Partitionnement et analyse de graphes.

Application à la construction et à la caractérisation d'un réseau de villes.

Pascal Eusebio, David Levy (INSEE PSAR Analyse Territoriale)
Jean Michel Floch (INSEE Département de l'action régionale)

Résumé

Analyser le réseau des villes a nécessité de s'éloigner des méthodes habituellement utilisées à l'Insee et de recourir à des représentations sous forme de graphes. Si les techniques sont inhabituelles, le problèms posé est assez classique : comment réaliser la partition d'une population en sous-populations. Le résultat attendu est de repérer des sous-populations homogènes (faible variance intra-classes) et assez différenciées (forte variance inter-classe). En utilisant les graphes, nous verrons que nous recherchons souvent des partitions qui conservent beaucoup de flux intra-zones et peu entre elles. Les solutions algorithmiques font apparaître des méthodes agglomératives ou divisives selon les cas, que nous pouvons rapprocher des méthodes ascendantes ou descendantes que nous connaissons en analyse des données. Des outils très différents, donc, mais des problèmes que nous avons l'habitude de rencontrer.

Mots clés

Théorie des graphes, réseau, modularité, degré, communauté

Abstract

To study the cities network, Insee needed to find new statistical methods and turned to the graph theory. Even though this method is unusual, the questions raised are rather classic : how to split a population into population groups. Actually, a classical issue in statistics is to find homogeneous (low intra-classes variance) and various (high inter-classes variance) population groups. We will show how the graphs method leads to identify communities with lots of inner flows and few links between each other. Algorithmic solutions can be classified into agglomerative and divisive approaches, that can be compared to ascendent and descendent methods in data mining. In short, innovative methods to deal with usual questions.

Keywords

Graph theory, network, modularity, degree, community

Introduction

L'objet de cette communication est de montrer comment a été construit un outil permettant de fournir, sur des maillages territoriaux divers (aires urbaines, EPCI, communes) à partir de données de flux disponibles (migrations résidentielles, navettes domicile-travail, flux d'établissement), un outil modulable permettant d'isoler au sein du réseau de ville des sous-réseaux ayant une logique interne de fonctionnement. Ce travail résultait de demandes locales de connaissance de l'armature urbaine et de demandes nationales émanant du CGET (DATAR à l'époque).

Pour construire des regroupements de villes, les méthodes de la théorie des réseaux se sont rapidement imposées. En effet, les méthodes pour rendre compte de relations préférentielles entre deux territoires sont très utilisées (flux majeur, intensité de lien ou modèle gravitaire), aucune de ces méthodes ne permet de prendre en compte l'ensemble des flux pour déterminer les relations privilégiées. Peu habituel chez les statisticiens, l'utilisation de graphes était mieux connue chez les géographes. Ainsi, les justifications sur l'intérêt des représentations sous la forme de graphe pour la connaissance et l'appréhension des réseaux de villes sont explicitées dans les travaux des géographes. Cela fera l'objet d'une brève première partie. Pour passer de ces grands graphes (petits cependant par rapport à ceux des réseaux sociaux ou celui du web) à des graphes plus petits et donc plus simples à caractériser, des méthodes de partitionnement doivent être utilisées. Elles feront l'objet de la deuxième partie, qui présentera quelques grands outils de partitionnement et d'analyse des graphes. La dernière partie fournira les principaux résultats obtenus sur le réseau des villes françaises.

1 - Les graphes et l'analyse géographique des réseaux de ville

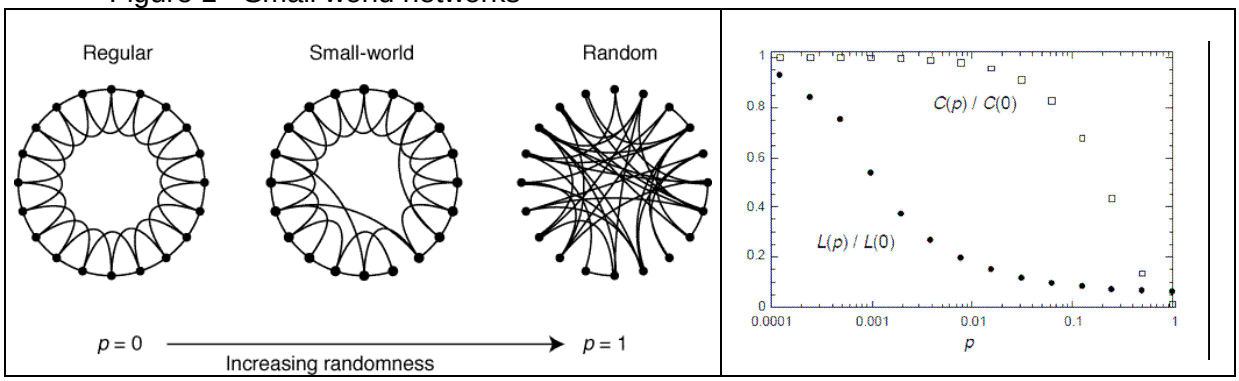
Les géographes se sont intéressés depuis longtemps à l'analyse des relations entre territoires. De nombreux travaux ont porté sur les hiérarchies urbaines. On peut citer parmi les exemples anciens la théorie des lieux centraux de W. Christaller (« *Les Lieux centraux dans le sud de l'Allemagne » 1933*). Les données disponibles et les outils de traitement ont longtemps limité l'analyse des flux. Les modèles gravitaires issus des travaux de Wilson ont constitué une façon simple de modéliser les interactions. C'est avec des développements spécifiques de la théorie des graphes, issus d'autres domaines que celui de la géographie que la situation a été considérablement modifiée (sociologie pour quelques intuitions, physique, informatique). Deux modèles de graphes ont eu une importance particulière : les graphes « petit-monde » (small world networks) et les graphes « invariants d'échelle » (free scale networks).

L'idée du petit monde trouve son origine (lointaine) dans les travaux de Stanley Milgram. L'expérience de Milgram consistait à demander à des habitants du Middle West de faire parvenir une lettre à un destinataire de la Côte Ouest, qu'ils ne connaissent pas, en utilisant comme intermédiaires des personnes de leur entourage. Milgram eu la surprise de constater que la moyenne des chaînes parvenues au destinataire n'était que de 5.6. Cette expérience a permis de confirmer la thèse F. Karinthy (1929) selon laquelle toutes les personnes du globe sont reliées par une chaine d'au plus 5 maillons, devenu dans sa version populaire les six degrés de séparation : en clair, seules cinq personnes nous séparent de n'importe quelle autre personne dans le monde.

Pendant longtemps, les spécialistes des graphes ne s'étaient intéressés qu'aux graphes aléatoires. Dans les années 1990, divers théoriciens des graphes ont proposé des modèles comme le petit monde (small world networks) et l'invariance d'échelle (free scale networks). Ces modèles n'ont pas été sans influence dans l'analyse géographique.

Les graphes de type petit monde ont été proposés par Watts et Strogatz dans un article de la revue nature. On trouve, dans la figure 1, la reproduction du schéma proposé par les deux auteurs pour illustrer la construction du graphe petit monde.

Figure 1 - Small world networks



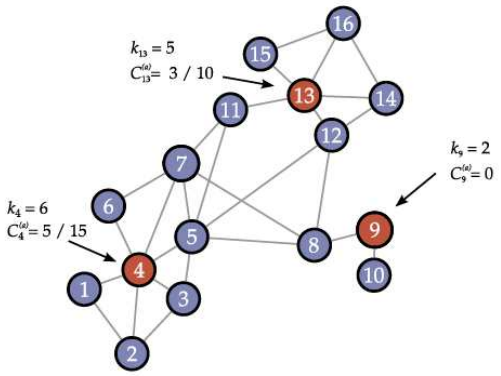
Source : Watts et Strogatz

Le graphe de départ est un graphe qualifié de k-régulier, c'est à dire dans lequel chaque sommet est lié à ses voisins proches, tous les sommets ayant le même degré. L'idée des

auteurs est de présenter une façon simple de transformer ce graphe régulier en graphe aléatoire. A chaque étape, un lien est supprimé de façon aléatoire avec une probabilité p, et on ajoute de la même façon un lien. Le processus est décrit de façon détaillée dans l'article fondateur. Watts et Strogatz proposent deux mesures.

L(p) désigne la longueur moyenne du plus court chemin entre les paires de sommet lorsque p varie. C(p) désigne le coefficient de clustering, dont on trouvera une illustration dans la figure 2. Ce coefficient est en rapport avec la notion de transitivité dans le graphe. L'idée de transitivité peut être traduite de façon simple par le fait que les amis de nos amis sont souvent nos amis. Une forte transitivité dans le graphe se traduit par le fait que, du point de vue topologique, on trouve beaucoup de triangles. Stogatz et Watts ont proposé des coefficients locaux (associés) à chaque nœud du graphe, et un coefficient global, qui est la moyenne arithmétique des coefficients locaux.

Figure 2 - Le coefficient de clustering



$$C(i) = \frac{Nombre\ de\ triangles}{C_k^2}$$

ex : $C(4) = \frac{5}{6 * \frac{5}{2}} = \frac{1}{3}$

Le coefficient du réseau est de 0,5208

Avec k le coefficient local ou le degré du nœud.

Source : Extrait de Young

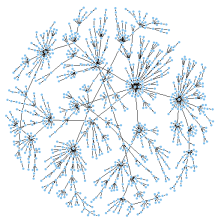
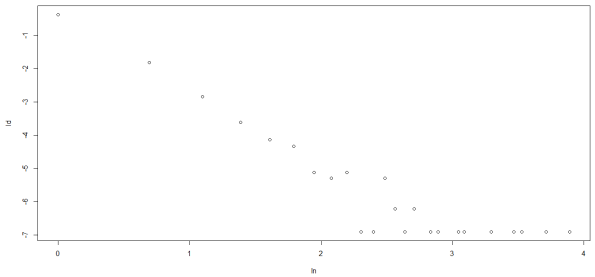
Les valeurs de C(p) et L(p) sont normées par les valeurs C(0) et L(0) correspondant à un graphe régulier. Les deux indicateurs évoluent de façon très différente. La distance moyenne entre les nœuds diminue rapidement tandis que celle du coefficient de clustering (rapport du nombre de triangle sur le nombre de triplets possibles) reste stable un moment et décroît plus rapidement. Watts et Strogatz estimaient que pour des valeurs intermédiaires de p, les réseaux restaient assez hautement structurés, à l'instar des graphes réguliers, mais avec une faible longueur moyenne des chemins, comme dans les graphes aléatoires. C'est ce qu'ils ont qualifiés de graphes « small world », dans une définition qui reste assez largement qualitative (grand nombre de sommets, nombre de liens existant loin de la saturation, degré important de clustering, faible distance moyenne). Des définitions mathématiques plus précises ont été proposées mais elles sont très techniques et dépassent notre propos.

Des réseaux de type petit monde peuvent être générés à l'aide de la commande watts.stogatz.game du package igraph de R.

```
#construction du réseau avec 100 noeuds
g <- watts.strogatz.game(1, 100, 5, 0.05)
#représentation du réseau
plot(g)
```

Un autre ensemble de graphes complexes est celui des graphes invariants d'échelle (scale free networks). Cette modélisation a été proposée initialement par Barabasi et Albert. On peut également générer ce type de graphe sous R avec la commande barabasi.game et on en trouvera une illustration dans la figure 3.

