# Classification spectrale par la laplacienne déformée dans des graphes réalistes

Lorenzo DALL'AMICO[1], Romain COUILLET[1,2], Nicolas TREMBLAY[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA lab, Grenoble, France

[2]L2S, CentraleSupelec, University of Paris Saclay, France

{lorenzo.dall-amico,romain.couillet,nicolas.tremblay}@gipsa-lab.fr

**Résumé –** Dans cet article, nous introduisons une analyse originale du comportement spectral de la matrice de *Laplace déformée* $D - rA$, au cœur des méthodes de détection de communautés spectrales dans des réseaux parcimonieux. Nous étudions tout d'abord le modèle statistique de réseaux dits "stochastique par blocs de degrés corrigés" (DC-SBM), et affirmons que ce modèle est adapté à la représentation de réseaux réels. Nous exploitons alors les propriétés spectrales asymptotiques du modèle DC-SBM afin d'ajuster finement le paramètre $r$ à une valeur rendant l'algorithme de classification spectrale basé sur $D - rA$ insensible à la distribution des degrés du réseau. Par le biais de simulations approfondies et de la comparaison avec des réseaux réels, nous validons notre algorithme et observons qu'il atteint des performances systématiquement meilleures que les méthodes concurrentes les plus avancées.

**Abstract –** In this article we introduce an original analysis of the spectral behavior of the *deformed Laplacian* matrix $D - rA$ used in spectral community detection methods for sparse networks. We first study the degree corrected stochastic block model (DC-SBM) and claim that it is appropriate from a statistical point of view as a generative model for real networks. We then exploit the asymptotic spectral properties of the DC-SBM model in order to finely tune the parameter $r$ to a value that makes the clustering algorithm based on $D - rA$ insensitive to the degree distribution of the network. Via extensive simulations and comparisons to real networks, we validate our algorithm and observe that it systematically outperforms the state-of-the-art competing methods.

## 1 Introduction

La théorie des graphes a fourni de nombreuses bases méthodologiques permettant de traiter de multiples problèmes d'inférence sur les réseaux. L'une des tâches les plus élémentaires à résoudre, notamment sur les grands graphes modernes, consiste à révéler les classes d'affinité au sein du réseau. Ce problème, dit de détection de communauté, exploite le fait que les nœuds d'une même communauté ont tendance à créer des interactions plus fortes ou plus fréquentes [1]. Il s'agit alors d'inférer ces classes d'affinité à partir d'une réalisation (potentiellement aléatoire) du graphe, à savoir à partir de sa matrice d'adjacence.

La solution la plus efficace à ce problème repose sur l'algorithme de *propagation des convictions* (*belief propagation* ou *méthode des cavités* pour les physiciens). Cependant, cet algorithme a de nombreuses limitations : il est mathématiquement difficile d'accès, son temps de convergence peut être long et, par ailleurs, aucune garantie de convergence n'est établie. En raison de ces limitations, d'autres méthodes ont été proposées pour résoudre le problème de détection de communautés. Nous nous intéresserons ici aux méthodes de classification spectrale qui consistent à inférer les communautés à partir des vecteurs propres de matrices représentatives du graphe [2]. Parmi ces dernières, on compte notamment la *matrice d'adjacence* $A$, [1] la matrice de *Laplace* $D - A$, [2] ainsi que les matrices de *Laplace normalisées* $D^{-1}A$, $D^{-1/2}AD^{-1/2}$.

Les vecteurs propres de ces matrices se sont révélés utiles pour déterminer efficacement les communautés dans les graphes dits *denses*, c'est-à-dire dont le degré moyen évolue avec la dimension globale du réseau, et de nombreux résultats théoriques permettent de finement comprendre les performances. Cependant, à ce jour, il existe encore peu de résultats théoriques permettant d'étudier des réseaux parcimonieux dont le degré moyen est indépendant de la taille du graphe. Les graphes réels sont pourtant souvent parcimonieux.

Dans cet article, nous proposons une étude originale d'une matrice moins connue : la matrice de *Laplace deformée* $D - rA$ avec $r \in \mathbb{R}$ [3, 4]. Nous démontrons qu'une classification spectrale basée sur $D - rA$, pour un choix pertinent de $r$, est particulièrement appropriée au cadre de graphes "réalistes" parcimonieux et de degrés hétérogènes. Nous nous limiterons ici au cas de deux communautés de taille égale, mais élaborerons sur les extensions possibles de ce modèle à des cas plus généraux.

