

UNIVERSITÉ PARIS DIDEROT (PARIS 7)
SORBONNE PARIS CITÉ

ÉCOLE DOCTORALE : FRONTIÈRES DU VIVANT

DOCTORAT

DISCIPLINE : SCIENCES INTERDISCIPLINAIRES DU VIVANT
Statistiques et écologie

Jean-Benoist LEGER

Modelling the topology of ecological bipartite networks with statistical models
for heterogeneous random graphs

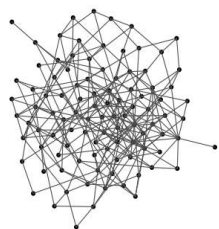
Thèse dirigée par : Jean-Jacques DAUDIN et Corinne VACHER
Soutenue le : 30 janvier 2014

Rapporteurs :

M. Christophe AMBROISE,
M. Eric KOLACZYK

Jury :

M^{me} Isabelle DAJOZ, Présidente,
M^{me} Élisabeth THÉBAULT, Examinatrice,
M. Christophe AMBROISE, Rapporteur,
M. Jean-Jacques DAUDIN, Co-directeur,
M^{me} Corinne VACHER, Co-directrice.



Remerciements

Il est une chose de se lancer dans une thèse, il est une chose de réaliser son travail de thèse, il est une chose de rédiger son manuscrit, mais il en est une autre de rédiger ses remerciements de thèse. Lors de la rédaction du manuscrit, j'ai souvent entendu me dire, ou dire à d'autres, ton manuscrit n'est pas ton testament ; autant je conçois que cela puisse être la vérité qu'il me semble que les remerciements de thèse puissent constituer un arrêt de mort.

Cet exercice, conventionnel et formel, revêt une importance capitale, et bien souvent, c'est la seule chose que liront de la thèse toutes les personnes non spécialistes du domaine de recherche. Ces remerciements sont ce qui restera aux yeux de la majorité. Ai-je pour autant envie d'écrire des remerciements ? Je ne le sais guère. Non que je n'aie personne à remercier, loin de là, mais je n'imagine pas que cette forme constitue la meilleure alternative pour remercier ceux qui ont contribué de manière directe et indirecte à cette thèse.

Comment devrais-je alors procéder ?

Une éventualité serait de rédiger les remerciements sous la forme d'une liste insipide et exhaustive. Une liste à énumération m'a paru être pendant une période une solution raisonnable. Mais le choix de la méthode de classement apparaît comme une problématique induite délicate à traiter, et il est exclu d'utiliser l'ordre alphabétique usuel, puisqu'il consiste à approuver un classement suivant les personnes tout au long de leur vie, ne leur permettant ni d'évoluer, ni d'être surpris. Beaucoup d'autres classements sont possibles, suivant un condensat cryptographique comme le SHA-1 par exemple. Je complique singulièrement la tâche de ceux qui voudraient vérifier leur position dans le classement. D'autres choix simples peuvent être envisagés, comme un classement par l'ordre alphabétique de la seconde lettre du prénom. Mais la justification de l'utilisation de la seconde lettre et non de la troisième s'avère délicate, et on ne pourrait pas exclure le choix volontaire de la seconde lettre pour faire apparaître un protagoniste particulier avant un autre. Bref, J'ai pris la décision de ne pas rédiger mes remerciements sous forme de liste.

J'ai également choisi de nommer la majorité des personnes par leurs prénoms, non qu'il soit naturel d'utiliser leurs prénoms pour tous, mais cela confère une homogénéité agréable, du moins à l'écriture de ces remerciements, je n'ai fait d'ex-

ception que pour le jury et le comité de thèse.

Tout d'abord, comme il se doit, je tiens à remercier mes encadrants de thèse, Jean-Jacques et Corinne. Ils m'ont apporté beaucoup et formé l'un et l'autre à la recherche, et je leur dois énormément. Jean-Jacques m'a appris à organiser mes idées, à les présenter, les écrire, et les hiérarchiser. Il m'a également appris à les abandonner lorsqu'il était nécessaire de le faire, et même si ce fût difficile, je dois l'en remercier. Corinne m'a été d'une très grande aide dans le processus d'écriture, et j'espère réussir un jour à décrire à l'écrit un concept comme elle le fait. Elle m'a également appris à relier mes idées à l'écologie, et à ne pas croire qu'au côté théorique des méthodes, je dois l'en remercier.

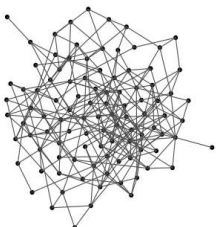
Il m'apparaît également judicieux de remercier Christophe Ambroise et Éric Kolaczyk, rapporteurs de ma thèse, dont les commentaires très intéressants ont permis de comprendre l'importance relative des points abordés, et les éléments à améliorer. Je remercie également Élisabeth Thébault et Isabelle Dajoz, pour avoir accepté de faire partie de mon jury.

Tom Snijders, Louis-Félix Bersier, Catherine Matias et Colin Fontaine, au sein de mes comités de thèse m'ont beaucoup apporté, aussi bien en termes d'idées que d'orientation, je les en remercie.

Il convient maintenant de remercier ceux, qui ont à chaque instant refréné leurs pulsions meurtrières, c'est-à-dire les thésards qui ont partagé mon bureau. Aurore, bien qu'ayant tenté au moyen de divers projectiles d'attenter à ma vie, n'a pas eu suffisamment de conviction pour arriver à ses fins, je l'en remercie, et je m'excuse de l'éventuel traumatisme que j'ai pu causer. Antoine, qui a tenté plusieurs fois de se débarrasser de mon encombrante personne en provoquant l'ire d'Aurore, il est ainsi de certaines personnes, agissant de manière détournée, subtiles et vicieuses. Alice, qui bien que n'ayant tenté à aucun moment, du moins à ma connaissance, de se débarrasser de moi, a été d'un calme et d'une patience spectaculaires.

Je profite de ces remerciements pour présenter mes excuses à Julie pour l'avoir vouvoyée pendant deux ans. Je remercie Pierre, Marie-Pierre et Tristan, dans le bureau voisin, qui nous a supportés. Pierre, qui a dû parcourir les affres de mon code, mérite une attention toute particulière. Je remercie également Tristan qui a été mon premier contact dans le labo. Marie-Pierre par son goût de l'étrange, comme l'orange, a su apporter de manière subtile de l'incongru à un bureau de thésard en rédaction.

Même s'ils sont membres de l'infâme, le perfide, l'ignoble *bureau des doctorants du bas*, ils méritent à titre personnel des remerciements pour l'ambiance apportée dans le labo, il s'agit de Frédéric, Jean-Baptiste, et Aurélien. Virginie Branier n'étant pas conviée dans ces remerciements. Ils ont su apporter la détente à ceux qui rédigeaient, et pour Frédéric et Aurélien, l'insouciance de ceux qui ne rédigeront pas dans l'année. Je dois aussi remercier Marie, du nouveau *bureau des doctorants*



du haut, à ce titre, et en particulier pour la mini randonnée à vélo, qui a été un très court mais utile bol d'air dans la rédaction. Même s'il ne s'agit pas de remerciement, j'en profite pour souhaiter bon courage à Jean-Baptiste pour sa rédaction.

Je remercie également Stéphane, couramment considéré comme un dieu, pour nous avoir montré un idéal, Liliane, qui m'a appris à doser avec précision mes critiques, Gabriel, Sarah, Marie-Laure, Maud, Celine, Pierre, Eleanna, Émilie, Anna, Loïc, Souhil, Éric, Guillaume, Julien, pour leur soutien.

Je remercie également Odile et Francine de m'avoir guidé dans les labyrinthes administratifs de l'INRA. Damien et Hamid, qui ont fait en sorte que les serveurs fonctionnent, et accueillent mes calculs, méritent un remerciement.

Je pense avoir énuméré toutes les personnes au labo qui ont contribué de manière indirecte à ce manuscrit, dans le cas où j'en aurais oublié, je présente mes excuses les plus sincères.

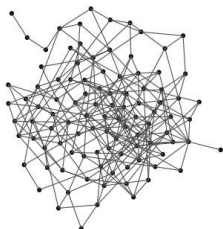
Lorsque j'ai commencé cette thèse je vivais dans une colocation que je qualifierais d'étrange, pour rester conforme à la réalité. Cette expérience m'a beaucoup marqué et j'ai entretenu dès lors des relations avec certaines personnes qui m'ont soutenu et ouvert l'esprit durant ma thèse. Je pense en particulier à Émilien, un mélange entre un sociologue et un physicien, et à Matthias un mélange entre un physicien et un geek pythoniste, me soutenir durant la rédaction. Au caractère étrange de ces deux personnages, je me dois de ne pas oublier Smaïl, qui, philosophe de son état, est capable de nous interroger sur le sens profond et le bien-fondé de nos recherches. Je me dois également de ne pas oublier de remercier Camille, Carole, Cécile, Félix et Pierre.

Je pense également à ceux qui ne m'ont soutenu, et qui considèrent que l'oubli est la meilleure des solutions tout en oubliant qu'une absence trop flagrante est plus marquante qu'une présence très discrète. Je ne les remercie pas.

Je me dois de penser à ceux qui m'ont permis d'entretenir ma folie durant ma thèse, à réfléchir et à imaginer, des choses diverses et variées. À ce titre, je remercie les anciens et nouveaux membres du CRANS avec qui j'ai pu aborder des thèmes qui m'étaient chers, me permettant d'avoir un pied dans un monde qui à défaut d'être normal, était extérieur à la thèse. Au même titre, je dois remercier les gens qui m'ont permis d'avancer sur mon moteur de recherche d'itinéraire cyclable, les membres d'OpenStreetMap France, la pratique cycliste était un moyen sûr de s'extraire de la thèse.

Enfin, il convient de ne pas oublier les membres de ma famille, qui ont su, chacun à leur manière, m'apporter un important soutien, Claudy et Robert, mes parents, Arlette ma grand-mère, et Francine ma tante.

Et pour terminer, je remercie Jean Boucasier.



Résumé

Un réseau écologique constitue une représentation de l'ensemble des interactions entre espèces dans un contexte donné. L'analyse de la structure topologique de ces réseaux permet aux écologues d'identifier et de comprendre les processus sous-jacents. La détection de sous-groupes d'espèces interagissant fortement ensemble, souvent nommés communautés ou compartiments, est un des principaux moyens pour interpréter la structure sous-jacente des réseaux.

La méthode de détection de communauté la plus utilisée dans les réseaux écologiques est la méthode de maximisation de la modularité. Toutefois, cette popularité semble plus fondée sur des raisons historiques, et en particulier le premier article publié sur ce thème dans ce contexte, que d'un choix rationnel avec de solides justifications.

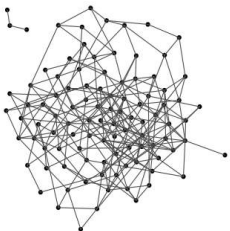
Il existe de nombreuses autres méthodes de détection de communauté, et de manière plus générale de classification non supervisée, qui peuvent être utilisées pour analyser des réseaux écologiques. L'analyse des réseaux est actuellement un sujet de recherche en pleine expansion avec des applications dans des domaines de recherches variés comme la génomique, les sciences sociales, l'informatique, ou la physique. À notre connaissance, il n'existe pas de comparaison des méthodes de classification non supervisée dans le cas des réseaux écologique.

Nous avons effectué une revue des méthodes disponibles de classification non supervisées des nœuds d'un réseau, et nous avons comparé une partie de ces méthodes dans un contexte écologique. Nous avons montré que la méthode de maximisation de la modularité produit des résultats satisfaisants pour détecter les sous-groupes d'espèces dans des réseaux bipartites, mais que cette méthode donne rarement les meilleurs résultats dans l'ensemble des méthodes comparés. Nos résultats montrent que l'algorithme *edge-betweenness* avec le critère de modularité pour sélectionner le nombre de groupes donne les meilleurs résultats dans le cas des réseaux d'interaction bipartite binaire. Dans le cas des réseaux valués, l'inférence du *stochastic block model* donne de très bons résultats, mais au prix d'un temps de calcul important.

Afin d'évaluer la contribution relative des différents processus pouvant expliquer la structure d'un réseau d'interaction, nous avons introduit de l'information extérieure au réseau (covariables) dans les méthodes de classification non

supervisée ; par exemple, nous avons utilisé l'effort d'échantillonnage et la fréquence de rencontre entre espèces afin d'expliquer la structure du réseau d'interaction. Après avoir développé un programme en C++ d'inférence du *stochastic block model* avec des covariables, nous avons analysé deux réseaux d'interaction arbre-champignon et arbre-insecte. Ces résultats sont préliminaires, mais l'application de la méthode semble ouvrir des perspectives intéressantes dans l'étude des réseaux écologiques.

De manière parallèle, nous avons également cherché des communautés dans un réseau écologique de nature différente, un réseau de reproduction entre arbres collecté sur deux espèces ayant la capacité de s'hybrider entre elles. Nous avons utilisé ces résultats pour discuter d'un concept central en écologie, le concept d'espèce.



Abstract

An ecological network is a representation of the whole set of interactions between species in a given context. Ecological scientists analyse the topological structure of such networks, in order to understand the underlying processes. The identification of sub-groups of highly-interacting species (usually called communities, or compartments) is an important stream of research.

The most popular method for the search of communities in ecological networks is the modularity optimization method. However this popularity is more due to the first paper published on this topic than to a rational choice based on solid grounds.

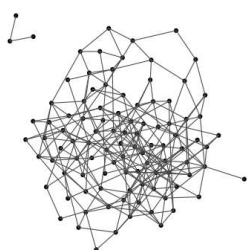
There are many other clustering methods that could be used to delimit communities in ecological networks. The analysis of complex networks is indeed a rapidly growing topic with many applications in several scientific fields, such as genomics, social, computer and physical sciences. To our knowledge, no comparison of different clustering methods is available in the case of ecological networks.

Here we reviewed the whole set of methods available for clustering networks and we compared them using an ecological benchmark. We showed that modularity maximization is a satisfactorily method for clustering species in ecological bipartite networks, but it is not the best. Our results showed that the edge-betweenness algorithm with modularity criterion for selecting group number is a good method for retrieving sub-groups of highly interacting species in binary bipartite networks. The stochastic block model gave very good results in the case of weighted bipartite networks, but it was very time consuming.

In order to assess the relative contribution of several processes to the network structure, we integrated exogenous information in the clustering model. For instance, we integrated the sampling effort and some ecological mechanisms such as the encounter probability between species. After having developed a C++ package based on the stochastic block model with covariates, we analysed two bipartite antagonistic networks with this method, a tree-fungus and tree-insect network. The results are still preliminary but the method seems to us very promising for future ecological studies.

Finally we searched communities in a different kind of network, a mating net-

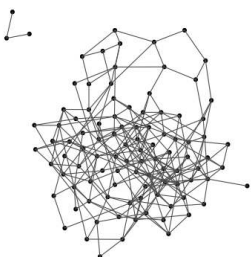
work between individuals belonging to two hybridizing tree species. We used our results to discuss a concept which is central in ecology, the species concept.



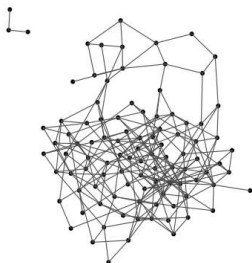
Contents

1	Introduction	15
1.1	Networks in ecology	15
1.1.1	Ecological networks topology	15
1.1.2	Ecological networks dynamics	16
1.2	Statistical methods for networks	16
1.3	Problematic and outline	17
2	Detection of structurally homogeneous subsets in graphs	23
2.1	Introduction	23
2.2	Basic notations and an example	26
2.2.1	A toy example	26
2.2.2	Transformation of the raw graph	27
2.3	Methods based on an algorithm	29
2.3.1	Markov Cluster algorithm (MCL)	30
2.3.2	Hierarchical agglomerative clustering algorithm	34
2.3.3	Spectral Clustering	35
2.3.4	Edge-Betweenness	37
2.4	Methods based on an optimization criterion	38
2.4.1	Modularity criterion	38
2.4.2	Cut	40
2.5	Methods developed with a model	41
2.5.1	Model-based clustering for social networks (MBCSN)	41
2.5.2	Random Dot Product Graphs (RDPG)	42
2.5.3	Stochastic Block Model (SBM)	43
2.5.4	Continuous Stochastic Block Model (CSBM)	45
2.6	Application to the Zachary's karate club	46
2.7	Conclusion	47
3	Graph clustering methods differ in their ability to detect patterns in bipartite ecological networks	57
3.1	Introduction	58

3.2	Materials and methods	59
3.2.1	Graph clustering methods	59
3.2.2	Simulation of ecological bipartite networks	60
3.2.3	Criteria used to compare the efficiency of the clustering methods	71
3.3	Results	72
3.3.1	Relative efficiency of the clustering methods when applied to ecological bipartite networks with average properties . . .	72
3.3.2	Effect of network properties on the hierarchy between clustering methods	75
3.4	Discussion	81
4	Wmixnet: Software for Clustering the Nodes of Binary and Valued Graphs using the Stochastic Block Model	93
4.1	Introduction	93
4.2	SBM model with covariates	94
4.2.1	Notations	94
4.2.2	The model	95
4.2.3	Probability laws	95
4.2.4	Analysis of groups when covariates are used	96
4.3	Estimation method	96
4.3.1	Initialization	97
4.3.2	Smoothing	97
4.3.3	Parallelism	97
4.4	wmixnet program	98
4.4.1	Sources availability and installation	98
4.4.2	Input format	98
4.4.3	Output format	99
4.4.4	Command line usage	99
4.4.5	Empirical complexity	100
4.4.6	Capacity of extension	100
4.5	Example	101
4.5.1	Projected networks	101
4.5.2	Covariate data	101
4.5.3	Example of command line	101
4.5.4	Results	102
5	Deciphering the mechanisms shaping ecological networks: a framework and a method	107
5.1	Introduction	107
5.2	Materials and methods	108



5.2.1	The network data	108
5.2.2	The statistical model	109
5.2.3	Application of the model to the data	111
5.3	Results	112
5.3.1	Relative importance of the three mechanisms	117
5.4	Ongoing work and discussion	117
A	Putting the Biological Species Concept to the Test: Using Mating Networks to Delimit Species	121



Chapter 1

Introduction

1.1 Networks in ecology

Ecology is the scientific study of interactions among organisms and their environment, such as the interactions organisms have with each other and with their abiotic environment. An ecological network is a representation of the whole set of interactions between species in a given context (in a given geographic area, over a particular period of time). Network nodes correspond to species and network links represent biotic interactions (predation, parasitism, mutualism).