Figure 3 - Scale free networks

	
Exemple de graphe de Barabasi	Évolution de la fréquence du nombre de voisins

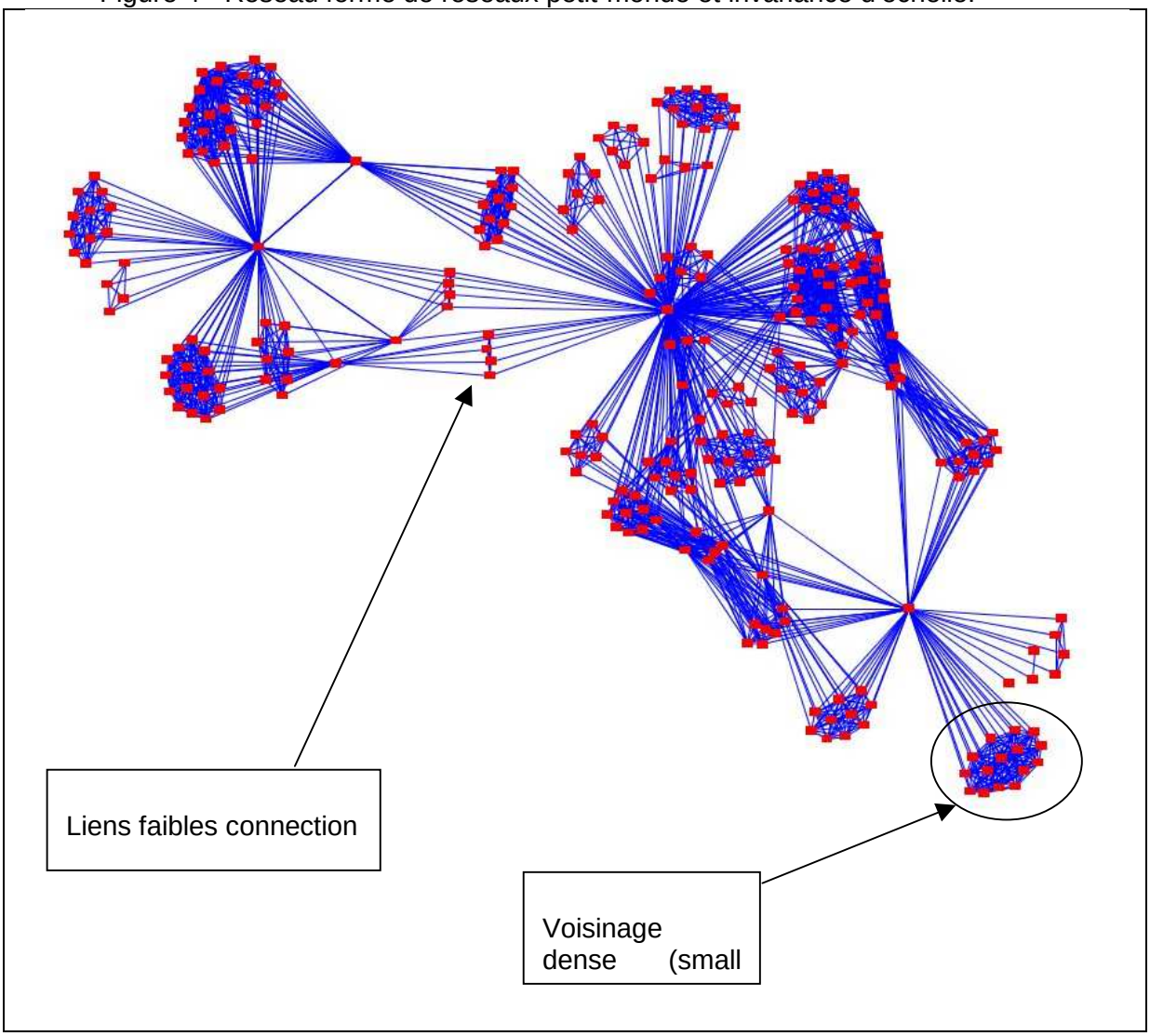
La logique de constitution de ce type de graphe est notablement différente de celle des petits mondes. Ces graphes font apparaître une distribution particulière des degrés qui est du type loi de puissance. Le réseau est dit invariant d'échelle si, lorsque k désigne le degré, P(k) la fréquence des sommets de degré k, l'estimation de la fonction $P(k)=k^{-\gamma}$ fait apparaître une valeur de γ supérieure à 2. Dans l'exemple que l'on présente ci-dessus, la valeur du coefficient γ est de 2,6. On retrouve - puisque l'on s'intéresse ici aux réseaux de ville - des idées qui avaient été émises par les géographes (Loi de Zipf).

Ces deux modèles, décrits ici sommairement n'épuisent pas la description des réseaux complexes. Dans un ouvrage Newman (auteur de plusieurs algorithmes de partitionnement de graphes), Barabasi (introducteur des graphes invariants d'échelle) et Watts (graphes petit monde) montrent que les graphes complexes combinent souvent des caractéristiques des deux types. C'est très net dans les réseaux urbains que nous allons aborder : on rencontre souvent des communautés de villes présentant de fortes interactions (caractéristiques petit monde) tandis qu'au niveau supérieur, les liens entre communautés relèvent plutôt d'une logique invariance d'échelle.

De nombreux travaux ont été menés sur les réseaux de ville. On peut citer ceux de Guimera & alii sur les réseaux de transport aériens, ceux de Ducruet et Rozenblat sur la combinaison des transports aériens et maritimes, ceux de Rozenblat sur les liens géographiques entre les firmes multinationales.

On trouvera ci-dessous un schéma (figure 4) illustrant les emboitements entre logique petit-monde et logique invariance d'échelle.

Figure 4 - Réseau formé de réseaux petit-monde et invariance d'échelle.



Source : C.Rozenblat (2014) -Présentation à la DR de Marseille

Certains auteurs (Beauguitte et Ducruet 2011) relativisent cependant l'apport des deux concepts à la géographie, estimant que celui de petit monde est généralement trivial tandis que celui d'invariance d'échelle est connu depuis longtemps. Par contre, l'utilisation des méthodes de partitionnement, issues de travaux de physiciens ont considérablement enrichi les possibilités d'analyse des réseaux complexes.

2 - Les méthodes de partitionnement de graphes

Ces méthodes sont depuis les années 2000 en plein développement, et il ne peut être question ici que d'en donner une vision introductive, en essayant de l'appuyer sur des intuitions. Elles forment une branche de la théorie des graphes, méthode assez ancienne d'analyse (problème d'Euler sur les ponts de Koenigsberg, problème de la coloration d'une carte, problème du voyageur de commerce). Les notions de la théorie des graphes « classique » ne seront mobilisées que lorsqu'elles seront indispensables et on se centrera sur les concepts spécifiques aux grands graphes et à leur partitionnement.

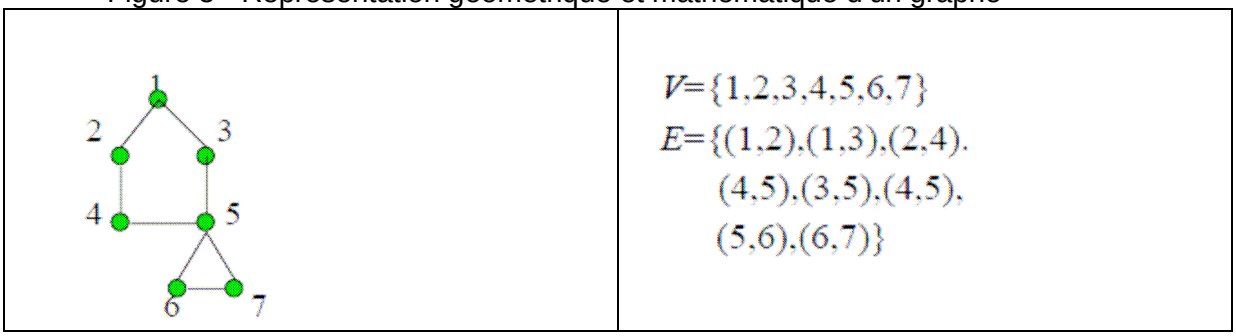
2.1 Notions de théorie des graphes

On appelle graphe un ensemble G={V,E} (figure 5) où V (de vertex) désigne les sommets et E (de edge) les arêtes. La taille du graphe est le nombre de liens, l'ordre le nombre de sommets. Le graphe présenté ci-dessous a pour taille 8 et pour ordre 7. Un graphe est dit vide lorsqu'il ne contient aucun lien, **complet** lorsque tous les sommets sont connectés à tous les autres. Il y a alors n(n-1)/2 liens dans un graphe complet d'ordre n.

Dans cette communication, on se limitera à des graphes non orientés, dans lesquelles les relations entre sommets sont de fait symétriques. Dans de nombreux cas, les résultats sont relatifs à des graphes non valués. Le passage à des graphes valués est plus simple que pour le passage à des graphes orientés.

Le **degré** d'un sommet est le nombre de sommets auquel il est relié. Dans un graphe d'ordre n, il est compris ente 0 et n-1. La séquence des degrés est la suite (d1,…dn).

Figure 5 - Représentation géométrique et mathématique d'un graphe



La **densité** d'un graphe est le rapport du nombre de liens au nombre maximum possible de liens soit le nombre de liens d'un graphe complet c'est à dire n(n-1)/2.

Si la formalisation de la théorie devient rapidement complexe, certains concepts sont assez faciles à appréhender. Comme dans les méthodes de statistique spatiale, on peut associer au graphe une **matrice d'adjacence** (Figure 6 à gauche). Elle vaut 0 sur la diagonale (c'est un graphe simple soit sans boucle) et elle est symétrique puisque issu d'un graphe non orienté.

Figure 6 - Matrices d'adjacence et Laplacienne associé au graphe de la figure 5.

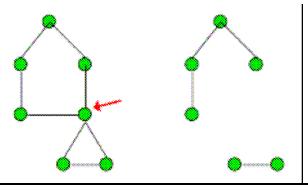
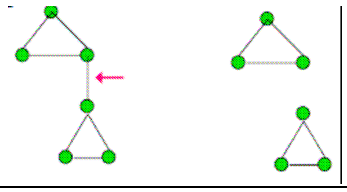
$\begin{matrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{matrix}$	$\begin{matrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{matrix}$	$\begin{matrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 2 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 2 \end{matrix}$
Matrice d'adjacence	Matrice de degrés	Matrice laplacienne

Si on soustrait cette matrice d'adjacence à la matrice des degrés (matrice diagonale), on obtient la **matrice laplacienne** (Figure 6 à droite) qui joue un rôle fondamental dans l'approche qualifiée de spectrale des graphes (méthodes de clustering).

Un **chemin** du sommet a vers le sommet b est une suite ordonnée de sommets dans laquelle chaque paire adjacente est reliée par une arête. Une **géodésique** entre deux points est le chemin de longueur minimale entre ces deux points. (1,3,5,7) est la géodésique entre les points 1 et 7, (1,2,4 ,5,7) étant un chemin et non une géodésique. Un point a est **atteignable** depuis un point b lorsqu'il existe un chemin entre les deux points. Si chaque point d'un graphe est atteignable depuis n'importe quel point, on dit que le graphe est **connecté ou connexe**.

Toutes les questions que l'on se posera par la suite tourneront autour de la possibilité de déterminer au sein de notre graphe des sous-graphes ou des communautés. Cela conduit à s'intéresser aux sommets et aux liens qui jouent un rôle particulier, ainsi qu'aux indicateurs qui permettent de mesurer cela. Les points de coupure et les ponts renvoient respectivement aux nœuds et aux liens dont la suppression diminue la connectivité globale du graphe (figure 7).

Figure 7 - Suppression de nœud ou de lien

	
Point de coupure(nœud)	Pont(lien)

La **connectivité** d'un graphe est le nombre de sommets qu'il faut enlever pour supprimer la propriété connexe du graphe. On définit de façon duale une connectivité de liens qui correspond au nombre de liens à supprimer pour que la connectivité disparaisse.

