# Un test d'ajustement basé sur les degrés pour des modèles de graphes aléatoires hétérogènes

Sarah Ouadah [1], Stéphane Robin [2] & Pierre Latouche [3]

[1] *AgroParisTech, UMR 518, MIA, Paris, France*
[2] *INRA, UMR 518, MIA, Paris, France*
[3] *Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, France*

**Résumé.** La topologie d'un réseau peut être étudiée à travers les degrés observés, c'est à dire le nombre de connexions par nœud. Nous nous intéressons ici aux degrés afin d'étudier l'ajustement à un modèle de graphe aléatoire hétérogène, dans lequel les arêtes ont leur probabilité propre d'existence. Un test d'ajustement est alors mis en place en montrant la normalité asymptotique du carré moyen des degrés sous ce modèle. Dans le même esprit, nous utilisons la variance des degrés pour le modèle d'Erdös-Renyi. Pour ces deux modèles, nous étudions la puissance des tests d'ajustement proposés, et montrons la normalité asymptotique pour différents régimes de parcimonie des graphes. Enfin, ces tests sont illustrés sur des réseaux réels en écologie et sciences sociales. Plusieurs séries de simulations sont également mises à en place afin d'évaluer les tests proposés.

**Mots-clés.** Graphes aléatoires, degrés, test d'ajustement.

**Abstract.** The degrees are a relevant way to study the topology of a network. They can be used to assess the goodness-of-fit of a specific heterogeneous random graph model in which the edges have different connection probabilities. We prove the asymptotic normality of the degree mean square under this model which enables us to derive a formal test. Similarly, the degree variance is used for the Erdös-Renyi model case. We study the power of the proposed goodness of-fit tests, and also prove the asymptotic normality under specic sparsity regimes. The tests are illustrated on real networks from social sciences and ecology, and their performances are assessed via a simulation study.

**Keywords.** Random graphs, degrees, goodness-of-fit test.

# 1 Introduction

Les réseaux sont utilisés dans de nombreux domaines comme la biologie, la sociologie, l'écologie ou encore l'économie pour décrire les interactions entre un ensemble d'individus ou entités. Formellement, un réseau d'interaction peut être vu comme un graphe dont les nœuds représentent les individus et une arête entre deux nœuds est présente si ces derniers interagissent.

Ces dix dernières années, la distribution des degrés (i.e. le nombre de connections de chaque nœud) s'est imposé comme un moyen simple et parlant pour étudier la topologie d'un réseau (voir e.g. Snijders, 1981 et Barabási et Albert, 1999). La distribution des degrés permet également d'inférer des modèles de graphes complexes (voir e.g. Bickel et al., 2011 et Channarond et al. 2012). Par ailleurs, la variance des degrés est un outil privilégié depuis les premières études statistiques des réseaux (voir e.g. Snijders, 1981). Dans nos travaux, nous utilisons la notion de carré moyen des degrés qui mesure l'écart entre les degrés observés et leurs espérances sous le modèle d'Erdös-Renyi hétérogène que nous définissons ci-après.

Considérant un graphe non dirigé $\mathcal{G} = (\{1, \dots n\}, \mathcal{E})$ sans boucle (connexion d'un nœud à lui même), nous notons $Y$ la matrice d'adjacence $n \times n$ correspondante. Les éléments $Y_{ij}$ de $Y$ valent donc 1 si $(i, j) \in \mathcal{E}$, et 0 sinon. Étant donné les propriétés de $\mathcal{G}$, $Y_{ij} = Y_{ji}, \forall i \neq j$ et $Y_{ii} = 0$, pour tous les $i$. Le degré du nœud $i$ est défini par $D_i = \sum_{j \neq i} Y_{ij}$. Dans ce travail, nous considérons deux modèles de graphes aléatoires, notés $ER(p)$ et $HER(\mathbf{p})$. $ER(p)$ désigne le modèle d'Erdös-Renyi, dans lequel toutes les arêtes $(Y_{ij})$ sont des variables de Bernoulli indépendantes de même paramètre $p$. $HER(\mathbf{p})$ désigne le modèle d'Erdös-Renyi hétérogène dans lequel les arêtes $Y_{ij}$ sont indépendantes avec des probabilités respectives $p_{ij}$. La matrice $n \times n$, notée $\mathbf{p}$, constituée des $p_{ij}$, est symétrique et de diagonale nulle.