1.1.1 Ecological networks topology

The topology of ecological networks is often described by using the following properties:

- Connectance, i.e. the proportion of realized ecological interactions among the potential ones (see Dunne et al., 2002).
- Node degree distribution, i.e. the statistical properties of the distribution of number of interactions per species (see Jordano et al., 2003; Bascompte and Jordano, 2007)
- Nestedness. A nested network displays both asymmetric specialization — *i.e.* species with few interactions (‘specialist’ species) preferentially interact with species with many interactions (‘generalist’ species) — and a dense core of interactions created by symmetric interactions between generalist species (see Almeida-Neto et al., 2008; Ulrich et al., 2009)
- Modularity, or compartmentalization. Compartmentalization is characterized by recognizable subsets of interacting species, with species more likely to be linked within than across subsets (see Krause et al., 2003; Krasnov et al., 2012; Rezende et al., 2009; Guimera et al., 2010).

There are however alternative ways of grouping species together within a network (Allesina and Pascual, 2009). Instead of grouping together highly-interacting species, species can be grouped according to their interaction profile. In this way, the species belonging to the same group tend to have a similar structural role. For instance, the generalist species tend to be grouped together, and specialist species tend to be grouped together. Hereafter this type of group is called Structurally Homogeneous Subset (Leger et al., 2013).

It is noteworthy that the study of compartmentalization is not disconnected from the study of stability (McNaughton, 1978). For instance, following Stouffer and Bascompte (2011), compartmentalization increases persistence of species. Many authors studied the influence of the structural properties of ecological networks, such as nestedness and compartmentalization, on their dynamics.

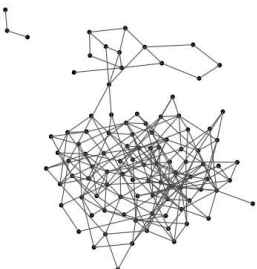
1.1.2 Ecological networks dynamics

The characterization of network dynamics allows to answer various questions about biodiversity conservation (see Traveset et al., 2013). Analysis of network dynamics allows to study the stability of the network. There are various definitions of stability used in ecology. The stability can be:

- Local stability and resilience. This stability index represents the robustness of the system to infinitesimal perturbation, and the ability to recover from this perturbation (i.e. to come back to the equilibrium point without considerations of time). This definition is used by Neubert and Caswell 1997, Chen and Cohen (2001) and Pimm and Lawton (1978).
- Reactivity. This stability index represents the short-term stability and the ability to rapidly recover from a perturbation. This stability index is described by Neubert and Caswell (1997) and Chen and Cohen (2001).
- Variability. This stability index measures the fluctuations of species abundances, when a perturbation is applied. This definition is used by Pimm (1984).
- Persistence. This stability index refers to the number of species which do not get extinct. This measure is used by Hofbauer and Sigmund (1998), Chen and Cohen (2001) and Thébault and Fontaine (2010).

1.2 Statistical methods for networks

Data sets structured as networks are present in many domains such as social science, engineering and physics, molecular and population biology, ecology and computer science. Many methods and models have been developed from three



communities: physics, statistics and computer science (Kolaczyk, 2009). Below we present a very brief introduction to this huge topic.

Statistical models for social networks defined for static networks include conditionally uniform models, latent space models and exponential random graph models (Snijders, 2011). The review made by Snijders (2011) underlines and discusses the probabilistic model. Computational and statistical estimation are more underlined in Hunter et al. (2012).

The community of computer science and statistical physics have developed closely related models from a different point of view. Some models assume a progressive building of the network node by node using simple rules and other ones underline general properties of networks such as scale-free probability distribution function of the degrees or the small world property (Kolaczyk, 2009; Strogatz, 2001).

The work done in this PhD thesis is inter-disciplinary. It lies at the crossroads between statistics, computer science and ecology. The identification of sub-groups of highly-interacting species in ecological networks is a main stream of research. Our work is thus aimed at comparing and improving the network clustering methods and models, for future applications to ecological networks.

1.3 Problematic and outline

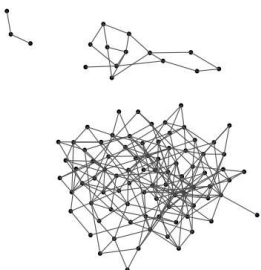
The most popular method for the search of groups of highly-interacting species in ecological networks is the modularity optimization method using the simulated annealing optimization approach (Guimera and Amaral, 2005) implemented in the software NETCARTO. However this popularity is more due to the first paper published on this topic than to a rational choice based on solid grounds. In the field of genomics the clustering for networks is often done using the MCL algorithm (Van Dongen, 2000), which has been found as the best by Vlasblom and Wodak (2009).

There are many other clustering methods that can be used and no comparison is available. The objective of our work is

1. to review the whole set of methods available for clustering networks in chapter 2.
2. to compare them using an ecological benchmark in chapter 3.
3. most of the above methods are not well suited to assess the relative contribution of different ecological mechanisms to the network structure. For that, we integrated exogenous information in the clustering model, such as sampling artefacts ecological mechanisms such as the encounter probability between species. After having developed (Chapter 4) a C++ package

based on the stochastic block model with covariates Mariadassou et al. (2010), we analysed two bipartite antagonistic networks with this method, a tree-fungus and tree-insect network. The results are still preliminary (Chapter 5).

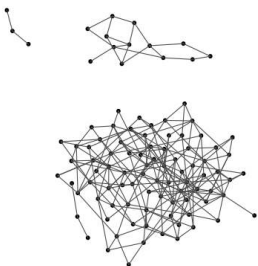
4. The assumption that species are grouped in communities is perhaps too stringent and not always pertinent in ecological studies. It is possible to use a more flexible model using continuous latent variables in place of the discrete ones associated with the clusters. In collaboration with Lelia Lagache and Remy Petit, we have used the model continuous stochastic block model to analyse a mating network (Lagache et al., 2013) (Appendix A).



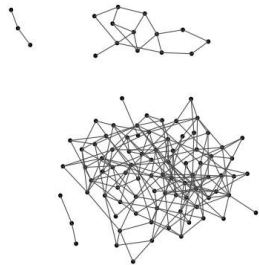
Bibliography

- S. Allesina and M. Pascual. Food web models: a plea for groups. *Ecology letters*, 12(7):652–662, 2009.
- M. Almeida-Neto, P. Guimaraes, P. R. Guimarães, R. D. Loyola, and W. Ulrich. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, 117(8):1227–1239, 2008.
- J. Bascompte and P. Jordano. Plant-animal mutualistic networks: the architecture of biodiversity. *Annu. Rev. Ecol. Evol. Syst.*, 38:567–593, 2007.
- X. Chen and J. E. Cohen. Global stability, local stability and permanence in model food webs. *Journal of Theoretical Biology*, 212(2):223–235, 2001.
- J. A. Dunne, R. J. Williams, and N. D. Martinez. Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, 5(4):558–567, 2002.
- R. Guimera and L. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005. ISSN 0028-0836.
- R. Guimera, D. Stouffer, M. Sales-Pardo, E. Leicht, M. Newman, and L. Amaral. Origin of compartmentalization in food webs. *Ecology*, 91(10):2941–2951, 2010.
- J. Hofbauer and K. Sigmund. *Evolutionary games and population dynamics*. Cambridge University Press, 1998.
- D. R. Hunter, P. N. Krivitsky, and M. Schweinberger. Computational Statistical Methods for Social Network Models. *Journal of Computational and Graphical Statistics*, 21(4):856–882, 2012.
- P. Jordano, J. Bascompte, and J. M. Olesen. Invariant properties in coevolutionary networks of plant–animal interactions. *Ecology letters*, 6(1):69–81, 2003.
- E. D. Kolaczyk. *Statistical analysis of network data*. Springer, 2009.

- B. R. Krasnov, M. A. Fortuna, D. Mouillot, I. S. Khokhlova, G. I. Shenbrot, and R. Poulin. Phylogenetic signal in module composition and species connectivity in compartmentalized host-parasite networks. *The American Naturalist*, 179(4): 501–511, 2012.
- A. E. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor. Compartments revealed in food-web structure. *Nature*, 426(6964):282–285, 2003.
- L. Lagache, J.-B. Leger, J.-J. Daudin, R. J. Petit, and C. Vacher. Putting the Biological Species Concept to the Test: Using Mating Networks to Delimit Species. *PLOS ONE*, 8(6):e68267, 2013.
- J.-B. Leger, C. Vacher, and J.-J. Daudin. Detection of structurally homogeneous subsets in graphs. *Statistics and Computing*, pages 1–18, 2013.
- M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, 4(2):715–742, 2010.
- S. McNaughton. Stability and diversity of ecological communities. *Nature*, 274 (5668):251–253, 1978.
- M. G. Neubert and H. Caswell. Alternatives to resilience for measuring the responses of ecological systems to perturbations. *Ecology*, 78(3):653–665, 1997.
- S. Pimm and J. Lawton. On feeding on more than one trophic level. *Nature*, 275 (5680):542–544, 1978.
- S. L. Pimm. The complexity and stability of ecosystems. *Nature*, 307(5949): 321–326, 1984.
- E. L. Rezende, E. M. Albert, M. A. Fortuna, and J. Bascompte. Compartments in a marine food web associated with phylogeny, body mass, and habitat structure. *Ecology Letters*, 12(8):779–788, 2009.
- T. A. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37:131–153, 2011.
- D. B. Stouffer and J. Bascompte. Compartmentalization increases food-web persistence. *Proceedings of the National Academy of Sciences*, 108(9):3648–3652, 2011.
- S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.



- E. Thébault and C. Fontaine. Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science*, 329(5993):853–856, 2010.
- A. Traveset, R. Heleno, S. Chamorro, P. Vargas, C. K. McMullen, R. Castro-Urgal, M. Nogales, H. W. Herrera, and J. M. Olesen. Invaders of pollination networks in the Galápagos Islands: emergence of novel communities. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758), 2013.
- W. Ulrich, M. Almeida-Neto, and N. J. Gotelli. A consumer’s guide to nestedness analysis. *Oikos*, 118(1):3–17, 2009.
- S. Van Dongen. Graph clustering by flow simulation. *University of Utrecht*, 275, 2000.
- J. Vlasblom and S. Wodak. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC bioinformatics*, 10(1):99, 2009.



Chapter 2

Detection of structurally homogeneous subsets in graphs

This article was published in *Statistics and Computing*.

Abstract

The analysis of complex networks is a rapidly growing topic with many applications in different domains. The analysis of large graphs is often made via unsupervised classification of vertices of the graph. Community detection is the main way to divide a large graph into smaller ones that can be studied separately. However another definition of a cluster is possible, which is based on the structural distance between vertices. This definition includes the case of community clusters but is more general in the sense that two vertices may be in the same group even if they are not connected. Methods for detecting communities in undirected graphs have been recently reviewed by Fortunato. In this paper we expand Fortunato's work and make a review of methods and algorithms for detecting essentially structurally homogeneous subsets of vertices in binary or weighted and directed and undirected graphs.

Keywords: Graphs, Clusters, Random Walk, Spectral Clustering, Stochastic Block Model, Bipartite Graphs

2.1 Introduction

The analysis of complex networks is a rapidly growing topic with many applications in different fields such as social sciences, physics, computer science, molecular biology and ecology. The size of the social and biological datasets and the size of the networks created by human-kind are growing with time. This is

an issue because networks with thousands of vertices are difficult to analyze as a whole object.

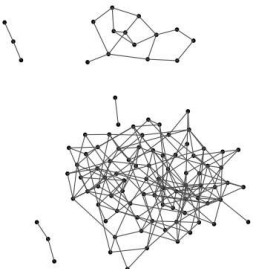
An obvious strategy consists in dividing the big network into smaller independent ones and analyzing each small network separately. Therefore at this time, one of the most important challenges is to build unsupervised classification of the vertices. Most of the current research is focused on the search for a community structure with high connectivity between vertices of the same cluster and low connectivity between vertices of different clusters. This strategy has been used in the field of molecular biology to obtain "independent modules" in metabolic or Protein-Protein interaction networks in the domain of molecular biology. It has also been used to extract the scientific communities from bibliometrics networks or social groups in social networks. Recently a very large and impressive review of community detection methods and algorithms in graphs has been made by Fortunato (2010). This paper describes many methods but also gives some elements for comparing them on benchmarks.

However this strategy has its own limits because in some cases connected vertices may be very different. A typical example is bipartite graphs such as host-parasite networks, where there is a connection between a host species and a parasite species if the species parasites the host species. Therefore the host species and the parasite species may be in the same community, putting in the same bag two very different species. Therefore there is a need for a more general definition of what constitutes a cluster of vertices in networks.

Another definition of a cluster is possible, which is based on the structural distance between vertices. Two vertices are in the same group if they have a similar profile of connection to the other vertices. This definition includes the case of community clusters but is more general in the sense that two vertices may be in the same group even if they are not connected. Moreover it is possible to obtain groups of vertices which are not connected within groups but are highly connected to another group of vertices. This notion of structural distance is related to the definition of the Structural Equivalence of Actors in a social network defined by Lorrain and White (1971): actors are structurally equivalent if they have identical relational ties to and from all the actors in a network. These two different approaches are introduced Burt (1978): *"There are several questions that can be posed for a specific project that might lead an individual to analyze subgroups in terms of cohesion versus structural equivalence. Here, considering a series of such questions, I conclude that subgroups based on structural equivalence are to be preferred to those based on cohesion."*

Two classes of methods for clustering the vertices of graphs can thus be defined with two different goals:

1. to obtain communities *i.e.* subsets of vertices strongly connected within



subsets and loosely connected between subsets,

2. to obtain structurally homogeneous subsets, *i.e.* subsets of vertices having the same or similar interaction profiles.

The concept of structurally homogeneous subsets generalizes the concept of community in the sense that a community is also a structurally homogeneous subset when the structure of the graph is represented by communities, because vertices in the same community share the same structural connectivity behavior. In Figure 2.1, in the same non-bipartite network, there is an example of the difference between communities and structurally homogeneous subsets with a hub structure. Even if hubs are within communities, they have a different behavior in the network structure, and they are classified in different structurally homogeneous subsets.

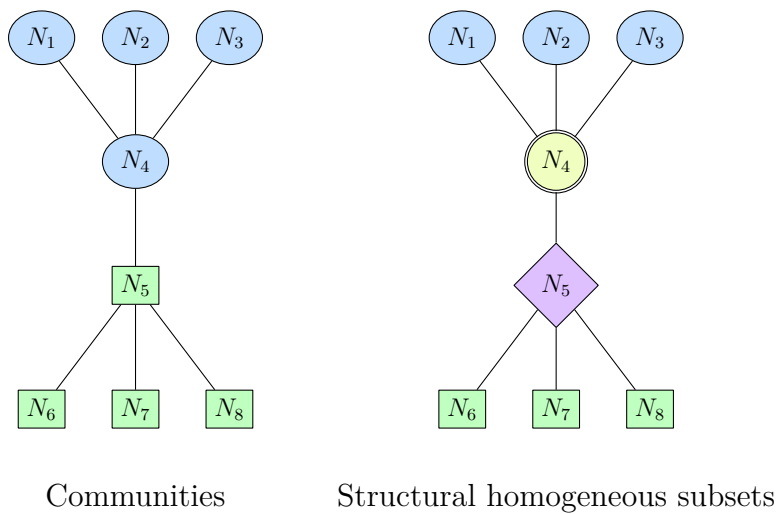


Figure 2.1: Difference between communities and structural homogeneous subsets in a hub structure network

Fortunato's review is focused on community detection for binary undirected graphs. In this paper we expand Fortunato's work and give a review of methods and algorithms for detecting essentially structurally homogeneous subsets of vertices in binary or weighted and directed or undirected graphs. Moreover we do not try to give an exhaustive list of methods. We prefer to limit the scope to what we have presumed to be the main methods, and to make a self-contained presentation of each of them. Note that we do not present the methods for very large graphs with more than 10^6 vertices, such as the world-wide web or telephone network.

The methods for detecting Structurally Homogeneous Subsets come from three different scientific fields: computer science, physics and statistics. Each scientific community has its own journals and there are few links between them. Statisticians

prefer to use probabilistic or statistical models whereas the other two communities use algorithmic or optimization methods. Optimization methods optimize a criterion which represents the quality of the partition of the graph. Algorithmic methods use a sequence of operations to build a partition of the graph. Probabilistic models are models of the process which are supposed to have generated the data and statistical methods are used to estimate the parameters of the probabilistic model. In this review a significant part is given to statistical models which had little space in Fortunato's review.

Section 2.2 gives the basic notations, some transformations of the base data and a toy graph that will be analyzed throughout the paper. Section 2.3 presents the methods of clustering based on an algorithm, Section 2.4 presents the methods based on an optimization criterion, Section 2.5 is devoted to statistical models for clustering graphs. Section 2.6 illustrates methods on the Zachary's Karate Club Network. The last section gives a summary of the methods and some links between them.

2.2 Basic notations and an example

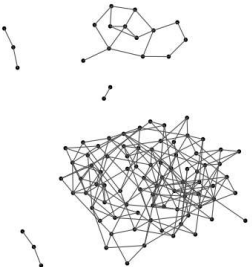
Let us consider a graph (or network) $G = (V, E)$ with V the set of n vertices (or vertices) and $E \subset V \times V$ the set of edges. The value of the edge from i to j is $w_{ij} \geq 0$. The self loops ($w_{ii} \neq 0$) may be accepted or not. The matrix W is the matrix of weights. Let $(d_W^o)_i = \sum_j w_{ij}$ be the out-degree of vertex i and $(d_W^i)_j = \sum_i w_{ij}$ be the in-degree of vertex j . The matrix of outgoing degrees D_W^o is the diagonal matrix composed of $(d_W^o)_i, i = 1 \dots, n$ (with a similar definition for D_W^i).

The graph is non-weighted (binary) when $w_{ij} \in \{0, 1\}$. In that case, W is called the adjacency matrix. The graph is undirected when W is symmetric. In that case, $(d_W^o)_i = (d_W^i)_i = (d_W)_i$ and D_W is called the degree matrix. A bipartite graph is a graph whose vertices can be divided into two disjoint sets A and B such that every edge connects a vertex in A to one in B.

Each method will be illustrated on the toy example presented in the following section. Searching for clusters is only relevant if the graph is void-free, so we suppose that no vertex of the toy graph is isolated.

