

## Détection de communautés statiques à partir de réseaux dynamiques.

M. Drevetton, K. AVRACHENKOV, L. LESKELÄ  
*Inria, Inria, Aalto University*

**Email** : maximilien.drevetton@inria.fr

**Mots Clés** : Réseaux temporels, Détection de communauté, Dynamic stochastic block model.

**Biographie** – Maximilien Drevetton est en troisième année de thèse à l’Inria Sophia-Antipolis, sous la direction de Konstantin Avrachenkov. Il travaille sur la détection de communauté dans des graphes. Il étudie en particulier des méthodes semi-supervisées ainsi que des méthodes pour les graphes dynamiques ou géométriques.

### Resumé :

De nombreuses disciplines étudient des entités interagissant entre eux par paires: interactions entre individus en sociologie, entre espèces animales en biologie, entre particules en physique, etc. Les individus sont souvent groupés en communautés, et deux individus appartenant à une même communauté ont plus de chances d’interagir entre eux. La *détection de communauté* réfère à la prédiction non supervisée des communautés de chaque individu à partir de la seule observation des interactions entre les paires d’individus [4].

Ces 20 dernières années, la détection de communauté a généré de nombreux intérêts, et de nombreuses méthodes et algorithmes ont été développés [3]. Plus récemment, l’attention des chercheurs s’est portée sur les réseaux dynamiques [5] (où les interactions varient au cours du temps) et sur les réseaux à plusieurs couches [6] (où différents types d’interactions peuvent co-exister).

Dans un réseau statique, les interactions sont représentées dans une matrice d’adjacence  $A$ , qui dans le cas le plus simple est une matrice symétrique et binaire, dont l’entrée  $(ij)$  est égale à 1 lorsque les individus  $i$  et  $j$  interagissent entre eux. De ce fait, les données d’un réseau dynamique peuvent consister en l’observation d’une liste de  $T$  matrices d’adjacence  $(A^{(1)}, \dots, A^{(T)})$ . Chaque matrice  $A^{(t)}$  représente une photographie du réseau à l’instant  $t$ . De manière équivalente, l’interaction  $A_{ij}$  entre deux individus  $i$  et  $j$  est un nombre binaire à  $T$  chiffres (*i.e.*, un élément de  $\{0, 1\}^T$ ).

De manière générale, nous étudions un modèle de graphe aléatoire dans lequel les  $N$  noeuds sont répartis dans  $K$  communautés, et les interactions entre paires de noeuds sont à valeurs dans un espace mesuré  $\mathcal{S}$ . Ce modèle est paramétré par le vecteur  $\sigma \in [K]^N$  représentant les communautés de chaque noeud et une collection  $f = (f_{k\ell})_{k,\ell \in [K]}$  de distributions de probabilité sur  $\mathcal{S}$ . On suppose que les interactions entre les paires de noeuds sont indépendantes, et la probabilité que deux noeuds  $i$  et  $j$  aient une interaction de type  $x \in \mathcal{S}$  vaut  $f_{\sigma_i \sigma_j}(x)$ . Cela définit la distribution de probabilité

$$P(A | \sigma) = \prod_{1 \leq i < j \leq N} f_{\sigma_i \sigma_j}(A_{ij})$$

d’un tenseur d’interaction  $A \in \mathcal{S}^{N \times N}$ . Nous nous restreindrons au cas homogène, où  $f_{k\ell}$  vaut  $f_{\text{in}}$  si  $k = \ell$  et vaut  $f_{\text{out}}$  sinon. Le problème d’inférence consiste donc à estimer  $\sigma$  étant donné l’observation de  $A$  et la connaissance du nombre de communautés  $K$ . Notre travail a deux contributions principales.

Tout d’abord, on commence par énoncer une borne inférieure au nombre d’erreurs faites dans l’estimation de  $\sigma$  par *n’importe quel* algorithme. Dans le cas de communautés de même taille, cette borne se comporte comme  $\exp(-\frac{N}{K}D)$ , où  $D = D_{1/2}(f_{\text{in}}, f_{\text{out}})$  est la divergence de Rényi

entre  $f_{\text{in}}$  et  $f_{\text{out}}$ . Cela élargit les récents résultats de [8] à un régime non asymptotique qui ne fait aucune hypothèse de régularité sur  $f_{\text{in}}$  et  $f_{\text{out}}$ , ni sur l'espace  $\mathcal{S}$ . Ensuite, nous montrons que cette borne peut être atteinte par un algorithme *ad-hoc*. Ainsi, dans le cas de communauté de même taille, la *détection presque exacte* (l'espérance de la proportion de noeuds faussement classifiés tend vers 0 quand le nombre de noeuds  $N$  tend vers l'infini) est possible si  $\lim ND = \infty$ , et impossible sinon. Cela donne une extension naturelle aux résultats bien connus du cas  $\mathcal{S} = \{0, 1\}$  [9, 1]. On applique ensuite ces résultats à des réseaux dynamiques où les noyaux d'interaction  $f_{\text{in}}$  et  $f_{\text{out}}$  sont markoviens, généralisant ainsi les résultats de [7] pour les réseaux à multicouches, où les couches sont indépendantes les unes des autres.

Dans une seconde partie, nous présentons divers algorithmes. En particulier, nous proposons un algorithme basé sur l'estimation de vraisemblance, qui met à jour la prédiction au fur et à mesure que les observations arrivent. Cet algorithme a une complexité de  $O(KN^2T)$ , qui peut être réduite à  $O(KmT)$  pour les réseaux diffus, où  $m$  est le nombre moyen d'interaction à un instant donné. Le code pour reproduire les simulations est disponible à <https://github.com/maximiliendrevetton/MarkovSBM>.

Pour de plus amples détails et preuves, nous référons à notre preprint [2].

## Références

- [1] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18:1–86, 2018.
- [2] Konstantin Avrachenkov, Maximilien Drevetton, and Lasse Leskelä. Estimation of static community memberships from temporal network data. *arXiv preprint arXiv:2008.04790*, 2020.
- [3] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010.
- [4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [5] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.
- [6] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.
- [7] Subhadeep Paul and Yuguo Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Statist.*, 10(2):3807–3870, 2016.
- [8] Min Xu, Varun Jog, and Po-Ling Loh. Optimal rates for community estimation in the weighted stochastic block model. *Annals of Statistics*, 48(1):183–204, 2020.
- [9] Anderson Y. Zhang and Harrison Huibin Zhou. Minimax rates of community detection in stochastic block models. *Annals of Statistics*, 44(5):2252–2280, 2016.