Partitionnement multi-échelle d'un graphe en communautés : détection des échelles pertinentes

Nicolas TREMBLAY, Pierre BORGNAT

Laboratoire de Physique de l'ENS de Lyon
CNRS UMR 5672
46 allée d'Italie, 69007, Lyon, France
prenom.nom@ens-lyon.fr

Résumé – La méthode de partitionnement multi-échelle d'un graphe récemment introduite [10] nous permet de partitionner le graphe en communautés à une échelle donnée. Le facteur d'échelle étant un paramètre continu, nous nous posons la question du choix des échelles pertinentes, et proposons une méthode pour les détecter qui se base sur des notions d'instabilité des partitions associées. Nous testons la méthode sur un exemple de graphe hiérarchique qui sert de banc d'essai et estimons avec succès ses différentes échelles pertinentes.

Abstract – The present work elaborates on [10] which introduces a procedure identifying community structures in a graph at different scales. We focus here on detecting the relevant scales of a graph. To this end, we propose a method based on several notions of instability of the partition associated to each scale: if it is stable with respect to a perturbation, then the associated scale is considered relevant. An example on a graph benchmark having hierarchical communities shows that we estimate successfully its relevant scales.

1 Introduction

Les graphes sont des outils de modélisation naturels des réseaux. Ils sont beaucoup utilisés dans tous les domaines où apparaissent des réseaux : télécommunication, biologie, épidémiologie, sociologie... [1] Afin de simplifier ces graphes de terrains dont certains atteignent de très grandes tailles, une des caractéristiques les plus largement utilisées est l'existence de communautés, c'est-à-dire des ensembles de nœuds plus connectés entre eux qu'avec le reste du graphe. Le partitionnement en communautés d'un graphe permet d'en faire une description simplifiée, et donne également des indications sur ses nœuds : en sachant à quelle communauté appartient un nœud, on peut en déduire certaines de ses propriétés. Le partitionnement automatique d'un graphe en communautés est une problématique de recherche très active [2]. Une méthode désormais classique est de chercher la partition qui maximise une fonction appelée modularité. Malgré de nombreux succès, certaines limites de la modularité ont été mises au jour, notamment sa résolution intrinsèque de description : la modularité favorise certaines tailles de communautés plutôt que d'autres. De plus, elle s'avère moins efficace face à une distribution hétérogène de la taille des communautés. C'est ce qui a motivé certains auteurs à s'intéresser à une description multi-échelle en communautés (par exemple [9] et [6]).

Inspirés par [3] qui introduit à la fois des ondelettes sur graphe et une définition intéressante d'un paramètre d'échelle, nous avons développé une méthode [10] basée sur la matrice de corrélation des ondelettes centrées en chaque nœud qui nous permet de partitionner le graphe à une échelle donnée. Une fois que nous avons le partitionnement du graphe en fonction de l'échelle (le paramètre d'échelle étant continu, nous pouvons obtenir –moyennant du temps de calcul– une description aussi résolue en échelle que souhaité), se pose la question des échelles pertinentes : quelles échelles de description sont les plus naturelles pour ce graphe ? Peut-on obtenir un classement automatique des « meilleures » échelles ?

Nous rappelons la définition des ondelettes sur graphe dans la partie 2, ainsi que la définition des fonctions d'échelles sur graphe introduites dans [10] qui s'avèrent plus stables que les ondelettes pour la détection de communautés. Dans la partie 3, nous résumons notre méthode de partitionnement multi-échelle. Finalement, nous proposons la méthode de détection des échelles pertinentes –la principale contribution de cette communication– dans la partie 4 ; avant de conclure dans la partie 5.