2.2.1 A toy example

The toy example (Fig.2.2) is a binary, undirected graph, with 10 vertices. Vertices 1 to 5 are of one type, and vertices 6 to 10 of another. The adjacency



matrix W and the degree matrix D are (only non-zero values are printed):

$$W = \begin{pmatrix} & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \end{pmatrix} \quad D = \begin{pmatrix} 3 & & & & & & & & & & \\ & 2 & & & & & & & & & \\ & & 3 & & & & & & & & \\ & & & 2 & & & & & & & \\ & & & & 1 & & & & & & \\ & & & & & 1 & & & & & \\ & & & & & & 2 & & & & \\ & & & & & & & 3 & & & \\ & & & & & & & & 2 & & \\ & & & & & & & & & 3 & \\ & & & & & & & & & & 3 \end{pmatrix}$$

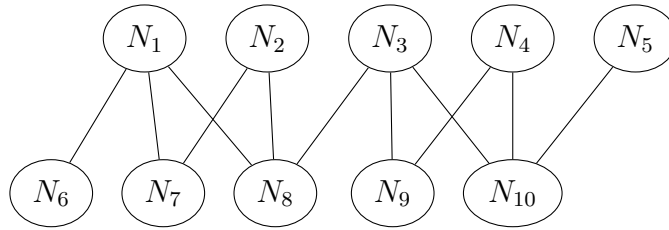


Figure 2.2: Bipartite toy-example

The difference between the community and structural homogeneity appears clearly in Figure 2.3: if one searches for communities, that is groups of vertices more connected within groups than between groups, one will find the clusters of the column titled "communities" including some vertices which have different connectivity behavior, with respect to bipartite network, in the same group. On the other hand the structurally homogeneous subsets are exclusively composed of one type of vertex. The bottom right case, with 4 clusters, decomposes each of the two clusters of the top left case into two subgroups, separating the two types of vertices. Note that this toy-example is proposed here for illustration purposes, showing that structurally homogeneous clusters may be composed of vertices that are not connected within-group. With real data things are not so clear-cut: in the same graph some clusters may be communities and other ones may be poorly connected within-group and characterized by a high connectivity to a given cluster. Some examples of this configuration of clusters are given in Daudin et al. (2008) for a metabolic network and in Picard et al. (2009) for different biological networks.

2.2.2 Transformation of the raw graph

The methods for detecting clusters of vertices can either be applied to the raw matrix of weights (or adjacency matrix) or to a weighted matrix obtained by a similarity transformation of the raw matrix. The same method used on the raw graph or the transformed one may give different results.

Therefore the process of clustering graphs may contain two successive steps:

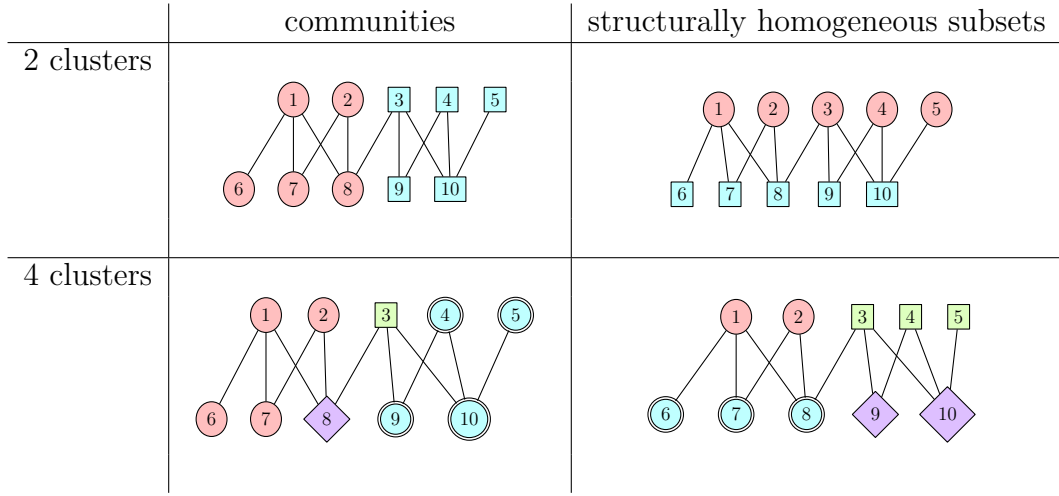


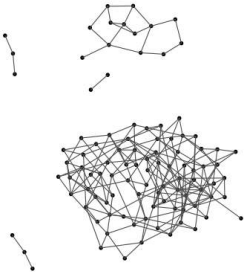
Figure 2.3: Various clusters obtained for the toy-example

1. The pre-treatment step, i.e. a transformation of the raw graph into a modified one, This step is *not* mandatory.
2. The application of a given clustering method to the modified graph.

In this paper we focus on the clustering methods, but the importance of the pre-treatment step, *i.e.* the transformation of the raw graph in a modified one, should not be underestimated in practice. Note that there are two types of transformation:

1. the transformation does not bring any new information concerning the vertices or the edges. In this case the transformation defines a specific similarity measure between vertices suited to answering a specific question. These transformations are generally not useful with generative statistical models that are supposed to model how the raw data have been generated. They are more often used in combination with algorithmic or optimization methods.
2. there is some more information available which is not included in the raw data W . This information may be included in the statistical model using co-variables on the edges or on the vertices. The algorithmic or optimization methods must include a pre-treatment using a transformation of W incorporating the new information.

The transformation of the raw data provides a weighted graph whose weights are a measure of similarity between each pair of vertices. Note that new edges may be produced by this procedure and old ones can be deleted. Note also that many similarity indices exist and that the similarity index should be chosen according to the scientific question. Let us consider the toy-example (Figure 2.2). If the



aim is to cluster the vertices with similar connectivity behavior, using the Jaccard similarity index may be a good choice. The Jaccard coefficient between two vertices i and j is the number of vertices connected to i and j divided by the number of vertices connected to i or j .

After this transformation, we obtain the following similarity matrix and the graph of Figure 2.4 which has two connected components, one per type of vertex.

For the particular case of bipartite graph note that two connected components are obtained, separating the two types of vertices. This is not the general case.

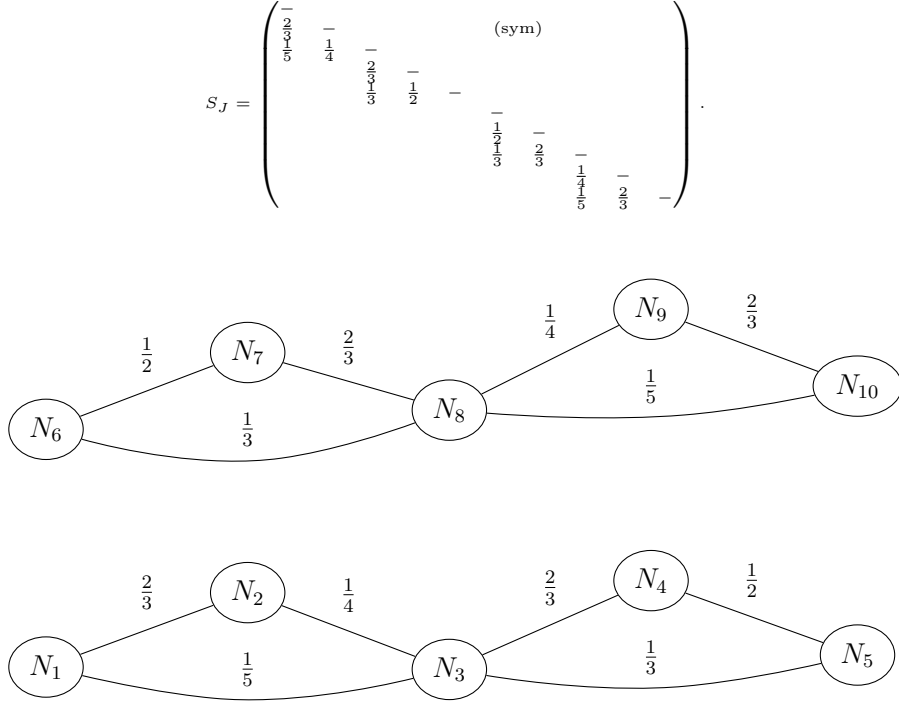


Figure 2.4: The toy example graph transformed with Jaccard's measure of similarity

Note that similarity transformation can change the meaning of groups. With the Jaccard similarity index, communities in the transformed graph are structurally homogeneous subsets in the original graph.

2.3 Methods based on an algorithm

The clustering methods that do not use a statistical model may be divided into two classes: both are defined by an algorithm, but this algorithm may be designed

or not to optimize a criterion. The following section presents the algorithmic methods that do not explicitly optimize any criteria.

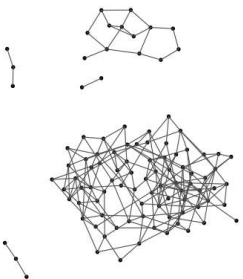
2.3.1 Markov Cluster algorithm (MCL)

Synopsis	
Name	Markov Cluster algorithm (MCL)
Type of method	Algorithm
Type of graphs	Undirected ^a , weighted or not, inducing an associated ergodic Markov Chain <i>a.</i> The condition of ergodicity is more problematic to obtain for directed graph
Type of clusters	Structurally Homogeneous Subsets
Summary	This method uses random walks on the graph and classifies in the same group the vertices whose associated random walk converge to the same state
Time complexity	The author claims a time complexity of $O(V ^3)$, but this complexity is obtained by considering the number of iterations as constant.

The MCL algorithm (Van Dongen, 2000) allows the search for structurally homogeneous subsets by considering a random walk on the graph.

A random walk on the graph is a sequence of moves at discrete time points, from one vertex to another, along graph edges. The probability of a move along an edge is proportional to its weight. Let E_{it} be the event of being in the set i in time t . For all t , $(E_{it})_{i=1\dots n}$ only depends on $(E_{it-1})_{i=1\dots n}$, therefore the random walk is a Markov chain. The behavior of a random walk from a starting vertex is determined by the set of probabilities of a move from this vertex to another vertex j in t steps for all (j, t) . A vertex is characterized by the behavior of the random walk starting from this vertex. The main idea of this method is to consider that vertices with the same random walk behavior are in the same cluster.

A standard random walk (with no inflation factor) on a connected graph converges to the same asymptotic state of the Markov chain for any starting vertex. The objective of the MCL algorithm is to build k clusters with $k > 1$, so the usual random walk has to be modified to achieve this goal. The idea of the MCL algorithm is to constrain the random walk to converge to a different state depending on the starting vertex. This is achieved using the *inflation* operation. The more important the *inflation* operation is, the more numerous the obtained asymptotic states are. The aim of the MCL algorithm is to group vertices whose associated random walks converge to the same state.



Let the transition matrix of the Markov chain be $T = (W_{sl})(D^o)_{W_{sl}}^{-1}$. T_{ij} is the probability of going from vertex j to vertex i in one step. Therefore T is a column-wise stochastic matrix ($\forall j, \sum_i T_{ij} = 1$). Note that in MCL notation, the transition matrix is the transpose of the usual notation for the transition matrix of the Markov chain, which is a row-wise stochastic matrix. T is the matrix of probabilities of transition in one step and T^k is the matrix of probabilities of transition in k steps.

Let $T^{(1)} = T$. MCL alternates two operations indexed by k starting at $k = 1$:

1. $T^{(2k)} = (T^{(2k-1)})^{e_k}$, is the *transition* operation which allows the progress of the random walk. The importance of this operation is larger when e_k is large.
2. $T^{(2k+1)} = \Gamma_{r_k}(T^{(2k)})$ is the *inflation* operation which allows the random walk to converge toward several stable states. Γ_{r_k} is a term by term r_k power operator followed by a column sum normalization. This operation inflates the high values of the matrix $T^{(2k)}$ and reduces the small ones. For example, $\Gamma_2([.5, .3, .2]) = [.25, .09, .04]/.38 = [.66, .24, .11]$. The importance of the inflation operation is larger when r_k is large.

The algorithm ends when $T^{(k)}$ is idempotent ($T^{(k+1)} = T^{(k)}$). Denote $T^{(\infty)}$ this idempotent matrix. The columns of $T^{(\infty)}$ correspond to the vertices of the graph. Each row of $T^{(\infty)}$ defines a cluster. Non-zero values within each row indicate the composition of the cluster. In the general case, several clusters are empty. Therefore there are fewer clusters than vertices. When a vertex belongs to several clusters, different affectation rules may be applied.

The Markov chain of the random walk must be ergodic. In particular the Markov Chain must be aperiodic and irreducible. Some graphs must be modified to satisfy this aperiodicity, generally by adding self-loops. The irreducible condition is always satisfied for undirected graphs, but can be not satisfied for directed graphs. For example in a bipartite graph, when all edges connect one type of vertex to another, there is one set of absorbent states, and consequently the Markov Chain is reducible. A directed bipartite graph must be transformed (for example by symmetrization) before using the MCL algorithm.

In any case, applying the MCL method means clustering a graph which is not the true one (because of the addition of self-loops). However the true graph can be approached with a graph with very low weighted self-loops.

Figure 2.5 shows that the MCL algorithm applied to the toy example with unitary self-loops, retrieves communities instead of structurally homogeneous subsets. This is because the vertices with different structural connectivity behavior have the same structural behavior in the random walk when self-loops are added to the graph. To illustrate this idea, let us imagine the random walk on the graph *without* self-loops. In this case, the random walk would alternate between the two types

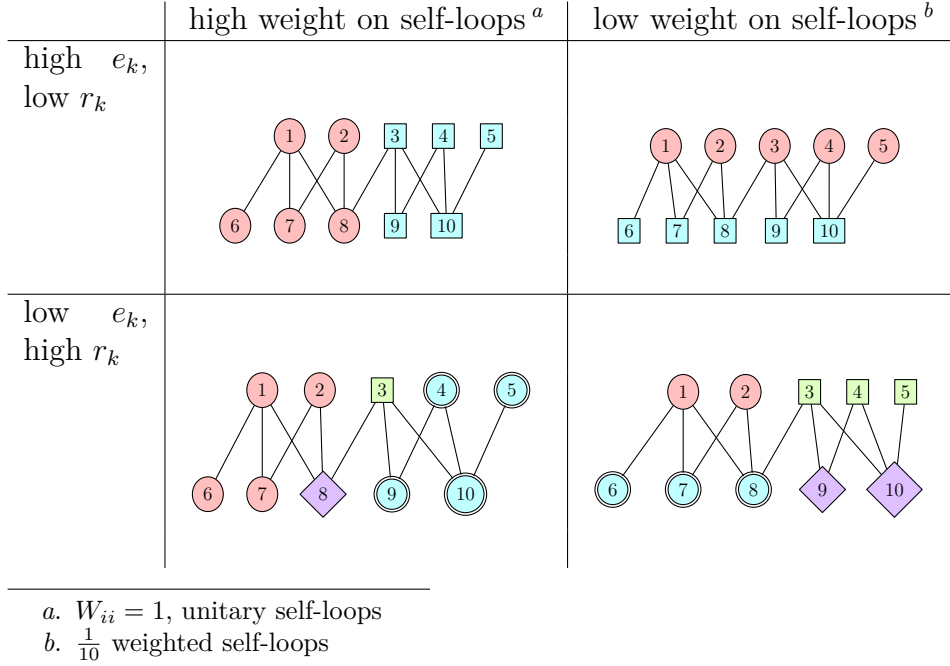
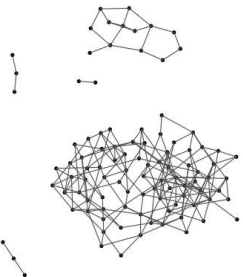


Figure 2.5: MCL applied to the toy example with 4 combinations of tuning parameters

of vertices. The algorithm would not converge and this is why the possibility of null self-loops is *a priori* excluded when using MCL. Nevertheless decreasing the weight (Δ) of self-loops is allowed. In the toy example with a decreased value of the self-loops (10 times less than unitary edges), structural homogeneous subsets are obtained (see Figure 2.5).

There are three tuning parameters, Δ , e_k and r_k . Their values have a great impact on the number of clusters of the final result, see Figure 2.5. There are few groups when $\frac{e_k}{r_k}$ is high and many groups when $\frac{e_k}{r_k}$ is low. The author does not give any default option for the choice of e_k and r_k . After a (limited) number of empirical trials we have found that values around 2 for these two parameters could be a good choice. Δ should be small in order to capture structurally homogeneous clusters.

MCL gives satisfactory results for dense graphs, and is less efficient for sparse graphs. To our knowledge, this method has been applied mostly in the domain of molecular biology. Brohee and Van Helden (2006) found that MCL gives satisfactory results for the extraction of complexes from protein-protein interaction networks.



Pons-Latapy distance. Pons and Latapy (2006) propose a distance based on a random walk on the graph. This distance is introduced for binary, undirected graphs but can be extended to the case of weighted, undirected graphs. Like MCL, this method requires the addition of self-loops if the Markov chain is not ergodic.

The method consists in stopping the random walk after a small number of steps, k . The transition matrix after k steps is T^k where T_{ij}^k is the probability of transition from vertex j to vertex i in k steps. As for MCL, a vertex is characterized by the behavior of the random walk starting at this vertex, but this method studies the behavior of a truncated random walk instead of the asymptotic behavior of a modified (by the inflation factor) random walk. Two vertices j_1, j_2 which have a similar structural behavior (in the graph with self-loops) spawn two random walks which have, for all i , a similar probability of going in k steps to the vertex i . Therefore the column vectors $T_{j_1}^k$ and $T_{j_2}^k$ are similar. To compare the structural behaviors of vertex j_1 and vertex j_2 , the distance $\|T_{j_1}^k - T_{j_2}^k\|_2$ between $T_{j_1}^k$ and $T_{j_2}^k$ can be computed. The issue is that this distance is also influenced by the vertex degree because the probability of going from vertex j to vertex i is affected by the degree of vertex i (a random walk has a higher probability of going to a vertex of high degree). Therefore a re-normalization is applied to vectors $T_{j_1}^k$ and $T_{j_2}^k$ by dividing their rows by the degree of the corresponding vertices.

The re-normalized vectors are $D_{W_{sl}}^{-\frac{1}{2}} T_{j_1}^k$ and $D_{W_{sl}}^{-\frac{1}{2}} T_{j_2}^k$. The Pons-Latapy distance between vertices j_1 and j_2 is defined as the distance between re-normalized vectors $D(j_1, j_2) = \left\| D_{W_{sl}}^{-\frac{1}{2}} T_{j_1}^k - D_{W_{sl}}^{-\frac{1}{2}} T_{j_2}^k \right\|_2$.

This distance is only an intermediate element in the algorithm and does not include the classification of the vertices. After computing the Pons-Latapy distances, a supplementary classification step is necessary. A hierarchical agglomerative algorithm is thus applied to the Pons-Latapy distance matrix in order to cluster the vertices of the graph.

