Comparaison computationnelle d'un réseau d'interaction d'étudiants à un réseau écologique

Marika Caouette,
Zachary Cloutier, Emma Couture,
Joannie Gagnon, Marie Jacques
et Marianne Mallette

24 avril 2020

Résumé

Comparaison computationnelle d'un réseau d'interaction d'étudiants à un réseau écologique via la dissimilarité de Whittaker. Cette étude révèle une corrélation négative entre le nombre d'étudiants dans un cours et la dissimilarité moyenne par rapport aux autres cours.

Introduction

La biodiversité est organisée en fonction de réseaux complexes d'interactions écologiques entre les espèces à l'échelle locale (Hagen et al., 2012). Ces réseaux sont composés de noeuds (*nodes*), représentant la composition en espèces, qui sont reliés entre eux par des liens (*edges*), représentant les interactions écologiques (Poisot et al., 2012).

Dans le cadre du cours *BIO500 - Méthodes en écologie computationnelle*, la question suivante a été soulevée : Est-ce que le réseau de collaborations entre les étudiants du baccalauréat en écologie a les mêmes propriétés que les réseaux écologiques ? Après avoir consulté la littérature à ce sujet et afin de préciser la question, nous avons décidé de nous intéresser au concept de la dissimilarité des réseaux d'interactions entre les espèces et de la mettre en relation avec le nombre d'étudiants par cours.

En effet, la composition en espèces et les interactions écologiques peuvent se renouveler dans le temps et l'espace, de façon corrélée ou non, ce qui contribue à la présence de dissimilarité dans ce type de réseaux locaux (Poisot et al., 2012). De plus, l'occurrence des interactions interspécifiques n'est pas

indépendante de la composition spécifique (Poisot et al., 2012). Nous avons donc décidé d'appliquer ce concept au réseau de collaboration des étudiants participant au cours, et ce, via l'élaboration computationnelle d'un tableau matriciel et de différentes figures afin de visualiser la complexité du réseau, ses dissimilarités et l'impact du nombre d'étudiants par cours.

Description de la méthode

Chaque étudiant du cours *BIO500* devait compiler tous les travaux faits en équipe durant ses années à l'Université, pour mettre le tout en commun. La compilation se trouve dans 3 fichiers : les collaborations entre étudiants (Data_collabo), les cours (Data_cours) et le nom des étudiants (Data_etudiants).

Nous avons fait ensuite un nettoyage dans le logiciel R. La fonction *STR_WHICH* nous a permis d'uniformiser les noms de colonne dans chacune des bases et de corriger les erreurs au niveau des sigles et des noms des étudiants. Avec les fonctions *ORDER* et *EDIT*, nous avons analysé les erreurs possibles dans des données entrées en double dans chaque fichier. Pour Data_cours, nous avons créé une boucle *FOR* pour s'assurer que chaque cours était unique une fois les erreurs retirées.

Ensuite, ce fut la création de trois tables (tbl_noms, tbl_collaborations et tbl_etudiants) et de deux requêtes (sql_requete) dans la connexion SQLite. La 1ère requête met en relation le nombre de collaborations entre chaque étudiant. La 2ème requête est la création des populations pour chaque cours. La requête a été répliquée pour les 30 cours, donnant 30 populations.

Les fonctions *SUBSET* et *%IN%* ont été utilisées pour faire une matrice carrée sous la requête des collaborations entre étudiants. Une autre matrice a été créée avec une itération du calcul de dissimilarité pour mettre en relation chaque population. Chaque matrice a été visualisée avec la fonction *GRAPH.ADJACENCY* et *PLOT* du package *plot.matrix* (Klinke, 2019).

C'est une itération de la combinaison des fonctions *ANTI_JOIN* et *SEMI_JOIN* qui nous a permis de calculer le degré de dissimilarité entre les populations. La visualisation de ce degré s'est fait avec la fonction *PLOT*.

