Simplification et partitionnement d'un graphe

César Ducruet

Simplification et partitionnement d'un graphe

César Ducruet, CNRS, UMR Géographie-cités
ducruet<at>parisgeo.cnrs.fr

Mars 2011 - Version 1

groupe f.m.r.

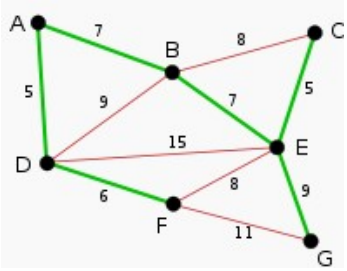
Introduction

Le partitionnement se définit comme le découpage d'un graphe en sous-ensembles disjoints et non vides dont l'union est égale au graphe d'origine. La simplification d'un graphe correspond elle à la suppression de certains sommets et/ou liens, mais n'aboutit pas forcément à la partition du graphe. Ces deux opérations visent deux buts principaux : a) faire ressortir la structure profonde du réseau, notamment pour ceux de grande taille dont la visualisation n'apporte rien ; et b) identifier des groupes cohérents au sein du réseau (régions, cliques, bassins). Certains aspects sont communs avec la recherche de cliques et dérivées dans le graphe (Beauguitte, 2011[3]), sur lesquels nous ne revenons pas, mais le partitionnement s'en distingue par le fait qu'il ne définit pas à l'avance la nature des sous-ensembles recherchés. Nous proposons une simple sélection de méthodes dont la plupart sont déjà connues (et appliquées par) des géographes, ou qui pourraient trouver en géographie un terrain favorable.

1 L'arbre d'étendue minimum

Selon la question posée au départ, on peut ne retenir du graphe que certains éléments sans toutefois remettre en question le principe de connexité. C'est le cas de la recherche et de l'extraction (au niveau visuel) d'un chemin spécifique dans le graphe ayant pour caractéristique d'être le plus court ou le plus long (distance kilométrique ou topologique), le plus dense ou le moins dense (trafic), de façon à ce que tous les sommets soient conservés. Par exemple, l'arbre d'étendue minimum (*minimum spanning tree*, ou *minimum weight spanning tree*) est un sous-graphe dont la somme des valeurs associées aux liens est la plus faible par rapport aux autres chemins possibles reliant

FIGURE 1 – Arbre d'étendue minimum par application de l'algorithme de Kruskal



On obtient ainsi une chaîne de sommets, le sommet E étant en position plus centrale que les autres. Les liens verts forment le chemin d'étendue minimum.

tous les sommets. L'un des algorithmes les plus couramment utilisés est celui de Kruskal (1956)[15], il est notamment intégré dans le programme TULIP. Les algorithmes de ce genre peuvent être classés selon leur angle d'attaque : recherche de l'arbre le plus court au niveau global (Kruskal, 1956[15] ; Roy, 1959[18] ; Warshall, 1962[20] ; Floyd, 1962[9] ; Johnson, 1977[14]) ou bien à partir d'un sommet et/ou d'une paire de sommets (Bellman, 1958[4] ; Dijkstra, 1959[6] ; Ford and Fulkerson, 1962[10]).

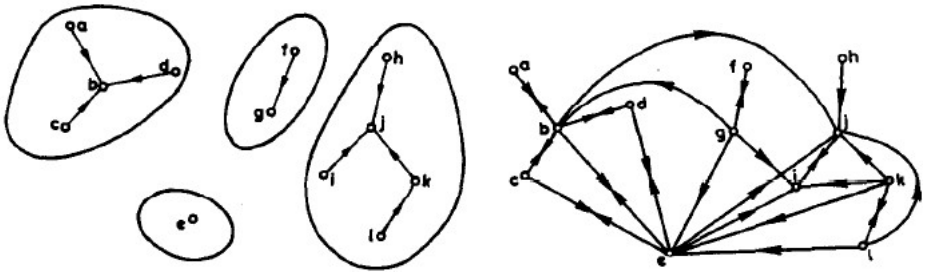
Depuis la formulation des algorithmes de base par ces mathématiciens de la théorie des graphes, une multiplicité de variantes a été proposée, notamment dans le but d'optimiser le temps (informatique) de la recherche de l'arbre le plus court dans le cas des très grands réseaux, comme l'algorithme A-Star (A^*), lui aussi faisant l'objet de nombreuses variantes. L'arbre résultant de cette recherche peut être, dans tous les cas, étudié comme objet à part, distinct du graphe d'origine dont les autres liens ont été supprimés pour l'analyse, d'où l'idée de la simplification. Les applications sont nombreuses : recherche du plus court chemin ou du flot minimum dans un réseau de transport, dans une ville, entre des villes, des pages Internet, des serveurs, des personnes, etc. On peut bien entendu appliquer ces algorithmes aux graphes orientés.