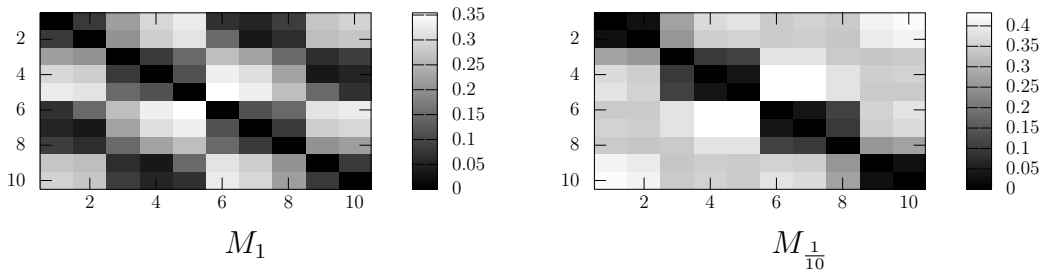


Figure 2.6: Pons-Latapy distance matrices for $k = 4$ corresponding to the toy example (Figure 2.2) with unitary (M_1) and $\frac{1}{10}$ -weighted self-loops ($M_{\frac{1}{10}}$). Vertices are ordered as $N_1 \cdots N_5, N_6 \cdots N_{10}$

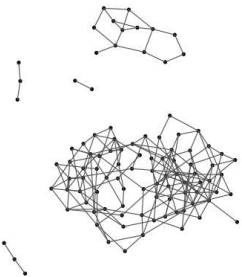
There is a strong influence of the weight of self-loops. In the case of the toy example, the Figure 2.6 shows that the results are completely changed when these weights are changed: with unitary self-loops, the distance matrix does not separate vertices which have a different structural linkage behavior. For example, N_1 and N_2 are closer to N_6, N_7, N_8 , than to N_3, N_4, N_5 . As with MCL, decreasing the weight of self-loops increases the ability of the method to separate vertices which have different structural behaviors.

There are three tuning parameters, k , Δ and the specific hierarchical algorithm used in the classification step, such as UPGMA, Ward or maximum linkage algorithms. The authors do not give any advice about their choice. Our empirical trials suggest that $k = 3$ could be a good choice. As for MCL, Δ should be small in order to capture structurally homogeneous clusters.

2.3.2 Hierarchical agglomerative clustering algorithm

Synopsis	
Name	Hierarchical Agglomerative clustering algorithm
Type of method	Algorithm
Type of graphs	Graph with a dissimilarity between vertices
Type of clusters	depends on the dissimilarity
Summary	This method groups vertices into meta-vertices recursively
Time complexity	Basically $O(V ^3)$, with an additive cost in memory, a time complexity of $O(V ^2 \log(V))$ can be reached

This algorithm is useful for clustering graphs once a distance between vertices has been defined. Note that we have two graphs, the original one and the weighted graph of dissimilarities. This algorithm gives communities of the graph of dissimilarities, but the clusters obtained can be structurally homogenous subsets or communities for the original graph, depending on the distance used in the algorithm. The result depends on the local or global building of the dissimilarities between vertices: for instance, if the dissimilarity between vertices is the Jaccard or Pons-Latapy (see section 2.3.1) dissimilarity measures, then one obtains structurally homogeneous subsets. Conversely if the dissimilarity between two vertices equals one when two vertices are not linked and 0 when there is an edge between them, then one obtains communities. The usual hierarchical agglomerative algorithms are well known (Hartigan, 1975): Ward, single linkage, complete linkage or UPGMA (Unweighted Pair Group Method with Arithmetic Mean). A classification algorithm such as a hierarchical agglomerative algorithm or a k-means method are the necessary final step for some methods that only compute a distance between vertices or continuous latent positions for vertices. Therefore the results



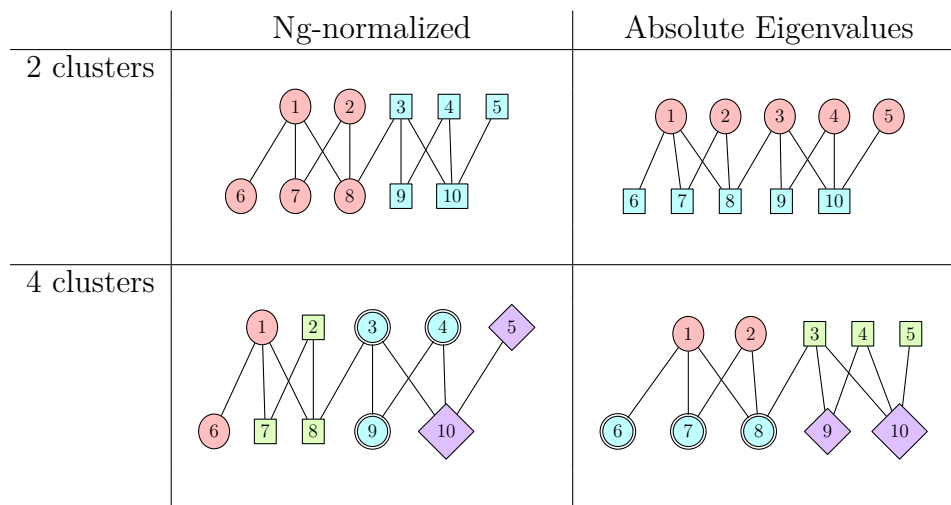


Figure 2.7: Clusters obtained with Ng-normalized and Absolute Eigenvalue Spectral Clustering, with $k \in \{2, 4\}$

of these methods depend not only on their own tuning parameter but also on the peculiar classification algorithm used to cluster the vertices.

2.3.3 Spectral Clustering

Synopsis	
Name	Spectral Clustering
Type of method	Algorithm
Type of graphs	Undirected, weighted or not
Type of clusters	Communities or Structurally Homogeneous Subsets
Summary	This method computes continuous latent variables using eigenvectors of the Laplacian matrix of the graph and classifies the vertices using a k-means algorithm based on the most important latent variables.
Time complexity	Time complexity depends mainly on the computation of the eigen decomposition, basically $O(V ^3)$.

This algorithm first proposed by Donath and Hoffman (1973) allows the search for communities by considering the Laplacian matrix of the graph, $L = D_W - W$. This algorithm spawned a family of algorithms which are well described by Von Luxburg (2007). It applies only to undirected graphs, but there is some work in progress to extend it to directed graphs.

It is known that if a graph has k connected components, the Laplacian matrix has a null eigenvalue with multiplicity k (Von Luxburg, 2007). Each eigenvector

associated with the null eigenvalue is composed of zero and non-zero values. A non-zero value for the j^{th} eigenvector and the row i means that vertex i is in connected component j . If the graph has k communities, the Laplacian matrix has k eigenvalues close to zero. The idea of Spectral Clustering is to determine the composition of the k communities by considering the k eigenvectors associated with the k lowest eigenvalues.

Let $L = D_W - W$ be the unnormalized Laplacian and $L_N = I - D_W^{-1/2} W D_W^{-1/2}$ the normalized Laplacian. The Spectral Clustering algorithm has several variants. The first three are described in Von Luxburg (2007) and the last one is more recent:

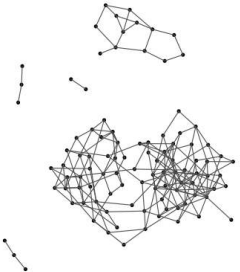
1. the unnormalized Spectral Clustering computes the first k eigenvectors sorted by the eigenvalues in ascending order $U = [u_1, \dots, u_k]$ of L ,
2. the Shi-normalized Spectral Clustering computes the first k eigenvectors sorted by the eigenvalues in ascending order $U = [u_1, \dots, u_k]$ of $D_W^{-1} L$,
3. the Ng-normalized Spectral Clustering computes the first k eigenvectors sorted by the eigenvalues in ascending order, $V = [v_1, \dots, v_k]$ of L_N and U is the V -matrix row-norm normalized,
4. the Absolute Eigenvalue Spectral Clustering computes the first k eigenvectors sorted by the *absolute value* of eigenvalues in descending order $U = [u_1, \dots, u_k]$ of $I - L_N$, see Rohe et al. (2011). In contrast to the other three variants, it allows the search for structurally homogeneous subsets.

Then the clusters are obtained by a k -means algorithm with k clusters on the n row vectors of matrix U . Each vertex is associated to a point in a k -Euclidean space. The coordinates of vertex i in this space are given by row i of the matrix U .

The toy example allows us to show the differences between the Absolute Eigenvalue Spectral Clustering and the Ng-normalized Spectral Clustering, (Figure 2.7). One can see that the first one detects the bipartite structure and the second one does not.

The tuning parameters of the Spectral Clustering algorithm are the choice of a specific method among the four (or more) possible ones, the number of latent variables and the number of clusters k .

Correspondence Analysis (CA). The CA developed by Hirschfeld (1935); Benzecri (1973) is a general method to analyze contingency tables. For undirected graphs, it can be described as a variant of Spectral Clustering considering the square of the normalized Laplacian, $L_{CA} = [D_W^{-1/2} W D_W^{-1/2}]^2$. Let k be the number of clusters. The CA clustering computes the first k eigenvectors sorted by the eigenvalues in descending order $V = [v_1, \dots, v_k]$ of L_{CA} and $U = D_W^{1/2} W V$. Note that the eigenvalues of L_{CA} are the square of the eigenvalues of $I - L_N$ with the



same associated eigenvectors. Therefore the k first eigenvectors of CA (sorted by the eigenvalues of L_{CA}) are the k first eigenvectors of $I - L_N$ sorted by the absolute values of the eigenvalues of $I - L_N$. As for Spectral Clustering, the clusters can be obtained by a k -means algorithm with k clusters on the n row vectors of matrix U . Therefore the Correspondence Analysis is equivalent to the Absolute Eigenvalue Spectral Clustering. This confirms the fact observed by Von Luxburg (2007) that many Spectral Clustering methods developed in different scientific communities are actually identical.

2.3.4 Edge-Betweenness

Synopsis	
Name	Edge Betweenness
Type of method	Algorithm
Type of graphs	Undirected Unweighted
Type of clusters	Communities
Summary	This method introduces a measure of the importance of a link to connect communities and it cuts edges with high values until the communities are disconnected from each other
Time complexity	No complexity is given by authors. <code>igraph</code> implementation (Csardi and Nepusz, 2006) complexity is $O(V E)$.

This algorithm, proposed by Girvan and Newman (2002), allows the search for communities. The main idea is to remove edges from the network until the communities are disconnected from each other. The edges to be removed are chosen as a function of a criterion called edge-betweenness. For illustrating the concept of betweenness, let us imagine that one should go from one "side" the network to the other by following edges. An edge with a high betweenness is an edge that is included in most paths between the two "sides" of the network. For instance, on the toy example, the edge with the highest betweenness is the edge between N_3 and N_8 in Figure 2.2.

More formally, the betweenness of one edge is equal to the number of shortest paths, using this edge, for all the pairs of vertices of the graph. The algorithm alternates the following steps:

1. the betweenness of all existing edges in the network is computed,
2. the edge with the highest betweenness is removed,
3. the betweenness of all edges affected by the removal is computed.

The method iterates as long as an edge remains. At the end of the algorithm, one obtains a classification tree showing the sequence of divisions of the network. Communities are obtained by truncating the classification tree.

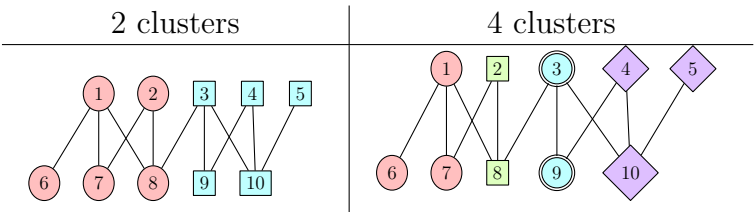


Figure 2.8: Clusters obtained with Edge-Betweenness with $k \in \{2, 4\}$

This algorithm needs to choose the level for stopping the classification tree, which is equivalent to choosing the number of groups. Applied to the toy example, it gives the communities of Figure 2.8. By definition, this algorithm cannot detect the structurally homogeneous clusters that are not communities.

2.4 Methods based on an optimization criterion

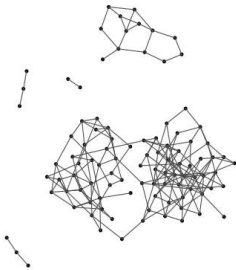
The following section presents the algorithmic methods that explicitly optimize a criterion.

2.4.1 Modularity criterion

Synopsis	
Name	Modularity
Type of method	Optimization
Type of graphs	Directed or not, weighted or not
Type of clusters	Communities
Summary	This method optimizes the Modularity which represents the quality of partition as the difference between the expectation of edges inside and outside communities.
Time complexity	The algorithm provided by Guimera is very time expensive and in many cases not usable in practice. An example of greedy implementation (Clauset et al., 2004) complexity is $O(E d \log(V))$, where d is the depth of the dendrogram describing the community structure.

The modularity, proposed by Newman and Girvan (2004), is a global quality measure of a partition.

The modularity measures the difference between the actual and expected within-community edges relative to a null model assuming a connectivity between vertices that is proportional to their degrees. Let $C = (C_1, \dots, C_k)$ be a partition of G .



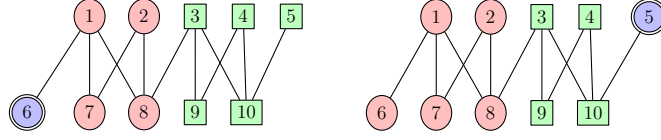


Figure 2.9: Clusters obtained by maximizing the Modularity (each of them have the same modularity)

There are two equivalent definitions of the modularity of C in the undirected graph G :

1. $\mathcal{M}_C = \sum_q (e_{qq} - a_q^2)$ with $e_{ql} = \frac{1}{2m} \sum_{ij} W_{ij} \delta_q(i) \delta_l(j)$ where m is the total number of edges, $\delta_q(i) = \mathbb{1}_{i \in C_q}$ is equal to one if i is in the class q and zero if not and $a_q = \sum_l e_{ql}$,
2. $\mathcal{M} = \frac{1}{2m} \sum_{ij} (W_{ij} - \frac{(d_W)_i (d_W)_j}{2m}) \delta(i, j)$ with $\delta(i, j) = \sum_q \delta_q(i) \delta_q(j)$ equal to one if i and j are in the same class.

For directed graphs the modularity is defined by: $\mathcal{M}_C = \frac{1}{2m} \sum_{ij} (W_{ij} - \frac{(d_W^o)_i (d_W^i)_j}{2m}) \delta(i, j)$.

The partition with the best (maximum) modularity is obtained using an optimization algorithm such as greedy algorithms or simulated annealing algorithms. Obtaining the best partition is NP-hard.

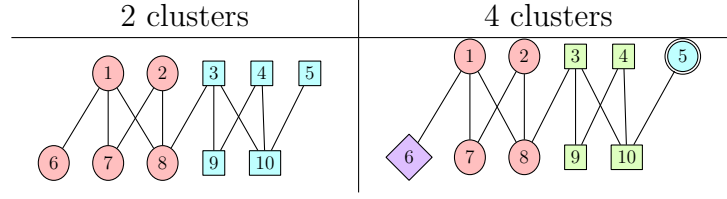
The optimization can be done conditionally to a fixed number of groups, or not.

Guimera et al. proposed the following algorithm. Optimization is done by a Simulated Annealing (SA), with levels of temperature decreasing exponentially. Three movements are possible:

1. individual movement of a vertex from one module to another
2. merging of two modules
3. splitting of one module into two, choice of modules being made by another SA at the level of the module

This algorithm does not need any proper tuning parameter, but there are optimization parameters in the simulated annealing. This algorithm is highly computationally intensive; therefore one may have to modify some optimization parameters in order to obtain a result with a reasonable time. This algorithm, applied on the toy example, gives the communities clusters of the Figure 2.9. By construction, this algorithm cannot detect the structurally homogeneous clusters that are not communities.

Since the Guimera algorithm is usable only for small graphs, greedy algorithms exist to optimize the modularity, see Clauset et al. (2004).

Figure 2.10: Clusters obtained with Cut cost with $k \in \{2, 4\}$

2.4.2 Cut

Synopsis	
Name	Cut criterion
Type of method	Optimization
Type of graphs	Undirected, unweighted
Type of clusters	Communities
Summary	This method minimizes the number of edges between communities by removing edges from the network until the <u>communities are disconnected</u> .
Time complexity	Depends on greedy implementation

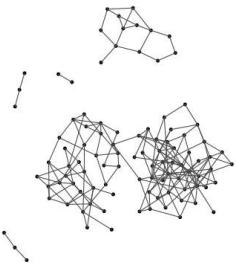
The idea (see Raj and Wiggins, 2010 for a recent reference) is to suppress some edges from G to obtain an unconnected partition of vertices with a minimum modification cost. The cut between two subsets (V_1, V_2) of V from the (V, E) graph is $\text{Cut}(V_1, V_2) = \sum_{v_1 \in V_1, v_2 \in V_2} W_{v_1, v_2}$. There are three cut criteria on a partition C , the cut $\text{Cut}(C)$, the ratio cut, $\text{rCut}(C)$ and the normalized Cut, $\text{nCut}(C)$:

1. $\text{Cut}(C) = \sum_{q < l} \text{Cut}(C_q, C_l) = \frac{1}{2} \sum_{q=1}^k \text{Cut}(C_q, V \setminus C_q)$.
2. $\text{rCut}(C) = \sum_{q=1}^k \frac{\text{Cut}(C_q, V \setminus C_q)}{\sum_{i,j \in V_q} W_{ij}}$.
3. $\text{nCut}(C) = \sum_{q=1}^k \frac{\text{Cut}(C_q, V \setminus C_q)}{\text{Cut}(C_q, V)}$.

The partition with the best (minimum) cut is obtained using an optimization algorithm such as heuristics, greedy algorithms or simulated annealing algorithms. Obtaining the best Cut partition is NP-hard for the three criteria. In practice only approximated methods can be used.

Applied to the toy example, these methods give the community clusters of the first column of Figure 2.10. By definition, these algorithms cannot detect the structurally homogeneous clusters that are not communities.

Tuning parameters are the choice of the cut criteria and the number of groups.



2.5 Methods developed with a model

Statisticians propose probabilistic models that are supposed to take into account the random variability in the data. These models are generative models in the sense that they mimic the real data generation. This section presents a synthetic summary of the more detailed review made in Daudin (2011).

All of these models use latent variables. These latent variables may be discrete and give directly the classification of the vertices such as in the Stochastic Block Model (SBM, Section 2.5.3). Alternative models such as the Model-Based Clustering for Social Networks (MBCSN, Section 2.5.1) and Random Dot Product Graph (RDPG, Section 2.5.2) use continuous latent variables; therefore a supplementary step of classification is necessary.

The above models assume that a vertex pertains to only one class, but there are alternative models such as the Continuous Stochastic Block Model (CSBM, section 2.5.4), which allows each vertex to pertain to several classes, these models are also known under the name of Grade of Membership (see Manton et al., 1994 and Erosheva, 2005).

