13. Partitionnement et analyse de graphes

PASCAL EUSEBIO, JEAN MICHEL FLOCH, DAVID LEVY
Insee

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 ces techniques sont encore peu répandues dans la statistique publique, le problème posé est assez classique : réaliser la partition d'une population en sous-populations. On cherche à repérer des sous-populations homogènes (faible hétérogénéité intra-classe) et assez différenciées (forte hétérogénéité inter-classe). En utilisant les graphes, nous verrons que nous recherchons souvent des partitions qui conservent beaucoup de flux intra-zones et peu de flux entre elles. Les solutions algorithmiques reposent sur des méthodes "agglomératives" ou "divisives" selon les cas, que nous pouvons rapprocher des méthodes ascendantes ou descendantes que nous connaissons en analyse des données. Elles utilisent la notion de modularité, fondée sur le comparaison du graphe étudié à un graphe aléatoire.

Ce chapitre n'a pas vocation à balayer l'ensemble des méthodes de la théorie des graphes qui ont connu de fortes évolutions depuis leur apparition dans les années 1930. Ces méthodes ont été élaborées dans des domaines très divers (géographie, analyse des réseaux sociaux, biologie, informatique). Les méthodes présentées sont issues essentiellement du monde de la physique (autour du concept clé de modularité) mais un encadré fournit quelques compléments sur les méthodes de *blockmodelling* et sur la prise en compte de l'espace dans les réseaux.

13.1 Les graphes et l'analyse géographique des réseaux de villes

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 CHRISTALLER 2005. 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 (WILSON 1974). 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 et les graphes invariants d'échelle.

Définition 13.1.1 — Graphe. Un **graphe** est une représentation graphique d'un ensemble de sommets reliés par des arêtes.
Une **arête** est un lien entre deux éléments distincts.
Un **sommet** ou **nœud** est un élément relié par des arêtes. Le **degré** d'un sommet est le nombre de sommets auquel il est relié.

■ **Exemple 13.1** Un graphe de villes représente des villes (les sommets) qui échangent des populations : les navettes domicile-travail (les arêtes, aussi appelées liens dans la suite). ■

13.1.1 Petit-monde

Pendant longtemps, les spécialistes des graphes ne s'intéressaient qu'aux graphes aléatoires encore aujourd'hui très utilisés. Dans les années 1990, divers théoriciens des graphes ont proposé des modèles comme le petit-monde et l'invariance d'échelle. 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* (WATTS et al. 1998). On trouve, dans la figure 13.1, la reproduction du schéma proposé par les deux auteurs pour illustrer la construction du graphe petit-monde.

Définition 13.1.2 — Graphe aléatoire. Graphe dont la distribution des arêtes est aléatoire.

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 connaissaient pas, en utilisant comme intermédiaires des personnes de leur entourage. Milgram eut la surprise de constater qu'en moyenne les chaînes parvenues au destinataire n'étaient composées que de 5,6 individus. Cette expérience a permis de confirmer la thèse de KARINTHY 1929 selon laquelle toutes les personnes du globe sont reliées par une chaîne d'au plus 5 maillons, devenue 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.

Le graphe de départ est un graphe qualifié de k-régulier.

Définition 13.1.3 — Graphe k-régulier. Graphe dans lequel chaque sommet est lié au même nombre k de sommets (BATTISTON et al. 2014). En d'autres termes, tous les sommets ont le même degré k.

L'idée des auteurs est de présenter une façon simple de transformer ce graphe régulier en graphe aléatoire. À 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 ont combiné deux mesures : $L(p)$ et $C(p)$ pour caractériser un type de réseau.

$L(p)$ désigne la longueur moyenne du plus court chemin entre les paires de sommets lorsque

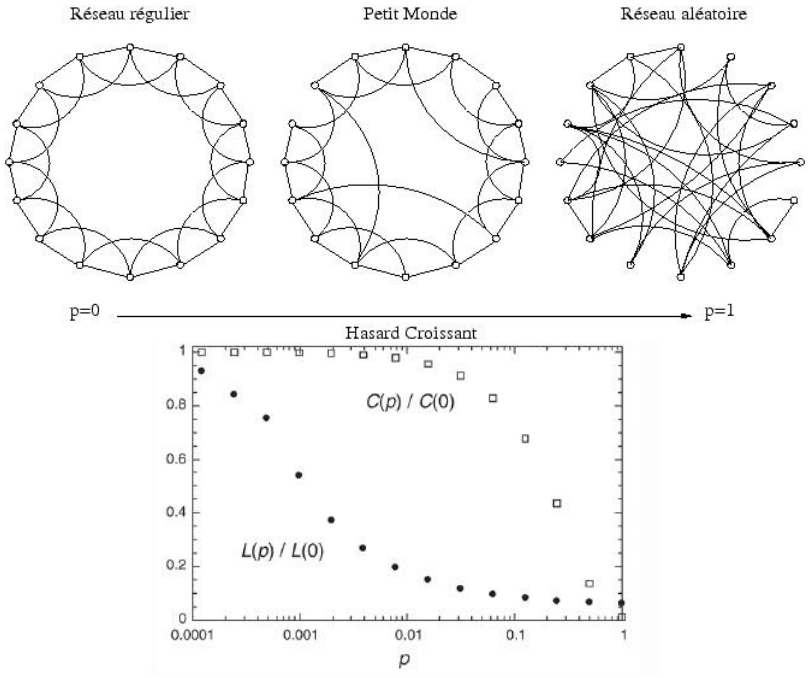


FIGURE 13.1 – Réseaux petit-monde
Source : WATTS *et al. 1998*

p varie. $C(p)$ désigne le coefficient de *clustering*, dont on trouvera une illustration dans la figure 13.2. Ce coefficient est en rapport avec la notion de transitivité dans le graphe, notion connue des sociologues depuis les années 1970. 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. Strogatz 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.