Pour un graphe aléatoire donné avec $\mathbf{p}^0$ pour matrice de probabilités de connexion, nous considérons la statistique appelée carré moyen des degrés et définie de la manière suivante

$$W_{\mathbf{p}^0} = \frac{1}{n} \sum_i (D_i - \mu_i^0)^2,$$

où $\mu_i^0 = \sum_{j \neq i} p_{ij}^0$ est l'espérance du degré du nœud $i$ sous $HER(\mathbf{p}^0)$. Nous mettons en place un test d'ajustement pour le modèle $HER$ en montrant la normalité asymptotique de cette statistique et en fournissant sa puissance sous l'alternative d'un modèle $HER(\mathbf{p})$ avec $\mathbf{p} \neq \mathbf{p}^0$.

De plus, nous établissons des résultats analogues sous un modèle $ER$ contre un $HER$ en se basant cette fois-ci sur la variance des degrés. Par ailleurs, étant donné que les gros réseaux sont souvent parcimonieux, nous étudions pour quels régimes de parcimonie nos résultats de normalité asymptotique restent valides.

# 2 Normalité asymptotique

Nous établissons la normalité asymptotique de $W_{\mathbf{p}^0}$ sous le modèle $HER(\mathbf{p})$. La preuve repose sur la décomposition de Hoeffding (voir e.g. le chapitre 11 de van der Vaart (1998)) de $W_{\mathbf{p}^0}$. Nous calculons toutes les projections de cette décomposition auxquelles nous appliquons le théorème de Lindeberg-Lévy (voir e.g. Billingsley (1968) , Theorem 7.2, p.42). Ce type de stratégie a déjà été utilisé pour des études de graphes, par exemple pour

prouver la normalité asymptotique de la variance des degrés sous un modèle $ER(p)$ (voir Bloznelis (2005)), et dans Nowicki et Wierman (1988) pour montrer celle des comptages de sous-graphes dans les graphes aléatoires.

**Théorème 1** *Sous le modèle $HER(\mathbf{p})$, la statistique $W_{\mathbf{p}^0}$ est asymptotiquement normale:*

$$(W_{\mathbf{p}^0} - \mathbb{E}_{HER(\mathbf{p})} W_{\mathbf{p}^0}) / \mathbb{S}_{HER(\mathbf{p})} W_{\mathbf{p}^0} \xrightarrow{D} \mathcal{N}(0,1),$$

*où $\mathbb{S}$ désigne l'écart-type et*

$$\mathbb{E}_{HER(\mathbf{p})} W_{\mathbf{p}^0} = \frac{2}{n} \left( \sum_{1 \le i < j \le n} (\sigma_{ij}^2 + \delta_{ij}^2) + \sum_{1 \le i < j < k \le n} (\delta_{ij}\delta_{ik} + \delta_{ij}\delta_{jk} + \delta_{ik}\delta_{jk}) \right).$$

*où $\sigma_{ij}^2 = p_{ij}(1 - p_{ij})$ où $\delta_{ij} = p_{ij} - p_{ij}^0$. De plus,*

$$\begin{aligned} \mathbb{V}_{HER(\mathbf{p})} W_{\mathbf{p}^0} &= \frac{4}{n^2} \sum_{1 \le i < j \le n} \sigma_{ij}^2 (\Delta_i + \Delta_j + 1 - 2p_{ij})^2 \\ &\quad + \frac{4}{n^2} \sum_{1 \le i < j < k \le n} \left( \sigma_{ij}^2 \sigma_{ik}^2 + \sigma_{ij}^2 \sigma_{jk}^2 + \sigma_{ik}^2 \sigma_{jk}^2 \right), \end{aligned}$$

*avec $\Delta_i = \sum_{j \neq i} \delta_{ij}$.*

## 2.1 Cas des graphes parcimonieux

Nous nous intéressons à la validité du Théorème 1 dans le cas de graphes parcimonieux. La parcimonie peut être définie de deux manières. Soit chacune des probabilités de connexion tend vers 0 lorsque $n$ croît, soit la fraction de connexions non nulles décroit quand $n$ croît. La proposition suivante tient compte de ces deux cas de figure.