2 La méthode des flux majeurs

L'algorithme ayant connu la plus grande diffusion en géographie est certainement celui dit des « flux majeurs » proposé par Nystuen et Dacey (1961[16]) et proposé en version française par Dumolard (1975[8]). Pratique et simple d'utilisation, cet algorithme est également remis en question pour la perte d'information qu'il implique. Son principe est le suivant : ne retenir pour chaque sommet que le lien le plus fort (en valeur) vers un autre sommet,

FIGURE 2 – Exemple d'application des flux majeurs et dominants (source : Puebla, 1987[17])

To centre:	a	b	c	d	e	f	g	h	i	j	k	l
From centre:												
a	00	75*	15	20	28	02	03	02	01	20	01	00
b	69*	00	45*	50*	58*	12	20	03	06	35*	04	02
c	05	51*	00	12	40*	00	06	01	03	15	00	01
d	19	57*	14	00	30*	07	06	02	11	18	05	01
e	07	40*	48*	26*	00	07	10	02	37*	39*	12	06
f	01	06	01	01	10	00	27*	01	03	04	02	00
g	02	16*	03	03	13*	31*	00	03	18*	08	03	01
h	00	04	00	01	03	03	06	00	12	38*	04	00
i	02	28	03	06	43	04	16	12	00	98*	13	01
j	07	40	10	08	40	05	17	34	98*	00	35	12
k	01	08	02	01	18*	00	06	05	12*	30*	00	15*
l	00	02	00	00	07*	00	01	00	01	06*	12*	00
Total	113	337	141	128	290	71	118	65	202	311	91	39
Rank order:	8	1	5	6	3	10	7	11	4	2	9	12



En bas à gauche : flux majeurs (soulignés dans la matrice) ; en bas à droite : flux dominants (*). Le sommet e apparaît isolé à gauche étant « indépendant » (taille supérieure à c), mais très central à droite, où en revanche on perd la partition en sous-graphes...

et supprimer tous les autres. Le résultat est un arbre (*tree*) dans lequel les sommets dont le flux majeur relie un sommet de moindre poids sont « indépendants » (*independent node*) et les sommets dont le flux majeur relie un sommet de poids plus grand sont « subordonnés » (*subordinate nodes*). Les sommets subordonnés reliés à d'autres sommets subordonnés restent sous la domination du sommet indépendant en remontant la hiérarchie selon le principe de transitivité (*transitivity*[1]), et les sommets de degré 1 sont des « terminaux » (*terminal nodes*).

Selon les cas, l'algorithme produit une partition ou non du graphe d'origine en plusieurs sous-graphes ou régions dites « nodales » (*nodal regions*),

1. Le terme est utilisé dans l'article original ; le sens actuel du terme en analyse des réseaux est différent, voir [7].