2 Ondelettes et fonctions d'échelle sur graphe

Dans [3], des ondelettes sont définies par analogie au cas classique, en filtrant des impulsions de Dirac centrées en chaque nœud par un filtre passe-bande plus ou moins étiré par un paramètre d'échelle s. Pour ce faire, la transformée de Fourier sur graphe est utilisée. Nous rappelons ici les notations et principales équations. Soit $\mathcal{L}$ le laplacien normalisé [1] défini par $\mathcal{L} = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}$ où A est la matrice d'adjacence du

1. Nous choisissons ce laplacien plutôt que le laplacien non-normalisé pour son lien avec les méthodes de détection de communautés par modularité.

graphe et D la matrice diagonale des forces [2] des nœuds du graphe. $\mathcal{L}$ est réelle symétrique, donc diagonalisable, et nous notons χ_i son i-ième vecteur propre de valeur propre λ_i. Les valeurs propres sont ordonnées et vérifient : $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{N-1} \leq 2$ où N est le nombre de nœuds du graphe. La transformée de Fourier d'un signal f défini sur le graphe s'écrit alors : $\hat{f} = \chi^\top f$ où $\chi = (\chi_0|\chi_1|\ldots|\chi_{N-1})$ est la matrice des vecteurs propres.

Soit g un noyau de filtre passe-bande qui sera étiré par le paramètre d'échelle $s > 0$. Nous notons sa représentation matricielle à l'échelle s : $\hat{G}_s = \text{diag}(g(s\lambda_0), \ldots, g(s\lambda_{N-1}))$. La base des ondelettes à cette échelle s'écrit :

$$\Psi_s = (\psi_{s,0}|\psi_{s,1}|\ldots|\psi_{s,N-1}) = \chi \hat{G}_s \chi^\top, \qquad (1)$$

où $\psi_{s,a}$ est l'ondelette centrée autour du nœud a.
Nous avons introduit dans [10] les fonctions d'échelle associées. Soit h le noyau de filtre passe-bas qui vérifie :

$$h(x) = \left(\int_x^\infty \frac{|g(x')|^2}{x'} dx' \right)^{1/2}. \qquad (2)$$

Soit $\hat{H}_s = \text{diag}(h(s\lambda_0), h(s\lambda_1), \ldots, h(s\lambda_{N-1}))$ la matrice filtre passe-bas à l'échelle s, la base des fonctions d'échelle s'écrit :

$$\Phi_s = (\phi_{s,0}|\phi_{s,1}|\ldots|\phi_{s,N-1}) = \chi \hat{H}_s \chi^\top, \qquad (3)$$

où $\phi_{s,a}$ est la fonction d'échelle centrée autour du nœud a. La forme du noyau de filtre g ainsi que les bornes du paramètre d'échelle s sont importants pour l'efficacité de la détection multi-échelle de communautés. Ils sont discutés dans [10].

3 Méthode de partitionnement en communautés à une échelle donnée

A un paramètre d'échelle s donné, notre méthode de partitionnement se résume comme suit :

1. Vecteurs caractéristiques. A chaque nœud a est associé un vecteur caractéristique. Selon ce qui est recherché, nous utilisons soit l'ondelette centrée en a, $\psi_{s,a}$, soit la fonction d'échelle centrée en a, $\phi_{s,a}$.

2. Distance entre deux nœuds. La distance entre deux nœuds est la distance de corrélation [3] (1 - le coefficient de corrélation) entre leurs vecteurs caractéristiques.

3. Algorithme de partitionnement. Nous utilisons un algorithme de partitionnement hiérarchique avec la méthode de chaînage moyenné (*average linkage* [4]). Sur l'exemple considéré dans la partie 4, utiliser cette méthode plutôt que la méthode de chaînage complet (*complete linkage*) ou simple (*single linkage*) ne change rien aux résultats. Nous préférons utiliser de base la méthode de chaînage moyenné qui est réputée plus stable. Cet algorithme produit un dendrogramme.

4. Choisir où couper le dendrogramme. Chaque coupe de ce dendrogramme définit une partition possible. Quelle coupe est la meilleure ? Nous proposons de le couper au niveau du plus grand saut entre deux de ses nœuds. Couper au niveau du plus grand saut est une idée connue mais nécessite pour être justifiée un rééchantillonnage et un test statistique (cf. *gap statistics* [4]). Ici, la justification est donnée par la notion de stabilité de la partition associée que nous discutons dans la suite.

