale d'un réseau social découpé en communautés.

4 Applications

4.1 Représentation du réseau social issu du livre « Les Misérables »

Dans ce premier exemple, nous traitons un exemple simple et très connu pour montrer comment la combinaison d'outils de classification et de visualisation permet de donner une représentation simple des structures principales d'un graphe. Il s'agit d'un réseau social issu du livre « Les Misérables » de Victor Hugo et décrit dans [16] qui dénombre sous la forme d'un graphe pondéré, les apparitions simultanées des 77 personnages du roman.
La première phase de l'analyse consiste à mettre en œuvre l'algorithme de recuit simulé pour la classification des sommets. Celui-ci a été testé avec :

- différentes valeurs du nombre de classes (variant de 5 à 10). En effet, la modularité n'est pas une fonction monotone du nombre de classes et peut donc permettre un choix optimal (au sens de ce critère) du nombre de classes.
- différentes valeurs du paramètre γ de l'Algorithme 1 .

Par ailleurs, plusieurs itérations de l'algorithme ont été effectuées : l'algorithme étant basé sur un processus stochastique, chaque itération produit une solution potentiellement différente, les répétitions ayant pour but d'approcher au mieux l'optimum en temps fini.

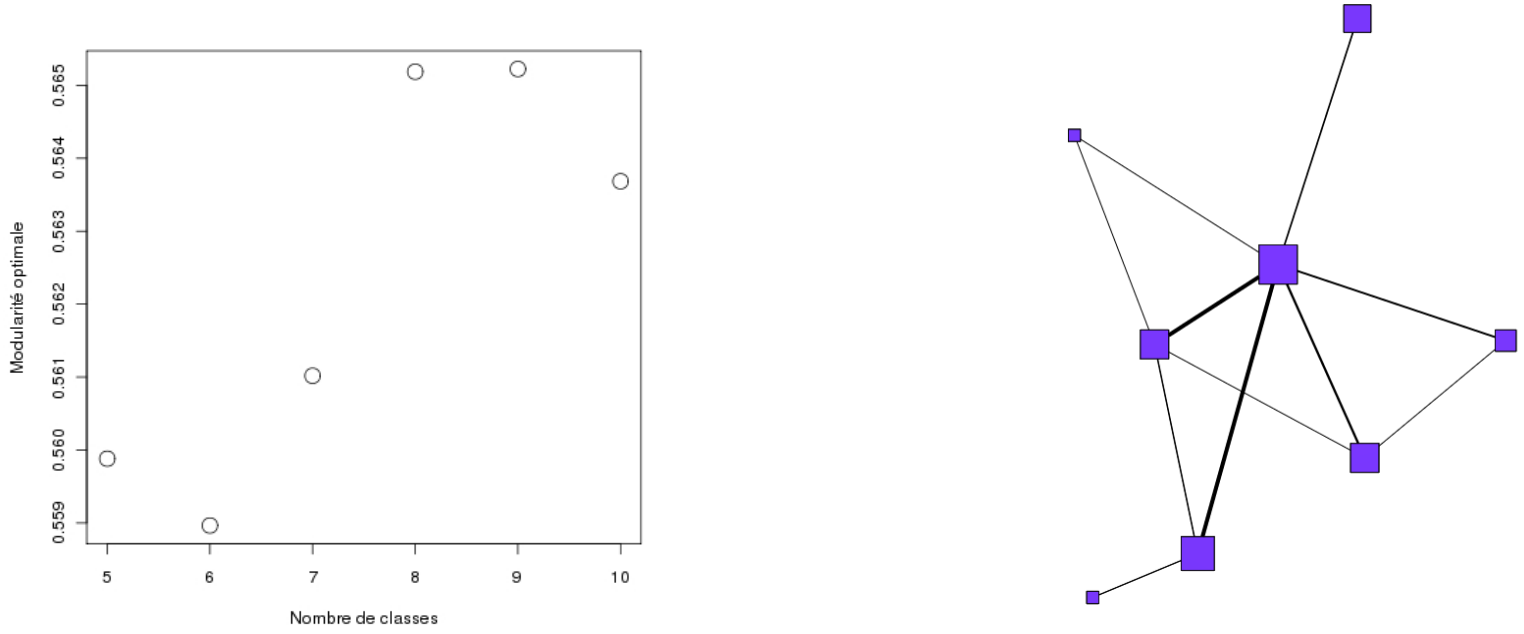

Figure 3 :Evolution de la modularité optimale trouvée en fonction du nombre de classes (à gauche) et représentation du graphe des classes par un algorithme de forces (la taille des carrés est proportionnelle à l'effectif de chaque classe ; l'épaisseur des arêtes est proportionnelle au nombre d'arêtes entre les deux classes).