Les indicateurs de **centralité** jouent un rôle très important dans l'analyse et le partitionnement d'un graphe. Plusieurs ont été définis :

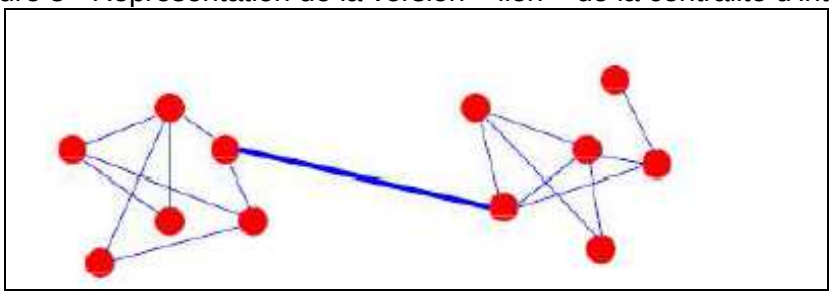
- la **centralité de degré** (degree centrality) est tout simplement le degré, c'est à dire le nombre de liens depuis un sommet. Dans notre exemple, c'est le sommet 5 qui a la plus forte centralité de degré. Cette centralité peut être normée en la rapportant au nombre de sommets moins un. C'est la notion la plus simple. Elle est utilisée fréquemment en sociologie, mais elle ne prend pas en compte la structure du graphe.

- La **centralité de proximité** (closeness centrality) indique si le sommet est situé à proximité de l'ensemble des sommets du graphe et s'il peut rapidement interagir avec ces sommets. Il s'écrit formellement : $C_C(v) = \frac{1}{\sum_{u \in V/\{v\}} d_G(u,v)}$

- La **centralité d'intermédiarité** (betweeness centrality) est un des concepts les plus importants. Il mesure l'utilité du sommet dans la transmission de l'information au sein du réseau. Le sommet joue un rôle central si beaucoup de plus courts chemins allant d'un sommet à un autre doivent emprunter ce sommet. Elle s'écrit : $C_B(v) = \sum_{i,j,i \neq j \neq v} \frac{\sigma_{ij}(v)}{\sigma_{ij}}$

- Il existe aussi une version « lien » de la centralité d'intermédiarité, qui rend compte aussi du nombre de géodésiques (plus court chemin) qui empruntent ce lien. La figure 8 montre un lien (trait foncé) ayant une forte centralité d'intermédiarité. Ainsi la suppression de ce lien conduit à la formation de deux sous-graphes. Cette propriété est utilisée dans le partitionnement de graphe.

Figure 8 - Représentation de la version « lien » de la centralité d'intermédiarité



- La **centralité de vecteur propre** ou centralité spectrale est définie par P. Bonacich à partir de la matrice d'adjacence. Elle correspond pour un sommet à la somme de ses connections avec les autres sommets, pondérée par la centralité de ces sommets.

$$C(v) = \frac{1}{\lambda} \sum_{u \neq v} A(v,u) C(u)$$

qui peut s'écrire

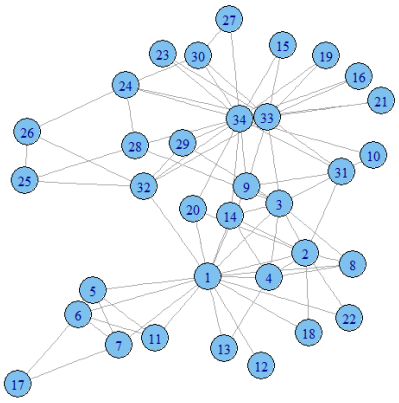
$$\lambda C = AC$$

Pour résoudre cette équation, Bonacich montre que le vecteur de centralité spectrale correspond en fait au vecteur propre dominant (ou principal) de la matrice d'adjacence.

On peut illustrer ces concepts et montrer dans quelle mesure ils diffèrent en utilisant une des bases de données les plus classiques, celle de Zachary sur le club de karaté (figure 9). Le package ***igraph*** du logiciel R permet de représenter le graphe et de calculer les indicateurs précédents.

```
> kar<-read.graph("karate.gml",format="gml")

> plot(kar)
```

figure 9 - Réseau de Zachary



Liens d'amitié entre 34 membres d'un club de karaté dans une Université américaine

```
> ## Centralité de degré

>cd<- degree(kar)

> ## Centralité de proximité

>cp<- closeness(kar)

> ## Centralité d'intermédiarité

> ci<-betweenness(kar)

> ## Centralité de vecteur propre

> ce<-graph.eigen(kar)[c("values", "vectors")]
```

Le tableau ci dessous montre le classement des individus selon les différents critères de centralité. Le classement est assez concordant pour les premiers du classement. Six individus partagent les cinq premières places de chaque indicateur. L'individu 1 est toujours dans les deux premières positions, notamment pour la proximité et l'intermédiarité. Il doit cette position au fait qu'il a un grand nombre de liens (centralité de degré élevé) et qu'il est l'intermédiaire obligé pour un petit groupe d'individus (centralité d'intermédiarité forte) qui sont eux-mêmes peu liés aux autres. Ainsi il est proche de tous les autres membres du club soit une forte centralité de proximité. La centralité de vecteur propre résume ces notions.

Classement pour chaque indicateur	**Degré**	**Proximité**	**Intermédiarité**	**Vecteur propre**
Premier	34	1	1	34
Deuxième	1	3	34	1
Troisième	32	34	33	3
Quatrième	3	32	3	33
Cinquième	2	33	32	2

2.2 Les méthodes de partitionnement

Si l'on revient à nos problèmes de réseaux de ville, on va être confronté à la détermination de sous-ensembles, que l'on appellera en général communauté. Dans le premier chapitre, on a vu que les réseaux de villes combinent souvent des aspects petit monde, avec de forts liens en intra et des aspects invariants d'échelle, avec des sous-groupes assez fortement différenciés. On emploiera pour désigner les éléments des partitions le terme de communauté qui a été importé de l'analyse des réseaux sociaux. On s'appuiera largement dans cette partie sur les synthèses réalisées par Newman et Fortunato, ainsi que sur les thèses francophones de Pons et Seifi.

2.2.1 Définition et qualité d'une partition

Le premier problème du partitionnement de graphes est celui de la définition d'une communauté. Aucune définition n'est universellement acceptée. Ce qui unifie les approches, sans déboucher sur une définition précise, c'est qu'il doit y avoir plus de liens au sein de la communauté que de liens vers le reste du graphe. Cela ne peut se produire que si les graphes sont peu denses, clairsemés (sparse), si le nombre de liens reste dans les mêmes ordres de grandeur que celui des sommets.

Les graphes associés aux réseaux sociaux, ou certains graphes décrivant des structures biologiques atteignent de très grandes tailles, contrairement à ceux que l'on a présentés jusqu'à présent. Le partitionnement de ces graphes en communautés nécessite des algorithmes très performants. Leur nombre est croissant. Ils utilisent des méthodes issues souvent de la physique (méthodes gloutonnes, spinglass).

Comme en classification, on sera confronté aux problèmes d'optimalisation du nombre de communautés, de hiérarchie et d'emboitement.

Les communautés peuvent être appréhendées d'un point de vue local, c'est à dire en faisant le plus possible abstraction du graphe perçu comme un tout. Dans cette perspective, on privilégie les indicateurs qui mesurent de cohésion interne, qu'on pourrait traduire dans le langage des réseaux sociaux par le fait que tout le monde est ami de tout le monde. Dans ces communautés, on doit voir apparaître beaucoup de cliques (sous-graphes maximaux complets comprenant au moins trois sommets). On s'intéresse aussi de ce point de vue à la densité des liens au sein de la communauté et à celle des liens qui la relient au reste du graphe.

Elles peuvent aussi être définies en considérant le graphe comme un tout. Une des idées essentielles est de comparer la structure d'un graphe présentant des communautés à celle d'un graphe aléatoire. Ces graphes, souvent qualifiés de graphes d'Erdos-Renyi ont été les premiers étudiés. Si l'on cherche encore une fois des analogies avec les méthodes statistiques, on cherche un « modèle nul » auquel comparer notre graphe réel. Ce modèle nul doit être un graphe aléatoire, bien sûr, mais qui respecte, pour qu'il soit comparable un certain nombre de contraintes. La version la plus utilisée est celle qui a été proposée par Girvan et Newman. Elle consiste en une version « randomisée » du graphe original, c'est à dire où les liens sont modifiés de façon aléatoire, sous la contrainte que le degré attendu de chaque sommet corresponde à celui du graphe original. Cette approche a permis à ces auteurs de proposer une des notions les plus fécondes en théorie du partitionnement, celle de **modularité**.

La modularité est une mesure de la qualité d'un graphe. Elle permet donc de justifier la pertinence des sous graphe obtenus après un partitionnement. L'hypothèse forte de la modularité est la comparaison avec un graphe aléatoire, ce qui sous-entend qu'un graphe ayant une structure complètement aléatoire doit avoir une modularité proche de 0. Cette comparaison permet donc de mettre en évidence des relations plus denses que la moyenne soit une structure communautaire ou à l'inverse si les relations sont moins denses des structures isolées.

Si l'on considère P la partition en p clusters, soit $P = \{C_1,....,C_k,...C_p\}$ du graphe G=(V,E). La modularité peut être introduite de façon assez simple de la façon suivante, en se référant à l'idée de Newman.

$$Q(P) = \sum_i \left(e_C - a_C{}^2\right)$$

avec e_C la part des liens d'un cluster C_i sur le total

et a_C la probabilité qu'un sommet se trouve dans le cluster C_i

et donc $a_C{}^2$ la probabilité que les deux sommets d'un lien se trouve dans le même cluster.

Cette expression générale est transformée dans la première forme usuelle de présentation de la modularité. $a_C{}^2$ est la probabilité que deux sommets liés sous la contrainte de conservation des degrés de sommet appartiennent à la même partition. On montre alors (Fortunato) que cette probabilité peut s'écrire sous la forme $a_c{}^2 = \frac{d_i d_j}{4m^2}$.

Et

$$Q(P) = \frac{1}{2m} \sum_{i,j \in V} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(C_i, C_j).$$

où A désigne la matrice d'adjacence du graphe et A_{ij} le poids des liens entre les sommets i et j.

d_i est la somme des degré de i

soit $d_i = \sum_j A_{ij}$

$\delta(C_i, C_j)$ une fonction de Kronecker qui vaut 1 si les deux sommets appartiennent à la même communauté et 0 sinon.

On peut montrer qu'une façon alternative d'écrire cette expression est la suivante :

