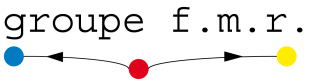
Cliques, communautés et dérivées

Laurent Beauguitte

Cliques, communautés et dérivés

Laurent Beauguitte - CNRS, UMR Géographie-cités
beauguittelaurent<at>parisgeo.cnrs.fr

Janvier 2011 - Version 1



Introduction

L'un des objectifs essentiels de l'analyse de réseaux est de repérer, à l'intérieur d'un graphe, des groupes d'acteurs fortement liés les uns aux autres. Les termes varient en fonction des disciplines : si les praticiens de la 'SNA' parlent de la recherche de *cohesive subgroups*([16], chap.7 et 8), les physiciens parlent de *communities* et les informaticiens de *clusters* (à ne pas confondre avec le *clustering coefficient* détaillé dans [8]).

Il convient donc d'être vigilant lorsqu'on parcourt la littérature... et d'adapter son vocabulaire en fonction de l'auditoire visé (conférence ou revue). Il apparaît peu pertinent de proposer un quatrième terme lorsque ces méthodes sont utilisées dans une perspective géographique (régions ? régions centrales ? pôles fonctionnels ? lieux centraux ? ? ? [1]) mais de reprendre les termes correspondants à la méthode choisie.

Précision importante : ce papier ne traite pas des méthodes de partitionnement en général mais uniquement de celles visant à mettre en évidence des sous-groupes aux relations denses. Cet aspect est en effet susceptible d'intéresser les géographes notamment en géographie économique ou en géographie politique pour mettre en évidence des structures de type centre - périphérie.

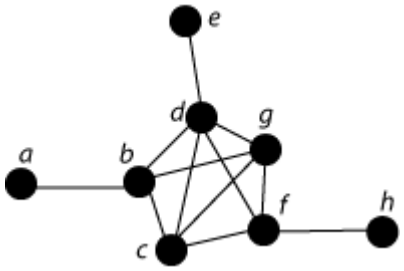
1 Cliques et dérivées

1.1 Cliques

La définition d'une clique est brève et précise : une clique est un sous graphe maximal complet comprenant au minimum trois sommets[1]. Formulée de façon plus explicite, une clique est un ensemble de sommets (minimum

1. Certains auteurs (physiciens) n'hésitent pas à parler de cartographie de graphes et de recherche de régions[10].

FIGURE 1 – Limite des cliques



3) entre lesquels tous les liens possibles sont présents (complet), et il n'est pas possible d'ajouter un sommet sans que la propriété précédente ne disparaisse (d'où l'adjectif maximal). Un même sommet peut être membre de plusieurs cliques distinctes. Cette mesure peut donc être adaptée aux graphes planaires (remplacer densité par indice γ)

Toutes les mesures suivantes sont des dérivées de celle-ci. Si ces mesures sont apparues, c'est pour une raison simple : il est rare de rencontrer des cliques dans des graphes issus de données empiriques en sociologie ; la rigidité de la définition exclut certains sommets pourtant fortement connectés. Pour illustrer le deuxième aspect, examinons la figure 1. À première vue, deux types de sommets existent : des sommets périphériques (*aeh*) et un noyau central fortement connexe (*bcdfg*). Ce dernier noyau ne sera pas mis en évidence pas la recherche de clique car un lien et un seul manque (*bf*). Si repérez cette configuration est aisée dans un graphe de cette taille, la tâche est évidemment plus compliquée quand le nombre de sommets dépasse plusieurs centaines.

La mesure s'applique de préférence à des graphes simples non valués et non orientés. Si le graphe est valué, il conviendra de choisir un seuil afin de dichotomiser la matrice et de chercher ensuite les cliques présentes.

Si le graphe est orienté, pour cette mesure comme pour toutes les suivantes, on choisit de prendre en considération ou de négliger l'orientation des liens. Par ordre décroissant d'exigence, on distingue la connectivité récursive (*recursively n-connected*), forte (*strongly n-connected*), unilatérale (*unilaterally n-connected*) et faible (*weakly n-connected*). Dans le premier cas, pour tout chemin entre deux sommets i et j, il existe un chemin de j à i parcourant les mêmes sommets dans l'ordre inverse. La connectivité est dite forte, quand pour tout chemin entre deux sommets i et j, il existe un chemin de j vers i qui parcoure éventuellement des sommets différents. La connectivité unilatérale suppose qu'entre deux sommets i et j, il existe un chemin de i vers j ou un chemin de j vers i. Enfin, la connectivité faible ne considère pas la direction et il suffit que deux sommets soient liés par une chaîne.