La Figure 3 présente les résultats de cette première simulation : à gauche, est représentée l'évolution de la modularité optimale trouvée par l'algorithme en fonction du nombre de classes initial. La modularité optimale est obtenue pour une initialisation à 9 classes et la classification optimale comporte seulement 8 classes. Le graphique et l'obtention d'une solution optimale dont le nombre de classes est inférieur au nombre de classes initial montre que la recherche d'une bonne classification dont le nombre de classes serait supérieur à 10 est inutile. Une fois la classification obtenue, on peut représenter celle-ci en remplaçant chaque classe par un symbole (ici un carré) dont la surface est proportionnelle à l'effectif de la classe. Un algorithme de forces permet ensuite de placer ces classes de façon lisible en mettant en valeur les relations existant entre classes et la structure globale du réseau : la Figure 3 (droite) est le résultat d'une telle démarche. Celle-ci a été réalisée à l'aide du logiciel Tulip[1] (voir [1]). On y remarque un groupe important, central, autour duquel gravitent les autres personnages du roman et on identifie également aisément 3 autres groupes de taille importante et qui permettent de structurer les relations autour du groupe principal.

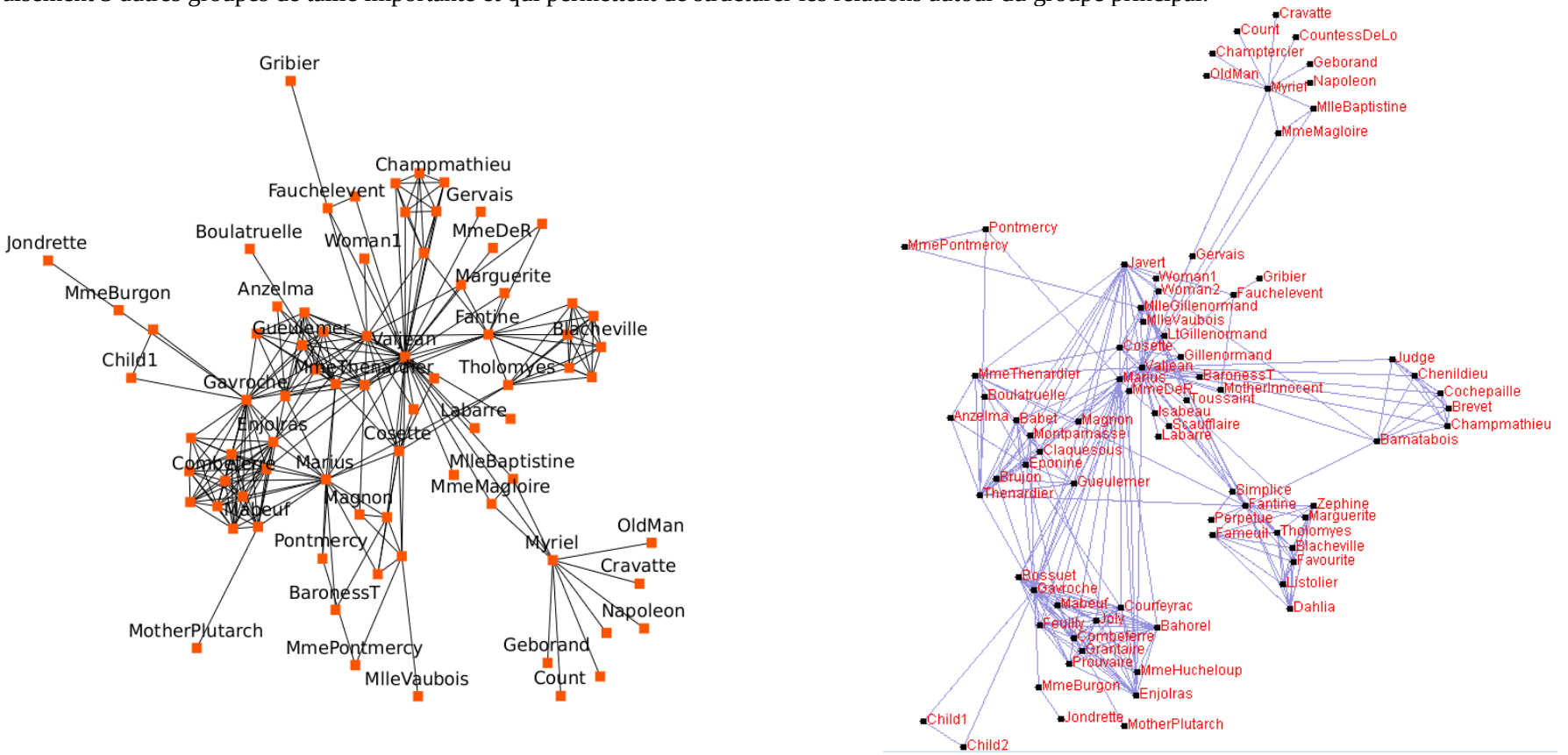

Figure 4 : Représentation initiale du graphe obtenue par algorithme de forces avec le logiciel Tulip (à gauche) et représentation par algorithme de forces contraint du graphe clusterisé (à droite)

[1] Logiciel libre de représentation de graphes, disponible à l'adresse http://tulip.labri.fr/.

Pour représenter l'intégralité du graphe à partir de l'algorithme décrit dans la partie 3, nous utilisons les positions des classes données dans la Figure 3 par l'algorithme de forces utilisé pour la représentation du graphe des classes. La représentation finale du graphe clusterisé est donné dans la Figure 4 (droite) où l'on peut la comparer à la représentation initiale obtenue par un algorithme de forces (à gauche). Si certains groupes denses sont facilement identifiables dans les deux représentations (comme par exemple celui qui est organisé autour de Myriel), la représentation du graphe clusterisée permet d'identifier plus clairement des groupes dans la partie la plus dense du graphe. Elle met bien en valeur la position centrale de certains personnages du groupe