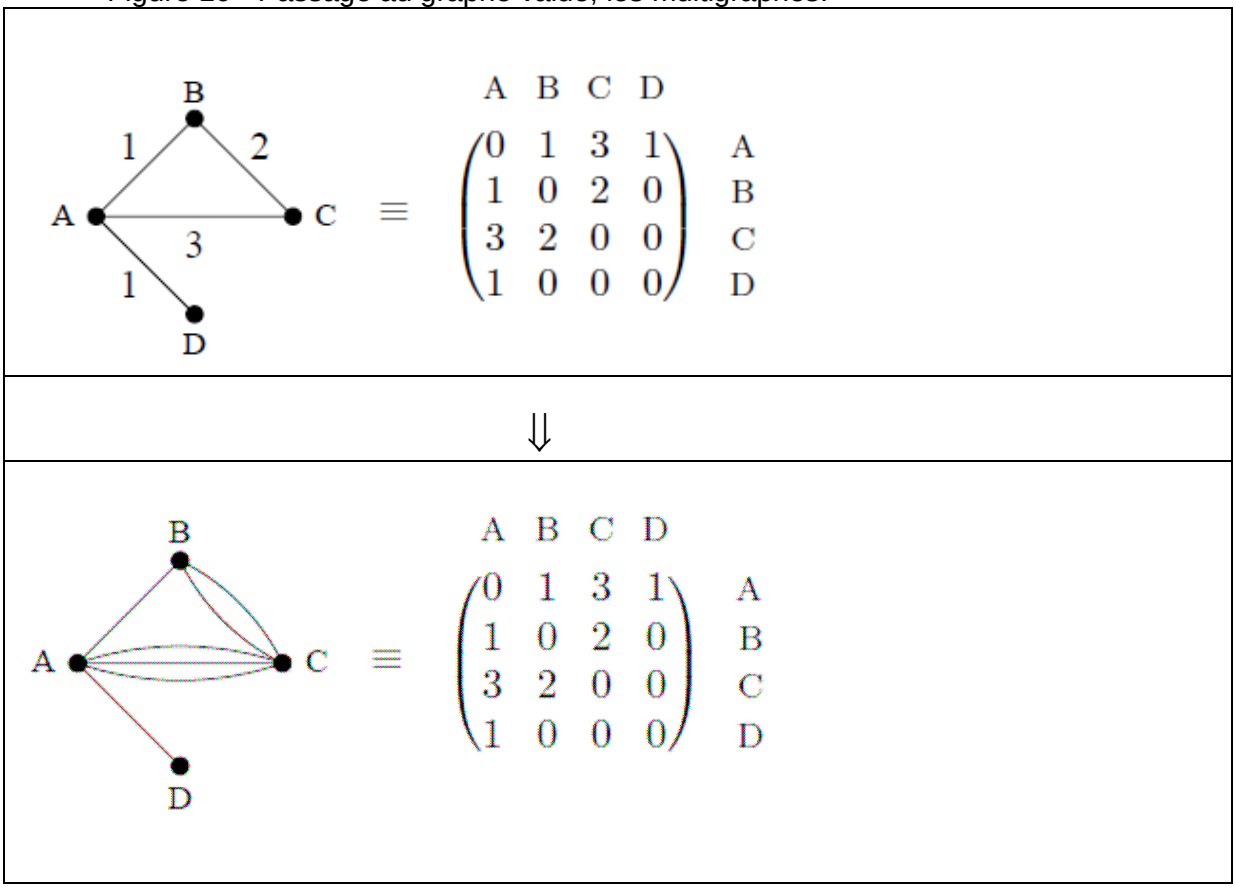
$$Q = \frac{1}{2m} \sum_{k=1} \sum_{i,j \in Cp} \left(A_{ij} - \frac{d_i d_j}{2m} \right) = \sum_{k=1}^{p} \left[\frac{l_k}{m} - \left(\frac{d_k}{2m} \right)^2 \right]$$

l_k désignant le nombre de liens joignant les sommets de la communauté k, d_k la somme des degrés de la communauté k.

Le terme $A_{ij} - \frac{d_i d_j}{2m}$ correspond à la différence de liens entre notre graphe et un graphe aléatoire dont la contrainte est la conservation des degrés de sommets.

Les définitions de la modularité ont d'abord été développées dans le contexte de graphes non valués. Ils ont été étendus aux graphes valués. Dans un graphe valué, la valeur de A_{ij} correspondant au lien entre les sommets, et qui vaut 1 les sommets sont liés, et 0 sinon, est à présent la valeur du flux si il y a un lien, la valeur 0 sinon. Dans un article de 2004, Newman donne une façon très simple de passer des graphes non valués aux graphes valués, en introduisant ce qu'il appelait des multigraphes (figure 10).

Figure 10 - Passage au graphe valué, les multigraphes.



Cette représentation permet de généraliser aux graphes pondérés les résultats présentés précédemment. Les Aij correspondent aux poids associés aux liens ou de façon équivalente au nombre de liens du multigraphe. M est le nombre de liens du multigraphe, ou la somme des pondérations.

La modularité est un des concepts les plus puissants de la théorie des partitions de graphe, et malgré les critiques émises à son encontre le plus utilisé. Il est utilisé comme fondement de certaines méthodes, et comme mesure de la qualité de partitions produites par d'autres méthodes. On l'utilisera à plusieurs reprises dans les exemples que l'on donnera..

Les travaux (Guimeras, Reichart et Bornholdt) ont montré que des partitionnement conduisant à des valeurs significatives de la modularité pouvaient être mis en évidence dans des graphes aléatoires. Une modularité importante n'est donc pas toujours le signe d'une structure modulaire. Newman a proposé quelques modifications qui sont lourdes en calcul. Le deuxième problème, mis en avant par Fortunato est un problème de « résolution ». Si le nombre de liens dans le graphe devient très grand et que le nombre de liens attendu (cf. formule de la modularité) est inférieur à 1, un seul lien entre les deux groupes suffit à entrainer leur fusion.

2.2.2 Panorama général des méthodes de partition

Une fois défini le schéma général d'une partition, reste à la réaliser pratiquement. Pratiquement dans ce cas implique de trouver des façons de faire, des algorithmes donc, qui permettent d'une part de résoudre le problème, et ensuite de le résoudre dans un temps acceptable. Les graphes des réseaux de ville sont déjà conséquents mais restent très petits si on les compare à ceux des réseaux sociaux ou même à ceux qui sont utilisés dans l'étude des protéines ou du génome. La complexité des algorithmes (problèmes NP difficiles ou Np complets) peut être trouvée dans Fortunato. On cherche souvent à mesurer la complexité des algorithmes en les notant $O(n^a m^b)$.

On va retrouver dans les méthodes des questions bien connues en analyse des données : combien de classes ?, doit-on les déterminer au préalable ?, doit-on appliquer des méthodes ascendantes ou descendantes ?, comment déterminer des critères d'arrêt ?

On se limitera ici à la présentation de quelques familles de méthodes testées dans le cadre des travaux du Psar-AT, en se centrant sur les méthodes qui sont implémentées dans le logiciel R. Les méthodes sont en pleine expansion, font l'objet de controverses au sein des spécialistes. Beaucoup de celles qui sont présentées ici sont issues des travaux de Mark Newman, introducteur entre autre de la notion de modularité présentée dans le paragraphe précédent.

La complexité algorithmique des questions a fait que beaucoup de travaux initiaux ont porté sur la bipartition des graphes (travaux de Kernighan et Lin notamment). D'autres méthodes s'inspiraient aussi de ce qui était fait en analyse des données (dendrogrammes de classification, méthodes de type k-mean)

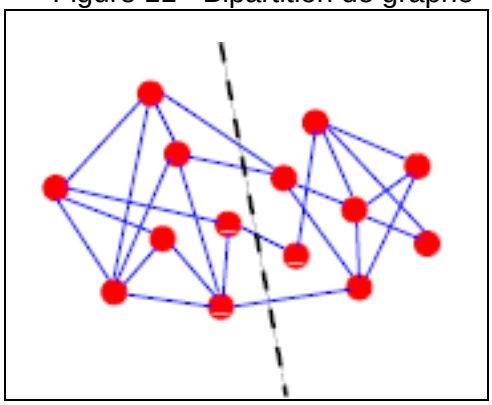
Ces méthodes reposeront sur des propriétés des graphes, ou sur le traitement de la matrice d'adjacence.

2.2.3 Méthodes classiques

On ne présentera que quelques méthodes :

Les **méthodes fondées sur la bissection de graphes** (figure 11). Elles sont assez simples à présenter. L'idée est de chercher la ligne qui partage le graphe en coupant le moins de lien (cut size)

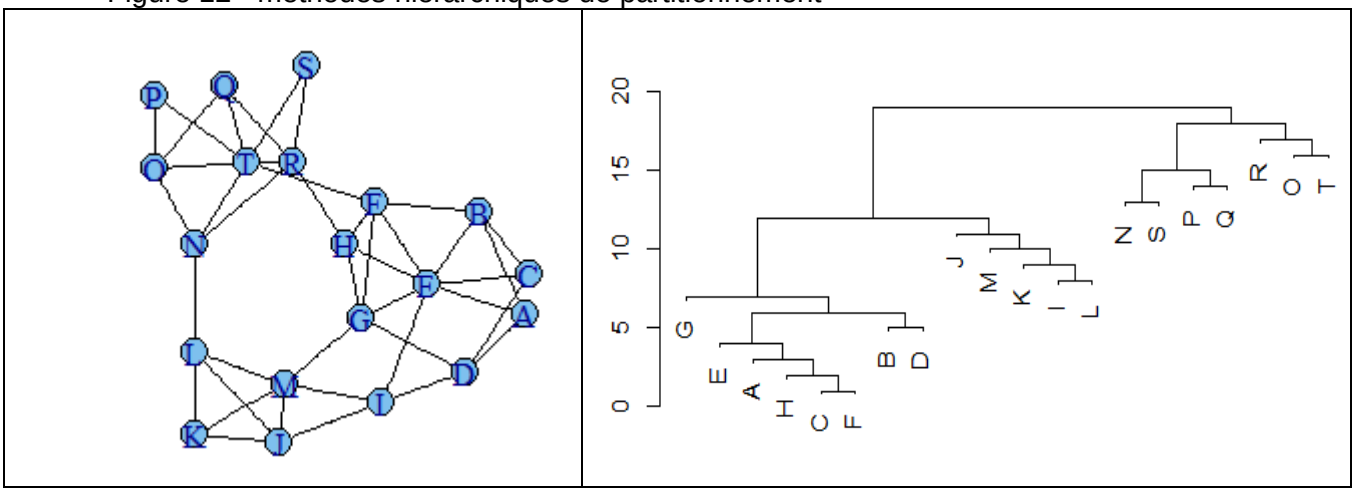
Figure 11 - Bipartition de graphe



Cette méthode dans sa version la plus simple risque cependant de ne faire apparaître que des solutions triviales (un sommet isolé). Des méthodes plus élaborées de bissection reposent sur des méthodes spectrales (propriétés du spectre de la matrice laplacienne) que l'on présentera en (2.2.5)

Les méthodes hiérarchiques (figure 12) : elles reposent sur des mesures de similarité entre les sommets. Lorsqu'on a calculé cette similarité pour chaque paire de sommets (matrice de similarité), on peut construire par exemple un dendrogramme par des méthodes assez classiques.

Figure 12 - méthodes hiérarchiques de partitionnement



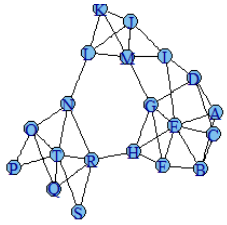
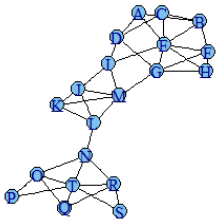
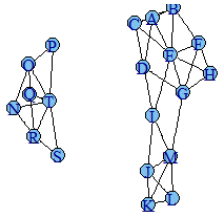
Les méthodes de clustering : elles sont bien connues en analyse des données. Dans ces méthodes, le nombre de classes est prédéterminé. On définit une distance entre couples de points, d'autant plus grandes que les sommets sont dissemblables. On va chercher à minimiser une fonction de coût basé sur les points et les centroïdes. Dans le minimum-kmean-clustering, par exemple la fonction de coût est la plus grande distance entre deux points de la classe. On cherche à trouver la partition qui rende minimale la plus grande des k classes (recherche de classes compactes). La méthode de MacQueen repose elle sur la minimisation du total des distances intraclasses.

2.2.4 La méthode divisive

Cette méthode est une des plus intuitive à présenter. Elle repose sur le concept présenté en 2.1 de centralité d'intermédiarité, avec un schéma qui expose assez bien dans un cas simple cette idée. Lorsque beaucoup de géodésiques allant d'un point quelconque du graphe à un autre passe par un sommet ou par un lien, la suppression de ceux ci est plus à même de faire apparaître des communautés.

Dans l'exemple présenté plus haut (figure 12), c'est le lien entre les sommets T et F qui a la plus forte centralité d'intermédiarité. Si on supprime ce lien, c'est le lien RH qui a alors la plus forte centralité, puis le lien NL

Figure 13 - Partitionnement d'un graphe avec la méthode divisive

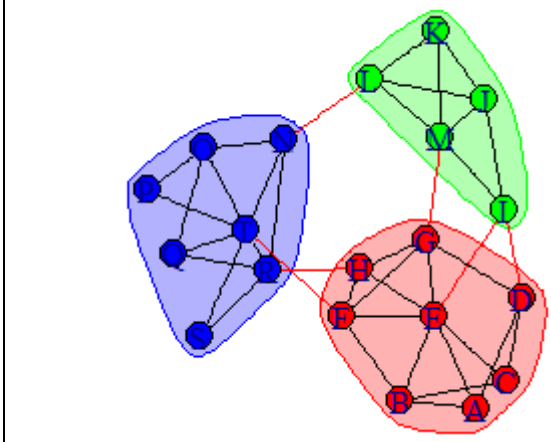
		
Suppression du lien T-F	Suppression du lien R-H	Suppression du lien N-L

Après la suppression de ces trois liens, le graphe n'est plus connexe et une communauté apparaît. Le processus peut se poursuivre.