2. La force d'un noeud est la somme des poids de ses liens

3. La distance de corrélation n'est pas une vraie distance : en particulier, elle ne vérifie pas l'inégalité triangulaire.

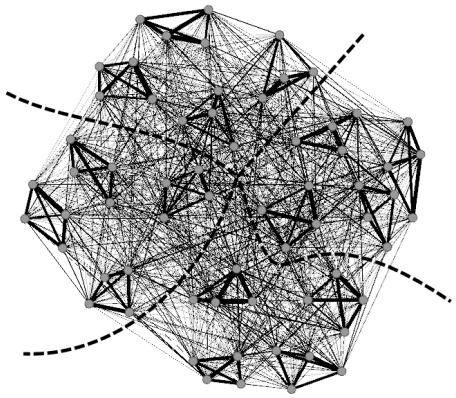


FIGURE 1 – Une réalisation d'un graphe hiérarchique discuté dans la section 4. Pour la clarté de la représentation, chaque communauté de 10 nœuds est dessinée comme un seul nœud. Les tirets marquent le découpage en quatre communautés.

4 Détection des échelles pertinentes

Dans [10], nous donnons les bornes s_{min} et s_{max} du paramètre d'échelle s au-delà desquelles il est illusoire de penser trouver des informations intéressantes sur les communautés. Notons $S = \{s_1 = s_{min}, s_2, \ldots, s_M = s_{max}\}$ l'ensemble des échelles auxquelles nous allons nous intéresser. Nous les choisissons espacées logarithmiquement car la densité des valeurs propres sur l'intervalle $[0, 2]$ n'est pas uniforme : elles sont beaucoup plus regroupées autour de 1 que de 0 pour des graphes complexes avec communautés [7]. Ainsi, une petite différence de paramètre à petite échelle (une petite échelle prend plus en compte les grandes valeurs propres) a un impact plus important sur le partitionnement qu'une même différence de paramètre à grande échelle. Pour le choix du nombre d'échelles étudiées, nous nous inspirons du cas classique (discret 1D) : $M = K \log_2(N)$ où K est typiquement inférieur à 10.

Afin d'illustrer les notions introduites dans la suite, nous choisissons, comme d'autres auteurs [6], d'appliquer la méthode à une réalisation d'un banc d'essai hiérarchique [8] : 64 communautés de 10 nœuds (niveau de description 3) sont imbriquées dans 16 communautés de 40 nœuds (niveau de description 2), elles-mêmes imbriquées dans 4 communautés de 160 nœuds (niveau de description 1). Ce banc d'essai a deux paramètres : $\bar{k}$, le degré moyen des nœuds, contrôle la densité du graphe ; et un deuxième paramètre ρ contrôle le ratio des liens intra et extra communautaires. Nous choisissons, comme dans [6] : $\bar{k} = 16$ et $\rho = 1$. Une illustration de ce graphe est donnée sur la Fig. 1. Nous discrétisons ici l'intervalle des échelles possibles en $M = 50$ échelles ($\log_2(640) \sim 10$).

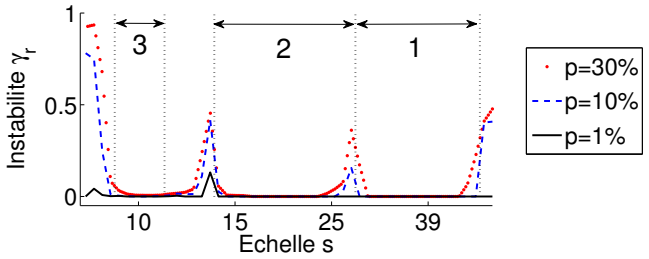


FIGURE 2 – Instabilité γ_r en fonction de l'échelle s pour trois paramètres différents : $p = 1, 10$ et 30%. Dans l'intervalle d'échelles numéro 1 (resp. 2, 3) délimité par les lignes pointillées verticales, l'algorithme trouve exactement la partition théorique à grande (resp. moyenne, petite) échelle.