principal : Valjean, Cosette, Marius, Javert (…), ainsi que la proximité de ces personnages centraux. On rappelle que la trame principale du roman est organisée autour de la vie de Jean Valjean, ancien forçat, qui mourra dans les bras de sa jeune protégée Cosette et du jeune étudiant tombé amoureux d'elle, Marius, après avoir été poursuivi durant des années par l'intransigeant Javert. Or, hormis pour Valjean, cette position centrale n'est pas aussi claire dans le graphe inital. Les trois sous-groupes principaux se présentent comme autant d'histoires secondaires gravitant autour de cette histoire principale : le groupe important situé le plus en bas du graphe est organisé autour de Gavroche, gamin des rues, capable de gestes de générosité envers Mabeuf (aussi dans ce groupe) et deux enfants perdus, appelés ici « Child 1 » et « Child 2 ». Un second groupe principal se situe au-dessus de celui-ci, à droite. Il est organisé autour de Fantine, ouvrière obligée de confier sa fille, Cosette à des inconnus. On retrouve justement ces inconnus dans le troisième groupe important, situé à gauche du groupe principal : ce sont les Thénardier ; c'est leur présence dans ce groupe qui influence la position de Cosette dans la partie gauche de son groupe.
Ainsi, la représentation du graphe clusterisé est facilement mise en relation avec l'histoire du roman. Elle permet d'en identifier les personnages principaux et les histoires secondaires. Sur cet exemple simple, on comprend l'apport d'une phase préalable de classification à la représentation du graphe afin de diriger la structuration du dessin et simplifier l'interprétation.

4.2 Représentation d'un réseau de collaborations scientifiques

Nous traitons, dans cette partie un autre exemple simple mais plus réaliste destiné à montrer que l'approche est utilisable pour des graphes de plusieurs centaines de sommets. L'exemple traité ici est issu de [17][2] : il s'agit d'un réseau de collaborations scientifiques d'individus travaillant autour de la thématique des réseaux sociaux, modélisé sous la forme d'un graphe pondéré (non orienté). La plus grande composante connexe du graphe contient 379 scientifiques. Une représentation basée sur un algorithme de forces et obtenue à l'aide du logiciel Tulip est donnée dans la Figure 5.

[2] Les données sont disponibles à l'adresse http://www-personal.umich.edu/~mejn/netdata/netscience.zip.