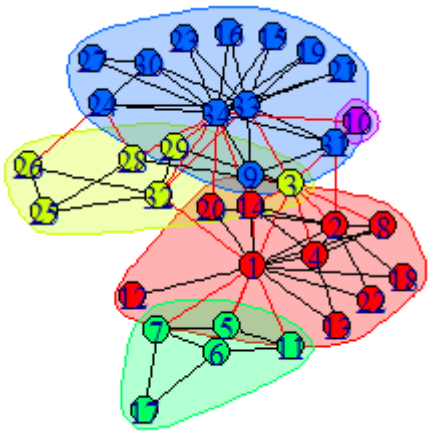
Une commande de R produit le résultat final .

```
> karate<-read.graph("karate.gml",format="gml")

> plot(karate,vertex.size=2)

> betkar<-edge.betweenness.community(karate)

> plot(betkar,karate)
```

Figure 14 - Résultat de la méthode divisive sur un graphe et sur le club de karaté



Graphe de démonstration
3 communautés



Club de Karaté
5 communautés

Le résultat sur ce graphe très simple est assez trivial et on peut voir ce qu'il produit sur un graphe encore lisible mais plus complexe comme celui du club de karaté.

La méthode divisive la plus connue est celle de Girvan-Newman. Elle marque d'ailleurs l'entrée des physiciens dans le domaine de l'étude des graphes. L'algorithme illustré précédemment est le suivant :

-1 Calcul de la centralité d'intermédiarité pour tous les liens

-2 suppression du lien ayant la plus forte centralité

-3 Recalcul de la centralité

-4 Itération du cycle à l'étape 2

Ce processus itératif peut se poursuivre jusqu'à l'isolement de tous les sommets et produit ainsi une hiérarchie de partitions emboitées. Le choix de la partition peut se faire à partir du critère de modularité. Cet algorithme nécessite à chaque étape le calcul des centralités d'intermédiarité et sa complexité est en $O(m^2n)$ ce qui le rend inexploitable sur de très grands graphes.

D'autres algorithmes divisifs ont été proposés. Fortunato a proposé un algortihme qui utilise la centralité d'information plus performant, mais de complexité plus grande que celui de Girvan-Newman. Ce dernier reste donc très utilisé, notamment à titre de comparaison des communautés détectées.

2.2.4 Les méthodes agglomératives fondées sur la modularité

Cette famille de méthode est très riche et très importante. Au contraire de la précédente, on part de l'ensemble des sommets, que l'on va progressivement agréger entre eux.

La méthode « optimale »

Elle repose sur l'exploration de toutes les communautés possibles et sur la maximisation de la modularité. On peut trouver dans Fortunato une valeur approchée du nombre de ces partitions, nombre qui explose avec la taille du graphe et la rend inexploitable, même pour des graphes de taille moyenne. Le calcul des communautés dans cette optique utilise une méthode issue de la physique appelée « recuit simulé » (simulated annealing) souvent utilisée dans les problèmes d'optimisation.

Elle est implémentée en R dans le package igraph par la commande optimal.community.

La méthode de Clauset et Newman

C'est un algorithme qualifié de « glouton » qui permet la constitution d'une partition à partir d'un critère de modularité. Il a d'abord été proposé par Newman en 2003 puis par Clauset, Newman et Moore dans une deuxième version. Il utilise la modularité sous la forme suivante : $Q = \sum_i (e_i - a_i^2)$. On définit une grandeur notée ΔQ_{ij} correspondant à la variation de modularité lorsqu'on fait un lien entre la communauté i et la communauté j. Le détail de l'algorithme, avec les indications liées au stockage de l'information peuvent être trouvées dans Clauset & alii 2004. Le schéma général est le suivant :

1- On part de n communautés (chaque sommet étant une communauté)

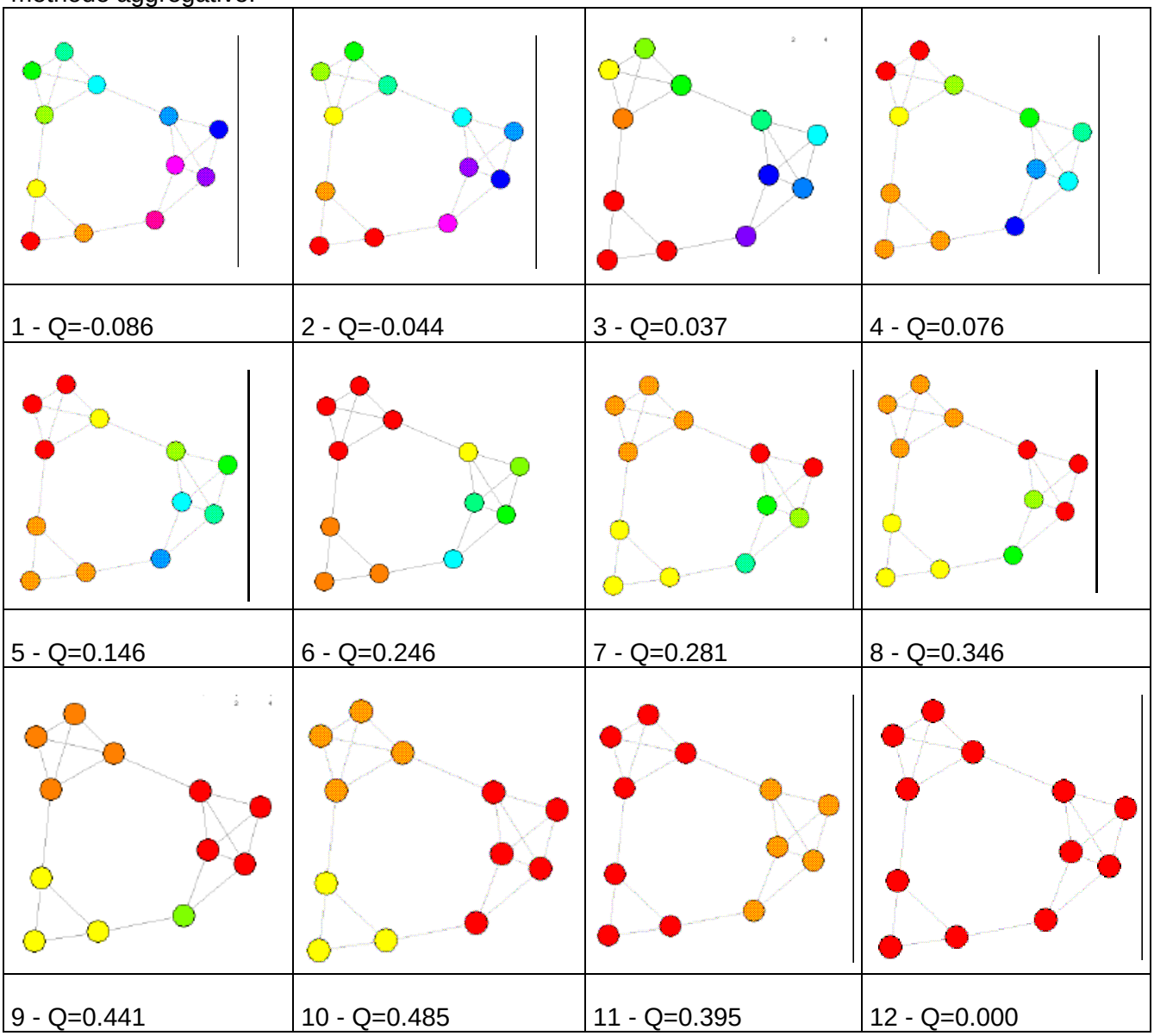
2- On calcule ΔQ_{ij} pour toutes les paires

3- On fusionne les paires qui accroissent le plus la modularité

4- On répète les phases 2 et 3 jusqu'à ce qu'on obtienne une seule communauté

5- On coupe le dendrogramme à la valeur correspondant à la plus forte modularité

Figure 15 - Les 12 étapes du partitionnement d'un graphe à 12 sommets avec la méthode aggrégative.



Dans cet exemple très simple, on peut voir que la modularité augmente jusqu'à l'étape 10 où les 3 communautés assez visibles sont identifiées. A l'étape 11, deux des communautés fusionnent et la modularité diminue, celle-ci devenant nulle lorsque les trois communautés sont regroupées. Le résultats est donc un partitionnement en 3 communautés avec une modularité de 0,485.

Cet algorithme est implémenté en R dans le package **igraph** par la fonction **fastgreedy.community**. La caractéristique de cet algorithme est sa grande vitesse d'exécution qui lui permet de traiter de grands graphes. Des critiques ont été exprimées sur ses résultats. L'algorithme est de complexité O(mn).

L'utilisation de méthodes spectrales

Newman a proposé une version spectrale du partitionnement fondé sur la modularité. Dans cette version, on introduit une matrice qui fait apparaître l'expression de la modularité :

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

Dans le cas initial d'une bipartition, généralisé ultérieurement, Newman introduisait un vecteur s valant +1 si le sommet appartenait au premier groupe, -1 au second. Il montre que la maximisation de la modularité en fonction du vecteur s se ramène à un problème que l'on peut formaliser par : $Bs = \lambda Ds$ dans lequel λ est un multiplicateur de Lagrange, et D une matrice diagonale contenant les degrés des sommets.

Lorsqu'on résout ce problème matriciel, compte tenu de la structure de la matrice sur laquelle on travaille, on obtient une solution triviale avec une valeur propre égal à 0 et un vecteur composé de 1, soit le regroupement de tous les sommets dans une seule communauté. Pour effectuer la partition, on utilise le vecteur propre associé à la deuxième plus grande valeur propre (Newman 2006).

On trouve dans le package **igraph** la fonction **leading.eigenvector.community** qui met en œuvre cette méthode.

Algorithme de Louvain

En 2008, trois chercheurs de l'université de Louvain ont proposé une autre méthode « gloutonne », plus rapide que la majorité des autres approches. Sa particularité est de se baser sur une approche locale de la modularité.

Dans une première phase, une communauté différente est attribuée à chaque sommet. On s'intéresse ensuite aux voisins de chaque sommet i, et on calcule le gain de modularité en retirant le sommet i et en le plaçant dans la communauté j. On recherche un gain positif et maximum pour déplacer i. On effectue cette opération de façon séquentielle jusqu'à ce qu'aucune amélioration ne soit possible. On trouvera dans Blundel l'expression du gain en modularité qui est utilisé dans l'algorithme.

La deuxième phase de l'algorithme consiste en la construction d'un nouveau réseau dont les sommets sont les communautés repérées dans la première phase, les poids des liens entre les communautés étant déterminés par la somme des poids des liens des sommets du graphe initial.

Une fois cette deuxième phase terminée, on réapplique l'algorithme à ce nouveau réseau pondéré. Une combinaison des deux phases est une « passe », et ces passes sont itérées jusqu'à ce qu'un maximum de modularité soit atteint.

On trouve dans le package **igraph** la fonction **multilevel.community** qui met en œuvre cette méthode.

2.2.5 Autres méthodes

Marches aléatoires(Walktrap)

L'algorithme vise au final, comme tous les autres à produire des distances entre les sommets du graphe. L'idée est d'aboutir à cette distance en se fondant sur l'idée de marche aléatoire. Le temps est discrétisé. A chaque instant, un marcheur se déplace aléatoirement d'un sommet vers un sommet choisi parmi ses voisins. La suite des sommets visités est alors une marche aléatoire. La probabilité d'aller du sommet i au sommet j est $P_{ij} = \frac{A_{ij}}{k_i}$.