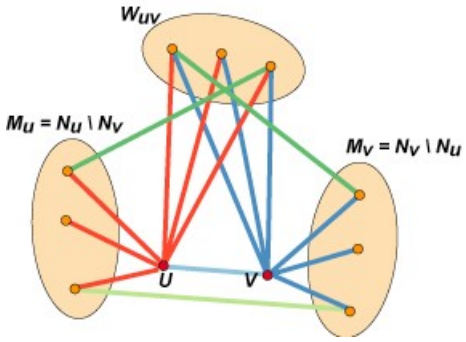
chacun étant polarisé par un seul sommet indépendant. Les nombreuses applications de cet algorithme montrent l'influence de la distance géographique et les effets de barrière (ex : administrative, douanière) sur la formation des régions ainsi obtenues, notamment dans le cas des flux ferroviaires et aériens (Cattan, 1995[5]) et migratoires dans l'étude des pôles d'emploi par l'INSEE. Gleyze (2008[12]) propose à partir des plus courts chemins au sein du réseau routier français une analyse de la régionalisation sous-jacente, notamment en agrégeant les sommets de chaque région nodale pour observer le phénomène à un niveau supérieur. La faiblesse principale de cette méthode est d'ignorer les liens ayant une valeur forte mais inférieure au flux majeur (*single linkage analysis*). Elle trouve en partie sa solution dans l'approche par les flux dominants (*multiple linkage analysis*) que l'on définit à partir d'un certain seuil : les 10 liens majeurs de chaque sommet, ou encore les liens dont la valeur totale concentre au moins 50% de la valeur totale du sommet (degré pondéré), etc. Si la méthode des flux majeurs permet de gagner en visibilité, celle des flux dominants en revanche court le risque de brouiller l'analyse par l'enchevêtrement des liens conservés.

3 Le filtrage par l'indice de cohésion des liens

Afin d'obtenir des sous-groupes sans toutefois perdre une partie importante de l'information originelle contenue dans le graphe, Amiel *et al.* (2005[1]) proposent d'appliquer un indice de cohésion aux liens entre villes aéroportuaires, qui se définit comme étant le « ratio entre le nombre de liens effectifs entre les voisins des sommets incidents à l'arête, et le nombre maximal de liens qui pourraient être observés », et reste proche de la transitivité et de l'indice de force (Ducruet, 2010[7] [2]) avec la différence que cet indice est rapporté à la valeur des liens (trafic de passagers). La méthode ne retient donc que les sommets ayant autant ou plus de voisins communs que de voisins non communs à partir du graphe d'origine. Selon les cas, les résultats permettent de révéler différents niveaux d'organisation dans le graphe, notamment selon des logiques régionales comme dans l'aérien, ainsi que des hiérarchies entre niveaux. D'autres applications en géographie ont été proposées, comme sur les migrations quotidiennes pour mettre en valeur l'organisation polycentrique ou monocentrique des aires urbaines françaises (Rozenblat et Tissandier, 2007[19]).

2. Pour un aperçu méthodologique, voir : http ://mappe-monde.mgm.fr/num7/articles/ENCA1.html

FIGURE 3 – Filtrage des liens par l'indice de cohésion (source : Amiel *et al.*, 2005)