Enfin, la recherche de clique est inappropriée aux graphes *two-mode* (liens entre deux groupes d'acteurs). Dans ce cas précis, on parle de *biclique* quand tous les liens possibles entre deux groupes d'acteurs sont présents et qu'il n'est pas possible d'ajouter un acteur dans l'un des groupes sans que la première propriété ne disparaisse (mesure proposée par Borgatti et Everett[7]).

La figure 2 montre un graphe *one-mode* (en haut) et un *two-mode* (en bas) et liste les cliques et bicliques présentes [2].

1.2 Variantes liées à la distance

Certains auteurs ont proposé des variantes supplémentaires.

La *n-clique* est un sous graphe maximal où la distance géodésique maximale entre deux sommets est égale à n. L'inconvénient est que les résultats sont des ensembles de sommets chaînés plus qu'interconnectés. On peut en trouver pas moins de 10 dans la figure 1 si $n = 2$.

Les *n-clans* cherchent à limiter cet inconvénient en imposant aux *n-cliques* présents dans un graphe un diamètre maximal égal à n (9 présents dans la figure 1 avec $n = 2$)

Un *n-club* est un sous-graphe maximal de diamètre n. Un *n-club* n'est pas nécessairement un *n-clique* mais sont toujours inclus dans une *n-clique*.

Notons cependant que peu de logiciels proposent un tel choix pour la recherche de sous-graphes.

1.3 Variantes liées au degré

Si les définitions précédentes sont des ajustements de la clique utilisant le diamètre et la distance, les deux mesures présentées ici sont basées sur le degré des sommets. Dans un cas, la tolérance concerne le nombre de liens manquants autorisé, dans l'autre, le nombre minimum de liens exigé.

Un *k-core* est un sous-graphe dans lequel tous les sommets sont reliés à au moins k autres sommets de ce sous-graphe.

Un *k-plex* est un sous-graphe maximal dans lequel tous les sommets envoient au minimum $x - k$ liens vers les autres sommets de ce sous-graphe. Si on fixe n à 4 et k à 2, on trouve 12 *k-plexes* dans la figure 1.

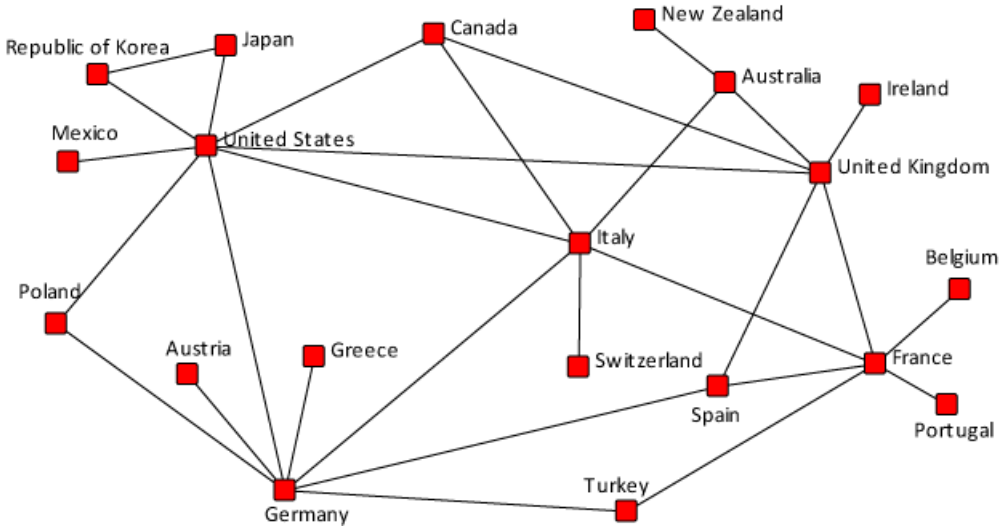
Ces mesures sont principalement employées dans le cas de graphes non planaires, non orientés et binaires. La valeur de k ainsi que la taille minimale du sous-graphe recherché sont fixées par le chercheur et dépendent de la taille et de l'ordre du graphe étudié.