L'article est organisé comme suit : dans la Section 2, nous introduisons formellement le problème de *détection de communautés* dans les graphes réels, en expliquant que le *modèle stochastique en blocs de degrés corrigés* (DC-SBM) constitue un modèle génératif de graphes parcimonieux suffisamment géné-

1. $A_{ij} = 1$ si les nœuds $i$ et $j$ sont connectés ; $A_{ij} = 0$ sinon.

2. $D = \mathrm{diag}(d_1, \ldots, d_n)$ avec $d_i = \sum_j A_{ij}$.

ral. À partir de ce modèle, nous expliquerons en quoi la matrice de *Laplace deformée* permet d'identifier les communautés d'un graphe DC-SBM, et donc d'un graphe réel. La Section 3 propose une comparaison des performances de notre algorithme à des simulations sur graphes artificiels et réels. Nous concluons l'article par la Section 4.

# 2 Modélisation de réseaux réels

Les réseaux réels sont caractérisés par deux propriétés importantes qui rendent le problème de détection de communautés particulièrement délicat en pratique : la *parcimonie* et l'*hétérogénéité* des degrés du graphe. Il a été notamment observé que la distribution des degrés de nombreux réseaux suit une *loi de puissance* [5]. Grâce à quelques approximations raisonnables, nous montrons ci-dessous que, dans le régime considéré, le modèle DC-SBM est un modèle génératif assez général. La propriété fondamentale de ce modèle est de considérer la prédisposition de chaque nœud à se connecter comme indépendante des classes d'affinité.

## 2.1 Le modèle DC-SBM

Pour générer le graphe, nous supposons un modèle dans lequel les connexions entre les différents nœuds $(i, j)$ sont créées en fonction de variables aléatoires indépendantes. En écrivant $n$ la taille du graphe, cette probabilité peut s'exprimer par :

$$\mathbb{P}(A_{ij} = 1) = \mathcal{F}_{ij}(Q, \sigma)$$

où $Q \in \mathbb{R}_+^n$ est le vecteur des "probabilités intrinsèques" de connection propre à chaque nœud, $\sigma \in \mathbb{Z}^n$ le vecteur des étiquettes des communautés de chaque nœud (on suppose ici que chaque nœud appartient à une unique communauté) et $\mathcal{F}_{ij}$ est une fonction décrivant la probabilité de connexion pour chaque paire de nœuds . Dans la classe des modèles génératifs considérés, cette fonction n'implique aucune autre hypothèse particulière. Elle permet notamment un comportement différent pour chaque paire de nœuds et permet aussi de maintenir une métrique à l'intérieur du graphe (par exemple prenant en compte la distance physique entre les nœuds). Un autre cas en phase avec cette expression est celui de l'existence d'autres étiquettes $(\sigma', \sigma'', \dots)$ transversales aux deux classes, qui rendent l'affinité des individus plus ou moins grande quelle que soit l'étiquette principale (par exemple, si nous divisons la population entre démocrates et libéraux, un sous-groupe transversal est celui des fumeurs et non-fumeurs). A ce stade, nous voulons simplifier ce modèle génératif, pour nous ramener au DC-SBM.

(a) La probabilité de connexion des nœuds $i, j$ ne dépend pas de la configuration complète des étiquettes, mais uniquement de $i, j$, de sorte que

$$\mathbb{P}(A_{ij} = 1) = \mathcal{F}_{ij}(Q_i, Q_j, \sigma_i, \sigma_j).$$

(b) Bien que nous ayons décrit un modèle génératif particulièrement complexe, notre objectif principal est de pouvoir déduire les propriétés statistiques de notre graphe, donc son comportement "émergent". Il est bien connu en physique que, dans la limite de $n \to \infty$ (*limite thermodynamique*) les caractéristiques émergentes d'un système ne sont pas la somme des contributions individuelles et ne dépendront pas des détails, comme expliqué dans [6]. Par conséquent, nous prévoyons qu'en utilisant la même fonction de génération pour chaque paire de nœuds, le comportement collectif ne sera pas largement modifié. Par cette hypothèse, nous demandons

$$\mathcal{F}_{ij}(\cdot) = \mathcal{F}(\cdot).$$

(c) Nous imposons les conditions de régularité suivantes :
- $\mathcal{F}(\cdot) = 0$ si $Q_i Q_j = 0$ ;
- $\partial_{Q_i} \mathcal{F}(Q_i, Q_j, \sigma_i, \sigma_j) > 0$ ;
- $\mathcal{F}(\cdot)$ est invariant par permutations des indices $i, j$.