On a ainsi la matrice de transition de la chaine de Markov correspondante, et on peut calculer la probabilité de passer du sommet i au sommet j en un temps t, $P_{ij}(t)$.

Lors d'une marche aléatoire suffisamment longue dans un graphe, la probabilité de se trouver sur un sommet donné est directement (et uniquement) proportionnelle au degré de ce sommet.

La probabilité d'aller de i à j et celle d'aller de j à i par une marche aléatoire de longueur fixée ont un rapport de proportionnalité qui ne dépend que des degrés des sommets de départ et d'arrivée.

$$k_i P_{ij}(t) = k_j P_{ji}(t)$$

La façon de comparer deux sommets i et j doit s'appuyer sur les constatations suivantes :

Si deux sommets ***i*** et ***j*** sont dans une même communauté, la probabilité $P_{ij}(t)$ est certainement élevée.

Par contre si $P_{ij}(t)$ est élevée, il n'est pas toujours garanti que ***i*** et ***j*** soient dans la même communauté.

La probabilité $P_{ij}(t)$ est influencée par le degré k_j du sommet d'arrivé : les marches aléatoires ont plus de chances de passer par les sommets de fort degré (dans le cas limite d'une marche aléatoire infinie, cette probabilité est proportionnelle au degré).

Les sommets d'une même communauté ont tendance à voir les sommets éloignes de la même façon, ainsi si ***i*** et ***j*** sont dans la même communauté et ***k*** dans une autre communauté il y a de fortes chances que $P_{ik}(t) = P_{jk}(t)$.

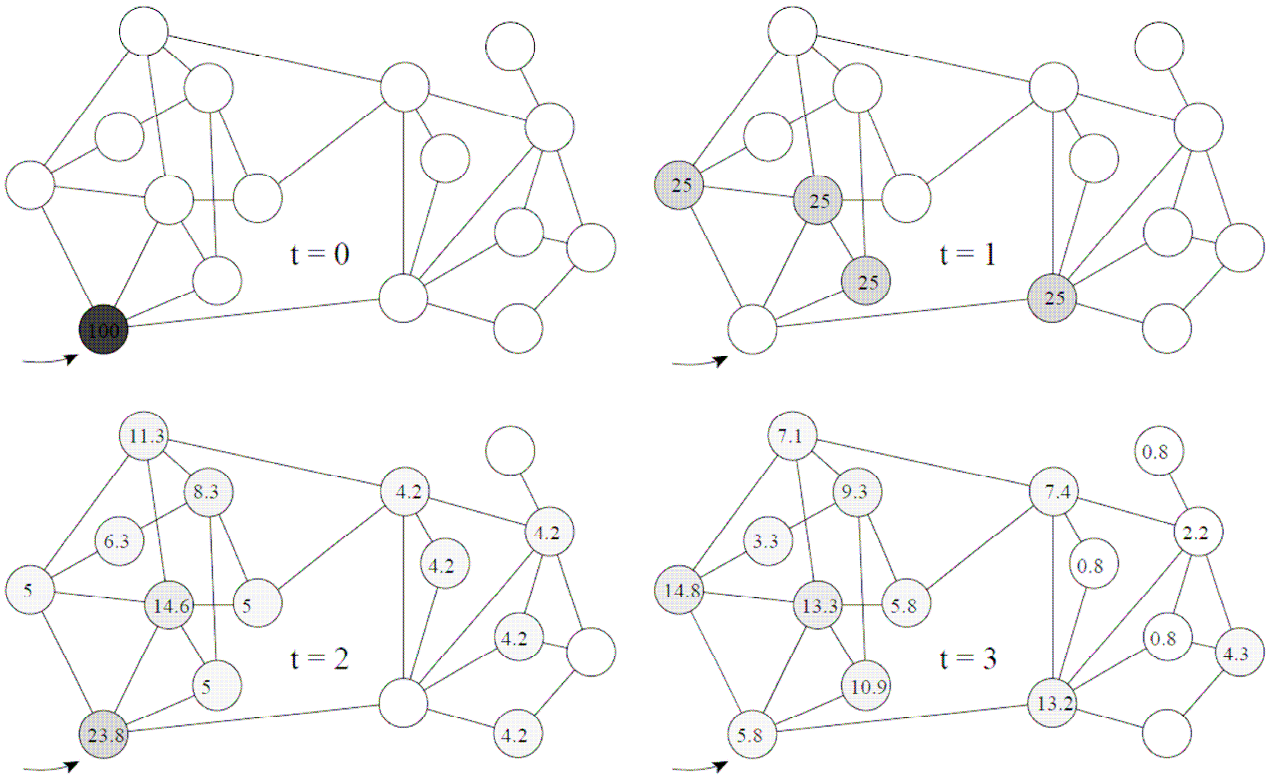
On définit ainsi une distance, qui doit être plus faible lorsque les deux sommets appartiennent à la même communauté.

$$t_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{ik}(t) - P_{jk}(t)\right)_2}{k_k}}$$

Dans cette méthode, le choix de t est très important. Si t est trop petit, les communautés sont minuscules. S'il est trop grand, les probabilités tendent vers la même valeur.

Une fois déterminée la matrice de distance, l'algorithme est assez classique : on part de n communautés et on agrège ensuite. On obtient un arbre et on utilise la modularité pour trouver la partition adaptée. On trouvera les détails dans Pons.

Figure 16 - Illustration de la marche aléatoire sur un graphe



Dans l'exemple de la figure 16, on a, jusqu'à t=3, représenté graphiquement la matrice de probabilité qui sera utilisée pour faire le partitionnement (par analyse spectrale).

On trouve dans le package **igraph** la fonction **walktrap.community** qui met en œuvre cette méthode.

Verres de spin

Avec cette méthode, on s'éloigne des méthodes usuelles. Elle s'inspire des verres de spin, qui sont des alliages correspondant à des impuretés, un spin étant associé à chaque impureté. Le couplage entre les différents spins peut être plus ou moins intense. Cette méthode est utilisée en physique théorique. Les paires de spins sont associées dans un graphe. On définit un hamiltonien et une distribution de probabilité des couplages.

Potts, puis Richart et Bornholdt ont utilisé cette approche. Chaque sommet est caractérisé par un spin prenant q valeurs possibles, et les communautés correspondent aux valeurs de sommets ayant des valeurs de spin égales. On définit l'énergie du système (un hamiltonien faisant intervenir la matrice d'adjacence du graphe). La minimisation de cette expression se fait par recuit simulé.

On trouve dans le package **igraph** la fonction **spinglass.community** qui met en œuvre cette méthode.

2.3 Comparaison des méthodes

Comme on a pu le voir ces méthodes sont assez diverses et aucune ne fait l'unanimité. On présente ci-dessus les résultats obtenus sur un cas simple, celui du club de karaté.

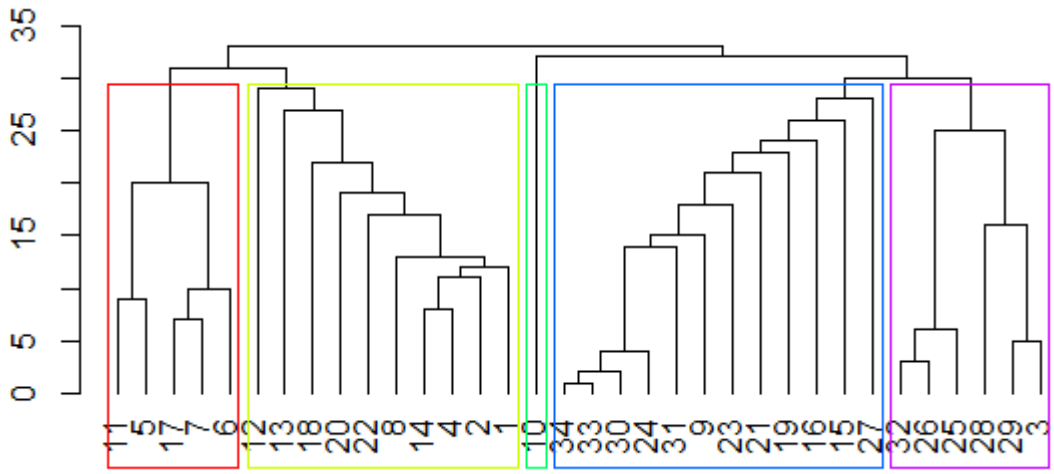
On utilise tous les algorithmes de détection des communautés proposés par le package igraph de R.

Des commandes de ce type ont été passées pour toutes les fonctions déterminant des algorithmes de partitionnement de graphe.

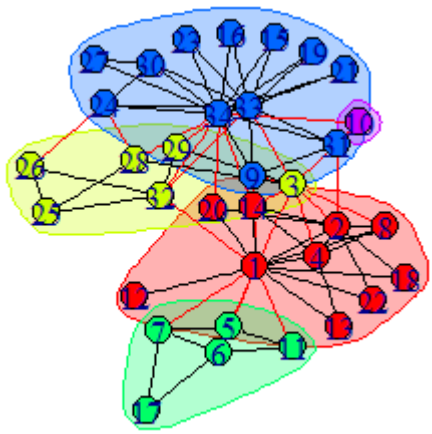
```
> betkar<-edge.betweenness.community(karate)

> plot(betkar,karate)

> membership(betkar)

 [1] 1 1 2 1 3 3 3 1 4 5 3 1 1 1 4 4 3 1 4 1 4 1 4 4 2 2 4 2 2 4 4 2 4 4

> modularity(betkar)

[1] 0.4012985

> dendPlot(betkar)
```

On obtient la valeur de la modularité et le résultat de la partition. On peut réaliser des cartes et des dendrogrammes. On trouvera ci-dessous les résultats obtenus grâce à l'algorithme de Girvan-Newman

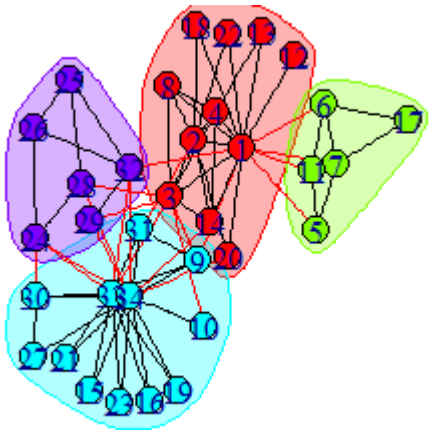
Figure 17 - Résultat du partitionnement et modularité de différents algorithmes