**Proposition 1** *Considérons le modèle $HER(\mathbf{p})$, lorsque $p_{ij} = p_{ij}^* n^{-a}$, $a > 0$, $p_{ij}^* \in [0,1]$ et lorsqu'une fraction $1 - n^{-b}$, $b \geq 0$, des $p_{ij}$ est mise à zéro. Les $p_{ij}^0$ suivent exactement les mêmes hypothèses. Alors, du moment que $a + b < 2$, la statistique $W_{\mathbf{p}^0}$ est asymptotiquement normale.*

# 3 Test et puissance

Nous étudions maintenant le test de $H_0 = HER(\mathbf{p}^0)$ contre $H_1 = HER(\mathbf{p})$. Les corollaires qui suivent donnent la distribution nulle de la statistique de test $W_{\mathbf{p}^0}$ ainsi que la puissance du test associé.

**Corollaire 1** *Sous le modèle $HER(\mathbf{p}^0)$ la statistique $W_{\mathbf{p}^0}$ est asymptotiquement normale de moments:*

$$\begin{aligned}\mathbb{E}_{HER(\mathbf{p}^0)}W_{\mathbf{p}^0} &= \frac{2}{n}\sum_{1\leq i<j\leq n}\sigma_{ij}^2,\\ \mathbb{V}_{HER(\mathbf{p}^0)}W_{\mathbf{p}^0} &= \frac{4}{n^2}\left(\sum_{1\leq i<j\leq n}\sigma_{ij}^2(1-2p_{ij})^2+\sum_{1\leq i<j<k\leq n}\left(\sigma_{ij}^2\sigma_{ik}^2+\sigma_{ij}^2\sigma_{jk}^2+\sigma_{ik}^2\sigma_{jk}^2\right)\right).\end{aligned}$$

Ce résultat est une conséquence directe du Théorème 1 dans le cas particulier du modèle $HER(\mathbf{p}^0)$ pour lequel tous les $\delta_{ij}$ sont nuls ($\delta_{ij}=p_{ij}-p_{ij}^0$).

À présent, en se basant sur le Corollaire 1, nous pouvons construire le test de niveau asymptotique $\alpha$, qui rejette $H_0$ dès que $W_{\mathbf{p}^0}$ excède $\mathbb{E}_{HER(\mathbf{p}^0)}W_{\mathbf{p}^0}+t_\alpha\mathbb{S}_{HER(\mathbf{p}^0)}W_{\mathbf{p}^0}$, où $t_\alpha$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite. La puissance du test est donnée dans le corollaire suivant.

**Corollaire 2** *La puissance asymptotique du test $H_0=HER(\mathbf{p}^0)$ contre $H_1=HER(\mathbf{p})$ est*

$$\pi(\mathbf{p})=1-\Phi\left(\left(\mathbb{E}_{HER(\mathbf{p}^0)}W_{\mathbf{p}^0}+t_\alpha\mathbb{S}_{HER(\mathbf{p}^0)}W_{\mathbf{p}^0}-\mathbb{E}_{HER(\mathbf{p})}W_{\mathbf{p}^0}\right)/\mathbb{S}_{HER(\mathbf{p})}W_{\mathbf{p}^0}\right),$$

*où $\Phi$ désigne la fonction de répartition de la loi normale centrée réduite et $t_\alpha=\Phi^{-1}(1-\alpha)$.*

# 4 Illustration

Nous considérons plusieurs réseaux pour illustrer le test proposé.

Réseaux écologiques : nous disposons de deux réseaux écologiques introduits par Vacher et al. (2008) et étudiés dans Mariadassou et al. (2010). Chacun de ces réseaux décrit une interaction entre une série de $n=51$ arbres et $n=154$ champignons respectivement. Dans le réseau d'arbres, deux arbres interagissent s'ils partagent au moins un parasite champignon. En ce qui concerne le réseau des champignons, deux champignons sont liés s'ils sont hôtes d'au moins une espèce commune d'arbre.

Réseau de blogs politiques : nous disposons également d'un réseau décrivant la blogosphère politique française (projet d'observatoire présidentiel). Ce dernier est constitué de $n=196$ blogs politiques français déjà étudiés dans Latouche et al. (2011). Deux blogs sont connectés si l'un contient un hyperlien vers l'autre.