De cette hypothèse nous déduisons que

$$\mathbb{P}(A_{ij} = 1) = \mathcal{F}(Q_i Q_j, \sigma_i, \sigma_j).$$

(d) En imposant un degré moyen fixe par rapport à la taille $n$ du graphe (condition de parcimonie), nous demandons que $Q_i = \frac{q_i}{\sqrt{n}}$ pour $q_i = O(1)$. Le développement de $\mathcal{F}(Q_i Q_j, \sigma_i, \sigma_j)$ en serie de Taylor donne alors

$$\mathbb{P}(A_{ij} = 1) = \frac{1}{n} \left. \frac{\partial \mathcal{F}(q_i q_j, \sigma_i, \sigma_j)}{\partial (q_i q_j)} \right|_{q_i q_j = 0} q_i q_j + o(n^{-1})$$

qui, au premier ordre, correspond au modèle DC-SBM

$$\mathbb{P}(A_{ij} = 1) = q_i q_j C(\sigma_i, \sigma_j). \tag{1}$$

Ce calcul est pertinent pour anticiper les sources d'erreurs induites par une comparaison de traitements effectués sur des réseaux synthétiques de type DC-SBM versus réalistes. Tout d'abord, les approximations (a) et (c) sont relativement raisonnables dans un grand nombre d'applications. L'approximation (b) a quant à elle été largement couverte. L'approximation (d) est plus cruciale : lorsque le réseau est parcimonieux, $\mathbb{P}(A_{ij} = 1)$ ne dépend pas de l'expression complète de $\mathcal{F}$ mais de sa dérivée en zéro. Cette observation implique qu'indépendamment de la complexité du modèle génératif $\mathcal{F}$, pour un réseau *parcimonieux*, ce modèle s'approxime bien par un modèle DC-SBM au premier ordre en $\frac{1}{n}$, avec pour conséquence majeure le découplage des dépendances en $q$ et $\sigma$. Ce phénomène n'apparait pas dans le cas de graphes denses où le modèle DC-SBM est alors moins justifié.

## 2.2 Une propriété clé du modèle DC-SBM

Dans cette section, nous exploitons une propriété fondamentale du modèle DC-SBM, propre aux graphes parcimonieux. Cette propriété est issue d'une simple estimation par inférence bayésienne : pour des communautés de même taille (à savoir, pour $\mathbb{P}(\sigma)$ constant), désignent avec $\mathbb{P}_q$ la probabilité à $q$ fixe on peut écrire :

$$\begin{aligned}\mathbb{P}_q(\sigma|A) &= \frac{\mathbb{P}_q(A|\sigma)}{Z} \\ &= \frac{1}{Z} \prod_{i,j} \left[ \frac{q_i q_j C(\sigma_i, \sigma_j)}{n} \right]^{A_{ij}} \left[ 1 - \frac{q_i q_j C(\sigma_i, \sigma_j)}{n} \right]^{1 - A_{ij}}\end{aligned}$$

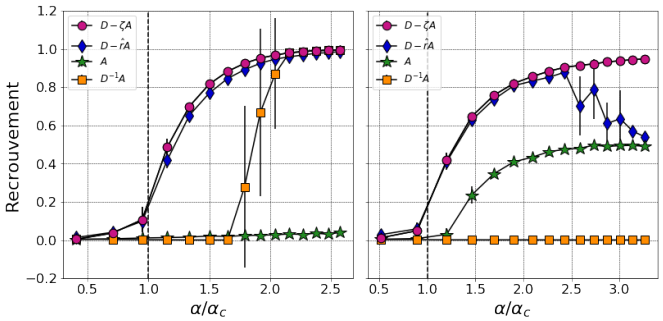


FIGURE 1 – Comparaison des méthodes de clustering spectral pour $q_i$ suivant la loi de puissance (à gauche) et la distribution " step " (à droite). Le taux de recrouvement varie entre 0 (assignation aléatoire) et 1 (reconstruction parfaite des classes) et est tracé en fonction du paramètre de contrôle $\alpha$. Quand $\alpha < \alpha_c$ il n'y a pas (asymptotiquement) de solution au problème de détection de communautés. Ici, $n = 5000$, $c_{\text{out}} = 1$, $c_{\text{in}} = 2 \to 16$. Moyenne sur 10 échantillons.