Finalement, une autre itération des combinaison des fonctions *LM* et *PLOT* nous a permis de montrer une régression linéaire des valeurs de dissimilarité moyenne en fonction du nombre d'étudiants.

Avant la réalisation des fonctions, les gestionnaires de paquets suivants ont été installés : *plot.matrix* (Klinke, 2019), *knitr* (Xie, 2020), *reshape2* (Wickham, 2007), *stringr* (Wickham, 2019), *dplyr* (Wickham et al., 2020), *RSQLite* (Müller et al., 2020) et *IGraph* (Csardi and Nepusz, 2006).

Description des résultats

Dans cette section, nous présentons les figures et le tableau permettant de visualiser la complexité du réseau, ses dissimilarités et l'effet du nombre d'étudiants par cours.

La figure 1 illustre le réseau de collaboration des étudiants du cours *BIO500* avec leurs collègues des autres cours, et schématise les valeurs de collaborations par un gradient de taille. Ces valeurs, représentant le nombre d'interactions différentes faites par chaque étudiant, vont de 1 (petit cercle) à 15 (grand cercle).

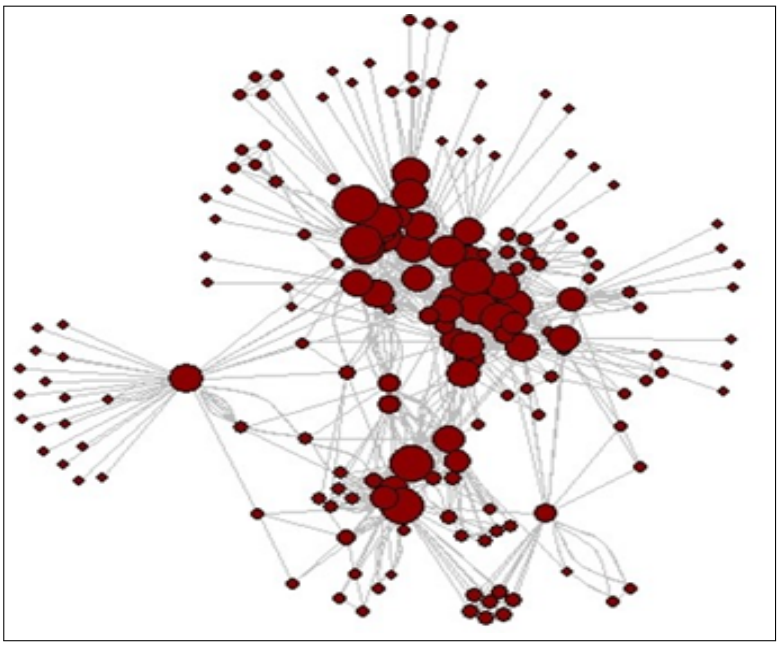

Figure 1 – Réseau de collaboration des étudiants ayant en commun le cours *BIO500 - Méthodes en écologie computationnelle*

Les figures 2 à 4 ci-dessous illustrent le réseau de collaboration entre les étudiants des cours *BCM113*, *BIO500* et *ECL406*. Chaque cercle représente un étudiant. Chaque ligne représente un lien entre étudiants.