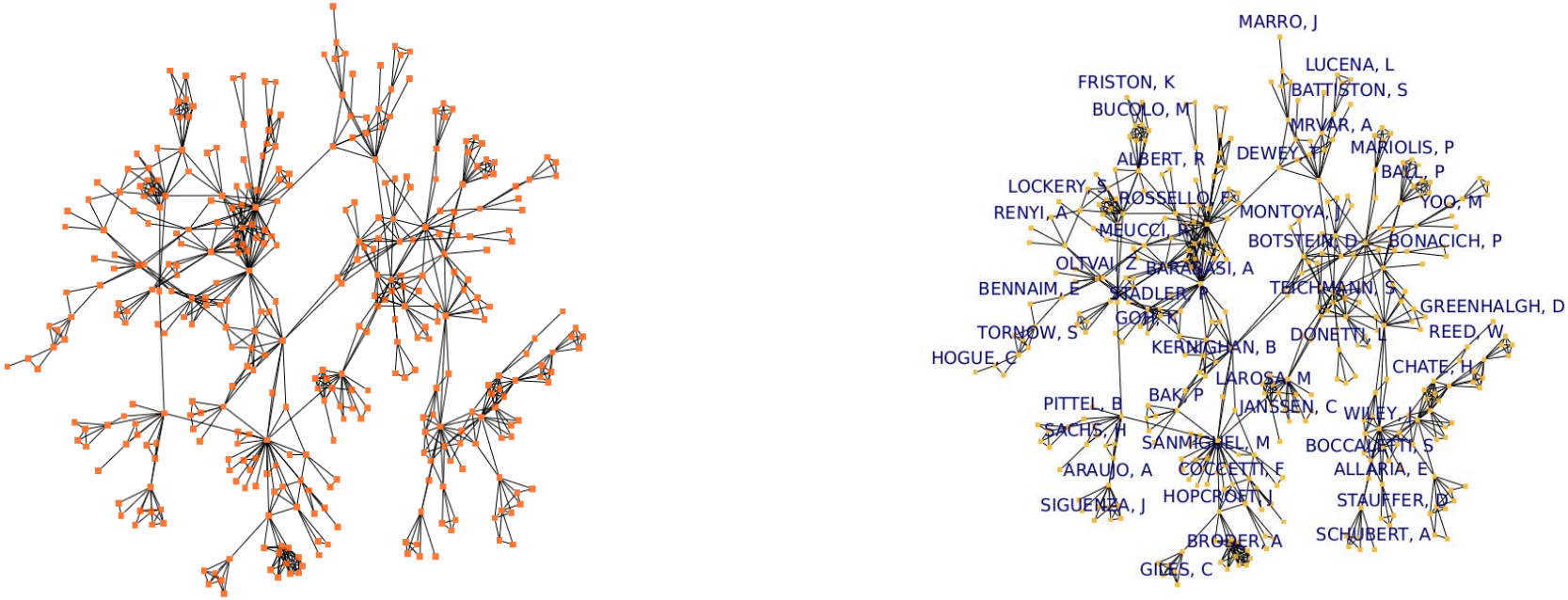

Figure 5 : Représentation initiale du réseau de collaboration scientifique sans légende (à gauche) et avec légende (à droite).

Sur ce graphe, nous avons, tout d'abord, mis en œuvre l'algorithme de recuit simulé pour optimisation de la modularité. Afin d'obtenir des classes de petites tailles, nous avons opté pour un nombre de classes égal à 20 ; plusieurs initialisations aléatoires ont été effectuées et plusieurs valeurs du paramètre γ de l'algorithme de recuit simulé ont été testées. La meilleure classification obtenue a une modularité finale égale à 0,816, ce qui est une valeur optimale attendue pour ce type de réseau (voir [18]). La Figure 6 (gauche) montre la vitesse de convergence de l'algorithme de recuit simulé et la Figure 6 (droite) est la représentation, par algorithme de forces, du graphe des classes. On y remarque des groupes isolés et des groupes très centraux autour desquels le réseau est structuré en étoile. Ce phénomène était déjà perceptible sur la représentation initiale du graphe mais il est, sur la représentation de la classification, plus évident et lisible, démontrant ainsi la validité et l'utilité d'une telle approche pour l'utilisateur.