Définition 13.1.4 — Le coefficient de *clustering*.

$$C(p) = \sum_{i=1}^{n} \frac{C_i}{n} \tag{13.1}$$

avec

$$C_i = \frac{\text{nombre de triangles dont un des trois sommets est le nœud i}}{\binom{k}{2}} \tag{13.2}$$

où k est le coefficient local ou le degré du nœud et n le nombre de nœuds du graphe.

■ **Exemple 13.2** Avec le graphe présenté en figure 13.2,

$$C_4 = \frac{5}{\binom{6}{2}} = 1/3$$

et le coefficient de *clustering* du réseau est :

$$C(p) = \sum_{i=1}^{n} \frac{C_i}{n} = 0,5208.$$

■

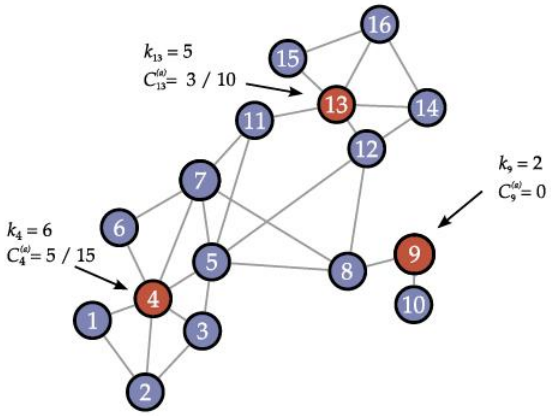


FIGURE 13.2 – Le coefficient de *clustering*

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 le 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 petit-monde, dans une définition qui reste assez largement qualitative (grand nombre de sommets, nombre de liens existants 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 fonction `sample_smallworld` du package *igraph* de R. Dans un tel réseau, on fait l'hypothèse que chaque sommet puisse être relié à n'importe quel autre.

Application avec R

```
# Package necessaire
library(igraph)

# Generation du graphe avec 100 noeuds
g <- sample_smallworld(dim = 1, size = 100, nei = 5, p = 0.05)


# Representation du graphe
plot(g, vertex.size=4,vertex.label.dist=0.5,
                vertex.color="green",
        edge.arrow.size=0.5)
```

```
# calcul des coefficients du graphe
## le coefficient  local
q=transitivity(g,type = "local")

## le coefficient global
transitivity(g,type = "average")
```

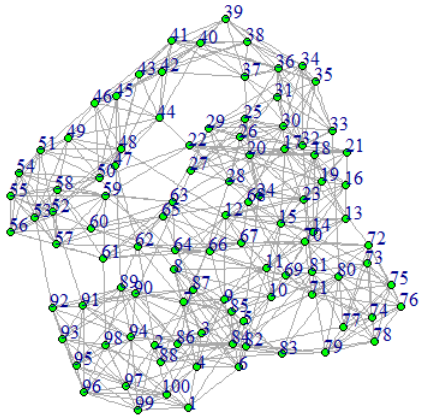


FIGURE 13.3 – Graphe petit-monde
Source : *Simulation à partir du package igraph*

```
# qui est bien egal a la moyenne des coefficients locaux
mean(q)
```

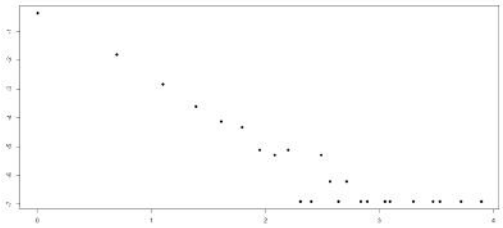
13.1.2 Réseaux invariants d'échelle

Un autre ensemble de graphes complexes est celui des graphes invariants d'échelle. Cette modélisation a été proposée initialement par BARABÁSI et al. 1999. On peut générer ce type de graphe sous R avec la fonction `barabasi.game` du package *igraph* et on en trouvera une illustration dans la figure 13.4.

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 (BARABÁSI et al. 1999). Chaque nouveau nœud aura une probabilité de se lier à un nœud d'autant plus forte que le degré de ce nœud est élevé. Ils sont appelés invariants, car un zoom sur n'importe quelle partie du graphe ne change pas sa forme. À chaque niveau de grossissement, le réseau contiendra quelques nœuds avec beaucoup de connexions et un grand nombre de nœuds avec très peu de connexions. Ainsi, le réseau est dit **invariant d'échelle** si, lorsque k désigne le degré, et $P(k)$ 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 a présenté en figure 13.4, la valeur du coefficient γ est de 2,6.

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 (introducteur des graphes petit-monde) montrent que les graphes complexes combinent souvent des caractéristiques des deux types (NEWMAN et al. 2011). Cela 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 villes. On

(a) Exemple de graphe de Barabasi



(b) Évolution de la fréquence du nombre de voisins

FIGURE 13.4 – Exemple de réseaux invariants d'échelle
Source : *Graphes simulés par la fonction barabasi.game du package igraph*

peut citer divers travaux de ROZENBLAT et al. 2013 sur les réseaux de transport aérien, sur la combinaison des transports aériens et maritimes, ou sur les liens géographiques entre les firmes multinationales. On trouvera en figure 13.5 un schéma illustrant les emboîtements entre logique petit-monde et logique invariance d'échelle.