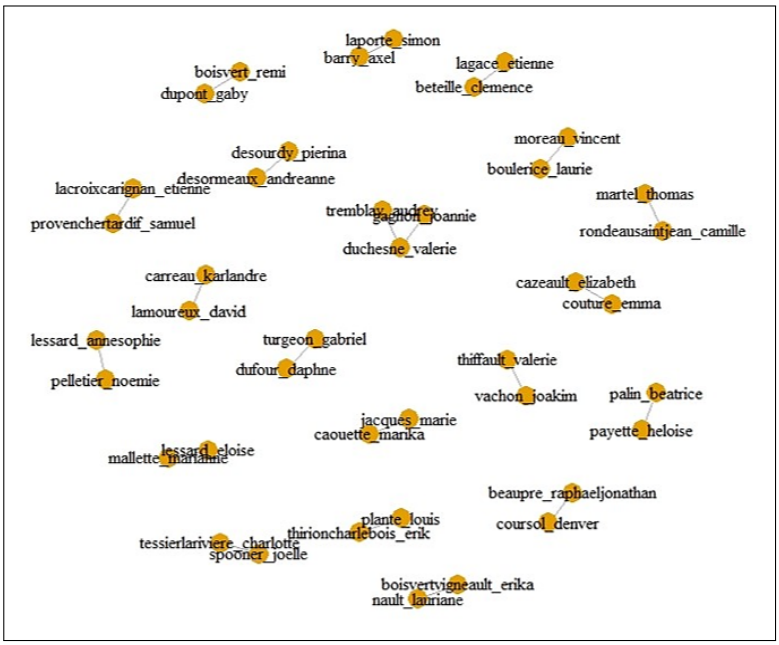

Figure 2 – Réseau de collaboration entre les étudiants de la population du cours *BCM113 - Biochimie générale - Travaux pratiques*

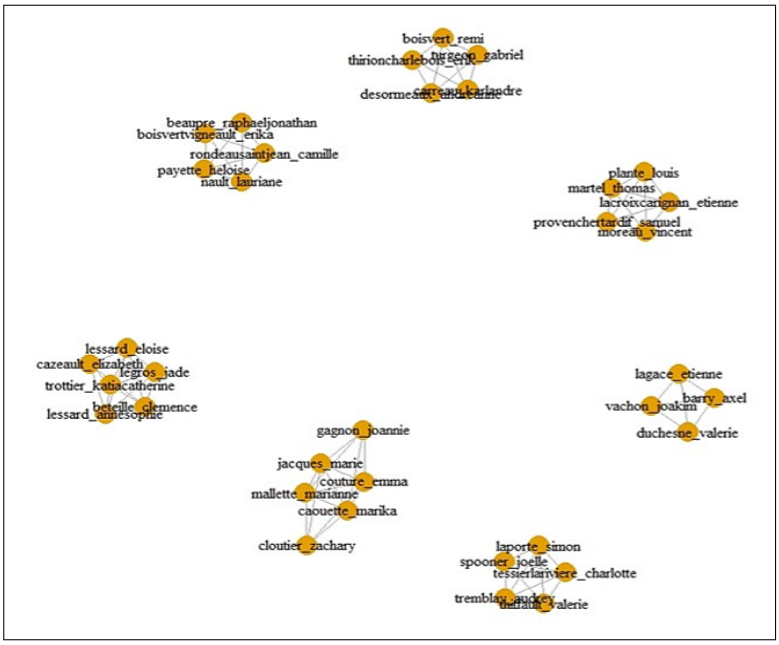

Figure 3 – Réseau de collaboration entre les étudiants de la population du cours *BIO500 - Méthodes en écologie computationnelle*