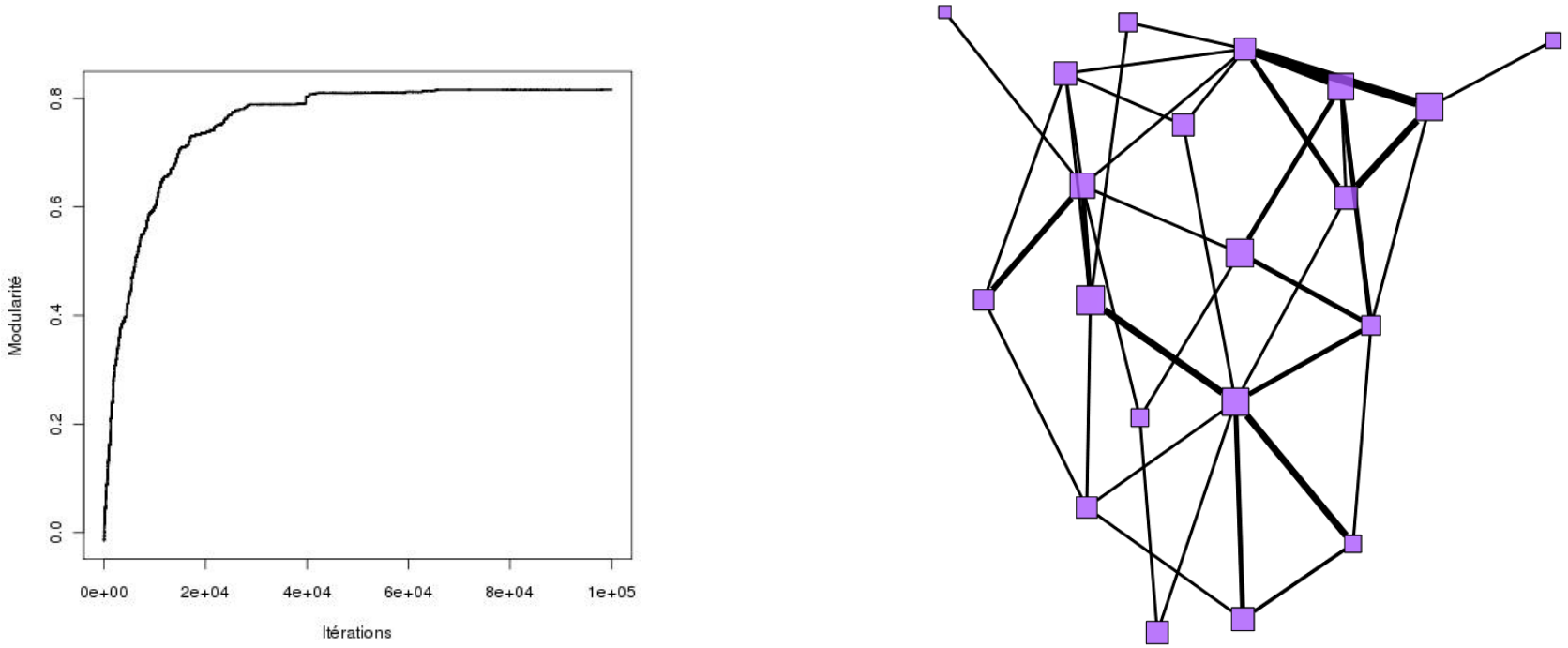

Figure 6 : Convergence de l'algorithme de recuit simulé pour la classification du réseau de collaboration scientifique en 20 groupes (à gauche) et représentation des groupes et de leurs relations par un algorithme de forces (la taille des carrés est proportionnelle à l'effectif de chaque classe ; l'épaisseur des arêtes est proportionnelle au nombre d'arêtes entre les deux classes).

Pour représenter l'intégralité du graphe à partir de l'algorithme décrit dans la partie 3, nous utilisons les positions des classes données dans la Figure 6 par l'algorithme de forces utilisé pour la représentation du graphe des classes.

La Figure 7 est le résultat final de la représentation du graphe clusterisé. On y voit clairement l'avantage de la représentation du graphe entier plutôt que du graphe de classes : les densités locales d'arêtes sont plus facilement identifiables, que ce soit à l'intérieur des classes ou entre celles-ci. Par rapport à la représentation initiale, la représentation peut sembler perdre légèrement en clarté mais elle présente l'avantage de mettre en valeurs des communautés et les relations entre celles-ci. C'est une information d'importance pour l'utilisateur (sociologue, biologiques, …), qui peut ainsi directement visualiser sa classification sur le graphe. L'information donnée par la visualisation directe des sommets permet de comprendre les raisons de la formation des classes et de leur organisation dans le plan.