2.5.1 Model-based clustering for social networks (MBCSN)

Synopsis	
Name	Model-based clustering for social networks (MBCSN)
Type of method	Model-based method
Type of graphs	Undirected and unweighted
Type of clusters	Communities
Summary	This method assumes that the graph is a realization of a generative model and infers its parameters. The model assumes that each vertex has a position in a continuous latent space, and linking behavior of each pair of vertex is determined by the distance between the vertices in the latent space.
Time complexity	Not known

The model Handcock et al. (2007) assumes that, conditionally to d -multidimensional latent variables z_i attached to the vertices and observed variables $x_{i,j}$ attached to the edges, the $W_{i,j}$ are independent and

$$P_{ij} = P(W_{ij} = 1) = \frac{e^{\beta_0 x_{i,j} - \beta_1 |z_i - z_j|}}{1 + e^{\beta_0 x_{i,j} - \beta_1 |z_i - z_j|}}$$

The probability of connection between vertices i and j is greater for vertices whose latent variable values are similar. The distribution of the latent variables is a mixture of k multivariate Gaussian distribution. The parameters of the mixture model and β_0 and β_1 are estimated with Bayesian or Maximum-Likelihood methods using MCMC and the values of the latent variables are predicted for each vertex. An R Package *latentnet* is available. Applied to the toy example with no covariate $X_{i,j}$ on the edges, $d = 2$ and $k = 2$, the package *latentnet* gives the community clusters of the top left corner of Figure 2.2. With $d = 4$ and $k = 4$, one obtains only two clusters that are the same ones as with $k = 2$ and $d = 2$.

Tuning parameters are d the dimension of the latent space, and k the number of distributions which is the number of wanted groups k .

2.5.2 Random Dot Product Graphs (RDPG)

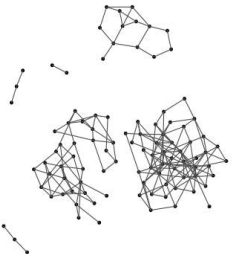
Synopsis	
Name	Random Dot Product Graphs (RDPG)
Type of method	Model-based method
Type of graphs	Undirected and unweighted
Type of clusters	Communities
Summary	This method assumes that the graph is the realization of a generative model and infers its parameters. The model assumes that the vertices lie in a continuous latent space and the linking behavior of each pair of vertices is determined by the position of the vertices in the latent space. Then a classification method such as the k -means algorithm classifies the vertices
Time complexity	Not known

The multidimensional scaling (MS) method, applied to the similarity matrix P , consists in positioning each vertex in a metric space of latent variables so that the similarity between vertices is approximately kept. The underlying model is $P = TT'$, where the (n, d) -matrix T contains the coordinates of the vertices in a d -dimensional metric space. The naive MS method is not well suited for modeling P , with two major drawbacks: TT' does not lie in $[0, 1]^{n^2}$ if $T \in \mathbb{R}^d$ and TT' is symmetric so it is not suited for the modeling of directed graphs.

The Random Dot Product Graph defined in Marchette and Priebe (2008) is

$$P_{ij} = f(t'_i t_j) \quad \text{with } t_i \in \mathbb{R}^d \quad \text{and} \quad f(x) \in [0, 1].$$

f is a simple threshold in Marchette and Priebe (2008): $f(x) = 0$ if $x < 0$, $f(x) = x$ if $0 \leq x \leq 1$ and $f(x) = 1$ if $x > 1$.



To get around the second drawback, the RDPG model is extended with two vectors for each vertex, an in-vector V and an out-vector U , such that the model becomes $P_{ij} = f(u'_i \cdot v_j)$. Another way to get around the symmetry of P , called DEDICOM, was proposed by Harshman (1978) and well described in Trendafilov (2002). This model uses only one vector for each vertex but inserts a non-symmetric (d, d) -matrix A in the dot product. The model is

$$X = TAT' + E$$

the matrix T is constrained by $T'T = I$ and T and A are obtained by minimizing $\|X - TAT'\|^2$. Several algorithms have been proposed to achieve this task (see Kiers et al., 1990).

Tuning parameters are d the dimension of the latent space, and the number of groups.

2.5.3 Stochastic Block Model (SBM)

Synopsis	
Name	Stochastic Block Model
Type of method	Model-based method (SBM)
Type of graphs	Directed or not, weighted or not
Type of clusters	Structurally Homogeneous Subsets
Summary	SBM is a mixture model where each vertex is supposed to pertain to only one structurally homogeneous subset. The assignment of vertices to subsets is done by inferring the model parameters.
Time complexity	Each iteration of VEM is $O(V ^2)$. The number of iterations depends on the number of nodes. For sparse graphs the inference can be made in $O(E)$.

The first probabilistic model which explicitly integrates heterogeneity in the network topology, the Stochastic block model, has been proposed by mathematicians and statisticians working in the domain of social science such as White et al. (1976), Holland et al. (1983) and Snijders and Nowicki (1997). These authors have developed this model in concordance with the notion of Structural Equivalence in a graph. Therefore the SBM is built to detect structurally homogeneous subsets.

The intuitive idea developed by White et al. (1976) (see also Arabie et al., 1978 and Winship and Mandel, 1983) is that the vertices of a graph may be classified into groups. Two vertices of the same group are connected in the same way to the other vertices. Therefore the adjacency matrix, sorted by the number of the group

in row and column, appears to be partitioned in homogeneous blocks composed of 0 or alternatively of 1.

BLOCKER and CONCOR (White et al., 1976) were historically the first algorithms for clustering vertices. More recently Snijders and Nowicki (1997) used the Markov Chain Monte Carlo method for estimating the parameters.

The modern version of the Stochastic Block Model is a mixture model, using discrete latent variables giving the assignment of each vertex to a group, where each vertex is supposed to pertain to only one group. The model for a binary directed network is the following:

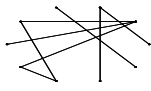
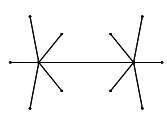


$i = 1, n$ vertices pertains to $q = 1, k$ classes. The class of each vertex is defined by a hidden discrete latent variable $Z_i = q$, if vertex i pertains to class q , with Probability Distribution Function (pdf) given by $Z_i \sim \mathcal{M}(1, \alpha_1, \alpha_2, \dots, \alpha_k)$ and \mathcal{M} is a multinomial pdf.

$W_{ij} = 1$ if there is an edge from vertex i to vertex j and 0 if there is no edge, and conditionally to Z , W_{ij} are independent Bernoulli random variables with

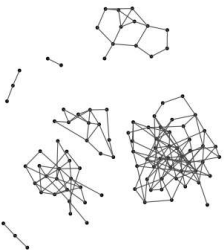
$$P(W_{ij} = 1 / Z_i = q, Z_j = l) = \pi_{ql}$$

Table 2.1 shows that the model is very flexible for it is able to modelize hubs, communities or hierarchical structures.

Table 2.1: Examples of SBM

Description	Graph	k	π
Erdos		1	p
Hubs		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
Communities		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$
Hierarchical		5	$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

Sinkkonen et al. (2008a) propose an alternative mixture model which allows



the analysis of large graphs. This model uses latent variables which operate not on the vertex level but on the edge level.

Daudin et al. (2008) used a variational method of estimation allowing the analysis of network up to 3000 vertices. Hofman and Wiggins (2008) use a Bayesian variational approach for a particular case of SBM with two parameters, $\pi_{q,q} = a$ and $\pi_{q,l} = b$ for $q \neq l$. Mariadassou et al. (2010) have extended the variational estimation method to weighted networks with probability distribution of the weights pertaining to the exponential family. Identifiability and consistency results have been obtained by Celisse et al. (2011), Rohe et al. (2011), Ambroise and Matias (2011), Bickel and Chen (2010) and Choi et al. (2012). A frequent criticism against SBM is that the number of clusters is not a fixed number for real-life networks and large networks have more clusters than small ones. However this point has been taken into account and some consistency results are obtained in an asymptotic framework that allows the number of clusters to increase with the number of vertices.

There is a C-Package *Mixnet* that uses the variational method. Applying this package to the toy example allows the retrieval of the subsets of the column "Structurally homogeneous subsets" in Figure 2.3.

For sparse graphs, Decelle et al. (2011) use belief propagation to infer parameters of the stochastic block model, and achieve a complexity linear to the number of edges.

There is only one tuning parameter, the number of clusters, k . It is possible to infer it from the data, see Daudin et al. (2008) and Hofman and Wiggins (2008), using a penalized or a Bayesian criterion. Sinkkonen et al. (2008b) presents Bayesian non-parametric methods which address the criticism of a fixed number of clusters.

2.5.4 Continuous Stochastic Block Model (CSBM)

Synopsis	
Name	Continuous Stochastic Block Model (CSBM)
Type of method	Model based method
Type of graphs	Directed or not, weighted or not
Type of clusters	Structurally Homogeneous Subsets
Summary	This method assumes that the graph is a realization of a generative model and infers its parameters. The model assumes that each vertex is a mixture of virtual vertices whose connectivity properties are known.
Time complexity	Not known

The Stochastic Block Model model can be written under the form

$$P_{ij} = P(W_{ij} = 1) = \sum_{q,l=1,k} z_{iq} a_{ql} z_{jl}$$

where $z_{iq} = 1$ if the vertex i is in class q , and 0 if not, which gives the matrix relation $P = ZAZ'$,

with Z the (n, k) -matrix containing the z_{iq} . If we allow z_{iq} to be in $[0, 1]$ (and not in $\{0, 1\}$) then each vertex does not pertain to only one group, which bears more flexibility to the model. This leads to the CSBM (Continuous-SBM) developed in Daudin et al. (2010).

This model displays the vertices in a continuous space. Therefore a supplementary step of clustering must be applied for obtaining groups. There is a MATLAB package *C-Mixnet* for this model. Applying this package, followed by a k -means clustering, to the toy example allows to retrieve the subsets of the column "Structurally homogeneous subsets" in Figure 2.3.

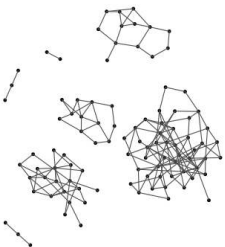
There is only one tuning parameter, the number of clusters, k . However it is possible to infer it from the data, see Daudin et al. (2008).

Other models allow each vertex to pertain to several classes such as the Mixed Membership Stochastic Block Model (Airoldi et al., 2008) and the Overlapping Stochastic Block Model (Latouche et al., 2011).

2.6 Application to the Zachary's karate club

The Karate Club network, introduced by Zachary (1977) is one of the most famous data set from the social science literature. The members of a karate club at a US University in the 1970s are the vertices of the network. Edges represent friendship relations between the members. This example is highly interesting because shortly after the observation, the club was split into two components. The resulting two groups are represented in Figure 2.11. Therefore one may easily compare the groups obtained by any clustering method to the exact split. A close examination of the graph shows that some members are highly connected to other members. These members, such as numbers 1 for the group on the left side of Figure 2.11 and 33, 34 for the other group have probably played a leading role in the split. Therefore one may consider that there are four groups: the leadings members and the satellite members of each group. Therefore we have decided to show the results of each method with 2 and 4 groups.

The results of 8 methods (EB, Pons-Latapy, Modularity, MCL, unnormalized SC, Absolute SV, SBM and CSBM) are given in Figures 2.12 to 2.18 presenting the graph with colored vertices. Each color corresponds to a cluster obtained by



the method. For all the methods excepted MCL we give the result with 2 groups (on the left side of the figure) and with 4 groups (on the right side).

In the case of 2 groups there are 3 classes of results:

1. The real split with one (MCL) or two (EB) errors of classification.
2. A somewhat isolated subgroup from the left group (members 5, 6, 11, 16 and 17), connected to no other member excepted member 1 on one side and all the other members on the other side (Pons-Latapy, SC unnormalized).
3. A group composed of high degree vertices (members 1, 2, 3, 33, 34) and a group composed of low degree vertices (SC absolute, SBM, CSBM).

In the case of 4 groups there are 3 classes of results:

1. EB and Modularity split the left group in one subgroup composed of the isolated subgroup (members 5, 6, 11, 16 and 17), and split the right group in two subgroups,
2. SBM and CSBM give the leadings members and the satellite members of each group, with one error of classification (member 9),
3. SC absolute gives groups separated by connectivity behavior, in leading behavior classes and satellite behavior classes.

This simple comparison from one small data set cannot be taken as a benchmark, but is only an illustration of two points:

1. the results may be very different from one method to another.
2. the result obtained with a given method are coherent with its objective. EB, Modularity and MCL (with $\frac{1}{10}$ -weighted self-loops) find communities and SC(absolute), SBM and CSBM find structurally homogeneous subsets.

2.7 Conclusion

There are mathematical relations between some of the methods:

1. The Markov Chain Clustering and the Spectral Clustering are two ways to study the behavior of the Markov Chain associated with a random walk along the graph. This behavior is controlled by the transition matrix and the asymptotic behavior of its power. The power of a matrix is related to its spectral decomposition which is studied in the Spectral Clustering method. Therefore the two methods are linked even if they do not necessarily give exactly the same results. Von Luxburg (2007) gives a detailed analysis of the connections between Spectral Clustering and Random Walks.

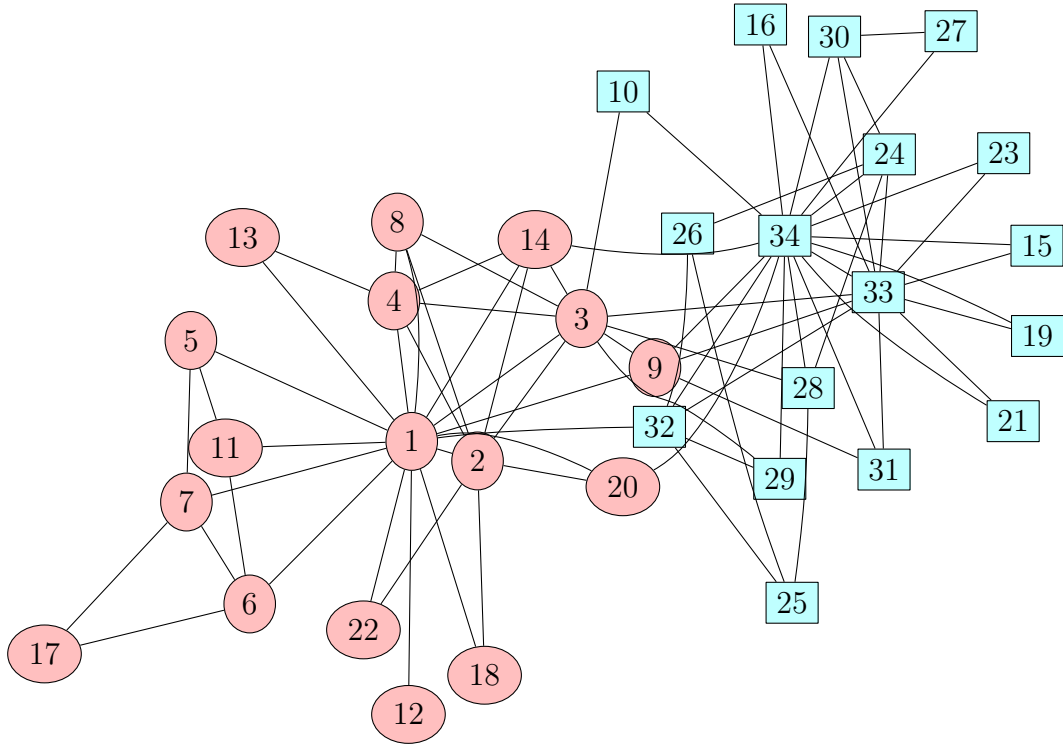


Figure 2.11: Zachary's Karate Club network. Colors show real fission of the club

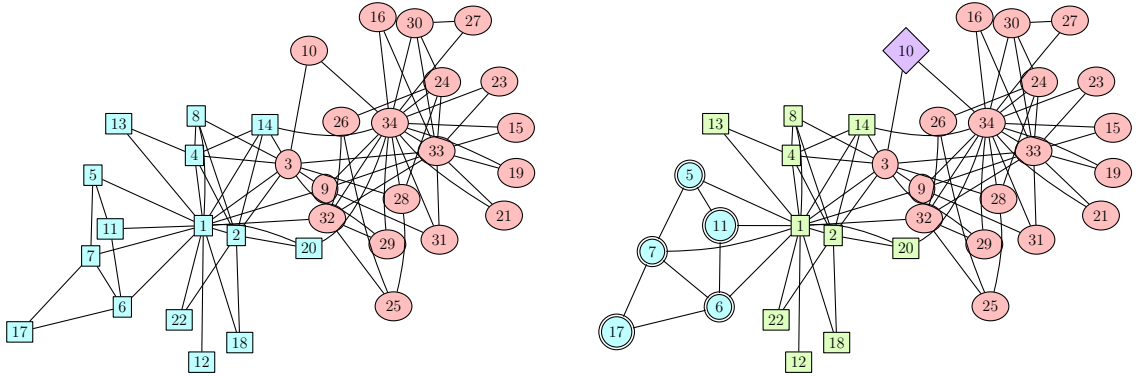
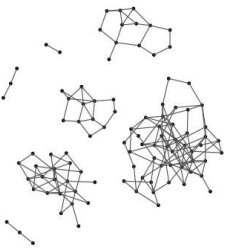


Figure 2.12: Edge-betweenness for 2 and 4 groups method applied to the Zachary's Karate Club network

2. The Spectral Clustering and the cut-methods are also linked by the following relations: let x be a vector with $x_i = 1$ if $i \in V_1$ and $x_i = -1$ if $i \in V_2$. Then $\text{Cut}(V_1, V_2) = \frac{1}{2}x' Lx$. Von Luxburg (2007) also gives a detailed analysis of the connections between spectral clustering and Cut criteria.
3. Rohe et al. (2011) proved that for undirected graphs, the Absolute Eigen-



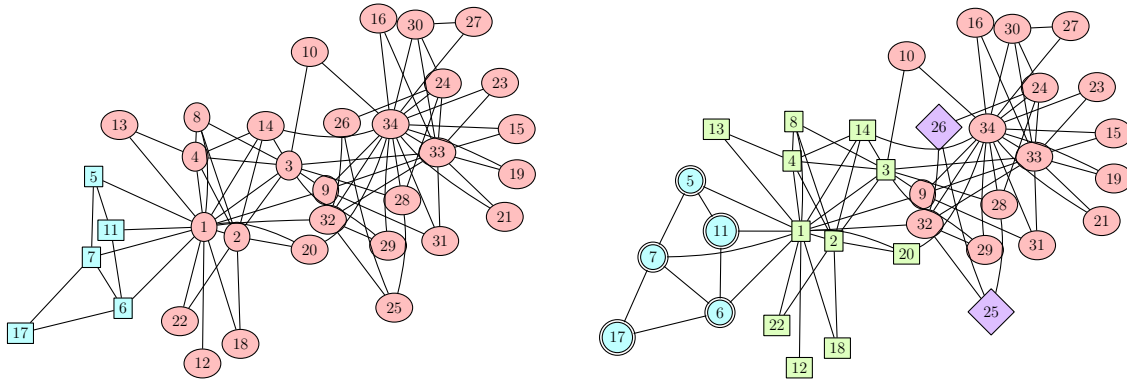


Figure 2.13: Hierarchical clustering method to Pons-Latapy distance for 2 and 4 groups applied to the Zachary's Karate Club network

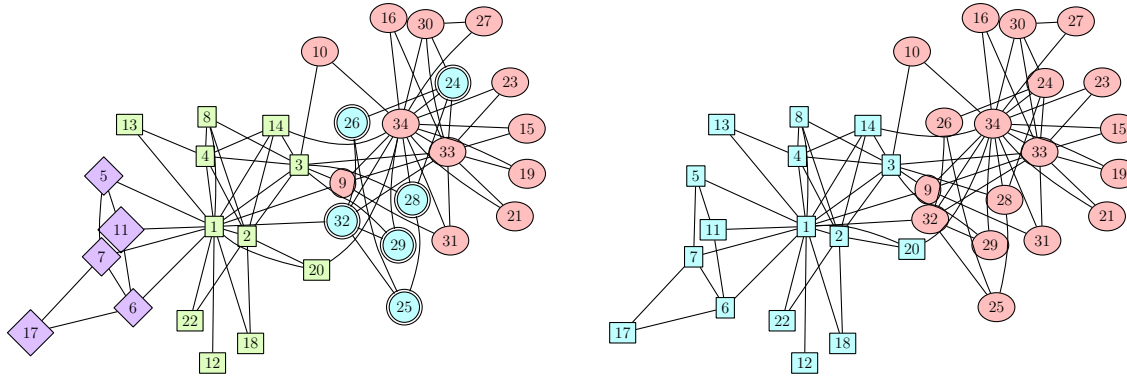


Figure 2.14: Modularity method without choice of number of groups (*left*) and MCL method without choice of number of groups (*right*) applied to the Zachary's Karate Club network

value Spectral Clustering is asymptotically able to approximately retrieve the clusters if the data are generated by SBM. Some work is in progress from the same authors to extend these results to directed graphs.