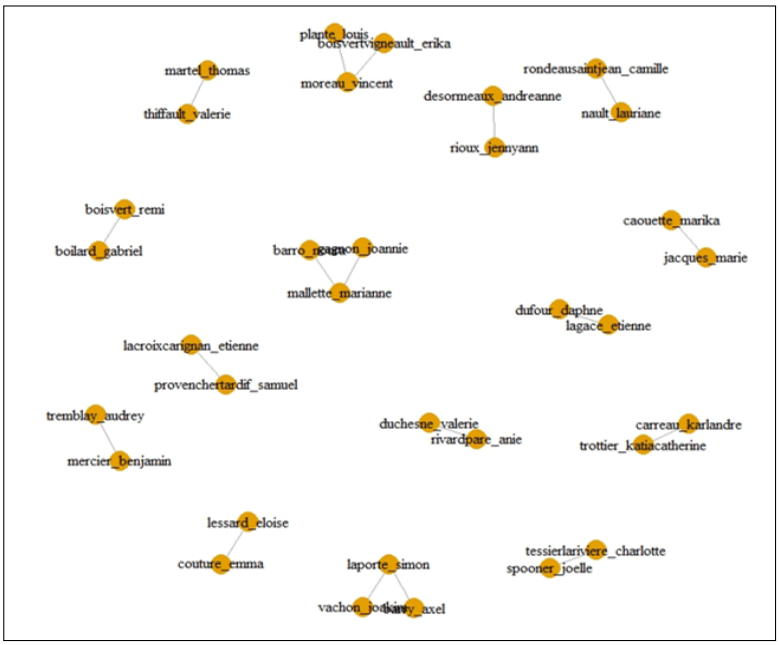

Figure 4 – Réseau de collaboration entre les étudiants de la population du cours *ECL406 - Tendances évolutives des plantes terrestres*

Une valeur de dissimilarité a été calculée entre chaque paire de cours. La dissimilarité représente le degré auquel deux cours diffèrent par leur composition en individus ou leurs interactions. Celle-ci est comprise entre 0 (dissimilarité nulle) et 1 (dissimilarité complète). La figure 5 (p.6) représente sous forme de matrice 30x30 la distribution de cette valeur pour chaque paire de cours. Les cours comparés avec eux-mêmes ont une dissimilarité nulle, 4 paires de cours ont une dissimilarité moyenne, 62 une dissimilarité élevée, et 804 une dissimilarité très élevée comprise entre]0.80, 1].

La figure 6 (p.7) représente une régression linéaire de la dissimilarité moyenne observée pour un cours par rapport aux autres cours en fonction de son nombre d'étudiants. On observe une faible corrélation négative.