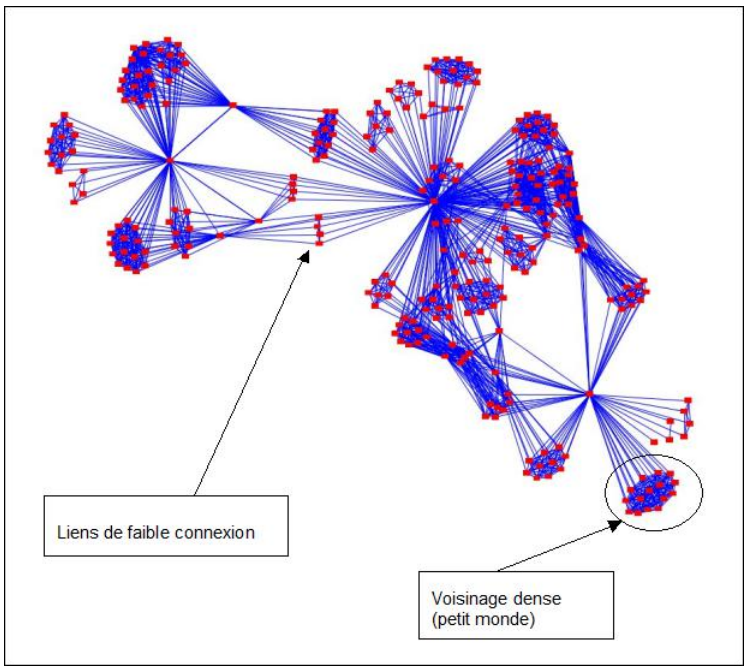


FIGURE 13.5 – Réseau formé de réseaux petit-monde et invariance d'échelle
Source : ROZENBLAT *et al. 2013*

Certains auteurs (BEAUGUITTE et al. 2011) relativisent cependant l'apport des deux concepts à la géographie, estimant que celui de petit-monde est généralement trivial tandis que celui d'inva-

riance d'échelle est connu depuis longtemps. En revanche, l'utilisation des méthodes de partitionnement, issues de travaux de physiciens, ont considérablement enrichi les possibilités d'analyse des réseaux complexes.

13.2 Les méthodes de partitionnement de graphes

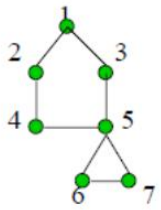
Si le graphe permet de représenter les échanges entre les sommets, le partitionnement de graphe met en évidence des groupes de sommets reliés préférentiellement. Ainsi, le partitionnement du graphe représentant les échanges commerciaux permet, par exemple, d'indiquer où implanter les plateformes de transport pour desservir au mieux le territoire : c'est au sein de chaque groupe obtenu par partitionnement. 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, etc.). 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.

13.2.1 Notions de théorie des graphes

Définition 13.2.1 — Caractériser un graphe. Le **graphe** est un ensemble $G = \{V, E\}$ (figure 13.6) où V (de l'anglais *vertex*) désigne les sommets et E (de l'anglais *edge*) les arêtes.
La **taille** du graphe est le nombre de liens.
L'**ordre** du graphe est le nombre de sommets.
Un graphe est dit **vide** lorsqu'il ne contient aucun lien.
Un graphe est dit **complet** lorsque tous les sommets sont connectés à tous les autres. Il y a alors $\frac{n(n-1)}{2}$ liens dans un graphe complet d'ordre n.
Un graphe **orienté** est un ensemble de sommets et d'arêtes, chaque arête étant un couple de sommets ordonnés. Ainsi, la relation entre les sommets x et y est différente de celle entre y et x.
Un graphe **valué**, par opposition à un graphe non valué, comporte des liens multiples (deux sommets sont liés plusieurs fois).
Dans cette communication, on se limitera à des graphes non orientés, dans lesquelles les relations entre sommets sont de fait symétriques.
Un graphe **simple** est un graphe non valué et sans boucle (sans arête d'un sommet vers lui-même).
Le **degré** d'un sommet est le nombre de sommets auquel il est relié. Dans un graphe simple d'ordre n, le degré d'un sommet est compris ente 0 et $n-1$. La séquence des degrés est la suite $d_1, ..., d_n$.
La **densité** d'un graphe est le rapport entre le nombre de liens observés et le nombre de liens d'un graphe complet. Ainsi, elle varie entre 0 pour un graphe vide et 1 pour un graphe complet.
Si chaque point d'un graphe est atteignable depuis n'importe quel point alors le graphe est **connecté** ou **connexe**.

■ **Exemple 13.3** Le graphe présenté en figure 13.6 est un graphe simple de taille 8 et d'ordre 7. ■

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 13.7). Une valeur supérieure à 0 indique qu'il existe un lien entre deux points. Si la matrice d'adjacence est symétrique alors elle est issue d'un graphe non orienté. Si sa diagonale vaut 0 alors le graphe associé est simple (sans boucle).