La figure 3 montre les *k-cores* présents dans un graphe où chaque lien symbolise un flux migratoire supérieur à 100 000 migrants entre pays de l'OCDE en 2005 (taille minimale 3 et $k = 2$).

2. Ces exemples sont tirés de [5] et [4]. Les matrices correspondantes sont téléchargeables sur le site www.eurobroadmap.eu

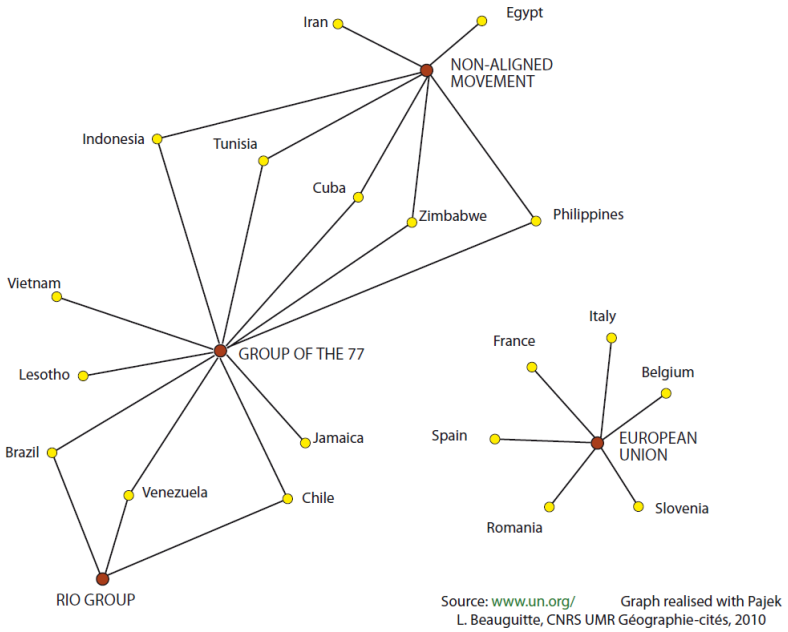
FIGURE 2 – Cliques dans un graphe simple et dans un graphe bipartite

Flux migratoires entre pays de l'OCDE en 2005



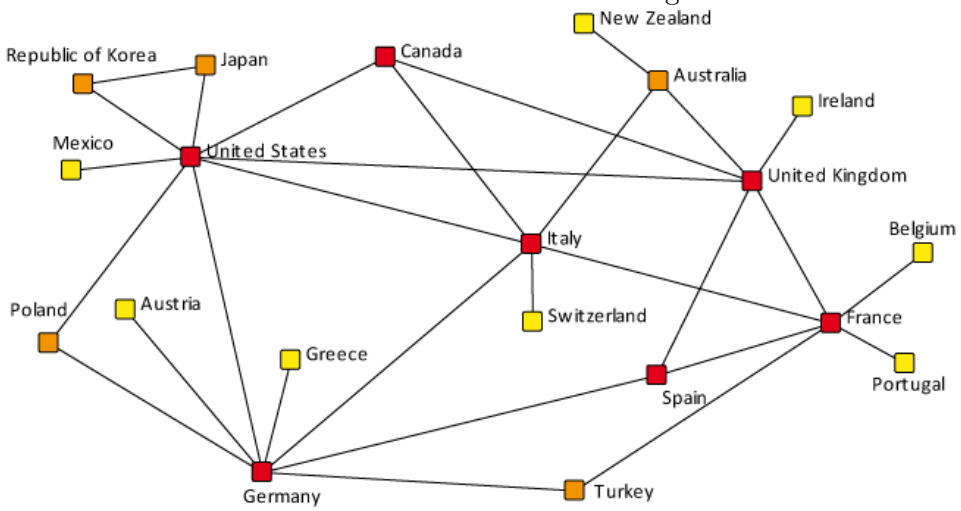
Les cliques présentes ont une taille maximale de 3 sommets et sont les suivantes : Canada - Italy - United States/ Canada - United Kingdom - United States / Germany - Italy - United States / Germany - Poland - United States / Japan - Republic of Korea - United States / France - Spain - United Kingdom.