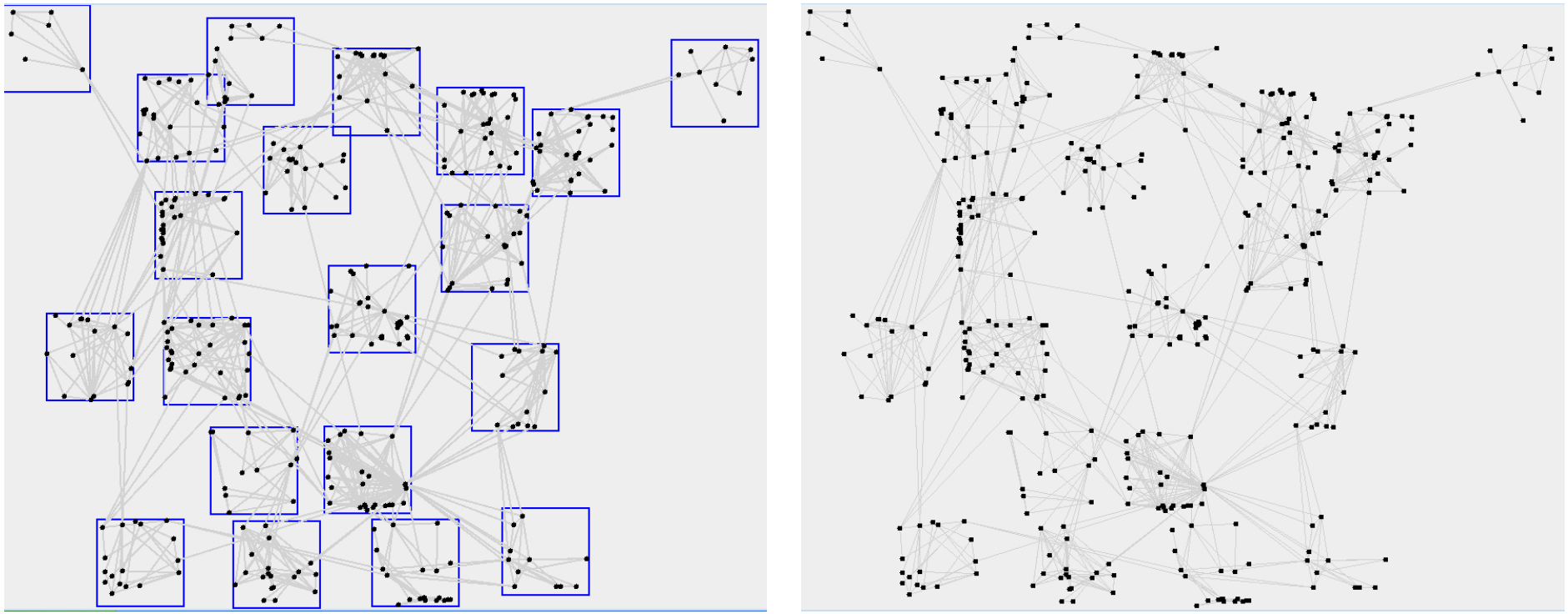

Figure 7 : Représentation du graphe clusterisé : les zones de localisation des clusters sont indiquées (à gauche) ou non (à droite)

5 Conclusions

Le travail présenté ici décrit donc la combinaison d'une méthode de classification et d'une méthode adaptée de visualisation de graphes pour donner une représentation simplifiée et lisible des grandes structures de relations existant dans un graphe. La méthode de classification des sommets, basée sur un algorithme stochastique et l'utilisation d'un critère de qualité spécifique au graphe permettent d'utiliser cette approche pour la représentation de graphes de taille importante. L'application dans des domaines tels que la classification de documents ou de réseaux biologiques est en cours d'investigation : ces domaines sont des challenges importants pour la fouille de graphes par la taille des données qu'elles mettent en œuvre. Enfin, une autre voie d'investigation consiste en l'utilisation, à la place d'une classification, d'un algorithme d'organisation des sommets du graphe (par exemple, les cartes auto-organisatrices de Kohonen) : de tels algorithmes adaptés à des données de type graphes ont été présentés dans [20] et [23]. Une approche similaire à celle que nous présentons ici pourrait être mise en œuvre pour la représentation de graphes, permettant d'éviter la phase de choix de la position des classes (ou de représentation du graphe des classes) en la remplaçant par une carte définie *a priori*. Cependant, l'avantage d'un choix automatique de la position des classes est contrebalancé par la contrainte que représente la forme de la carte qui doit être choisie *a priori* et qui peut être peu adaptée au plongement du graphe étudié.