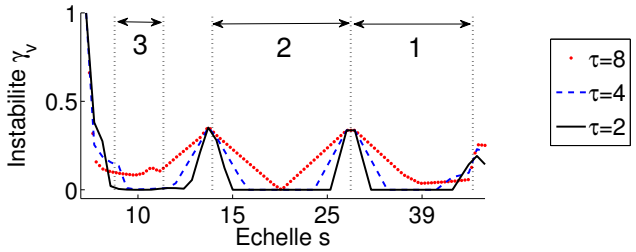


FIGURE 3 – Instabilité γ_v en fonction de l'échelle s pour trois paramètres différents : $\tau = 2, 4$ et 8. Les intervalles d'échelles numérotés sont les mêmes que ceux de la Fig. 2.

4.1 Deux indicateurs d'instabilité

Parmi ces M échelles, lesquelles sont les plus pertinentes ? Nous explorons deux méthodes inspirées par [6].

Méthode de rééchantillonnage Une première méthode est de rééchantillonner le graphe en modifiant aléatoirement de $\pm p\%$ la valeur de chaque lien et de refaire le partitionnement pour chacun des graphes rééchantillonnés (typiquement 50). Puis, à une échelle donnée s, nous caractérisons en calculant l'indice de Rand ajusté [5] la similarité de toutes les paires de partition à cette échelle, dont nous obtenons la moyenne $m(s)$ et l'écart-type $\sigma(s)$. Définissons un premier indicateur d'instabilité $\gamma_r(s) = 1 - (1 - \sigma(s)) * m(s)$ ($\gamma_r \in [0, 1]$). Plus $\gamma_r(s)$ est petit, plus la partition initiale est stable, et plus l'échelle s est jugée pertinente. Pour un p trop petit, les graphes rééchantillonnés seront trop similaires au graphe initial et on trouvera que toutes les échelles sont stables. Pour un p trop grand, les graphes rééchantillonnés seront trop différents du graphe initial et on trouvera que toutes les échelles sont instables. C'est ce que nous observons sur la Fig. 2 où nous traçons l'instabilité γ_r en fonction de l'échelle pour trois paramètres différents : $p = 1, 10$ et 30%. Plus p est grand et plus les échelles sont jugées instables. Un inconvénient de cette méthode est qu'elle privilégie les grandes échelles par rapport aux petites. En effet, $\pm p\%$ sur les liens d'une communauté de 10 nœuds a plus d'impact sur le partitionnement que sur les liens d'une communauté de 160 nœuds. Cet effet est observable pour $p = 1\%$ où l'instabilité est mesurable uniquement à petites échelles.

Méthode de comparaison avec les échelles voisines Une deuxième méthode est de calculer, à une échelle donnée s, la moyenne de la similarité de la partition trouvée à cette échelle avec les partitions des échelles voisines de s. Nous utilisons pour ce faire le même indice de similarité Rand ajusté. La notion de voisinage est paramétrée par τ et nous définissons un deuxième indicateur d'instabilité de la manière suivante :

$$\gamma_v(s) = 1 - \frac{1}{2 * \tau} \sum_{k \in [s-\tau, s+\tau], k \neq s} \text{simi}(P_s, P_k)$$

où P_k symbolise la partition trouvée à l'échelle k. Plus $\gamma_v(s)$ est petit, plus les partitions associées aux échelles voisines de s sont similaires à celle de l'échelle s, et plus l'échelle s est jugée pertinente. La Fig. 3 montre l'instabilité γ_v en fonction de l'échelle pour trois paramètres différents : $\tau = 2, 4$ et 8. On voit que $\tau = 8$ est trop grand car des échelles stables sont jugées instables (surtout à petites échelles). L'inconvénient de cette méthode est que le paramètre τ optimal va dépendre du nombre d'échelles que l'on considère. En effet, si on prend beaucoup d'échelles, alors les partitions associées aux échelles voisines auront plus de chance d'être similaires, et le τ optimal sera plus grand que si on prend un nombre d'échelles restreint.