pour un certain paramètre de normalisation $Z$. Au premier ordre en $1/n$ [7], nous obtenons

$$\begin{aligned}\mathbb{P}_q(\sigma|A) &\approx \frac{\prod_{(ij)\in\mathcal{E}} q_i q_j C(\sigma_i,\sigma_j)}{\sum_{\{\sigma\}}\prod_{(ij)\in\mathcal{E}} q_i q_j C(\sigma_i,\sigma_j)} \\ &= \frac{\prod_{(ij)\in\mathcal{E}} C(\sigma_i,\sigma_j)}{\sum_{\{\sigma\}}\prod_{(ij)\in\mathcal{E}} C(\sigma_i,\sigma_j)}\end{aligned}$$

qui est ainsi indépendant de la distribution des paramètres $\{q_i\}$. Pour un modèle complètement symétrique à deux classes, en dénotant $C(\sigma_i,\sigma_j) = c_{\text{in}}$ lorsque $i,j$ sont d'une même communauté et $c_{\text{out}}$ sinon, nous avons alors immédiatement :

$$\mathbb{P}(\sigma_i\sigma_j = 1|A_{ij}=1) = \frac{c_{\text{in}}}{c_{\text{in}}+c_{\text{out}}} \tag{2a}$$

$$\mathbb{P}(\sigma_i\sigma_j = -1|A_{ij}=1) = \frac{c_{\text{out}}}{c_{\text{in}}+c_{\text{out}}} \tag{2b}$$

une expression qui, comme souhaité, ne fait pas apparaitre de dépendance entre nœuds distants dans le graphe.

## 2.3 Classification spectrale pour le DC-SBM

Sur la base de ces résultats, nous étudions ici le spectre de la matrice de Laplace déformée $D - rA$ et démontrons qu'un choix précis du paramètre $r$ permet d'inférer les communautés du réseau sans être impacté par les degrés du graphe.

Considérons pour ceci l'action de $D - rA$ *en espérance* (sur l'allocation des classes) sur le vecteur des étiquettes $\sigma \in \{-1,1\}^n$, en indiquant par $|\partial_i^{(S/O)}|$ le nombre de voisins du noeud $i$ qui sont de même classe (S) ou de classe opposée (O) :

$$[[(D-rA)\sigma]_i|A] = d_i\sigma_i - r d_i\sigma_i\left[\frac{|\partial_i^{(S)}| - |\partial_i^{(O)}|}{d_i}\right].$$

En utilisant (2) et la propriété $|\partial_i^{(S)}| + |\partial_i^{(O)}| = d_i$, nous obtenons les estimés suivants :

$$|\partial_i^{(S/O)}| = \mathbb{E}[|\partial_i^{(S/O)}|] \pm \Delta \tag{3a}$$

$$\mathbb{E}[|\partial_i^{(S/O)}|] = d_i\frac{c_{\text{in/out}}}{c_{\text{in}}+c_{\text{out}}} \tag{3b}$$

$$\sqrt{\mathbb{E}(\Delta^2)} = \frac{\sqrt{c_{\text{in}}c_{\text{out}}}}{c_{\text{in}}+c_{\text{out}}}\sqrt{d_i} \tag{3c}$$

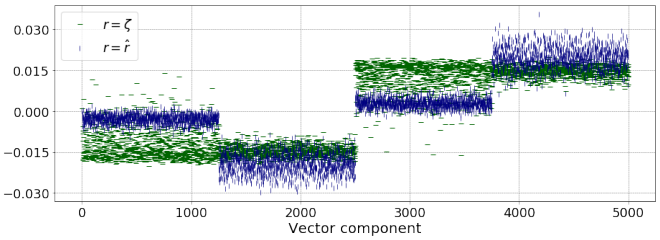


FIGURE 2 – Vecteurs propres informatifs de $D - rA$ pour $r = \zeta$ (vert) et pour $r = \hat{r}$. Ici nous avons $n = 5000$, $c_{\text{in}} = 12$, $c_{\text{out}} = 1$, et $q_i \in \{q_{\min}, q_{\max}\}$ avec $q_{\max} = 8q_{\min}$.