Déclarations de délégations nationales soutenant des déclarations faites par des groupes à l'Assemblée générale de l'ONU en 1996-1997.



Les bicliques présentent sont au nombre de deux : Brésil, Chili et Venezuela associés au Groupe des 77 et au Groupe de Rio ; Cuba, Indonésie, Philippines, Tunisie et Zimbabwe associés au Groupe des 77 et au Mouvement des Non-Alignés.

FIGURE 3 – K-cores dans un réseau migratoire



Signalons deux autres méthodes pour rechercher les sous-graphes. Les *LS sets* sont des ensembles de sommets qui ont de plus de relations à l'intérieur de chaque ensemble qu'avec l'extérieur. La définition d'un *lambda set* est basée sur la même logique mais utilise la robustesse des liens. Un ensemble de sommets N est un *lambda set* si toute paire de sommets dans N présente des liens plus robustes que toute paire de sommets composée d'un sommet appartenant à N et d'un sommet extérieur. La robustesse des liens est mesurée par le nombre de chemins entre deux sommets ne contenant aucun lien en commun (plus il est élevé, plus la liaison entre les deux sommets est robuste).

2 *Communities*

Comme signalé en introduction, les physiciens n'utilisent pas ou très peu le terme de clique pour désigner un sous-graphe fortement connexe mais celui de communauté (*community*). Les mesures proposées dans cette section sont issues des articles de Arenas *et al.* [3], Boccaletti *et al.*[6], Newman et Girvan [13] et Guimerà et Amaral [11]. D'autres méthodes existent et cette partie ne prétend pas à l'exhaustivité.

La définition donnée des communautés est la suivante : 'division of network nodes into group within which the networks connections are dense, but between which they are sparser'[13]. La méthode proposée par Newman et Girvan est la suivante : *a*) calculer l'intermédiarité des liens du graphe, *b*) supprimer celui qui a la plus forte (en cas de *betweenness* égale, en tirer un au sort), *c*) recalculer l'intermédiarité des liens restants et *b*) à nouveau. La pertinence du découpage obtenu est vérifiée *via* le calcul de la modularité

(*modularity*) :

$$M = \sum_{s=1}^{N_M} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right]$$

Dans cette équation tirée de Guimerà et Amaral [11], M représente la modularité, N_M le nombre de modules, L le nombre de liens du graphe, l_s le nombre de liens entre les sommets du module s et d_s la somme des degrés des sommets du module s. Il semble donc que cette méthode s'applique principalement aux graphes non orientés.

Si Arenas *et al.* reprennent le même principe, Guimerà et Amaral proposent une méthode sensiblement différente : utiliser la *simulated annealing*[3] pour trouver la partition donnant la modularité la plus élevée. Cela revient à rechercher, à partir d'un partitionnement aléatoire en communautés à déplacer des groupes de sommets entre celles-ci pour maximiser la modularité.

Si la méthode a été appliqué avec succès sur un certain nombre de graphes, elle est exigeante en temps de calcul et Newman a proposé peu de temps après une méthode qui vise à maximiser la modularité sans passer par l'étape itérative de suppression de liens [12].

Boccaletti *et al.*[6] liste les algorithmes suivants permettant une partition en communautés : le *spectral graph partitionning* basé sur les vecteurs propres de la matrice d'adjacence (p.276), les classifications hiérarchiques ascendantes (*agglomerative hierarchical clustering*) et descendantes (*divisive hierarchical clustering*), les nuées dynamiques (*k-means*) et diverses variations à l'algorithme de Newman et Girvan déjà évoqué.

En informatique, Schaeffer propose un bon état de la question[15] et propose une définition équivalente au terme de *cluster*[4]. On trouvera un panorama plus complet des méthodes utilisées en informatique dans la thèse (en français) de Pons[14]. Il faut noter que le vocabulaire est parfois flou et que la différence entre *clustering* (au sens strict, créer des groupes ayant de fortes relations internes et de faibles relations externes) et partitionnement de graphes en général (diviser un graphe en plusieurs sous graphes qui ne correspondent pas obligatoirement à des *clusters*) est parfois ténue.

Les physiciens ont également proposé une utilisation intéressante de la méthode des *k-cores* évoquée dans la section précédente : il s'agit lorsque le graphe est de grande taille de déterminer les régions centrales en éliminant pas à pas les sommets dont le degré est inférieur à k. La méthode est particulièrement intéressante pour la visualisation des sommets centraux [2].