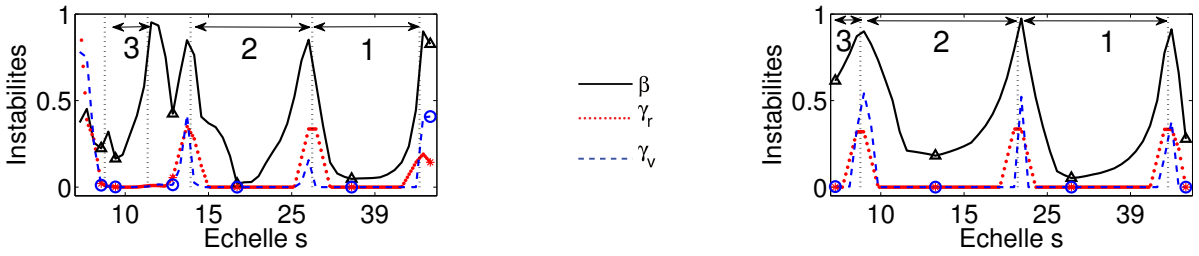
Les Fig. 2 et 3 sont tracées pour le calcul de partitionnement utilisant les ondelettes, mais les résultats sont similaires avec les fonctions d'échelles. L'objectif ici étant de retrouver les 3 niveaux de description du graphe, on voit que chercher les minima globaux des instabilités γ_r et/ou γ_v sont de bonnes solutions : les partitions associées à ces minima globaux sont les 3 partitions théoriques du graphe (les intervalles d'échelles qui donnent les partitions théoriques sont représentés par des lignes pointillées verticales sur les figures). Néanmoins, au lieu d'effectuer ces calculs d'instabilité à toutes les échelles et afin de réduire le temps de calcul, nous proposons de calculer ces instabilités uniquement pour un ensemble restreint d'échelles S^* potentiellement stables, que nous introduisons ci-dessous.

4.2 Présélection d'échelles potentiellement stables

Préselectionnons les échelles potentiellement stables de la manière suivante :

1. Considérons le dendrogramme à l'échelle s, et δ_1 et δ_2 le plus grand et deuxième plus grand sauts du dendrogramme. Notons $\beta(s) = \frac{\delta_2(s)}{\delta_1(s)}$ un troisième indicateur d'instabilité que nous introduisons. L'intuition derrière cet indicateur est la suivante : si le plus grand saut est nettement plus grand que le deuxième plus grand saut (si β est proche de 0), alors le choix de couper le dendrogramme au niveau de δ_1 est univoque : la partition est probablement stable. En revanche, si β est proche de 1, c'est-à-dire si $\delta_1 \simeq \delta_2$, alors le choix de couper le dendrogramme au niveau de δ_1 est moins pertinent : la partition est probablement peu stable.

2. Nous cherchons les échelles où la partition correspondante est la plus stable, c'est-à-dire les minima locaux de β. Parmi les



a) b)

FIGURE 4 – Instabilités β, γ_v et γ_r en fonction de l'échelle s. Les intervalles d'échelles numérotés sont les mêmes que ceux de la Fig. 2. Les abscisses des triangles noirs représentent S^*. Les ronds bleus représentent $\{\gamma_r(s), s \in S^*\}$ et les étoiles rouges $\{\gamma_v(s), s \in S^*\}$. Les résultats présentés dans la figure a (resp. b) ont été calculés avec les ondelettes (resp. les fonctions d'échelle).

partitions associées à ces minima locaux, certaines sont identiques. Pour une partition donnée, nous gardons uniquement le minimum local associé qui a la plus petite instabilité. De plus, nous ne considérons pas les échelles dont les partitions séparent le graphe en plus de $N/2$ communautés, ce qui impliquerait au moins une communauté à un seul nœud, ce qui n'est pas intéressant. Nous obtenons ainsi un ensemble S^* d'échelles potentiellement stables.