qui conduit alors à

$$\mathbb{E}\left[[(D-rA)\sigma]_i|A\right] = d_i\sigma_i\left[1 - r\frac{c_{\text{in}}-c_{\text{out}}}{c_{\text{in}}+c_{\text{out}}}\right]. \tag{4}$$

Ainsi, en choisissant $r = \zeta \equiv \frac{c_{\text{in}}+c_{\text{out}}}{c_{\text{in}}-c_{\text{out}}}$, nous obtenons une équation aux valeurs propres approximativement correcte. L'erreur $\Delta$ disparaît dans (4) mais demeure dans la fluctuation de $[(D-rA)\sigma]_i$ (à moins que $c_{\text{in}} \gg c_{\text{out}}$). Dans [8], nous fournissons une analyse plus détaillée de cette erreur lorsque $c_{\text{in}}$ et $c_{\text{out}}$ sont comparables.

De manière fondamentale, cette brève analyse montre que le vecteur propre associé à la valeur propre nulle de $D - \zeta A$ est indépendant de la distribution des degrés du réseau. Mais cette matrice a d'autres propriétés importantes. Dans le régime de graphes parcimonieux, les matrices de Laplace réduites exhibent un spectre de valeurs propres étalé qui "engloutit" les valeurs propres associées aux vecteurs informatifs (liés à $\sigma$). Pour la laplacienne deformée $D - \zeta A$, une première valeur propre (minimale) est négative, la seconde est (presque) nulle, et le reste du spectre est positif et isolé de zéro, garantissant ainsi un bon alignement du vecteur propre associé à zéro au vecteur $\sigma$.

Dans [8], nous proposons une étude fine du comportement du vecteur propre informatif de $D - \zeta A$ et avons établi une formule pour la probabilité asymptotique de classification correcte ainsi qu'une méthode pour estimer la valeur de $\zeta$. L'algorithme consiste ensuite à calculer le vecteur propre de $D - \zeta A$ correspondant à la deuxième plus petite valeur propre, puis à appliquer l'algorithme *k-means* aux composants de ce vecteur propre pour obtenir la partition des nœuds.

# 3 Simulations

Dans cette section, nous présentons une comparaison des performances de l'algorithme que nous proposons comparativement aux matrices d'adjacences et laplaciennes couramment utilisées dans la littérature. Dans la Figure 1 nous comparons les performances de ces différents algorithmes sur des graphes synthétiques générés avec le modèle DC-SBM. En écrivant $\hat{\sigma}$ l'estimation faite des différentes classes, les performances sont estimées au moyen du taux de *recouvrement* défini par

$$2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\sigma_i,\hat{\sigma}_i} - \frac{1}{2}\right). \tag{5}$$

| Graph | $A$ | $D-A$ | $D^{-1}A$ | $H_{\hat{r}}$ | $D-\zeta A$ |
|---|---|---|---|---|---|
| Karate | $\mathbf{1.0}$ | $0.65$ | $0.94$ | $\mathbf{1.0}$ | $\mathbf{1.0}$ |
| Dolphins | $0.65$ | $\mathbf{0.97}$ | $\mathbf{0.97}$ | $0.87$ | $\mathbf{0.97}$ |
| Adjnoun | $0.54$ | $0.03$ | $\mathbf{0.73}$ | $0.68$ | $0.68$ |
| Polblogs | $0.25$ | $0.03$ | $0.03$ | $0.32$ | $\mathbf{0.90}$ |

TABLE 1 – Performance du taux de recouvrement sur les graphe de référence. Le réseau 'Adjnoun' est désassortatif, *i.e.*, $c_{\text{in}} < c_{\text{out}}$ ; ainsi, le vecteur propre informatif est ici associé à la plus petite valeur propre et $\zeta$ est négatif.

Nous évaluons le recouvrement pour différents $\alpha = \frac{c_{\text{in}}-c_{\text{out}}}{\sqrt{c}}$. Ce paramètre s'avère être la bonne valeur de contrôle, comme déjà observé dans [7, 9, 10, 11]. Nous utilisons ici deux distributions de probabilités différentes pour le vecteur $\{q_i\}$ : dans un cas une loi de puissance, dans l'autre une fonction binaire telle que $q_i \in \{q_{\min}, q_{\max}\}$ avec $q_{\max} = 8q_{\min}$. Dans le régime parcimonieux étudié ici, la laplacienne déformée (en rose) est plus efficace que les matrices d'adjacence (en vert) et laplacienne réduite (en orange), ainsi que pour le choix initialement proposé dans [11] pour la laplacienne déformée mais avec $r = \hat{r} \equiv \sqrt{\sum_i d_i^2 / \sum_i d_i}$ (en bleu).