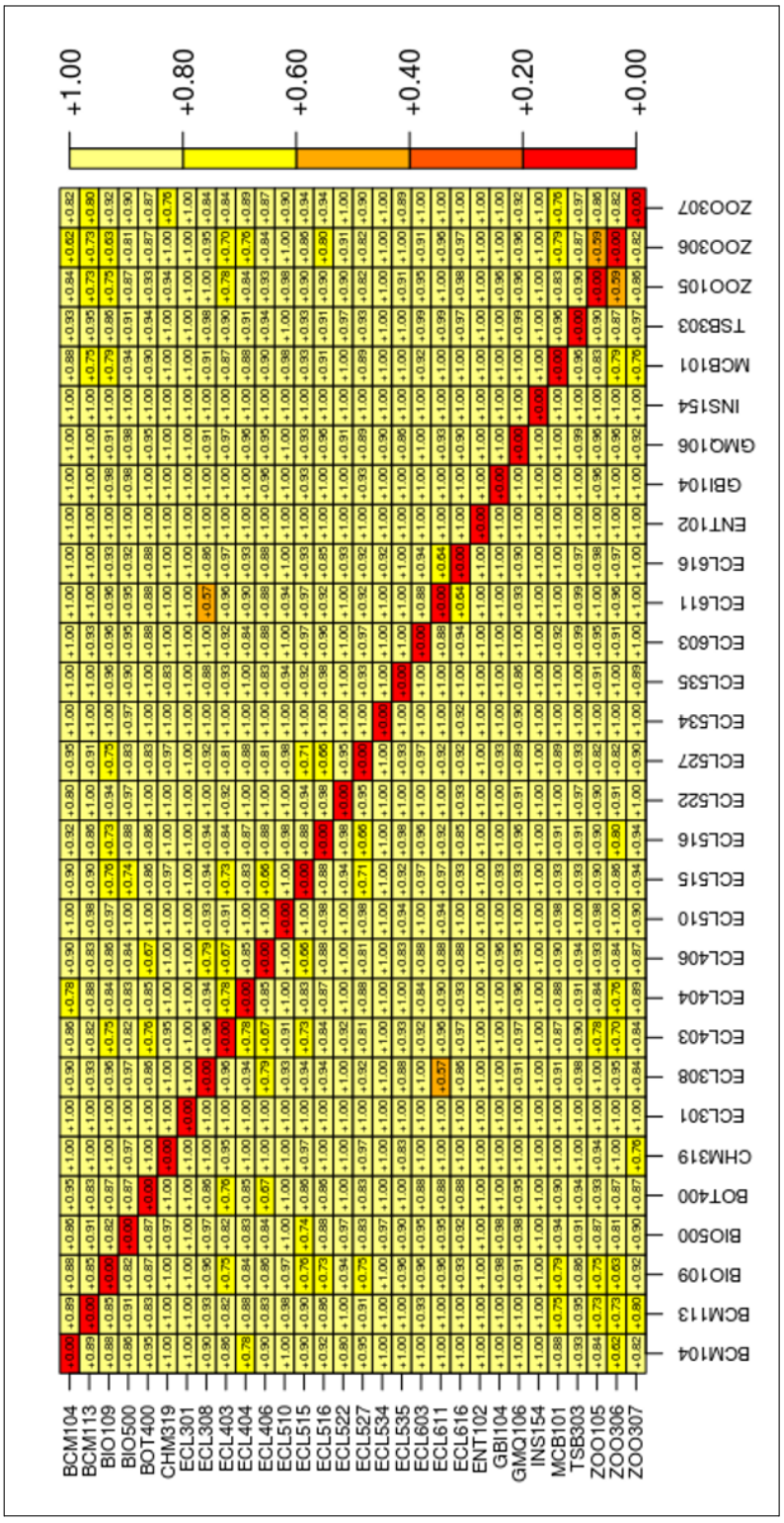

Figure 5 – Matrice de dissimilarité des collaborations entre chacun des cours

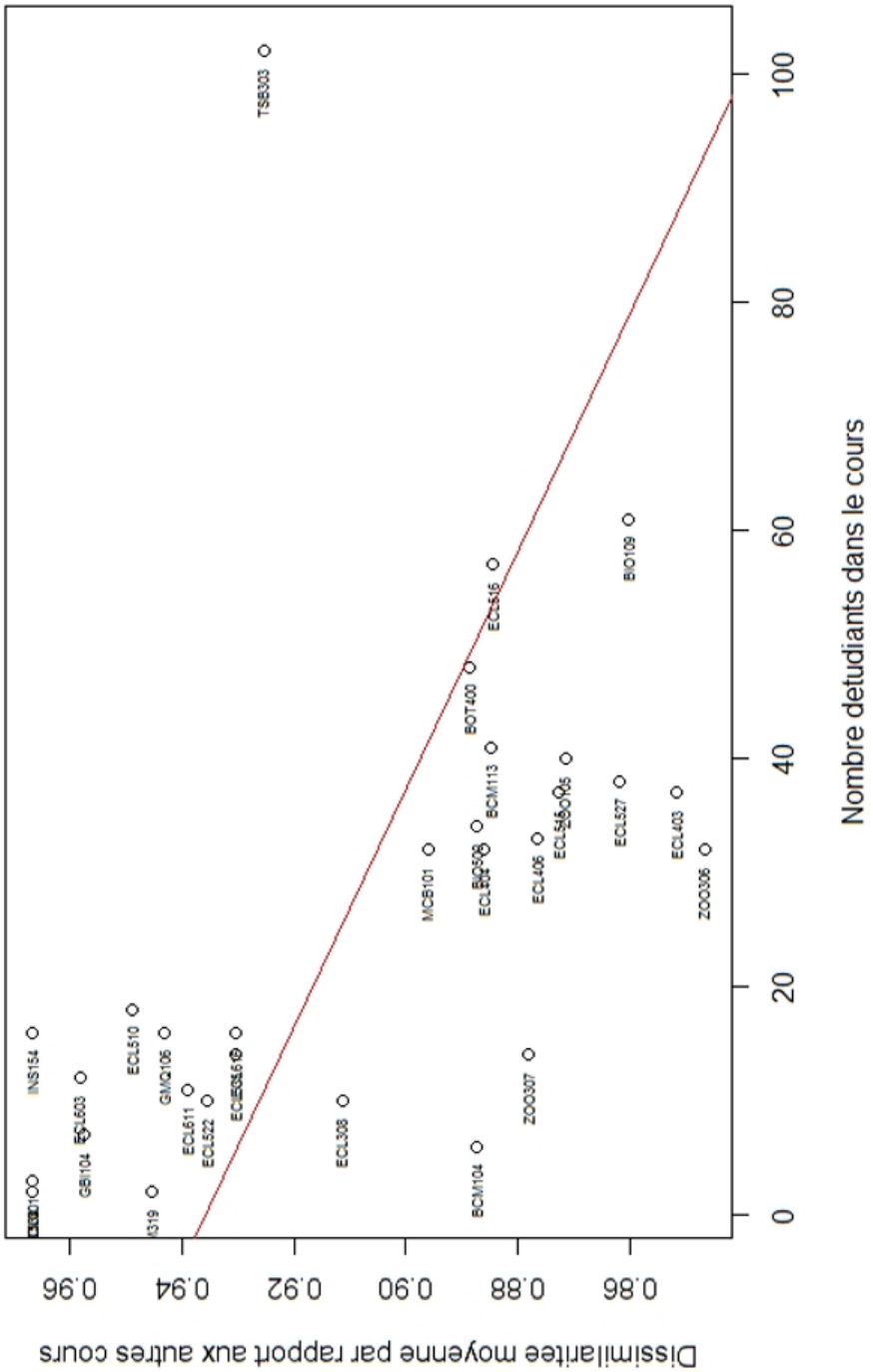

Figure 6 – Régression de la dissimilarité moyenne en fonction du nombre d'étudiants dans un cours

Discussion

Constat des résultats et interprétation

Réseaux

Dans la figure 1, il est possible de constater que le réseau est constitué d'individus avec des valeurs de collaborations variées. Les étudiants n'ayant fait que quelques collaborations semblent provenir d'une cohorte ou d'un programme qui diffère des étudiants centraux ayant des collaborations nombreuses dans ce réseau, ce qui explique cette différence dans le nombre d'interactions.

Nous avons décidé d'illustrer les réseaux de collaborations des cours *BCM113*, *BIO500* et *ECL406* car ce sont les cours qui représentent le mieux les étudiants centraux du réseau à travers les cours répertoriés. Ce sont ceux qui démontrent le mieux le comportement d'un large groupe d'étudiants en "population", et qui restent le plus constant dans le nombre et la variabilité des individus. Ceci nous permet de construire le reste de notre analyse car tous les liens de coopérations sont notamment tirés du groupe d'étudiants de *BIO500*.

Matrice