Pour chaque réseau, plusieurs covariables sont disponibles. Les distances génétiques, taxonomiques et géographiques entres espèces d'abres sont données, ainsi que les similarités nutritionnelles et les distances taxonomiques entre champignons (voir la description dans

Mariadassou et al. (2010). Pour le réseau des blogs, nous avons accès au parti politique de chaque blog et le statut de l'auteur (journaliste ou non).
La question est de savoir si ces covariables sont suffisantes pour expliquer l'hétérogénéité du réseau, essentiellement en terme de degrés. Pour y répondre, considérant chaque réseau un à un, posons le modèle de régression logistique $\text{logit}(p^0_{ij}) = x_{ij}^\intercal \beta$, où $\text{logit}(u) = \log(u/(1-u)), u \in \mathbb{R}$, $x_{ij} \in \mathbb{R}^d$ est le vecteur des covariables pour les $(i,j)$ et $\beta$ pour le vecteur des coefficients de régression.
Ce modèle de régression nous fournit une estimation de $\mathbf{p}^0$ la matrice des probabilités de connexion. Nous appliquons ensuite le test basé sur le carré moyen des degrés pour vérifier si les covariables considérées sont suffiantes pour expliquer l'hétérogénéité du réseau considéré. Comme indiqué dans la Table 1, l'hypoyhèse nulle $H_0 : Y \sim HER(\mathbf{p}^0)$ est rejetée pour tous les exemples. En ce qui concerne les réseaux écologiques, nos résultats concordent avec ceux de Mariadassou et al. (2010), qui ont détecté une hétérogénéité résiduelle dans les versions valuées de ces réseaux. Pour le réseau de blogs, la détection d'une hétérogénéité résiduelle concorde avec des résultats obtenus par les auteurs de cet article, pour une méthodologie très différente, liant modèles de $W$-graph et covariables.

| Réseau | moyenne($\mathbf{p}^0$) | écart-type($\mathbf{p}^0$) | $W_{\mathbf{p}^0}$ | $\mathbb{E}_{HER(\mathbf{p}^0)}$ | $\mathbb{S}_{HER(\mathbf{p}^0)}$ | $\frac{W_{\mathbf{p}^0} - \mathbb{E}_{HER(\mathbf{p}^0)}}{\mathbb{S}_{HER(\mathbf{p}^0)}}$ |
|---|---|---|---|---|---|---|
| Arbres | 0.540 | 0.192 | 136.8 | 10.57 | 2.09 | 60.4 |
| Champignons | 0.227 | 0.006 | 593.8 | 26.83 | 3.06 | 185.0 |
| Blogs | 0.075 | 0.119 | 78.7 | 10.7 | 1.17 | 57.8 |

Table 1: Test basé sur le carré moyen des degrés pour les réseaux écologiques et de blogs politiques.

Remarque : La qualité de la puissance du test proposé, ainsi que celle de la normalité de la statistique de test dans le cas parcimonieux ont été étudiées via des simulations.

# Bibliographie


[1] Barabási, A. - L. et Albert, R. (1999), Emergence of scaling in random networks, *Science*, 286, 509–512.
[2] Bickel, P. J. and Chen, A. et Levina, E. (2011), The method of moments and degree distributions for network models, *Ann. Stat.*, 39 (5), 2280–2301.
[3] Billingsley, P. (1968), Convergence of Probability Measures, *Wiley: New-York*.
[4] Bloznelis, M. (2005), Degree variance is asymptotically normal, *Rapport technique*, Vilnius university, Faculty of Mathematics and Informatics.
[5] Channarond, A., Daudin, J.-J. et Robin, S. (2012), Classification and estimation in the Stochastic Block Model based on the empirical degrees, *Elec. J. Stat.*, 6, 2574–601.

[6] Latouche, P., Birmelé, E. et Ambroise, C. (2011), Overlapping stochastic block models with application to the French political blogosphere, *Ann. Appl. Stat.* , 5 (1), 309-336.
[7] Mariadassou, M., Robin, S. et Vacher, C. (2010), Uncovering structure in valued graphs: a variational approach, *Ann. Appl. Statist.*, 4 (2), 715–42.
[8] Nowicki, K. et Wierman, J. C (1988). Subgraph counts in random graphs using incomplete U-statistics methods, *Discrete Math.*,72 (1), 299–310.
[9] Snijders, T. A. B. (1981), The degree variance: An index of graph heterogeneity, *Social Networks*, 3 (3), 163–174.
[10] Vacher, C., Piou, D. et Desprez-Loustau, M.-L. (2008), Architecture of an Antagonistic Tree/Fungus Network: The Asymmetric Influence of Past Evolutionary History, *PLoS ONE*, 3 (3):1740.
[11] van der Vaart, A. W. (1998), Asymptotic statistics, *Cambridge Series in Statistical and Probabilistic Mathematics*, 3, Cambridge University Press, Cambridge.