(a) Un graphe à 5 nœuds et 8 arêtes

V={1,2,3,4,5,6,7}
E={(1,2),(1,3),(2,4).
(4,5),(3,5),(4,5),
(5,6),(6,7)}

(b) Son écriture mathématique

FIGURE 13.6 – Représentations géométrique et mathématique d'un graphe

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

(a) Matrice d'adjacence

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

(b) Matrice des degrés

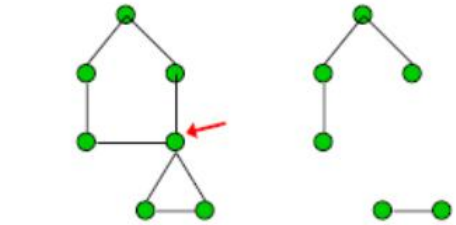
2	−1	−1	0	0	0	0
−1	2	0	−1	0	0	0
−1	0	2	0	−1	0	0
0	−1	0	2	−1	0	0
0	0	−1	−1	4	−1	−1
0	0	0	0	−1	2	−1
0	0	0	0	−1	−1	2

(c) Matrice laplacienne

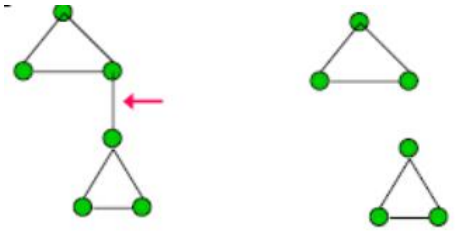
FIGURE 13.7 – Matrices d'adjacence, des degrés et laplacienne associées au graphe de la figure 13.6

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. Dans l'exemple de la figure 13.6, la suite de sommets $(1,3,5,7)$ est la géodésique entre les points 1 et 7, et la suite de sommets $(1,2,4,5,7)$ est 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 on soustrait cette matrice d'adjacence à la matrice des degrés (matrice dont la diagonale est constituée des degrés de chaque sommet), on obtient la matrice **laplacienne** (figure 13.7) qui joue un rôle fondamental dans l'approche qualifiée de spectrale des graphes (méthodes de *clustering*).

Toutes les questions que l'on examinera désormais tournent autour de la possibilité de déterminer au sein de notre graphe des sous-graphes appelés **communautés** ou **cliques**. 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 13.8).



(a) Point de coupure (nœud)



(b) Pont (lien)

FIGURE 13.8 – Suppression de nœuds ou de liens

Définition 13.2.2 — Quelques indicateurs de centralité. Mesures qui permettent d'appréhender les sommets (et les liens) les plus importants.
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 en figure 13.6, 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 :

$$C_c(v) = \frac{1}{\sum_{u \in V \setminus \{v\}} d_G(u,v)} \tag{13.3}$$

avec $d_G(u,v)$ la distance entre les sommets u et 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 entre deux sommets doivent emprunter ce sommet. Elle s'écrit :

$$C_B(v) = \sum_{\substack{i,j \\ i \neq j \neq v}} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \tag{13.4}$$

avec $\sigma_{ij}(v)$ le nombre de chemins entre i et j qui passent par v.

Il existe aussi la centralité d'intermédiarité de lien, qui rend compte du nombre de géodésiques (plus courts chemins) qui empruntent ce lien. La figure 13.9 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 graphes.

La **centralité de vecteur propre** ou **centralité spectrale** est définie par Bonacich à partir de la matrice d'adjacence. La centralité spectrale est une mesure de l'influence d'un nœud au sein d'un réseau. Elle correspond pour un sommet à la somme de ses connexions avec les autres sommets, pondérée par la centralité de degré de ces sommets. On peut l'écrire sous la forme :

$$C(v) = \frac{1}{\lambda} \sum_{u \neq v} A(v,u) C(u) \tag{13.5}$$

qui peut s'écrire $\lambda C = AC$.
Pour résoudre cette équation, BONACICH 1987 montre que le vecteur de centralité spectrale correspond en fait au vecteur propre dominant (ou principal) de la matrice d'adjacence.

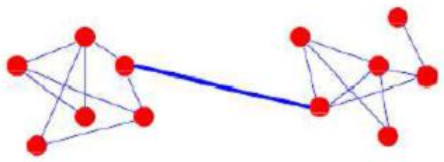


FIGURE 13.9 – Une forte centralité d'intermédiarité (trait foncé)

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 (ZACHARY 1977) sur le réseau social constitué par les membres d'un club de karaté universitaire (figure 13.10). Le package *igraph* du logiciel R permet de représenter le graphe et de calculer les indicateurs précédents.

```
# Centralite de degre
d<- degree(kar)
# Centralite de proximite
cp<- closeness(kar)
# Centralite d'intermediarite
```

```
ci<- betweenness(kar)
# Centralite de vecteur propre
ce<- graph. eigen(kar)[c("values", "vectors")]
```

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

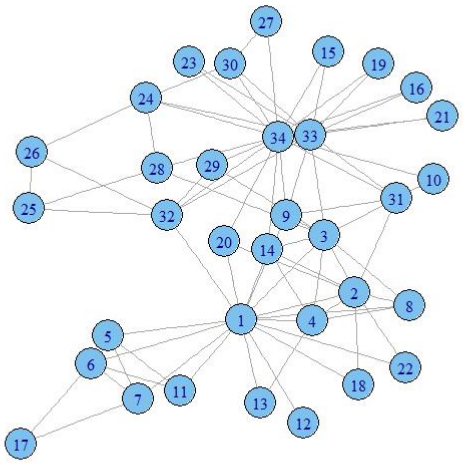


FIGURE 13.10 – Réseau de Zachary
Note : liens d'amitié entre 34 membres d'un club de karaté dans une université américaine

Le tableau ci-dessous montre le classement des individus du réseau présenté en figure 13.10 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

13.2.2 Les méthodes de partitionnement

Si l'on revient à nos problèmes de réseaux de villes, on va être confronté à la détermination de communautés. 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 s'appuiera largement dans cette partie sur les synthèses réalisées par NEWMAN 2006 et FORTUNATO 2010, ainsi que sur les thèses francophones de PONS 2007 et SEIFI 2012.

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*), et si le nombre de liens reste du même ordre de grandeur que celui de 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 souvent issues de la physique (méthodes "gloutonnes", *spinglass*).