Par ailleurs, dans la Figure 2, nous démontrons qu'en dépit des performances correctes observées pour $r = \hat{r}$, le vecteur propre de la laplacienne deformée est fortement affecté par les degrés du graphe. Ce n'est pas le cas pour notre choix $r = \zeta$, aiguillé par l'analyse précédente.

Enfin, dans le Tableau 1 nous comparons les performances des matrices précédentes pour la détection de communautés dans des réseaux réels. De nouveau les performances obtenues sont systématiquement égales ou supérieures à celles atteintes par les approches classiques. Grâce à la relation entre la matrice $D - rA$ et la matrice dite de *non-backtracking* [9, 10, 12], nous pouvons estimer la valeur de $\zeta$ directement à partir du graphe. Notre algorithme demeure ainsi purement non supervisé.

# 4 Conclusions et Perspectives

Dans ce travail, nous avons présenté une approche spectrale efficace pour la détection de communautés dans des réseaux hétérogènes parcimonieux basée sur l'exploitation des vecteurs propres de la matrice $D-\zeta A$ pour $\zeta = (c_{\text{in}}+c_{\text{out}})/(c_{\text{in}}-c_{\text{out}})$.

Bien que l'étude menée ici s'appuie sur des considérations bayésiennes élémentaires, la laplacienne normalisée a des liens étroits avec la physique statistique, et en particulier avec les matrices dites *hessienne de Bethe* et de *non-backtraking* [10, 11] qui découlent d'un parallèle établi entre les graphes parcimonieux et un système de particules en interaction. Ce parallèle peut être poussé bien plus loin afin de généraliser nos résultats à des modèles de détection de communautés multiple (et non réduite à deux classes) et non-symétrique (c'est-à-dire avec des classes de tailles distinctes).

Notre ligne d'argumentation et le parallèle avec la physique statistique demeurent jusqu'à présent essentiellement heuristiques ou basés sur des observations issues de simulations. Des outils mathématiques rigoureux, qui existent dans le cas de graphes denses, seraient indispensables à un meilleur discernement du cas parcimonieux. Progresser dans la mise en place de ces outils constitue clairement une prochaine étape naturelle.

# Remerciements

Ces travaux sont soutenus par le projet ANR RMT4GRAPH (ANR-14-CE28-0006) et par la Chaire IDEX GSTATS de l'Université Grenoble Alpes.

# Références

[1] Fortunato, *Community detection in graphs*, Physics reports, 486(3-5), 75-174, (2010)

[2] Von Luxburg, *A tutorial on spectral clustering*, Statistics and computing, 17(4), 395-416, (2007)

[3] Grindrod, Higham, Noferini, *The deformed graph laplacian and its applications to network centrality analysis* – SIAM Journal on Matrix Analysis and Applications – 39(1) :310–341, (2018)

[4] Morbidi, *The deformed consensus protocol* – Automatica – 49(10) :3049–3055, October (2013)

[5] Barabási, *Emergence of scaling in random networks* – Science – 286.5439, (1999)

[6] Anderson, *More is different* – Science – 177.4047 (1972)

[7] Decelle, Krzakala, Moore, Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications* – Physical Review E – 84(6), 066106, (2011)

[8] Dall'Amico, Couillet, Tremblay, *Optimized Deformed Laplacian for Spectrum-based Community Detection in Sparse Heterogeneous Graphs* – arXiv preprint 1901.09715, (2019)

[9] Massoulié, *Community detection thresholds and the weak Ramanujan property* – Proceedings of the forty-sixth annual ACM symposium on Theory of computing, (2014)

[10] Krzakala, Moore, Mossel, Neeman, Sly, Zdeborová, Zhang, *Spectral redemption in clustering sparse networks* – Proceedings of the National Academy of Sciences – 110(52), 20935-20940, (2013)

[11] Saade, Krzakala, Zdeborová, *Spectral clustering of graphs with the bethe hessian* – Advances in Neural Information Processing Systems, (2014)

[12] Gulikers, Lelarge, Massoulié,*Non-backtracking spectrum of degree-corrected stochastic block models* – Innovations in Theoretical Computer Science Conference (ITCS 2017).