Pour interpréter la matrice, prenons par exemple les cours suivant : *ECL308* et *ECL611*. Il est possible de constater que les interactions dans ces cours sont les plus similaires de l'analyse (coefficient de dissimilarité de 0.57). Puisque ce sont des cours offerts dans un cheminement irrégulier, les élèves y sont moins nombreux et sont "communs" aux deux groupes. Cela pourrait expliquer la faible dissimilarité entre ces deux cours, et appuyer notre constat global.

À l'inverse, on constate que les cours *BIO500* et *BCM113* ont un coefficient de dissimilarité de 0.91, ce qui témoigne d'une grande différence dans les interactions qu'on y retrouve. Cela pourrait être expliqué, entre autre, par le fait que *BCM113* est un cours suivi en début de cheminement alors que *BIO500* est plutôt suivi à la fin du cheminement. Ainsi, les habitudes sociales des étudiants ont pu changer entre temps.

Régression linéaire

En ce qui a trait à la question de départ, nous avons décidé de nous intéresser à la relation entre la dissimilarité moyenne des cours (réseaux d'interaction) et le nombre d'étudiants par cours. À partir de la régression linéaire obtenue, on constate une corrélation négative entre le nombre d'étudiants et la dissimilarité. Cela pourrait s'expliquer par le fait que les étudiants ont plus de possibilités de partenaires dans de grands groupes, tandis qu'ils se retrouvent plus restreints dans leur choix dans les petits groupes, au risque de former les mêmes équipes.