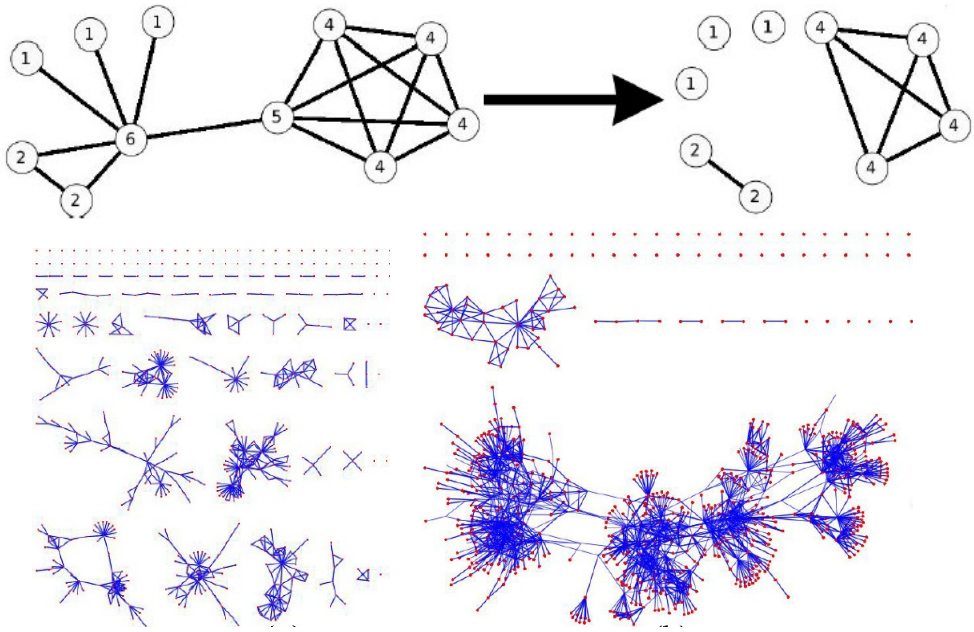


N_u et N_v : ensemble de leurs voisins respectifs
M_u et M_v : ensemble des sommets n'étant pas voisins communs
W_{uv} : ensemble des voisins communs des sommets u et v

4 La décomposition topologique (*topological decomposition*)

Cette méthode s'intéresse avant tout aux relations entre sommets de degré identique ou comparable, qui sont isolés du reste du graphe soit « par le haut » (sommets de fort degré) soit « par le bas » (sommets de faible degré), à partir d'un certain seuil jugé pertinent (Zaidi, 2010[21]). Elle part du principe que même dans un réseau très hiérarchique, des relations transversales existent en dehors du rôle central des grands *hubs*. Le seuil en question (*cutoff limit* ou *fragmentation threshold*) est décidé par l'utilisateur, il peut être le degré moyen (*average degree*) ou bien une valeur à laquelle un changement majeur se produit dans l'organisation du graphe. La suppression des sommets dont le degré est inférieur, égal ou supérieur au seuil choisi entraîne un partitionnement du graphe. Lorsque l'on supprime successivement les sommets de degré supérieur, l'émergence d'un composant géant (*giant component*) indique à quel moment la plupart des sommets se retrouvent dans un seul sous-graphe connexe (Janson *et al.*, 1993[13]). Ce phénomène est appelé percolation (ou *phase transition*) par les physiciens, il permet de fixer un seuil pertinent *a posteriori* à partir duquel le composant géant occupe un tel poids que l'on ne peut plus discerner d'autres sous-groupes cohérents (Barabási, 2002[2]). La densité (indice γ) moyenne des sous-graphes ou bien le pourcentage de sommets (et/ou de liens) présents dans le composant géant par rapport au total du graphe sont des indicateurs utiles pour mettre en valeur un seuil. Selon la méthode, on obtient des sous-graphes de degré induit maximum (MaxDIS) ou bien des sous-graphes de degré induit

FIGURE 4 – La décomposition topologique (source : Zaidi, 2010)



Haut : résultat de la suppression des sommets de degré supérieur ; bas : application au réseau aérien mondial. À droite, sous-graphe maximum induit de degré 10 (MaxDIS10) ; à gauche, sous-graphe maximum induit de degré 15 (MaxDIS15). On observe à gauche de nombreuses configurations très variées (hubs, cliques, combinaisons) et à droite l'émergence du composant géant

minimum (MinDIS).

La Figure 4 montre le résultat de cette méthode MaxDIS et son application à un réseau aérien. Le seuil pertinent c se caractérise en général par une densité moyenne forte des sous-graphes obtenus et une taille de composant géant ne dépassant pas les 50% du graphe d'origine. On peut aussi recourir au simple comptage du nombre de sous-graphes obtenus par valeur de k. De façon logique, le nombre de sous-graphes diminue et leur taille moyenne s'accroît à mesure que des sommets de degré supérieur sont inclus. La méthode, au final, permet de vérifier les configurations topologiques des sous-graphes obtenus à partir des liens transversaux entre les sommets, qu'elles soient des cliques, des chaînes, des étoiles, ou des sous-graphes complets, tout en mettant de côté la dimension hiérarchique, notamment dans les réseaux invariants d'échelle (*scale-free networks*). On peut aussi pointer les sommets grâce auxquels les sous-graphes se connectent, et donc identifier les « ponts » (*bridges*). L'inconvénient majeur de la méthode est l'aller-retour nécessaire

entre la décomposition et les résultats avant de fixer un seuil.