Comme en classification, on sera confronté au problème d'optimisation du nombre de communautés, à celui de hiérarchie et à celui d'emboîtement.

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 la 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 NEWMAN et al. 2004. 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 permet de justifier la pertinence des sous-graphes 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.

Définition 13.2.3 — La modularité. C'est une mesure de la qualité d'un partitionnement de graphe. Si l'on considère $\mathbf{P}$ la partition en p clusters du graphe $G = \{V, E\}$, alors : $\mathbf{P} = \{c_1, ..., c_n, ...c_p\}$

La modularité peut être introduite de façon assez simple de la façon suivante, en se référant

à l'idée de Newman.

$$Q(P) = \sum_i (e_{c_i} - a_{c_i}^2) \tag{13.6}$$

avec e_{c_i} la part des liens d'un cluster c_i sur le total, a_{c_i} la probabilité qu'un sommet se trouve dans le cluster c_i et donc $a_{c_i}^2$ la probabilité que les deux sommets d'un lien se trouvent dans le même cluster c_i.

Cette expression générale est transformée dans la première forme usuelle de présentation de la modularité. On montre (FORTUNATO 2010) que la modularité peut s'écrire sous la forme :

$$Q(P) = \frac{1}{2m} \sum_{i,j \in V} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(c_i, c_j) \tag{13.7}$$

avec

— m le nombre d'arêtes du graphe ;
— A la matrice d'adjacence du graphe ;
— A_{ij} le poids des liens entre les sommets i et j ;
— d_i la somme des degrés de i avec $d_i = \sum_i A_{ij}$;
— $a_{c_i}^2 = \sum_j \frac{d_i d_j}{4m^2}$;
— $\delta(c_i, c_j)$ 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 :

$$Q(P) = \frac{1}{2m} \sum_{k=1}^{p} \sum_{i,j \in V} \left(A_{ij} - \frac{d_i d_j}{2m} \right) = \sum_{k=1}^{p} \left[\frac{l_k}{m} - (\frac{d_k}{2m})^2 \right] \tag{13.8}$$

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

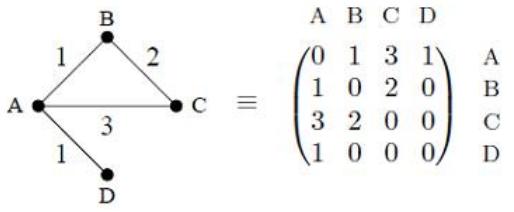
Le terme $A_{ij} - \frac{d_i d_j}{2m}$ correspond à la différence de liens entre notre graphe et un graphe aléatoire dont la contrainte est la conservation des degrés de sommets.

Les définitions de la modularité ont d'abord été développées dans le contexte de graphes non valués. Elles ont été étendues aux graphes valués. La valeur de A_{ij} correspond au lien entre les sommets, qui dans un graphe non valué, vaut 1 si les sommets sont liés et 0 sinon, et dans un graphe valué vaut la valeur du flux s'il y a un lien et 0 sinon. On trouve dans NEWMAN 2004 une façon très simple de passer des graphes non valués aux graphes valués, en introduisant ce qu'il appelait des multigraphes (figure 13.11).

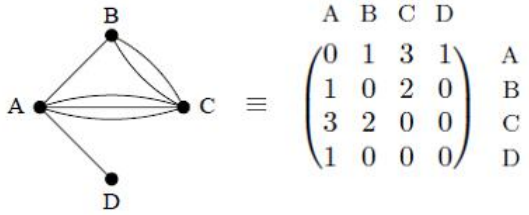
Cette représentation permet de généraliser aux graphes pondérés les résultats présentés précédemment. Les A_{ij} correspondent aux poids associés aux liens ou de façon équivalente au nombre de liens du multigraphe. M est le nombre de liens du multigraphe, ou la somme des pondérations.

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 de Guimaras, Reichart et Bornholdt, mis en avant par Fortunato (FORTUNATO 2010) se penchent sur le problème de "résolution". Si le nombre de liens dans le graphe devient très grand et que le nombre de liens attendu (voir formule de la modularité) est inférieur à 1, un seul lien entre les deux groupes suffit à entraîner leur fusion.



(a) Graphe valué



(b) Représentation multigraphe

FIGURE 13.11 – Passage au graphe valué, les multigraphes

Panorama général des méthodes de partition

Une fois défini le schéma général d'une partition, il reste à la réaliser en pratique. En pratique 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 villes 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) est présentée dans FORTUNATO 2010. On cherche souvent à mesurer la complexité des algorithmes en les notant $O(n^2m^2)$ avec n le nombre de liens, m le nombre d'arêtes. 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 pôle "Analyses Territoriales" de l'Insee sur les réseaux de villes (voir section 13.3) et en se centrant sur celles qui sont implémentées dans le logiciel R. Les méthodes sont en pleine expansion et 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 autres 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 (KERNIGHAN et al. 1970). D'autres méthodes s'inspiraient aussi de ce qui était fait en analyse des données (dendrogrammes de classification, méthodes de type *k-means*). Ces méthodes reposent sur des propriétés des graphes, ou sur le traitement de la matrice d'adjacence.

Méthodes classiques

On ne présentera que quelques unes des méthodes classiques :

Les méthodes fondées sur la bissection de graphes

Ces méthodes (figure 13.12) sont assez simples à présenter. L'idée est de chercher la ligne qui partage le graphe en coupant le moins de liens (*cut size*).