Comparaison des méthodes avec la littérature

Dans la littérature, on retrouve plusieurs méthodes utilisées pour comparer des réseaux écologiques. Certaines d'entre elles considèrent les ressemblances entre réseaux comme des caractéristiques mathématiques. Elles ont parfois recours à des comparaisons algorithmiques, des analyses multivariées ou de simples analyses statistiques (Faust and Skvoretz, 2002; Vermaat et al., 2009; Poisot et al., 2011; Baiser et al., 2011). Afin de répondre à notre objectif de déterminer la relation dissimilarité/taille des groupes, nous avons eu recours à la méthode de calcul de dissimilarité de Whittaker. Cette dernière considère le nombre d'items, soit dans notre cas, le nombre d'interactions, unique à chacune des populations comparées ainsi que le nombre d'items communs aux deux populations (Poisot et al., 2012).

Conclusion

Les résultats démontrent une haute dissimilarité des réseaux de collaborations entre l'ensemble des cours. Cette dissimilarité semble être due au nombre d'étudiants dans le cours. Il aurait été intéressant de pousser l'analyse en calculant un indice de fidélité pour chaque étudiant participant au cours. Cela permettrait de déterminer si les individus conservent les mêmes liens, donc les mêmes interactions interspécifiques. Nous pourrions poser l'hypothèse qu'un individu infidèle se retrouverait plus au centre du réseau de collaborations, puisqu'il possèderait plus de liens, et que, à l'inverse, un individu fidèle se retrouverait en périphérie du réseau, puisqu'il possèderait moins de liens. Ce serait un beau défi pour les prochains étudiants !

Bibliographie

Baiser, B., Ardeshiri, R. S., and Ellison, A. M. (2011). Species richness and trophic diversity increase decomposition in a co-evolved food web. *PLoS One*, 6(5).

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems :1695.

Faust, K. and Skvoretz, J. (2002). Comparing networks across space and time, size and species. *Sociological methodology*, 32(1) :267–299.

Hagen, M., Kissling, W. D., Rasmussen, C., De Aguiar, M. A., Brown, L. E., Carstensen, D. W., Alves-Dos-Santos, I., Dupont, Y. L., Edwards, F. K., Genini, J., et al. (2012). Biodiversity, species interactions and ecological networks in a fragmented world. In *Advances in ecological research*, volume 46, pages 89–210. Elsevier.

Klinke, S. (2019). *plot.matrix : Visualizes a Matrix as Heatmap*. R package version 1.4.

Müller, K., Wickham, H., James, D. A., and Falcon, S. (2020). *RSQLite : 'SQLite' Interface for R*. R package version 2.2.0.

Ognyanova, K. (2016). Network analysis with r and igraph : Netsci x tutorial. https ://kateto.net/networks-r-igraph. Consulté en mars 2020.

Plutniak, S. (2018). L'analyse de graphes avec r : un aperçu avec igraph. https ://hal.archives-ouvertes.fr/hal-01885485/document. Consulté en mars 2020.

Poisot, T., Canard, E., Mouillot, D., Mouquet, N., and Gravel, D. (2012). The dissimilarity of species interaction networks. *Ecology letters*, 15(12) :1353–1361.

Poisot, T., Lepennetier, G., Martinez, E., Ramsayer, J., and Hochberg, M. E. (2011). Resource availability affects the structure of a natural bacteria–bacteriophage community. *Biology letters*, 7(2) :201–204.

Vermaat, J. E., Dunne, J. A., and Gilbert, A. J. (2009). Major dimensions in food-web structure properties. *Ecology*, 90(1) :278–282.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12) :1–20.

Wickham, H. (2019). *stringr : Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.

Wickham, H., François, R., Henry, L., and Müller, K. (2020). *dplyr : A Grammar of Data Manipulation*. R package version 0.8.5.

Xie, Y. (2020). *knitr : A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.28.