Si l'objectif des physiciens (et des informaticiens) est le même que celui des sociologues - trouver des sous-graphes aux relations internes denses -, on

3. Si quelqu'un(e) peut proposer un équivalent francophone, j'en serais ravi.

4. '*Graph clustering* is the task of grouping the vertices of the graph into clusters taking into consideration the edge structure of the graph in such a way that there should be many edges *within* each cluster and relatively few *between* the clusters.'[15], p.27.

note la différence d'approche nette. Là où les seconds se basent sur le degré ou la distance, les premiers fonctionnent de façon itérative en supprimant progressivement certains liens ou sommets. Les deux approches gagnent à être testées en fonction des données étudiées et de la problématique choisie.

Conclusion

Si de nombreuses déclinaisons existent, le principe de base est toujours le même, et le choix de la méthode appropriée dépendra essentiellement de la nature des données. Ainsi, la recherche de cliques est rarement fructueuse dans le cas de graphes issus de données empiriques, d'où les nombreuses variations élaborées par les chercheurs.

Si cette méthode de partitionnement permet d'isoler aisément centre(s) et périphérie(s), elle ne tient compte que du volume de liens et non de leur nature ni même de leur direction dans leur utilisation la plus courante. D'autres outils existent pour mener à bien cette dernière tâche (équivalences), outils qui seront présentés dans un papier à venir. Il paraît également utile, dans un souci de clarté, de différencier la recherche de cliques (communautés, clusters) et les méthodes de partitionnement de graphes au sens large.

Enfin, certains auteurs proposent d'autres outils permettant de mettre en évidence les sous groupes cohésifs, notamment l'utilisation de modèles statistiques (p_1, $p*$ etc.)[9]. Ces modèles étant relativement simples dans leur principe mais délicats à mettre en œuvre, eux aussi seront abordés dans un papier fmr à venir.

Références

[1] R.D. Alba : A graph-theoretic definition of a sociometric clique. *The Journal of Mathematical Sociology*, 3(1):113–126, 1973.

[2] J.I. Alvarez-Hamelin, L. Dall Asta, A. Barrat et A. Vespignani : Large scale networks fingerprinting and visualization using the k-core decomposition. *Advances in neural information processing systems*, 18:41, 2006.

[3] A. Arenas, L. Danon, A. Díaz-Guilera, P.M. Gleiser et R. Guimerà : Community analysis in social networks. *The European Physical Journal B*, 38(2):373–380, 2004.

[4] L. Beauguitte : Basic notions on SNA : Dealing with 2-mode networks, September 2010 (http ://cnrs4.nfrance.com/ euro_drupal/node/26).

[5] L. Beauguitte : Basic notions on SNA : Looking for subgroups, March 2010 (http ://cnrs4.nfrance.com/ euro_drupal/node/26).

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez et D.U. Hwang : Complex networks : Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[7] S.P. Borgatti et M.G. Everett : Network analysis of 2-mode data. *Social Networks*, 19(3):243–269, 1997.

[8] C. Ducruet : Les mesures globales d'un réseau. *Groupe fmr*, 8p., 2010 (http ://halshs.archives-ouvertes.fr/halshs-00541902).

[9] K.A. Frank : Mapping interactions within and between cohesive subgroups. *Social networks*, 18(2):93–119, 1996.

[10] R. Guimerà et L.A.N. Amaral : Cartography of complex networks : modules and universal roles. *Journal of Statistical Mechanics : Theory and Experiment*, 2005(2), 2005.

[11] R. Guimerà et L.A.N. Amaral : Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[12] M.E.J. Newman : Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):66133, 2004.

[13] M.E.J. Newman et M. Girvan : Finding and evaluating community structure in networks. *Physical review E*, 69(2):26113, 2004.

[14] P. Pons : *Détection de communautés dans les grands graphes de terrain.* Thèse de doctorat, Paris 7, 2007(http ://psl.pons.free.fr/publi/these_pascal_pons.pdf).

[15] S.E. Schaeffer : Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[16] S. Wasserman et K. Faust : *Social Network Analysis. Methods and Applications.* Structural analysis in the social sciences. Cambridge University Press, 1994.

Table des matières