Dans leurs versions les plus simples, ces méthodes risquent cependant de ne faire apparaître que des solutions triviales (un sommet isolé). Des méthodes plus élaborées de bissection reposent

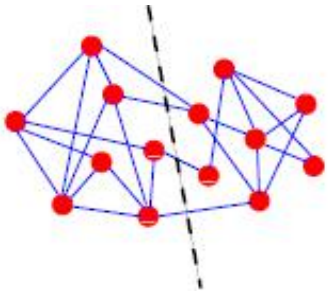


FIGURE 13.12 – Bipartition de graphe

sur des méthodes spectrales (propriétés du spectre de la matrice laplacienne) que l'on présentera plus loin.

Les méthodes hiérarchiques

Ces méthodes (figure 13.13) 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.

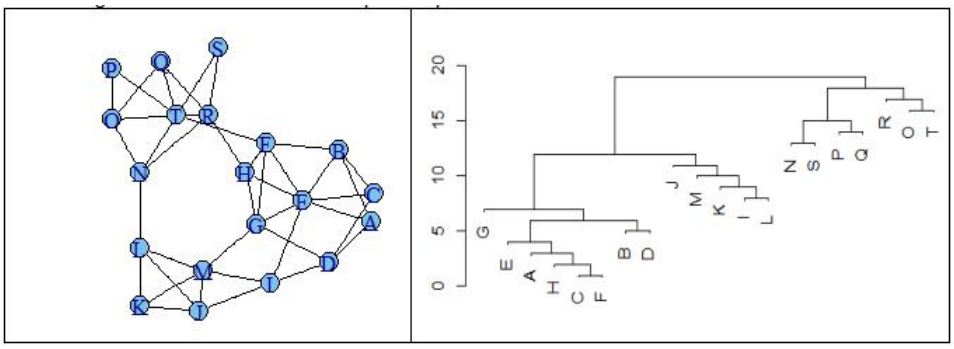


FIGURE 13.13 – Méthodes hiérarchiques de partitionnement

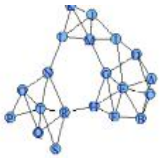
Les méthodes de *clustering*

Ces méthodes 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 grande que les sommets sont dissemblables. On va chercher à minimiser une fonction de coût basée sur les points et les centroïdes. Dans le minimum *k-means 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 intra-classes.

La méthode divisive

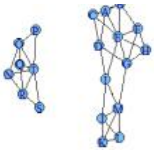
Cette méthode est une des plus intuitives à présenter. Elle repose sur le concept présenté en 13.2.1 de centralité d'intermédiarité, avec un schéma qui expose assez bien dans un cas simple cette idée. Lorsque beaucoup de géodésiques allant d'un point quelconque du graphe à un autre passent par un sommet ou par un lien, la suppression de ceux-ci est plus à même de faire apparaître des communautés. Dans l'exemple présenté plus haut, 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. Cette démarche est représentée en figure 13.14.

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.



(a) Suppression du lien T-F

(b) Suppression du lien R-H



(c) Suppression du lien N-L

FIGURE 13.14 – Partitionnement du graphe de de la figure 13.13 avec la méthode divisive

```
karate <- read.graph("karate.gml",format="gml")
plot(karate,vertex.size=2)
betkar<-  edge.betweenness.community(karate)
plot(betkar,karate)
```

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 NEWMAN et al. 2004. Elle confirme d'ailleurs l'attrait des physiciens pour 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. re-calcul 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 produire ainsi une hiérarchie de partitions emboîtées. Le choix de la partition peut se faire avec le 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.

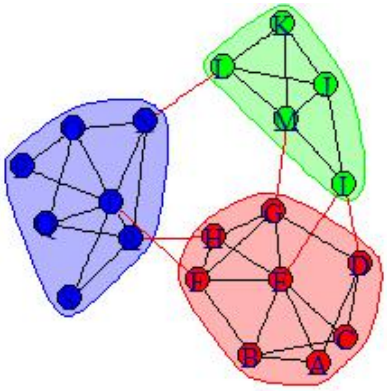
D'autres algorithmes divisifs ont été proposés. FORTUNATO 2010 a proposé un algorithme qui utilise la centralité d'information de lien définie comme étant la diminution relative de l'efficacité du réseau lorsque l'on retire ce lien du graphe. Cet algorithme est plus performant, mais de complexité plus grande que celui de Girvan-Newman. Ce dernier reste donc très utilisé, notamment à titre de comparaison des communautés détectées.

Les méthodes agglomératives fondées sur la modularité

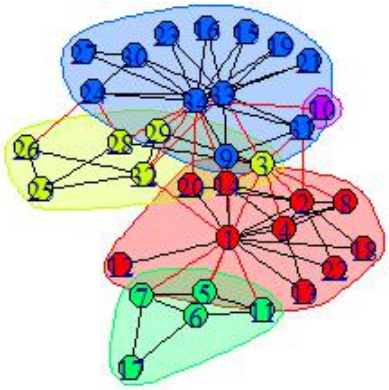
Cette famille de méthodes est très riche et très importante. Au contraire de la précédente, on part de l'ensemble des sommets, que l'on va progressivement agréger entre eux.

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 2010 une valeur approchée du nombre de ces



(a) Graphe de démonstration 3 communautés



(b) Club de Karaté 5 communautés

FIGURE 13.15 – Résultat de la méthode divisive sur un graphe et sur le club de karaté

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" (succession d'allers/retours à l'état d'équilibre) 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 - méthode aggrégative

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 et al. 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.