We can summarize the methods in Table 2.2. They are sorted by ascending level of generality. The first three methods cannot retrieve structural homogeneous subsets of vertices that are not communities. These three methods are devoted to one objective, the detection of communities and it seems difficult to generalize them to a more general objective. On the other hand the SBM model, which has been built around the concept of structural equivalence, is more general for it can detect every type of structurally homogeneous subset. Spectral Clustering and Random Walk methods have been used for a long time to detect communities. However these methods may be customized to be able to detect structurally

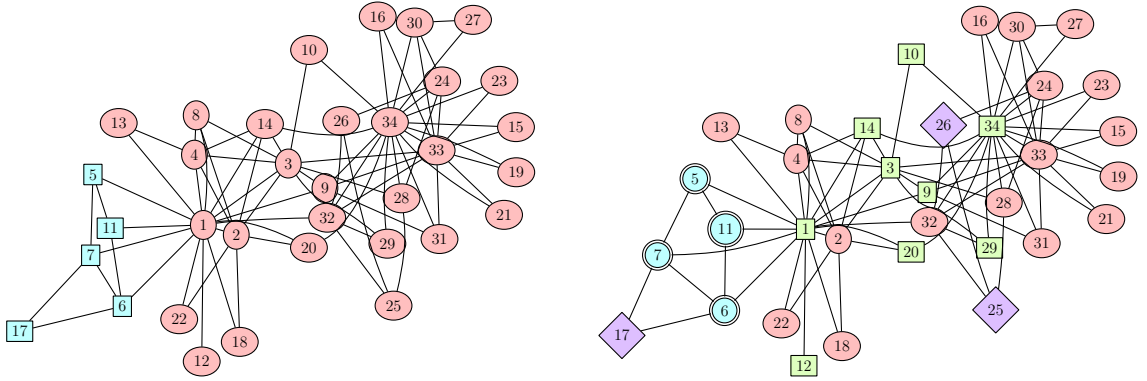


Figure 2.15: Spectral clustering (unweighted variant) method for 2 and 4 groups applied to the Zachary's Karate Club network

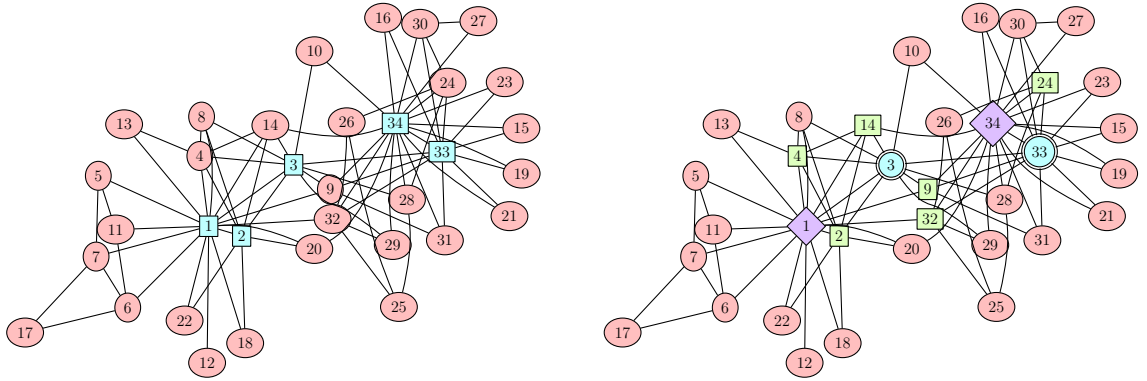
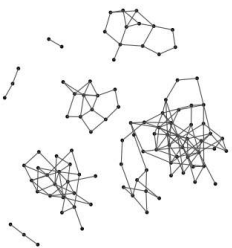


Figure 2.16: Spectral clustering (Absolute Eigenvalues variant) method for 2 and 4 groups applied to the Zachary's Karate Club network

homogeneous subsets as well (Rohe et al., 2011). The trick for random walk methods consists in decreasing the value of self-loops. The modification of the usual Spectral Clustering consists in keeping not only the eigenvectors corresponding to the higher eigenvalues but also the ones corresponding to the negative eigenvalues that have a high absolute value, see Rohe et al. (2011). It is quite surprising that this new method is equivalent to a very old one, Correspondence Analysis.

Comparing methods to relevant datasets of networks is not a trivial task, many points must be considered. To have a reference partition, simulated networks can be a solution. However, the simulation method must be carefully chosen not to advantage a subset of methods, and in particular, must be different from generative models used by some methods. On the other hand, real world networks, in general case, do not provide any reference partition. A comparison of methods in a particular case of simulated bipartite networks is under work.



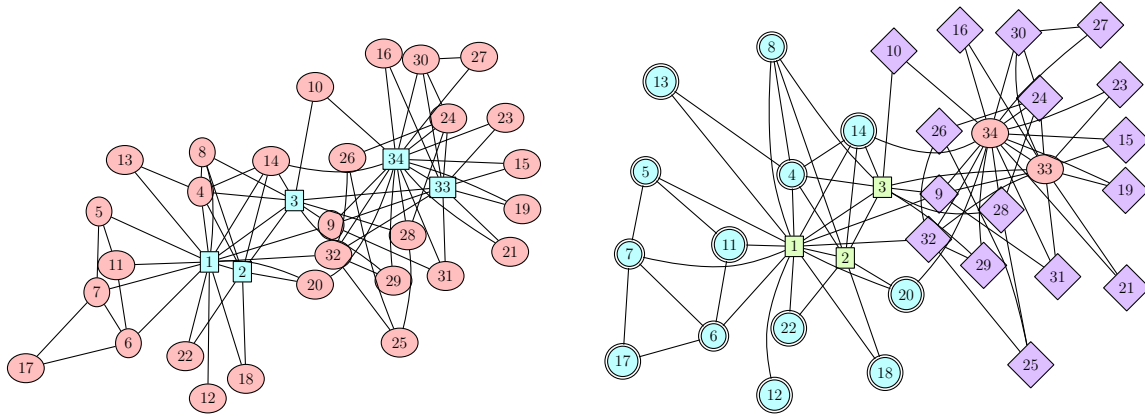


Figure 2.17: Stochastic Block Model method applied for 2 and 4 groups applied to the Zachary's Karate Club network

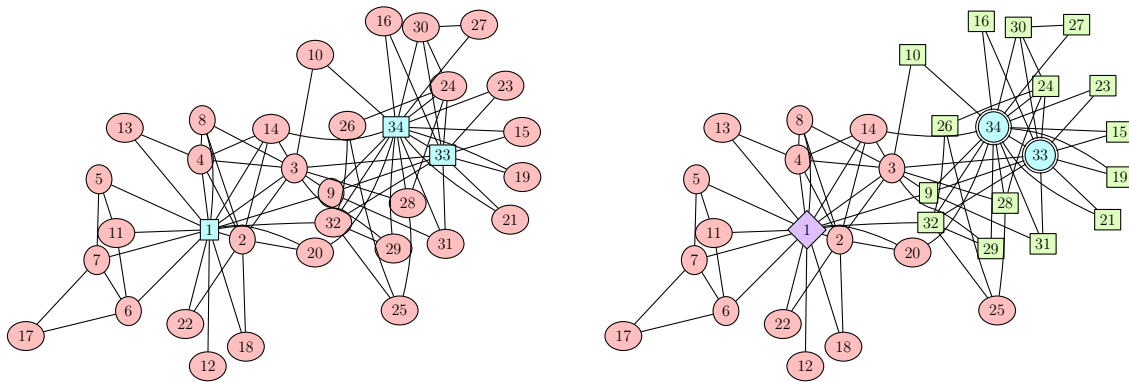


Figure 2.18: Continuous Stochastic Block Model method for 2 and 4 groups applied to the Zachary's Karate Club network

Table 2.2: Summary of the clustering methods

Method	Type of Method ^a	Directed ^b	Weighted ^c	Goal ^d	Tuning parameters
Edge-Betweenness	A	N	N	C	none
Cut	O	N	Y	C	Criteria
Modularity	O	Y	Y	C	none
Spectral Clustering	A	N	Y	C or SHS	method, k^e
Hierarchical Clustering	A	N	Y	C or SHS	method
Markov Chain Clustering	A	Y	Y	SHS	r^f, e^f, Δ^g
Pons-Latapy	A	N	Y	SHS	k^e, Δ^g
SBM	M	Y	Y	SHS	k^e or none
CSBM	M	Y	N	SHS	k^e or none
MBCSN ^h	M	N	N	C	d^i and k^e
RDPG	M	N	N	C	d^i

a. A for algorithm, O for optimization, M for probabilistic model

b. Y if the method can be applied to a directed graph, N otherwise

c. Y if the method can be applied to a weighted graph, N otherwise

d. C for Community research algorithm, SHS for Structural homogeneous subset research algorithm

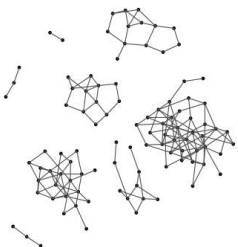
e. k is the number of groups

f. e and r are the importance of transition and inflation step, $\frac{e}{r}$ control the number of groups

g. Weight of self-loops added for ergodicity

h. Model-based clustering for social network

i. d is the dimension of the latent space



Bibliography

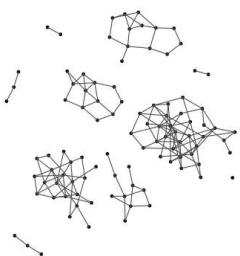
- E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.
- C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011.
- P. Arabie, S. Boorman, and P. Levitt. Constructing blockmodels: How and why. *Journal of mathematical psychology*, 1978.
- J. Benzecri. *L Analyse des Donnees. Volume II. L Analyse des Correspondances*. Dunod, 1973.
- P. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *PNAS*, pages 1–6, 2010.
- S. Brohee and J. Van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7(1):488, 2006. ISSN 1471-2105.
- R. Burt. Cohesion versus structural equivalence as a basis for network subgroups. *Sociological Methods & Research*, 7(2):189–212, 1978.
- A. Celisse, J. Daudin, and L. Pierre. Consistency of maximum likelihood and variational estimators in mixture models for random graphs. *Electronic Journal of Statistics* 2012, 6(0):1847–1899, 2011.
- D. Choi, P. Wolfe, and E. Airoldi. Stochastic blockmodels with growing number of classes. *Biometrika*, 99(2):273–284, 2012.
- A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

- G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL <http://igraph.sf.net>.
- J.-J. Daudin. A review of statistical models for clustering networks with an application to a PPI network. *JFSDS*, 2011.
- J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008. ISSN 0960-3174.
- J.-J. Daudin, L. Pierre, and C. Vacher. Model for Heterogeneous Random Networks Using Continuous Latent Variables and an Application to a Tree-Fungus Network. *Biometrics*, 66(4):1043–51, 2010.
- A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.
- W. E. Donath and A. J. Hoffman. Lower Bounds for the Partitioning of Graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.
- E. Erosheva. Comparing latent structures of the grade of membership, Rasch and latent class model. *Psychometrika*, 70(4):619–628, 2005.
- S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010. ISSN 0370-1573. doi: 10.1016/j.physrep.2009.11.002. URL <http://www.sciencedirect.com/science/article/pii/S0370157309002841>.
- M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821, 2002.
- R. Guimera, D. Stouffer, M. Sales-Pardo, E. Leicht, M. Newman, and L. Nunes Amaral. Origin of compartmentalization in food webs. *Ecology*. doi: 10.1890/09-1175.1. URL <http://www.esajournals.org/doi/abs/10.1890/09-1175.1>.
- M. S. Handcock, A. E. Raftery, and J. Tantrum. Model-Based Clustering for Social Networks. *Journal of the Royal Statistical Society: Series A*, 170(2):301–354, 2007.
- R. Harshman. Models for analysis of asymmetrical relationships among N objects or stimuli. In *First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology*, McMaster University, Hamilton, Ontario, August, 1978.



- J. Hartigan. *Clustering Algorithms*. Wiley, 1975.
- H. Hirschfeld. A connection between correlation and contingency. *Proc. Cambridge Philosophical Society*, 31:520–524, 1935.
- J. M. Hofman and C. H. Wiggins. Bayesian Approach to Network Modularity. *Phys. Rev. Lett.*, 100:258701, Jun 2008. doi: 10.1103/PhysRevLett.100.258701. URL <http://link.aps.org/doi/10.1103/PhysRevLett.100.258701>.
- P. Holland, K. Laskey, and K. Leinhardt. Stochastic blockmodels: some first steps. *Social Networks*, 5:109–137, 1983.
- H. Kiers, J. ten Berge, Y. Takane, and J. de Leeuw. A generalization of Takane’s algorithm for DEDICOM. *Psychometrika*, 55(1):151–158, 1990. ISSN 0033-3123.
- P. Latouche, E. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, 2011.
- F. Lorrain and H. White. Structural Equivalence of Individuals in Social Networks. *Journal of Mathematical Sociology*, 1:49–80, 1971.
- K. Manton, M. Woodbury, and H. Tolley. *Statistical Applications Using Fuzzy Sets*. 1994.
- D. Marchette and C. Priebe. Predicting unobserved links in incompletely observed networks. *Computational Statistics & Data Analysis*, 52(3):1373–1386, 2008. ISSN 0167-9473.
- M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *Ann Appl Stat*, 4:715–742, 2010.
- M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- F. Picard, V. Miele, J.-J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics*, 10, 2009.
- P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218, 2006.
- A. Raj and C. H. Wiggins. An Information-Theoretic Derivation of Min-Cut Based Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:988–995, 2010. doi: 10.1109/TPAMI.2009.124.

- K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional Stochastic Block Model. *The Annals of Statistics*, 39(4):1878–1915, 2011.
- J. Sinkkonen, J. Aukia, and S. Kaski. Inferring vertex properties from topology in large networks. *arXiv:0803.1628v1 [stat.ML]*, 2008a.
- J. Sinkkonen, J. Aukia, and S. Kaski. Component models for large networks. *arXiv preprint arXiv:0803.1628*, 2008b.
- T. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997. ISSN 0176-4268.
- N. Trendafilov. GIPSCAL revisited. A projected gradient approach. *Statistics and Computing*, 12(2):135–145, 2002. ISSN 0960-3174.
- S. Van Dongen. Graph clustering by flow simulation. *University of Utrecht*, 275, 2000.
- U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. ISSN 0960-3174.
- H. C. White, S. A. Boorman, and R. L. Breiger. Social Structure From Multiple Networks. *American Journal of Sociology*, 81:730–780, 1976.
- C. Winship and M. Mandel. Roles and positions: A critique and extension of the blockmodeling approach. *Sociological methodology*, 1983.
- W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.



Chapter 3

Graph clustering methods differ in their ability to detect patterns in bipartite ecological networks

Abstract

Network ecology has been an extraordinarily fertile field of research during the last 15 years. Simultaneously, a number of graph clustering methods have been developed by physicians, statisticians and computer scientists. However, only a couple methods are currently used by researchers in ecology. Here we compared several graph clustering methods in terms of their ability to retrieve species groups in ecological bipartite networks. The methods compared were the edge-betweenness algorithm, one method of modularity maximization, two spectral clustering methods, two Markov chain clustering methods and the stochastic block model. They were applied to the weighted and binarized version of 6400 bipartite networks, simulated with ecologically-relevant parameters. Our results showed that the edge-betweenness algorithm with modularity criterion for selecting group number is a good method for retrieving sub-groups of highly interacting species in binary bipartite networks. The stochastic block model gave very good results in the case of weighted bipartite networks, but it was very time consuming. In the case where thousands of weighted networks should be analyzed in a reasonable amount of time, an algorithmic efficient version of modularity maximization appears as a good alternative.

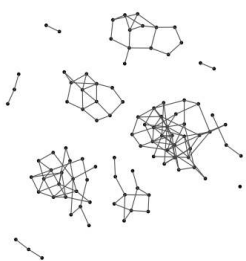
Keywords: ecological network, community, compartmentalization, nestedness, edge-betweenness, modularity maximization, spectral clustering, Markov chain clustering, stochastic block model

3.1 Introduction

Interactions between species form complex networks, called ecological networks. Most ecological networks contain several sub-groups of species, distinguishable by the higher density or the higher strength of interactions. These sub-groups of species have received various names in the ecological literature (reviewed by Dormann and Strauss, 2013), including compartment, module, cohesive group or community. Detecting these sub-groups of interacting species is important for ecologists for three reasons. First, it provides a simplified picture of the network (Allesina and Pascual, 2009). Second, it reveals the processes underlying the assembly of the network (e.g. Vacher et al., 2008b; Rezende et al., 2009; Krasnov et al., 2012). Third, it has implications for the functioning (especially, the stability) of the network (e.g. Fontaine et al., 2011; Stouffer and Bascompte, 2011).