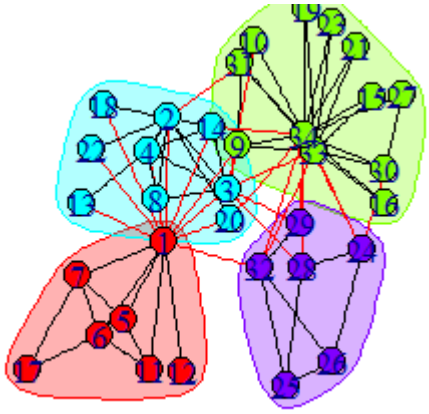
La carte est présentée ci-dessous



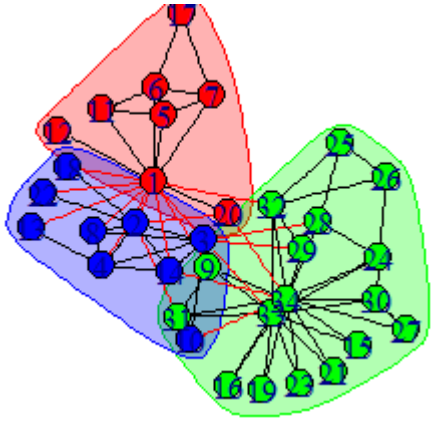
Edge.betweenness **Modularité=0.401**



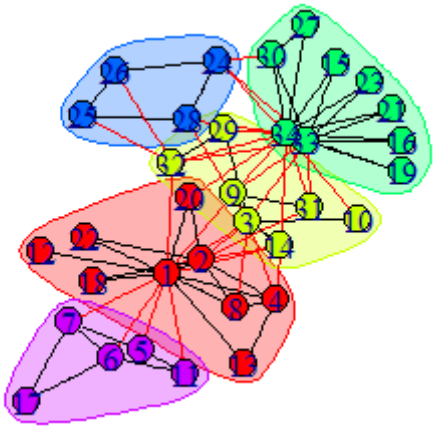
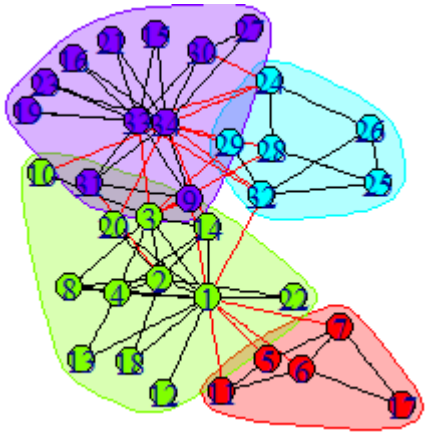
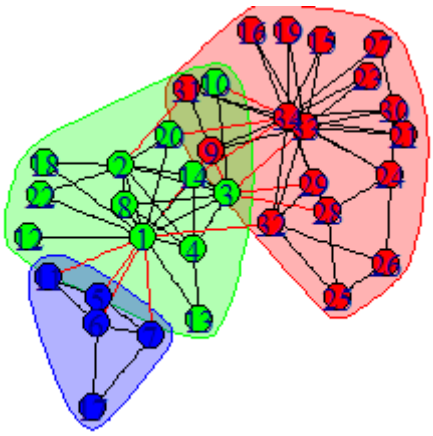
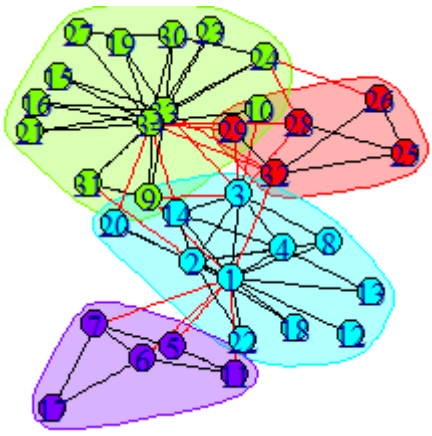
Optimal **Modularité=0.420**



Leading.eigenvector **Modularité=0.393**



Fastgreedy **Modularité=0.381**

	
Walktrap **Modularité=0.353**	Multilevel **Modularité=0.420**
	
Infomap **Modularité=0.402**	Spinglass **Modularité=0.418**

On n'a pas présenté les résultats du graphique issu de la commande label.propagation qui ne distinguait dans ce cas aucune structure communautaire.

Relativement au critère de modularité, les algorithmes de Louvain, ou le spinglass donnent de bons résultats et le fastgreedy apparaît peu performant. Louvain et spinglass font apparaître le même nombre de classes et répartissent les individus de la même façon.

3 Application à l'identification d'un réseau de villes

Un réseau de villes représente un enjeu territorial et politique important car il est le témoin d'une coopération entre des ensembles urbains. La politique d'aménagement du territoire tend à mettre les métropoles au coeur du développement des territoires. Considérées une par une, chaque grande ville, ou métropole, est en elle-même un système territorial complexe. Certaines agglomérations métropolitaines (Bordeaux, Toulouse) poursuivent leur croissance sur le mode de l'aire centrée. D'autres, parce qu'elles sont plus proches les unes des autres, ou parce qu'elles côtoient de grosses villes moyennes, tendent plutôt à poursuivre leur promotion métropolitaine sur le mode de la grappe, ou d'alliances de villes. La France est un pays de villes moyennes. Ces villes jouent un rôle essentiel dans l'armature urbaine et le maillage du territoire national. Malgré leur grande diversité, elles ont toutes la particularité de faire le lien entre les métropoles d'un côté et un arrière-pays périurbain et rural auquel elles sont fonctionnellement, culturellement et historiquement attachées. Tendre vers un optimum de cohésion sociale et territoriale est également un enjeu clé. Car si la métropole est un lieu de développement et de rayonnement économique, culturel et décisionnel, les enjeux sont aussi de maintenir un cadre de vie agréable pour les habitants, de protéger l'environnement et de préserver la cohésion sociale.

3.1 Présentations des données

L'étude des flux et des interactions spatiales permet de mettre en évidence le fonctionnement des territoires. Parmi ces flux, les déplacements domicile-travail jouent un rôle particulièrement important parce qu'ils sont structurants et qu'ils ne connaissent que très peu de déformation au cours du temps. De tels déplacements, parce qu'ils sont quotidiens, alternants et très nombreux, participent grandement à l'évolution de la géographie locale des territoires, même s'ils ne représentent qu'un tiers des déplacements quotidiens. Les flux utilisés sont issus du recensement de la population de 2011, le champ est réduit à la France métropolitaine.

Par définition, l'aire urbaine constitue un ensemble cohérent au regard de ces déplacements. Le graphe est ici constitué des aires urbaines qui forment des nœuds. Les liens entre ces nœuds représentent les flux domicile-travail. On cherche à décomposer ce graphes en communautés d'aires urbaines en répondant à une préoccupation : maximiser les liens interne à chaque communauté et minimiser les liens entre communautés. L'algorithme procède par étapes d'agrégations, à chaque étape on contrôle la qualité de la partition obtenue par la modularité

Deux indicateurs permettant de caractériser le résultat obtenu :

L'indice de popularité, qui permet de positionner chaque communauté - ou sous-réseau - au sein de l'ensemble du graphe. Plus l'indice est élevé et plus a communauté concernée sera importante dans le graphe ;

L'indice de centralisation, qui, pour chaque réseau, permet de déterminer son organisation monocentrique (dominée par une aire urbaine), ou polycentrique.

3.2 Construction des réseaux

Dix-sept réseaux de villes sont identifiées dans l'ensemble du graphe. A l'intérieur de chaque réseau, les villes entretiennent entre elles des relations mesurées par les déplacements domicile-travail des actifs. Ces ensembles de villes jouent un rôle particulier dans la structuration du territoire et caractérisent le rayonnement économique des espaces.

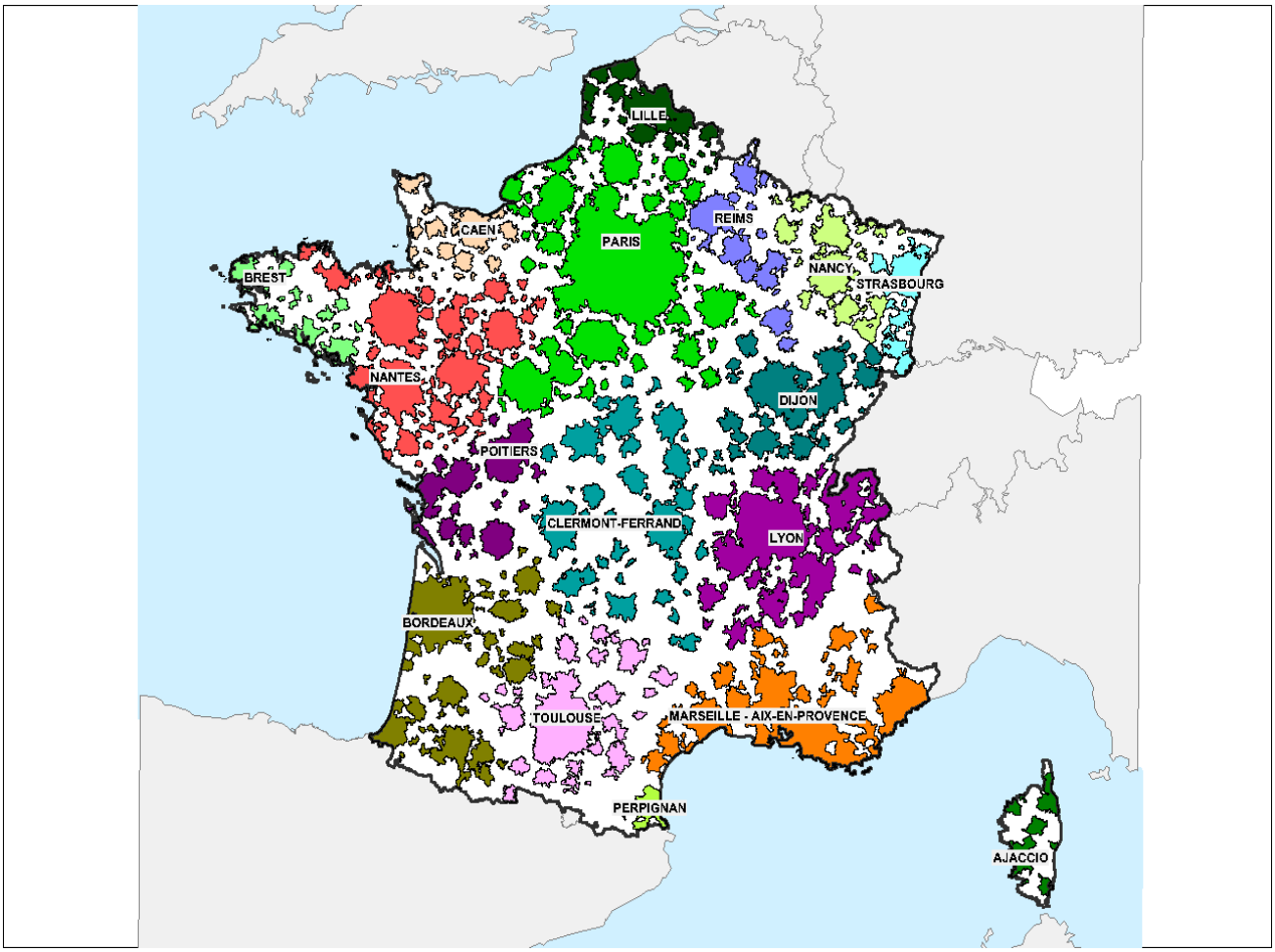
Les dix-sept aires urbaines têtes des réseaux qui structurent la France

Communautés	**Population 2011**
Paris	12 292 895
Lyon	2 188 759
Marseille - Aix-en-Provence	1 720 941
Toulouse	1 250 251
Lille	1 159 547
Bordeaux	1 140 668
Nantes	884 275
Strasbourg	764 013
Clermont-Ferrand	467 178
Nancy	434 565
Caen	401 208
Dijon	375 841
Reims	315 480
Brest	314 239
Perpignan	305 546
Poitiers	254 051
Ajaccio	100 621

Le territoire est organisé autour de dix-sept réseaux dont six de plus d'1 million d'habitants. Paris et son aire d'influence occupe une place à part avec plus de 12 millions d'habitants. Son influence s'étend au Sud jusqu'à l'aire urbaine de Tours et d'Auxerre, au Nord-Ouest jusqu'à Rouen et Amiens. En revanche, les aires de l'Est sont sous l'influence de Reims.

Rennes et Nantes sont les deux principaux nœuds d'un même réseau, rayonnent surtout sur leur côté est jusqu'au Mans. En revanche, les villes du sud de la Bretagne jusqu'à l'extrême Ouest forment un réseau distinct.

Les dix-sept réseaux d'aires urbaines qui structurent la France



Source Insee, RP2011

L'Est de la France est découpé en trois réseaux, autour des villes de Reims, Metz-Nancy et de Strasbourg. Un vaste réseau méditerranéen s'étend de Nice à Carcassonne, sans toutefois englober Perpignan.