S^* est très rapide à calculer, et nous avançons que les minima globaux des instabilités γ_v et γ_r se trouvent dans S^*. C'est ce que nous observons sur la Fig. 4, où nous traçons les résultats obtenus avec $p = 10\%$ pour γ_r et $\tau = 2$ pour γ_v. Pour le calcul fait avec les ondelettes (Fig. 4a), S^* ne contenant que 6 éléments, il suffit de calculer les instabilités qu'en 6 échelles au lieu de 50. De plus, les trois seules échelles d'instabilité nulle (quelle que soit l'instabilité considérée) correspondent exactement aux 3 partitions théoriques cherchées. En revanche, certaines partitions qui ne correspondent pas exactement aux bonnes partitions (typiquement un mélange de deux partitions théoriques) ont des valeurs d'instabilité très proches de zéro : même si ces valeurs ne sont pas strictement nulles, nous pouvons les considérer comme des fausses détections. Pour le calcul fait avec les fonctions d'échelles (Fig. 4b), les quatre seules valeurs sont nulles. Les trois premières correspondent aux 3 bonnes partitions et la dernière à un découpage en deux communautés. En réalité, des tests non présentés ici montrent que les ondelettes retrouvent mieux les 3 échelles (même pour des jeux de paramètres du banc d'essai plus difficiles) que les fonctions d'échelle, mais donnent lieu à plus de fausses détections.

5 Conclusion

Le travail présenté est un complément important de la communication [10] : nous proposons une méthode pour estimer la pertinence d'une échelle donnée basée sur deux notions d'instabilité d'une partition. Une première instabilité observée en perturbant le graphe (γ_r), et une autre instabilité observée en perturbant le paramètre d'échelle (γ_v). Aucune des deux n'est mieux que l'autre et considérer les deux permet de passer outre les inconvénients de chacune. Aussi, nous proposons une manière d'estimer un ensemble restreint d'échelles potentiellement stables en définissant une troisième notion d'instabilité (β), plutôt basée sur l'allure du dendrogramme. Cette approche permet d'avancer vers l'étude de la partition de graphes en communautés comme étant un problème statistique de détection.

Références

[1] A. Barrat, M. Barthlemy and A. Vespignani, Dynamical processes on complex networks. *Cambridge University Press*, 2008.

[2] S. Fortunato, Community detection in graphs. *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[3] D.K. Hammond, P. Vandergheynst, et R. Gribonval, Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[4] T. Hastie, R. Tibshirani et J. Friedman, Elements of statistical learning : data mining, inference, and prediction. *Springer*, 2001.

[5] L. Hubert et P. Arabie, Comparing partitions. *Journal of classification*, vol. 2, no. 1, pp. 193-218, 1985.

[6] R. Lambiotte, Multi-scale modularity in complex networks. In *Proc. IEEE WiOpt*, pp. 546–553, 2010.

[7] M. Mitrovic et B. Tadic, Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Phys. Rev. E*, vol. 80, no. 2, 026123, 2009.

[8] Sales-Pardo, M. and Guimera, R. and Moreira, A.A. and Amaral, L.A.N., Extracting the hierarchical organization of complex systems. *PNAS*, vol. 104, no. 39, pp. 15224–15229, 2007.

[9] M.T. Schaub, J.C. Delvenne, S.N. Yaliraki et M. Barahona, Markov dynamics as a zooming lens for multiscale community detection : non clique-like communities and the field-of-view limit. *PloS one*, vol. 7, no. 2, e32210, 2012.

[10] N. Tremblay et P. Borgnat, Multiscale community mining in networks using spectral graph wavelets. *EUSIPCO*, Marrakech (Maroc), Sept. 2013.