Since the seminal article by Olesen et al. (2007), modularity maximization (Newman and Girvan, 2004) with a simulated annealing optimization approach (Guimera and Amaral, 2005) is the most widely used method for detecting such sub-groups in ecological networks. Other methods (reviewed by Fortunato, 2010; Leger et al., 2013) have been neglected so far by ecologists. These methods originate from various fields of research (physics, mathematics, statistics, computer science) and are of two kinds (Allesina and Pascual, 2009; Leger et al., 2013): those that detect sets of highly interacting nodes (hereafter called *communities*) and those that detect sets of nodes that have similar interaction patterns (hereafter called *structural homogeneous subsets* (SHS) as in Leger et al., 2013). Although the two kinds of methods correspond to different lines of research (Allesina and Pascual, 2009; Leger et al., 2013), the difference between both is not so clear-cut because nodes within a community tend to have similar interaction patterns. Both kinds of methods may thus be used to detect sets of highly interacting species in ecological networks.

Our aim here is to compare the ability of several methods of graph clustering, including modularity maximization, to detect sub-groups of highly interacting species in ecological networks. Although it would be relevant to merge different types of interactions within the same ecological network (Fontaine et al., 2011), the ecological literature is traditionally divided between the study of food webs and the study of bipartite networks (such as plant-pollinator or host-parasite networks). Here we focused on the latter and simulated several thousands of bipartite weighted ecological networks, with known sub-groups. We then assessed the ability of each clustering method to retrieve the sub-groups, depending on the properties of the sub-groups and several other properties of bipartite networks (total number of species, ratio between the number of plant (or host) species and the number of pollinator (or parasite) species, total number of interactions, mean weight of the interactions, degree of nestedness). The simulated networks were then binarized



and the analyses were performed again, in order to investigate whether the accuracy of the clustering methods changes with binarization and whether the ranking of clustering methods differs between weighted and binary networks.

3.2 Materials and methods

3.2.1 Graph clustering methods

Seven methods, chosen according to their popularity in ecology but also in bioinformatics or social science, were compared in this study. They are briefly described below and more details are available in Leger et al. (2013).

Modularity maximization (Mod) is a method proposed by Newman and Girvan (2004) to detect communities. The modularity measures the mean difference between the actual and expected within-community edges under a null model assuming a connectivity between vertices that is proportional to their degrees. Several algorithms have been proposed to obtain the maximum (Clauset et al., 2004). The algorithm used in our study is the one proposed by Newman (2006), which is based on successive splits by spectral analyses of a modularity matrix.

The edge-betweenness (EB) algorithm has been proposed by Girvan and Newman (2002) to detect communities. The main idea is to remove edges from the network until the communities are disconnected from each other. The edges to be removed are chosen using a criterion called edge-betweenness.

The MCL algorithm (MCL) (Van Dongen, 2000) was initially developed to detect communities. The method builds clusters considering the behavior of a random walk on the graph. Vertices with the same random walk behavior are classified within the same cluster. The method requires that the network contains self-loops, so self-loops are added if they are not in the raw data. The standard weight of self-loops is equal to 1.

Leger et al. (2013) proposed to reduce the weight of self-loops to *e.g.* $\frac{1}{10}$. They showed that in that case, the algorithm (MCL- $\frac{1}{10}$) tends to detect SHS.

Spectral clustering methods, first proposed by Donath and Hoffman (1973), are based on the Laplacian matrix of the graph. They determine the composition of the k clusters by applying a k-means algorithm to the nodes contained in a vector space generated by the k eigenvectors associated with the k lowest eigenvalues of the Laplacian matrix of the graph. There are several variants described in Von Luxburg (2007). In this study, we used two variants: the Ng-normalized spectral clustering (NSC) which was developed to find communities and the Absolute Eigenvalues Spectral Clustering (ASC) (see Rohe et al. 2011) which was developed to detect SHS.

The Stochastic Block Model (SBM) (White et al., 1976; Mariadassou et al.,

2010) is a probabilistic model which explicitly integrates heterogeneity in the network topology. This model was developed to detect SHS structure. According to the model, two nodes belonging to the same group have the same probability of connection with a third node. This probability depends only on the group membership of the three nodes.

3.2.2 Simulation of ecological bipartite networks

In order to compare the efficiency of the graph clustering methods, we applied them to simulated networks with known communities. The simulation method was designed according to the two following requirements:

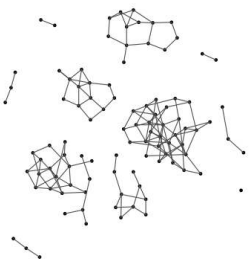
First, the simulation model should create networks similar to real ecological ones. To define the value of the parameters of the simulation model, we thus analyzed the structure of 23 unweighted and 24 weighted networks taken from the Interaction Web Database (Table 3.1). The ecological relevance of the simulated networks was then assessed by comparing simulated networks and real ecological networks for five topological properties: the cumulative distribution of degrees for each level of the network (Fig. 3.1 and 3.2), the frequency distribution of dependence for each level of the network (Fig. 3.3 and 3.4) and the frequency distribution of asymmetry values of dependences as defined in Bascompte et al. (2006) (Fig. 3.5). The distributions in simulated networks were similar to those observed by Jordano et al. (2003) and Bascompte et al. (2006) for real ecological networks.

Second, the simulation model should not favour one graph clustering method over others. We therefore chose simulation model developed by Thébault and Fontaine (2010), because it is not linked with any clustering method. We adapted it in order to simulate weighted networks.

The simulation of each weighted network was done in three steps. First, we defined the total number of nodes n_1 on the higher network level (usually corresponding to parasite species, pollinator species or herbivore species), the total number of nodes n_2 on the basal network level (usually corresponding to host species or plant species), the total number of links n_l , the total weight of all links n_w , the number of groups g , the compartmentalization index p_{comp} and the nest-edeness index p_{nest} .

Second, the attractivity of each node was randomly defined using a power law probability distribution with power equal to 2, and nodes were assigned to the g groups randomly with the same probability for each group. The groups were not constrained to have exactly the same size. Most of the groups contained both types of nodes (higher or basal level), but in some cases with n_2 low and g high, some groups contained only one type of node.

Third, the network links were simulated by using the following algorithm:



Dataset	Habitat type	Location	Data type	# of species in the lower trophic level	# of species in the higher trophic level
Ollerton et al. (2007)	Coral reefs	East coast of Africa and the Red Sea through the Indian Ocean to the western Pacific, and from southeastern Australia to the latitude of Tokyo	binary	10	26
Joern (1979)	Arid grasslands	Marathon, Trans-Pecos, Texas, USA	binary	54	24
Joern (1979)	Arid grasslands	Altuda, Trans-Pecos, Texas, USA	binary	52	22
Leather (1991)	Whole country	Finland	binary	5	64
Leather (1991)	Whole country	Finland	binary	6	88
Arroyo et al. (1982)	Andean scrub	Chile	binary	87	98
Arroyo et al. (1982)	Andean scrub	Chile	binary	43	62
Arroyo et al. (1982)	Andean scrub	Chile	binary	41	28
Clements and Long (1923)	Montane forest and grassland	USA	binary	96	276
Dupont et al. (2003)	High-altitude desert	Tenerife, Canary Islands	binary	11	38
Hocking (1968)	Arctic community	Canada	binary	29	86
McMullen (1993)	Multiple communities	Galápagos Islands	binary	106	54
Medan et al. (2002)	Xeric scrub	Laguna Diamante, Mendoza, Argentina	binary	21	45
Medan et al. (2002)	Woody riverine vegetation and xeric scrub	Río Blanco, Mendoza, Argentina	binary	23	72
RAMÍREZ and BRITO (1992)	Palm swamp community	Venezuela	binary	33	53
Robertson (1929)	Agricultural area dominated by crops, with some natural forest and pasture	USA	binary	456	1429
Barrett and Helenurm (1987)	Boreal forest	Canada	individuals caught	12	102
Blüthgen and Fiedler (2004)	rainforest	Australia	no. visits	51	41
Davidson et al. (1989)	rainforest	Peru	no. visits	8	18
Davidson and Fisher (1991)	tropical forest	Peru	no. visits	6	4
Elberling and Olesen (1999)	Alpine subarctic community	Sweden	no. visits	23	118
Fonseca and Ganade (1996)	Amazon rainforest	Brazil	no. visits	16	25
INOUE and PYKE (1988)	Montane forest	Australia	individuals caught	42	91
Mosquin and Martin (1967)	Arctic community	Canada	individuals caught	11	18
Motten (1986)	Deciduous forest	USA	no. visits	13	44
Olesen et al. (2002)	Coastal forest	Mauritius Island	no. visits	14	13
Olesen et al. (2002)	Rocky cliff and open herb community	Azores Islands	no. visits	10	12
Ollerton et al. (2003)	Upland grassland	KwaZulu-Natal region, South Africa	individuals caught	9	56
Schemske et al. (1978)	Maple-oak woodland	USA	no. visits	7	32
Small (1976)	Peat bog	Canada	individuals caught	13	34
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	10	29
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	9	33
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	9	27
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	10	29
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	8	35
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	8	26
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	7	24
Vázquez (2002)	Evergreen montane forest	Argentina	no. visits	8	27

Table 3.1: Table 3.1: List of the real antagonistic and mutualistic bipartite networks used in this study. Networks were extracted from the interaction web database (<http://www.nceas.ucsb.edu/interactionweb/>)).

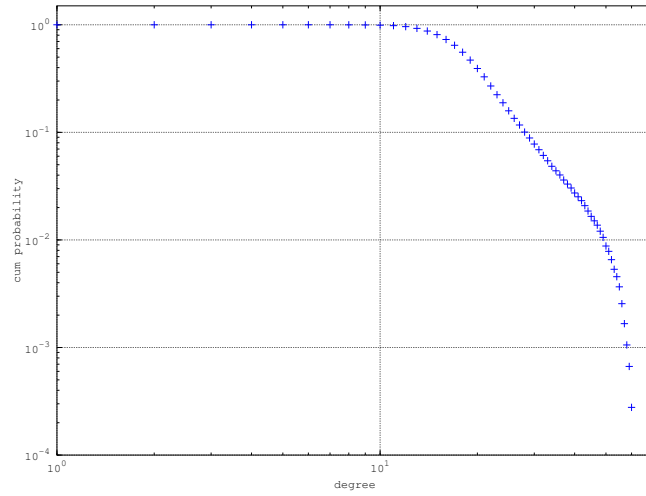


Figure 3.1: Complementary cumulative distribution of the degrees of the antagonist species for the simulated networks at the central point, on a log-scale. The linear trend is typical of the scale free distribution observed by Jordano et al. (2003) for real ecological networks.

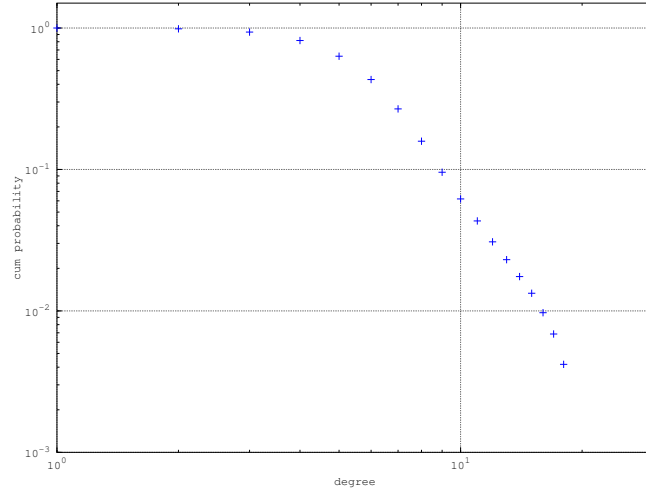
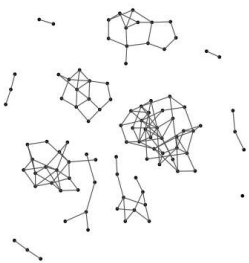


Figure 3.2: Complementary cumulative distribution of the degrees of the host species for the simulated networks at the central point, on a log-scale. The linear trend is typical of the scale free distribution observed by Jordano et al. (2003) for real ecological networks.



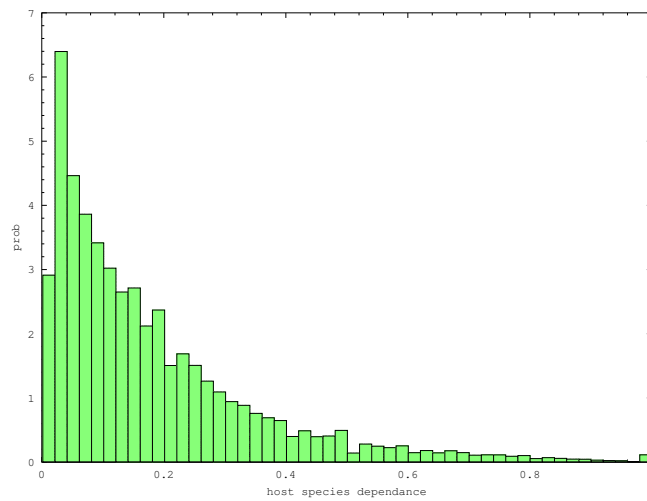


Figure 3.3: Frequency distribution of dependence values as defined by Bascompte et al. (2006), for antagonist species for simulated networks at the central point. The distribution is similar to those observed by Bascompte et al. (2006) for real ecological networks.

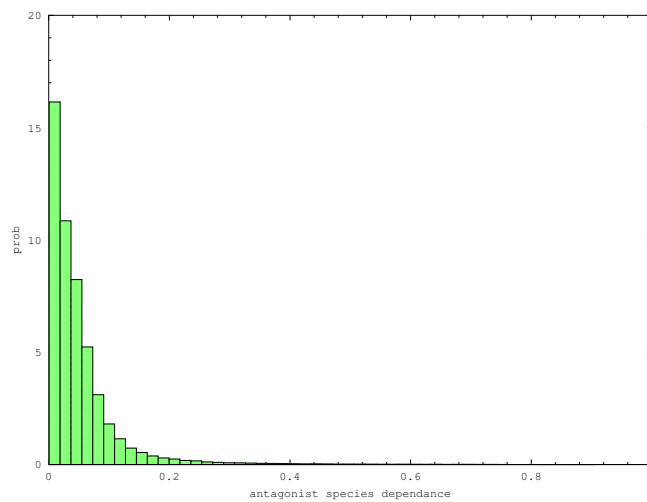


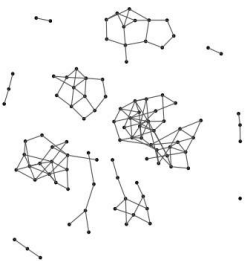
Figure 3.4: Frequency distribution of dependence values as defined by Bascompte et al. (2006), for host species for simulated networks at the central point. The distribution is similar to those observed by Bascompte et al. (2006) for real ecological networks.

1. initialization of the network with no link
2. while the total weight, n_W , is *not* reached
 - (a) one type of node is chosen randomly: the first type of node is chosen with probability equal to $\frac{n_1}{n_1+n_2}$ (the second type is thus chosen with probability $\frac{n_2}{n_1+n_2}$)
 - (b) one node is chosen randomly in the type chosen in (a)
 - (c) the set of the allowed partner nodes of the chosen node in (b) is:
 - with probability p_{comp} : the set of nodes which are in the other type and in the same group than the node chosen in step (b)
 - with probability $1 - p_{\text{comp}}$: the set of all the nodes of the other type, without taking account of their membership to a group
 - (d) the partner node is chosen in the set of allowed partner nodes defined in (c):
 - with probability p_{nest} : randomly with a probability proportional to its attractivity
 - with probability $1 - p_{\text{nest}}$: randomly with uniform probability
 - (e) the weight of the link between the two nodes chosen in steps (b) and (d) is modified as follows:
 - if the weight is stricly positive, it is incremented by one
 - if the weight is null and the total number of links, n_l , is not reached a link with a weight equal to one is added
3. If the total number of links n_l is not reached the network is discarded.

Based on the 47 ecological networks taken from the Interaction Web Database (Table 3.1), we found that some pairs of parameters among n_1 , n_2 , n_l and n_w were highly correlated. Allowing them to vary independently would lead to simulated networks very different from real ones. Therefore we used a reparametrization with almost independent new parameters that represent network size and connectivity.

The correlation between n_1 and n_2 (Fig. 3.6) was highly significant for ecological networks. The ratio $\frac{n_1}{n_2}$ and the product $n_1 n_2$ were less correlated (Fig. 3.7). Therefore we used these two new parameters in place of n_1 and n_2 . $n_1 n_2$ represents the size of the network (the number of possible links) and $\frac{n_1}{n_2}$ the ratio between the two types of nodes. Their distribution for real ecological networks are respectively shown in figures 3.8 and 3.9.

The number of edges n_l and the size of the network $n_1 n_2$ were correlated in ecological dataset networks (Fig. 3.10). The relationship between $\log(n_l)$ and $\log(n_1 n_2)$ was almost linear, with the slope of the linear regression equaled to .63 (Figs. 3.10). $\log(n_l) - .63 \log(n_1 n_2)$ is the residual of this regression. It is not correlated with the regressor $\log(n_1 n_2)$. The expression $\frac{n_l}{(n_1 n_2)^{.63}}$ is therefore not



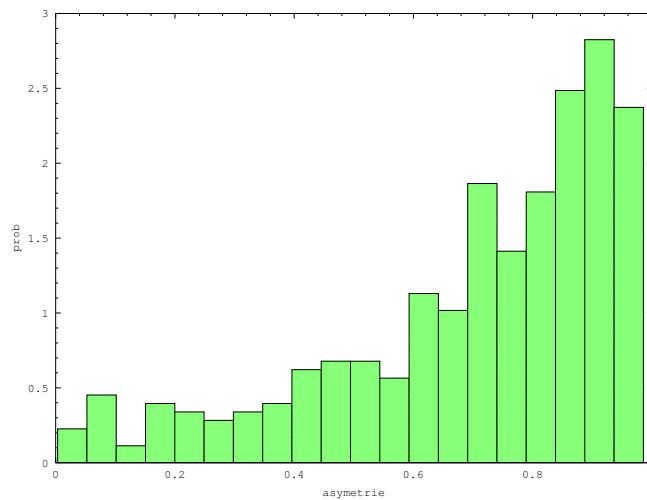


Figure 3.5: Frequency distribution of symmetry values as defined by Bascompte et al. (2006) for simulated networks at the central point. The distribution is similar to those observed by Bascompte et al. (2006) for real ecological networks.