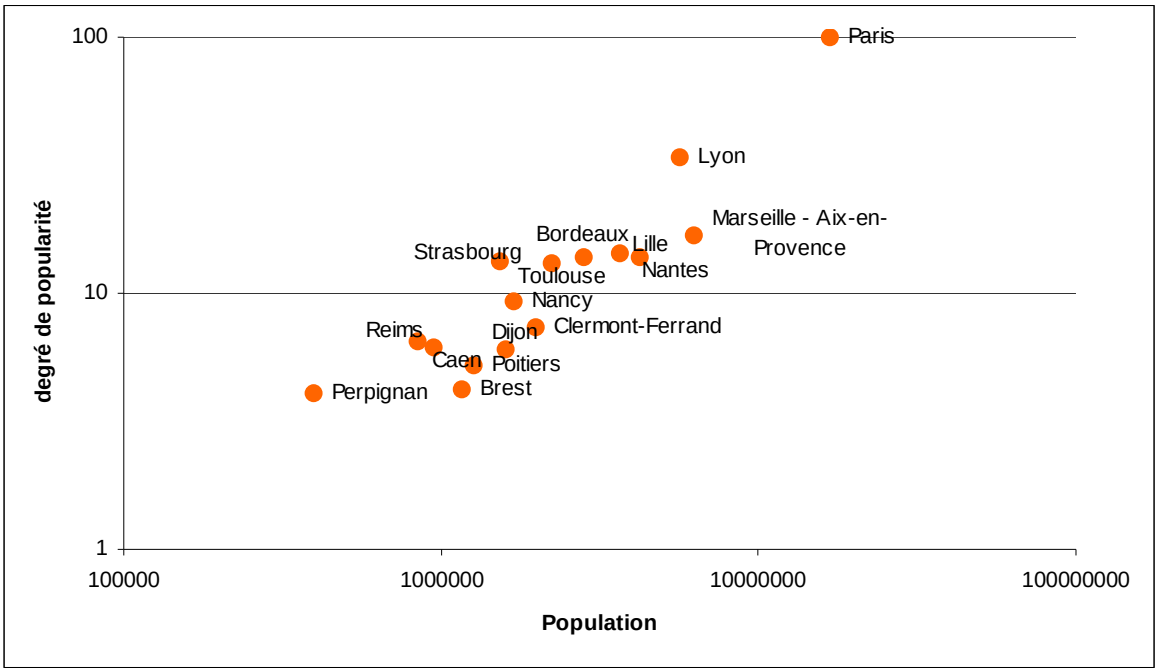
Toulouse rayonne des Pyrénées jusqu'à Rodez, Bordeaux rayonnant sur le reste de sud-ouest.

Le territoire français est caractérisé par l'importance d'un réseau dont le centre est Paris : ce vaste ensemble qui regroupe 12 millions d'habitants et des aires urbaines de la Normandie, Picardie et du Centre forme le nœud principal. Le graphe est clairement de type monocentrique. Vient ensuite le réseau formé autour de Lyon, qui regroupe des villes essentiellement au sud et à l'est.

Après Paris et Lyon, les six autres réseaux d'importance dans le graphe sont ceux de Marseille, Lille, Nantes, Bordeaux, Toulouse et Strasbourg. Ces huit réseaux structurent l'ensemble du graphe car ce sont ceux qui reçoivent le plus de flux.

Un autre indicateur, plus complet, confirme l'importance des huit premiers réseaux : l'indice de popularité, qui correspond au pouvoir de connectivité du sous-réseau (encadré 2).

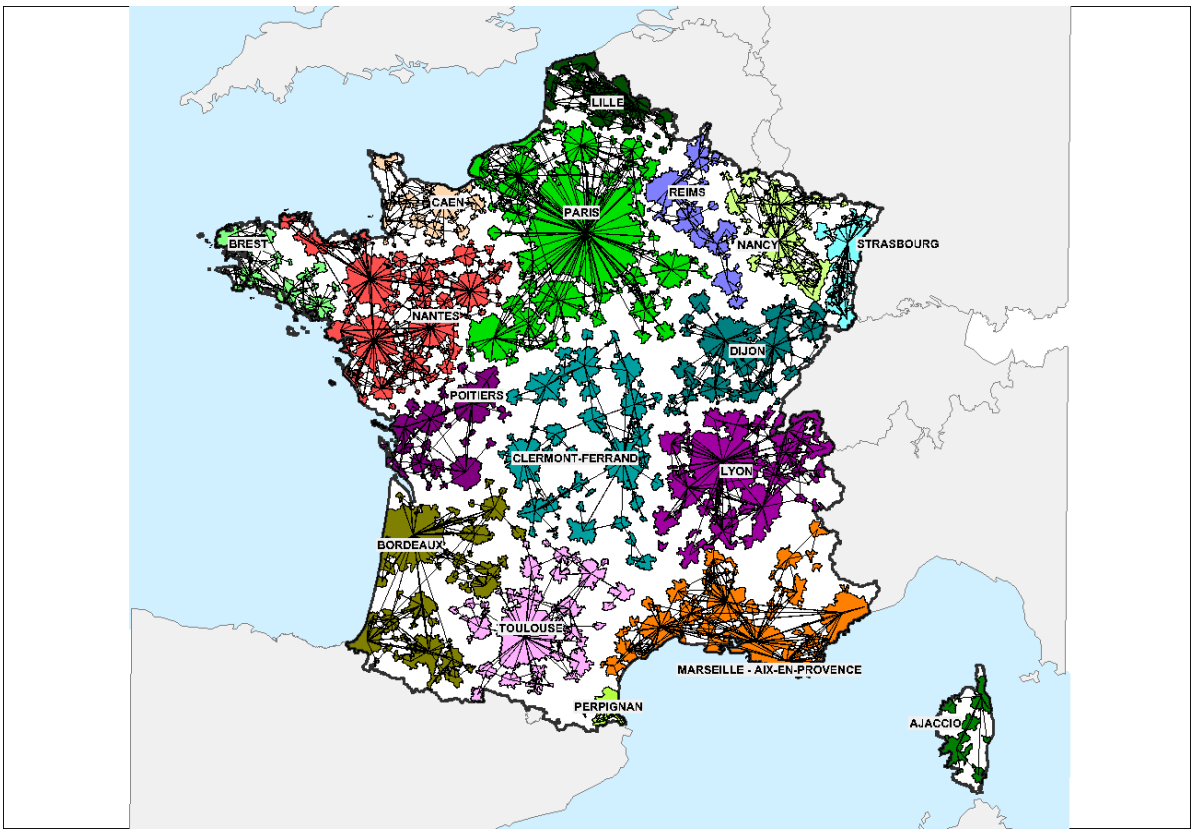
Indice de popularité et population des dix-sept réseaux (échelle logarithmique)



3.3 Caractérisation fonctionnelles des réseaux

Ces dix-sept réseaux d'aires urbaines se caractérisent par des fonctionnements internes différents. Certains sont animés essentiellement par une aire tête de réseau, on parlera de réseau monocentrique. D'autres renvoient à des fonctionnements organisés autour de quelques pôles, ce sont des espaces polycentriques.

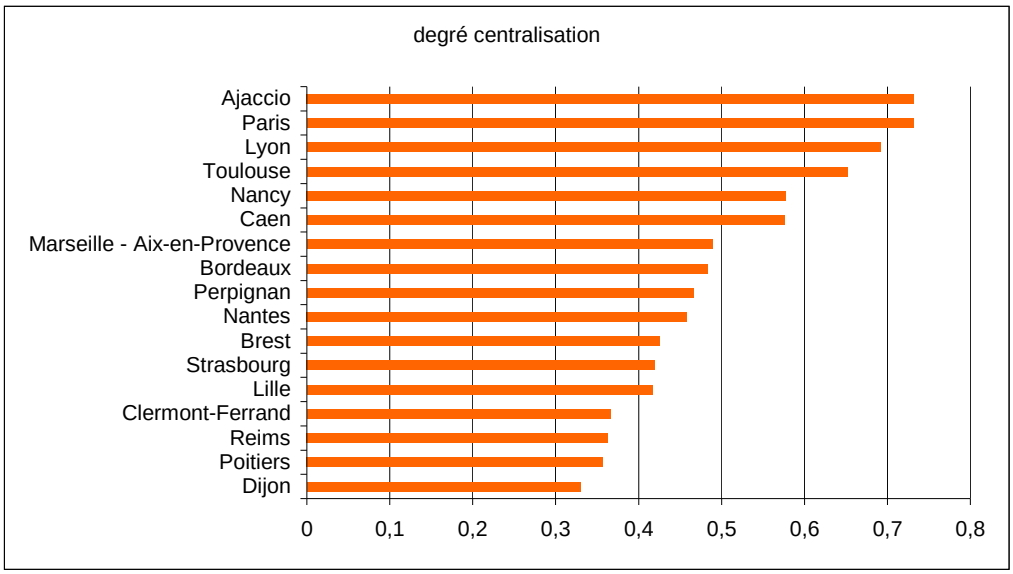
Les échanges entre aires urbaines au sein de chaque réseau.



Source Insee, RP2011

Ainsi, Toulouse et Paris sont les réseaux les plus monocentriques, la grande majorité de leur flux internes partent ou arrivent dans l'aire urbaine centrale. Ce n'est pas le cas de Rennes, dont les flux sont répartis entre Rennes, Nantes, Le Mans et Angers. On retrouve cette configuration polycentrique dans la plupart des autres réseaux.

L'indice de centralisation permet dévaluer la forme d'organisation spatiale de chaque communauté. Plus l'indice est élevé, plus la communauté est monocentrique.



La méthode utilisée, la théorie des graphes a répondu à nos attentes : identifier des réseaux et donner des indicateurs pour en évaluer la pertinence, l'organisation interne.

Des travaux restent à mener, notamment sur l'utilisation de plusieurs flux dans une même analyse. En effet, la question de la superposition des graphes pour ensuite constituer des communautés constitue encore défi en terme méthodologique. Des travaux de géographes (Berroir et al. 2006) ont essayé de cumuler des flux, mais le choix des pondérations permettant de les hiérarchiser pose de nombreuses difficultés. La méthode des graphs permet de s'affranchir des questions de pondérations des flux en repérant les aires urbaines centrales dans plusieurs réseaux.

Bibliographie

Fortunato S. **« Community detection in graphs ».** Physics Reports, vol. 486, no. 3-5, pages 75–174, 2010.

Newman M.E.J., Strogatz et Watts D.J. **« Random graphs with arbitrary degree distributions and their applications »**. Physical Review E, vol. 64, no. 2, page 026118, 2001. (Cité en page 57.)

Newman M.E.J. et Girvan M. **« Finding and evaluating community structure in networks»** Physical review E, vol. 69, no. 2, page 026113, 2004.

Reichardt J. et Bornholdt S. **« Statistical mechanics of community detection »** Physical Review E, vol. 74, no. 1, page 016110, 2006.

Seifi M.**« Coeurs stables de communautés dans les graphes de terrain ».**

Pons P. et Latapy M. **« Computing communities in large networks using random walks »** Computer and Information Sciences-ISCIS 2005, pages 284– 293, 2005.

Rozenblat C., Melancon G., **« Methods for Multilevel Analysis and Visualisation of Geographical Networks »** Methodos series 11.

Sandrine Berroir, Nadine Cattan, Marianne Guérois, Fabien Paulus, Céline Vacchiani-Marcuzzo **« Les systèmes urbains français - synthèse »,** Travaux en ligne n°10, Datar

« Des aires urbaines... aux systèmes métropolitains, une première approche », Fédération nationale des agences d'urbanisme, septembre 2006.

Pumain D. 1992**, « Les systèmes de villes »,** in Bailly A. Ferras R. Pumain D. (eds), Encyclopédie de géographie, Paris, Economica, chap. 34, p.645.

Brutel C. Levy D. 2011, **« Le nouveau zonage en aires urbaines de 2010 »,** .Insee Première 1374, octobre.

Beauguitte L, **« Graphes, réseaux, réseaux sociaux : vocabulaire et notation »** UMR Géographie-cités.

Clauset A., Newman M.E.J, et Moore C. **« Finding community structure in very large networks »**