Dans cet exemple très simple (figure 13.16), on peut voir que la modularité Q augmente jusqu'à l'étape 10 où les trois communautés assez visibles sont identifiées. À l'étape 11, deux des communautés fusionnent et la modularité diminue, celle-ci devenant nulle lorsque les trois communautés sont regroupées. Le résultat est donc un partitionnement en trois 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. L'algorithme est de complexité $O(mn)$.

Les méthodes spectrales

Newman a proposé une version spectrale du partitionnement fondée 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, généralisée 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 : $B_s = \lambda D_s$ 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 égale 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 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 fonder 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. La deuxième phase de l'algorithme consiste à construire un nouveau réseau dont les sommets sont les communautés repérées dans la première phase, les

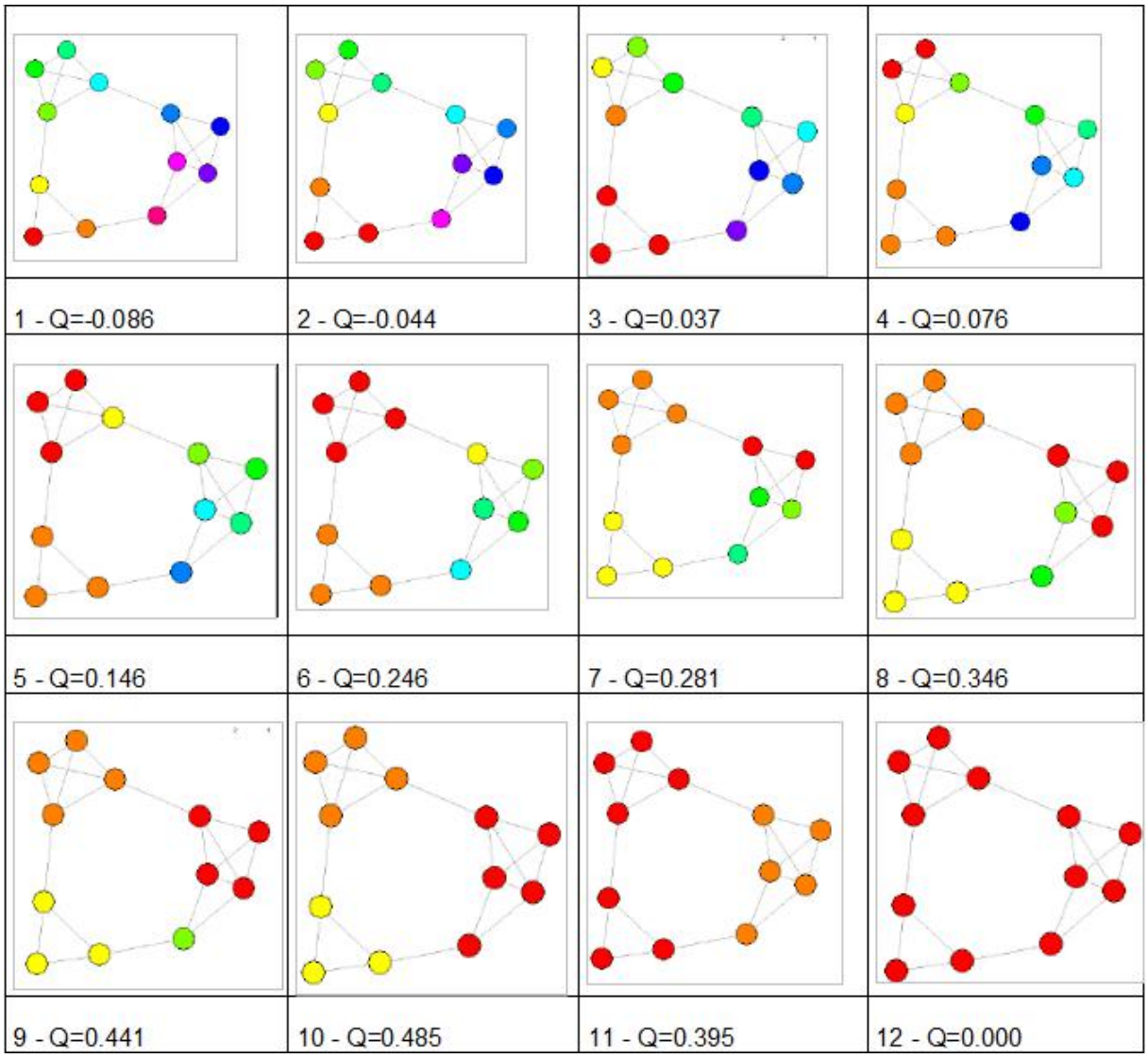


FIGURE 13.16 – Les 12 étapes du partitionnement d'un graphe à 12 sommets avec la méthode aggrégative

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 ré-applique 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. Elle est souvent présentée, notamment dans des articles récents de Newman comme la plus performante en temps et en qualité de partitionnement (NEWMAN 2016).

Autres méthodes

Marches aléatoires (*random walk*)

L'algorithme `walktrap.community` 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. À 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 :

$$P_{ij} = \frac{A_{ji}}{k_i}. \tag{13.9}$$

On a ainsi la matrice de transition de la chaîne 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 :

$$k_i P_{ij}(t) = k_j P_{ji}(t). \tag{13.10}$$

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. En revanche 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'arrivée : 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 éloignés 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 :

$$\sqrt{\sum_{k=1}^{n} \frac{(P_{ik}(t) - P_{jk}(t))^2}{k_k}}. \tag{13.11}$$

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 distances, 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 2007.

Dans l'exemple de la figure 13.17, 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.