6 Bibliographie

[1] AUBER D. *Tulip*. In P. Mutzel, M. Jünger, and S. Leipert, editors: 9th International Symposium on Graph Drawing, Lecture Notes in Computer Science, 2265, Springer-Verlag, 2001, p 335-337.

[2] CAPOCCIA A., SERVEDIOA V., CALDARELLIA G. and COLAIORIB F. *Detecting communities in large networks*, Physica A, Statistical Mechanics and its Applications, 352(2-4), 2005, p 669-676.

[3] CHUANG J.H., LIN C.C. and YEN H.C. *Drawing graphs with nonuniform nodes using potential fields*. In Liotta, Giuseppe (Eds.), Proceedings of the Graph Drawing, Perugia, 2004, 460-465.

[4] CONDON A. and KARP R.M. *Algorithms for graph partitioning on the planted partition model*. Random Structures and Algorithms, 18(2), 2001, p 116–140.

[5] DONG Y., ZHUANG Y., CHEN K. and TAI X. *A hierarchical clustering algorithm based on fuzzy graph connectedness*. Fuzzy Sets and Systems, 157(13), 2006, p 1760-1774.

[6] EADES P. *A heuristic for graph drawing.* Congressus Nutnerantiunt, 42, 1984, 149–160.

[7] EADES P and HUANG M.L. *Navigating clustered graphs using force-directed methods*. Journal of Graph Algorithms and Applications, 2000, 157-181.

[8] FRUCHTERMAN T. and REINGOLD E. *Graph drawing by force-directed placement*. Software Practice and Experience, 21, 1991, 1129-1164.

[9] HACHUL S. and JUGER M. *An Experimental Comparison of Fast Algorithms for Drawing General Large Graphs*. Proceedings of the Graph Drawing 2005, 3843, Lecture Notes in Computer Science, Springer-Verlag, 2006, pp. 235-250. Limerick, Ireland.

[10] KAMADA T. and KAWAI S. *An algorithm for drawing general undirected graphs*. Information Processing Letters, 31, 1989, 7-15.

[11] KIRKPATRICK S., GELATT C. D., VECCHI M. P., *Optimization by simulated annealing*. Science, 220(4598), 1983, May, p. 671-680.

[12] LAKROUM S., DEVLAMINCK V., TERRIER P., BIELA ENBERG P. and POSTAIRE J.G. *Clustering of the Poincare vectors.* In IEEE International Conference on Image Processing, 2, 2005.

[13] LEHMANN S. and HANSEN L.K. *Deterministic modularity optimization*. The European Physical Journal B, 60 (1), 2007, p 83-88.

[14] von LUXBURG U. *A tutorial on spectral clustering*. Statistics and Computing, 17, 2007, p 395-416.

[15] HASTINGS W.K. *Monte Carlo sampling methods using Markov chains and their applications*. Biometrika, 57(1):97-109, 1970.

[16] KNUTH D.E. *The Stanford GraphBase: A Platform for Combinatorial Computing*, Addison-Wesley, Reading, MA, 1993.

[17] NEWMAN M. *Finding community structure in networks using the eigenvectors of matrices*. Pysical Review E, 74: 036104, 2006.

[18] NEWMAN M. and GIRVAN M. *Finding and evaluating community structure in networks.* Physical Review E, 69:026113, 2004.

[19] NOACK A. *An Energy Model for Visual Graph Clustering*. In Liotta, Giuseppe (Eds.), Proceedings of the Graph Drawing, Perugia, 2004, 425-436.

[20] ROSSI F., VILLA N. *Topologically ordered graph clustering via deterministic annealing*. To appear in Proceedings of ESANN 2009, Bruges, Belgique.

[21] SCHAEFFER S.E. Graph clustering, Computer Science Review, 1, 2007, p 27-64.

[22] TRUONG Q.D., DKAKI T. and CHARREL P.J. *An energy model for the drawing of clustered graphs*. In Vème colloque international VSST, 21/25 octobre 2007, Marrakech, Maroc.

[23] VILLA N. and ROSSI F. *A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph*. In proceedings of WSOM 2007, Bielefeld, Allemagne, 3/6 Septembre 2007.