Conclusion

Bien que d'autres méthodes existent, et que celles présentées soient en évolution permanente, on peut reconnaître à ces dernières leur simplicité, parfois extrême. Si la méthode des flux majeurs met en avant l'existence de sous graphes fonctionnels fondés sur le principe de la hiérarchie, la décomposition topologique, au contraire, se focalise sur les liens transversaux entre sommets de degré comparable. On peut imaginer une décomposition topologique qui se baserait sur d'autres indicateurs, comme le poids démographique ou économique des villes, afin de vérifier les relations entre objets semblables, ou sur d'autres mesures topologiques, sachant que le degré est bien souvent une mesure-clé. Si tout graphe de taille conséquente peut inclure des sous-graphes intéressants, il serait souhaitable de démonter plus avant pour quel type de graphe telle ou telle méthode de partitionnement est plus à même de révéler des sous-ensembles pertinents (Fortunato, 2010[11]). Certaines méthodes s'appliquent mieux aux réseaux invariants d'échelle ou bien aux réseaux petit-monde (Zaidi, 2010[21]). Les mesures globales suffisent-elles à donner une telle indication ?

Références

[1] M. Amiel, G. Mélançon et C. Rozenblat : Réseaux multi-niveaux : l'exemple des échanges aériens mondiaux de passagers. *Mappemonde*, (3):12, 2005.

[2] A.L. Barabási : *Linked : The New Science of Networks*. Basic Books, 2002.

[3] L. Beauguitte : Blockmodeling et équivalences. *Groupe fmr*, 9 p., 2011 (http ://halshs.archives-ouvertes.fr/FMR/fr/).

[4] R. Bellman : On a Routing Problem. *Quarterly of Applied Mathematics*, 16(1):87–90, 1958.

[5] N. Cattan : Barrier effects : the case of air and rail flows. *International Political Science Review*, 16(3):237–248, 1995.

[6] E. W. Dijkstra : A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[7] C. Ducruet : Les mesures locales d'un réseau. *Groupe fmr*, 10 p., 2010 (http ://halshs.archives-ouvertes.fr/FMR/fr/).

[8] P. Dumolard : Région et régionalisation, une approche systémique. *L'Espace géographique*, (2):93–111, 1975.

[9] R.W. Floyd : Algorithm 97 : Shortest Path. *Communications of the ACM*, 5(6):345, 1962.

[10] L. Ford et D. Fulkerson : *Flows in Networks*. Princeton University Press, 1962.

[11] S. Fortunato : Community detection in graphs. *Physics Reports*, (486):75–174, 2010.

[12] J.F. Gleyze : La simplification des réseaux en géographie. L'évaluation des dépendances relationnelles induites par les chemins au sein du réseau et l'agrégation des sommets autour de plate-formes relationnelles. *Document de travail COGIT-IGN*, 2008.

[13] S. Janson, D.E. Knuth, T. Łuczak et B. Pittel : The birth of the giant component. *Random Structures & Algorithms*, 4(3):233–358, 1993.

[14] D.B. Johnson : Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, 24(1):1–13, 1977.

[15] J.B. Kruskal : On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[16] J.D. Nystuen et M.F. Dacey : A graph theory interpretation of nodal regions. *Papers in Regional Science*, 7(1):29–42, 1961.

[17] J.G. Puebla : Spatial structures of network flows : A graph theoretical approach. *Transportation Research Part B*, 21(6):489–502, 1987.

[18] B. Roy : Transitivité et connexité. *CR Acad. Sci. Paris*, 249:216–218, 1959.

[19] C. Rozenblat et P. Tissandier : Commuter graphs and cities' polycentric cohesion. *In Proceedings of the 15 th ECTQG, Montreux, Switzerland, September*, pages 7–11, 2007.

[20] S. Warshall : A theorem on boolean matrices. *Journal of the ACM*, 9(1):11–12, 1962.

[21] F. Zaidi : *Analysis, structure and organization of complex networks*. Thèse de doctorat, Thèse de Doctorat en Informatique, LABRI, Université de Bordeaux I, 2010.