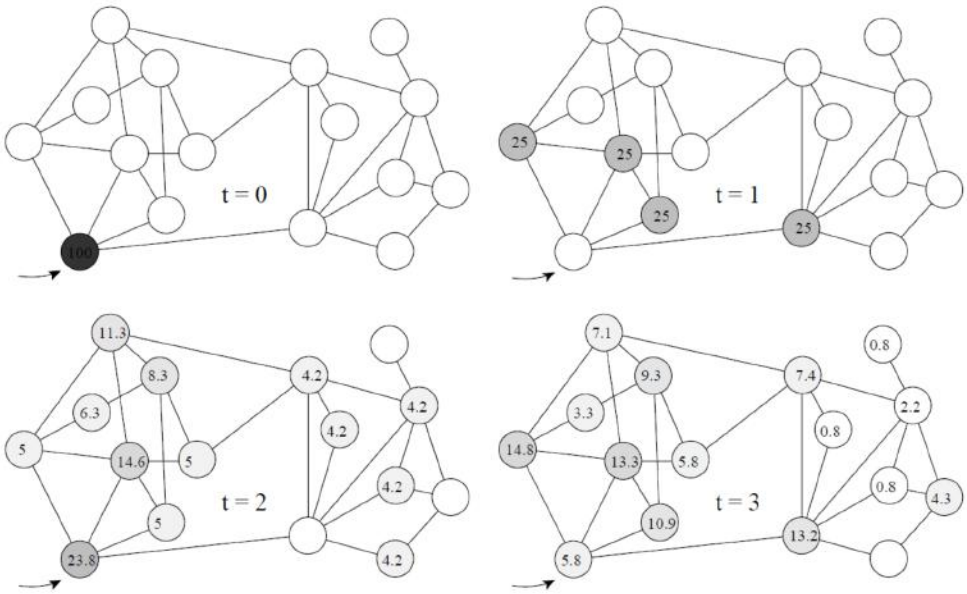


FIGURE 13.17 – Illustration de la marche aléatoire sur un graphe
Source : *d'après* PONS *2007*

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 **graphe hamiltonien** (graphe possédant au moins un cycle passant par tous les sommets une fois au plus) et une distribution de probabilité des couplages. REICHARDT et al. 2006 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 spins égales. On définit l'énergie du système par un hamiltonien faisant intervenir la matrice d'adjacence du graphe. La minimisation de cette expression se fait par recuit simulé, comme pour la méthode "optimale" présentée précédemment. On trouve dans le package *igraph* la fonction `spinglass.community` qui met en œuvre cette méthode.

Références - Chapitre 13

BARABÁSI, Albert-László et Réka ALBERT (1999). « Emergence of scaling in random networks ». *Science* 286.5439, p. 509–512.

BATTISTON, Federico, Vincenzo NICOSIA et Vito LATORA (2014). « Structural measures for multiplex networks ». *Physical Review E* 89.3, p. 032804.

BEAUGUITTE, Laurent et César DUCRUET (2011). « Scale-free and small-world networks in geographical research : A critical examination ». *17th European Colloquium on Theoretical and Quantitative Geography*, p. 663–671.

BONACICH, Phillip (1987). « Power and centrality : A family of measures ». *American journal of sociology* 92.5, p. 1170–1182.

CHRISTALLER, Walter (2005). « Les lieux centraux en Allemagne du Sud Une recherche économico-géographique sur la régularité de la diffusion et du développement de l'habitat urbain ». *Cybergeo : European Journal of Geography*.

CLAUSET, Aaron, Mark EJ NEWMAN et Cristopher MOORE (2004). « Finding community structure in very large networks ». *Physical review E* 70.6, p. 066111.

FORTUNATO, Santo (2010). « Community detection in graphs ». *Physics reports* 486.3, p. 75–174.

KARINTHY, Frigyes (1929). « Chain-links ». *Everything is the Other Way*, p. 25.

KERNIGHAN, Brian W et Shen LIN (1970). « An efficient heuristic procedure for partitioning graphs ». *The Bell system technical journal* 49.2, p. 291–307.

NEWMAN, Mark EJ (2004). « Analysis of weighted networks ». *Physical review E* 70.5, p. 056131.

— (2006). « Modularity and community structure in networks ». *Proceedings of the national academy of sciences* 103.23, p. 8577–8582.

NEWMAN, Mark EJ et Michelle GIRVAN (2004). « Finding and evaluating community structure in networks ». *Physical review E* 69.2, p. 026113.

NEWMAN, Mark, Albert-Laszlo BARABASI et Duncan J WATTS (2011). *The structure and dynamics of networks*. Princeton University Press.

NEWMAN, MEJ (2016). « Community detection in networks : Modularity optimization and maximum likelihood are equivalent ». *arXiv preprint arXiv :1606.02319*.

PONS, Pascal (2007). « Détection de communautés dans les grands graphes de terrain ». Thèse de doct. Paris 7.

REICHARDT, Jörg et Stefan BORNHOLDT (2006). « Statistical mechanics of community detection ». *Physical Review E* 74.1, p. 016110.

ROZENBLAT, Céline et Guy MELANÇON (2013). *Methods for multilevel analysis and visualisation of geographical networks*. Springer.

SEIFI, Massoud (2012). « Cœurs stables de communautés dans les graphes de terrain ». Thèse de doct.

WATTS, Duncan J et Steven H STROGATZ (1998). « Collective dynamics of'small-world'networks ». *nature* 393.6684, p. 440.

WILSON, Alan Geoffrey (1974). *Urban and regional models in geography and planning*. John Wiley & Sons Inc.

ZACHARY, Wayne W (1977). « An information flow model for conflict and fission in small groups ». *Journal of anthropological research* 33.4, p. 452–473.