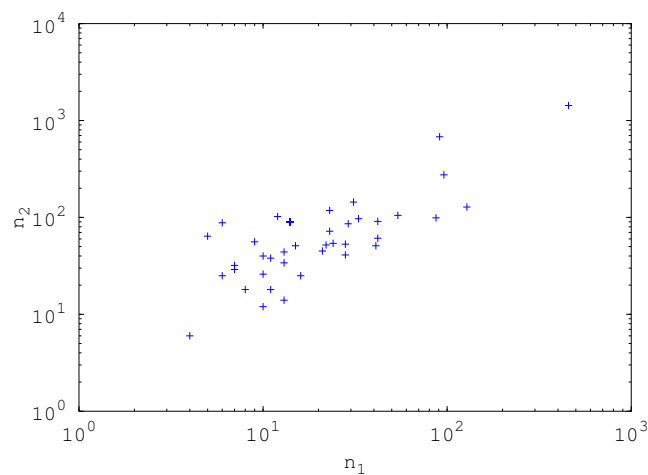


Figure 3.6: Relationship between the number of hosts species n_1 , and the number of antagonist species n_2 , on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_1 and n_2 are clearly linked.

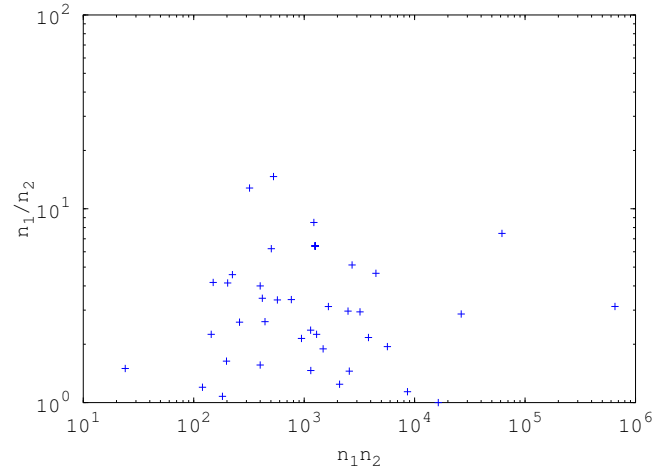


Figure 3.7: Relationship between the ratio of nodes types $\frac{n_1}{n_2}$, and the product $n_1 n_2$, on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_1/n_2 and $n_1 n_2$ are not linked

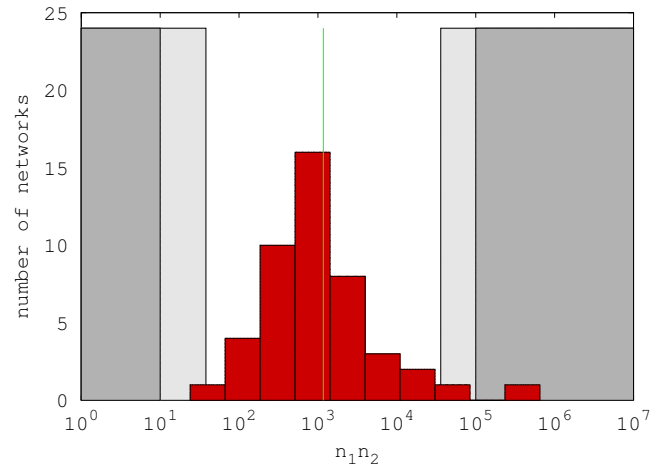
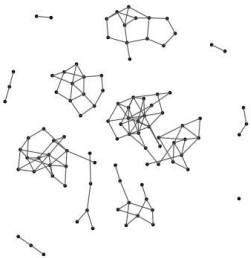


Figure 3.8: Empirical distribution of the product of number of species of both types, $n_1 n_2$, for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line.



correlated with $n_1 n_2$ (Fig. 3.11). We thus chose to use $\frac{n_l}{(n_1 n_2)^{.63}}$ as a new parameter instead of n_l . This new parameter represents the connectivity of the network. Its ranges were chosen using the ecological dataset (Fig. 3.12).

The total weight of the network n_w was correlated with the number of edges n_l in the weighted ecological networks dataset (Fig. 3.13). We chose the mean weight of the edges $\frac{n_w}{n_l}$ as new parameter and its studied range was chosen according to the ecological dataset (Fig. 3.14).

The ecologically relevant ranges for the four new parameters are given in Table 3.2. Ranges used for simulations (called studied parameter ranges in Table 3.2) were slightly larger than ecological ranges. It was not possible to give an ecological range for g , p_{comp} and p_{nest} , because these three parameters cannot be measured in real networks. Therefore we used a large studied range for them. The large range of the number of groups g constrained the studied range for the network size $n_1 n_2$. Indeed the number of species of each type must be greater or equal than the number of groups. Thus the minimum value for the studied range of network size (i.e. 56) was slightly greater than the minimum value found in the ecological dataset (i.e. 38).

The networks were simulated by keeping all the parameters fixed to a *central point*, except one that was allowed to vary within the studied range. In the case of size and connectivity parameters, the central point was defined as the geometrical mean value for real ecological networks. A network simulated with this central values had 64 species on the higher network level, 18 species on the basal network level, 120 edges and 1018 as total weight. We arbitrarily chose $g = 4$, $p_{\text{comp}} = .6$ and $p_{\text{nest}} = .5$ for central values. This choice corresponds to a rather strongly compartmentalized and nested network with four compartments. The varying parameter was allowed to take 10 different values (except for the number of groups that took only four values). One hundred networks were simulated for each parameter combination. This simulation design resulted in $(6 \times 10 + 4) \times 100 = 6400$ weighted networks. The binary version of the networks was obtained by replacing positive weights by one.

Number of groups

The number of groups in a real ecological network is usually unknown. All the clustering methods provide an estimate of this number. It is therefore logical to let the method estimate the number of groups and then compare the results to the true partition. However some methods may give an estimate of the number of groups that is far from the truth, and thus obtain poor results. Therefore we displayed two kinds of results: the results when the number of groups is estimated by the method (*unknown number of groups*), and the results when the true number of groups is given before the clustering (*fixed number of groups*).

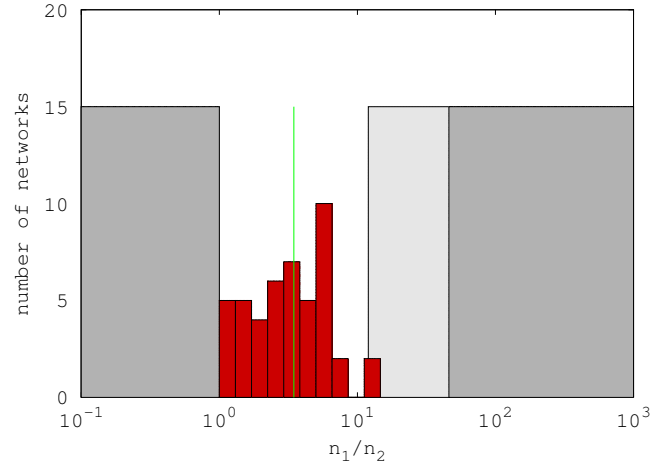


Figure 3.9: Empirical distribution of the ratio of the numbers of species of both types, n_1/n_2 , for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line.

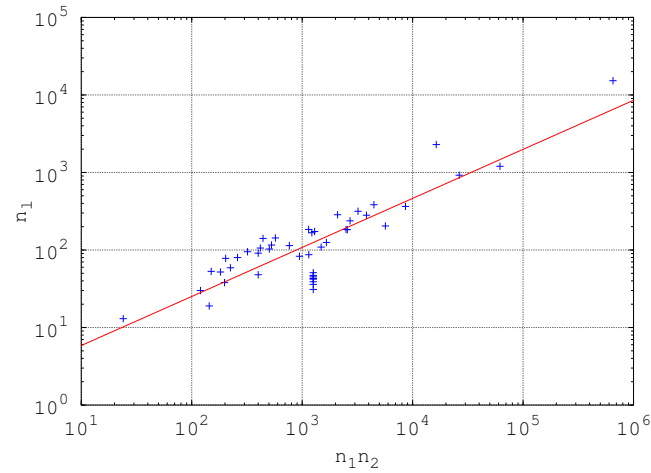
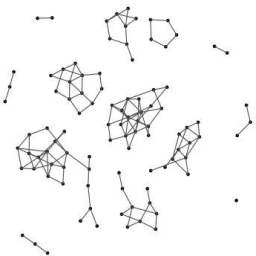


Figure 3.10: Relationship between the number of edges n_l , and the product of the numbers of species of both types $n_1 n_2$, on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_l and $n_1 n_2$ are clearly linked. The red line is the regression line.



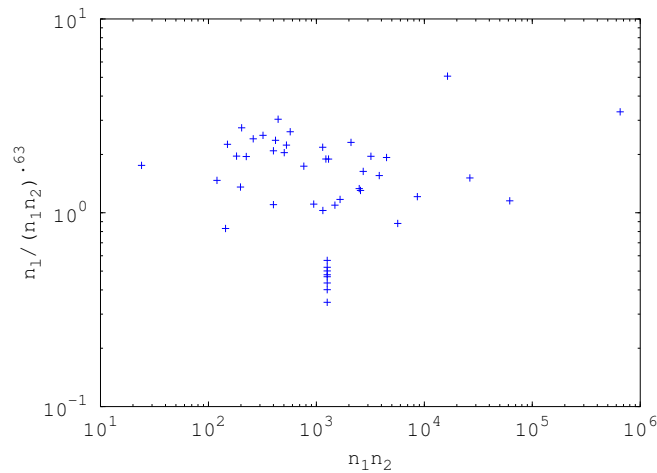


Figure 3.11: Relationship between the ratio $\frac{n_l}{(n_1 n_2)^{0.63}}$, and the product of the numbers of species of both types $n_1 n_2$, on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. There is no relation between them.

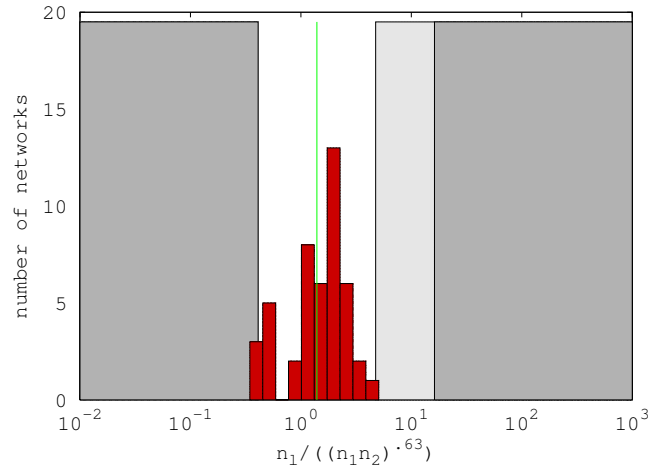


Figure 3.12: Empirical distribution of the ratio $\frac{n_l}{(n_1 n_2)^{0.63}}$, for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line.

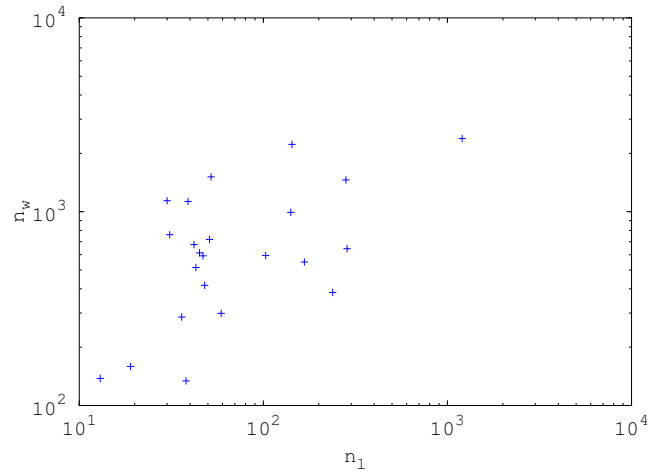


Figure 3.13: Relationship between the total weight of edges n_w , and the number of edges, n_l , on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_w and n_l are clearly related.

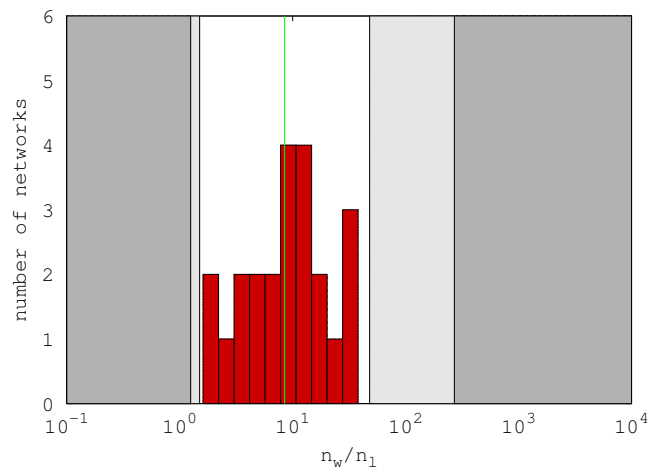
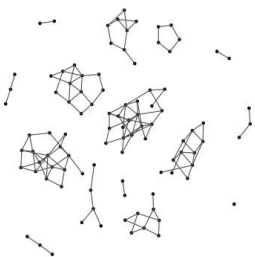


Figure 3.14: Empirical distribution of the mean weight by present edge, $\frac{n_w}{n_l}$, for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line.



The number of groups is not an input parameter for the methods MCL, $\text{MCL}_{\frac{1}{10}}$ and Mod. A few adjustments were thus made to obtain the results with a fixed number of groups. The MCL algorithm chooses automatically the number of groups, but this latter depends on the ratio of the expansion and the inflation coefficients (for a unknown number of groups used values are respectively 2 and 2.6). We dynamically changed this ratio to have a number of groups which is close to the fixed number of groups. It was not always possible to obtain exactly the fixed number of groups, but the difference was always very small. The modularity maximization algorithm selects the number of groups which maximizes the modularity for all partitions. In order to obtain a fixed number of groups, we have implemented a greedy algorithm which restricts the search to partitions with the fixed number of groups.

The number of groups is an input parameter for NSC, ASC, EB and SBM. In order to apply these methods to simulated networks in the case of an unknown number of groups, the number of groups was selected by using an additional method or criteria. In the case of EB, the number of groups was estimated by maximizing the modularity. In the case of SBM, we used the ICL criterion to select the number of groups (see Daudin et al. 2008). As long as we know, very few methods, if any, exist for estimating the number of groups for spectral clustering methods. We therefore used a method which is simple and works well in practice on the simulations. This method is based on the eigenvalues of the Laplacian matrix. The eigenvectors were chosen in relation with the eigenvalues in growing order conditionally to be lower than $\frac{1}{2}$.

3.2.3 Criteria used to compare the efficiency of the clustering methods

Quality of clustering

We assessed the ability of each clustering method to retrieve groups of species. For that, we compared the obtained partition with the true partition by using the adjusted rand-index (Rand, 1971) for each simulated network. The adjusted rand-index takes its values in the range $[-1, 1]$. An adjusted rand-index value equal to 1 corresponds to a perfect clustering. A value of 0 is expected for a random clustering. We calculated two adjusted rand indexes, one indicating the quality of clustering for the basal network level (the 'adjusted rand-index for host species', ariH) and the other indicating the quality of clustering for the higher network level (the 'adjusted rand-index for antagonist species', ariA).

Quality of prediction

We also assessed the ability of each clustering method to predict edges between the groups of species, by calculated the adjusted R^2 (R_a^2) for each simulated network. A method that does not associate correctly groups of species has a poor adjusted R^2 .

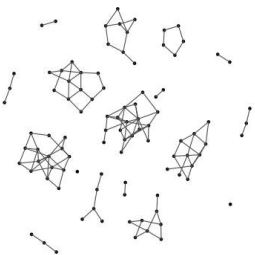
The R^2 is the square of the correlation coefficient between the observed and predicted edges. The predicted value for an edge between two nodes is defined as the mean value of the observed edges between the groups to which the two nodes have been classified. Prediction is restricted to the edges between the two levels of the bipartite network. The number of parameters for obtaining this prediction is thus equal to the product of the number of groups in the basal network level, g_H , with the number of groups in the higher network level, g_A . The R^2 is adjusted by taking into account this number, by using the expression $R_a^2 = 1 - (1 - R^2) \frac{n_1 n_2 - 1}{n_1 n_2 - g_H g_A - 1}$.

3.3 Results

3.3.1 Relative efficiency of the clustering methods when applied to ecological bipartite networks with average properties

We first compared the clustering methods for average ecological networks (simulated by fixing the network properties to the *central point*), in the case where the number of groups was estimated by the method itself. In the case of binary networks (Fig. 3.15), all the methods, except SBM, strongly over-estimated the number of groups and hardly retrieved their composition, with mean adjusted Rand indexes lower than .25. SBM under-estimated the number of groups and had the lowest adjusted Rand indexes. EB was the method which gave the best estimated number of groups and their composition. In the case of weighted networks (Fig. 3.16), all the methods gave better results than for binary networks. SBM was the method which estimated the most precisely the number of groups and it retrieved well their composition, with mean adjusted Rand indexes equal to .4 for the basal network level and .6 for the higher network level. It was also among the three best methods regarding the ability to predict associations between groups. EB also performed well, despite a high over-estimation of the number of groups.

The relative performance of the methods changed in the case where the true number of groups was given to the method before the clustering step. NSC was the method which retrieved the most precisely the composition of groups, both for



Parameter name	Parameter expression	Ecological parameter range	Studied parameter range	Central point
compartmentalization index	p_{comp}		[.1, .9]	.6
Nestedness index	p_{nest}		[.1, .9]	.5
Number of groups	g		{2, 4, 8, 16}	4
Size	$n_1 \times n_2$	$[38, 36 \cdot 10^3]$	$[56, 100 \cdot 10^3]$	$1.2 \cdot 10^3$
Ratio between types	$\frac{n_1}{n_2}$	[1, 12]	[1, 46]	3.5
Connectivity parameter	$\frac{n_l}{(n_1 n_2)^{0.63}}$	[.41, 4.75]	[.41, 16.2]	1.4
Mean links weight	$\frac{n_w}{n_l}$	[1.5, 48]	[1.25, 270]	8.5

Table 3.2: Names, expression, ecological and studied range of the seven parameters of the simulations. The central point corresponds to the central values for each parameter.

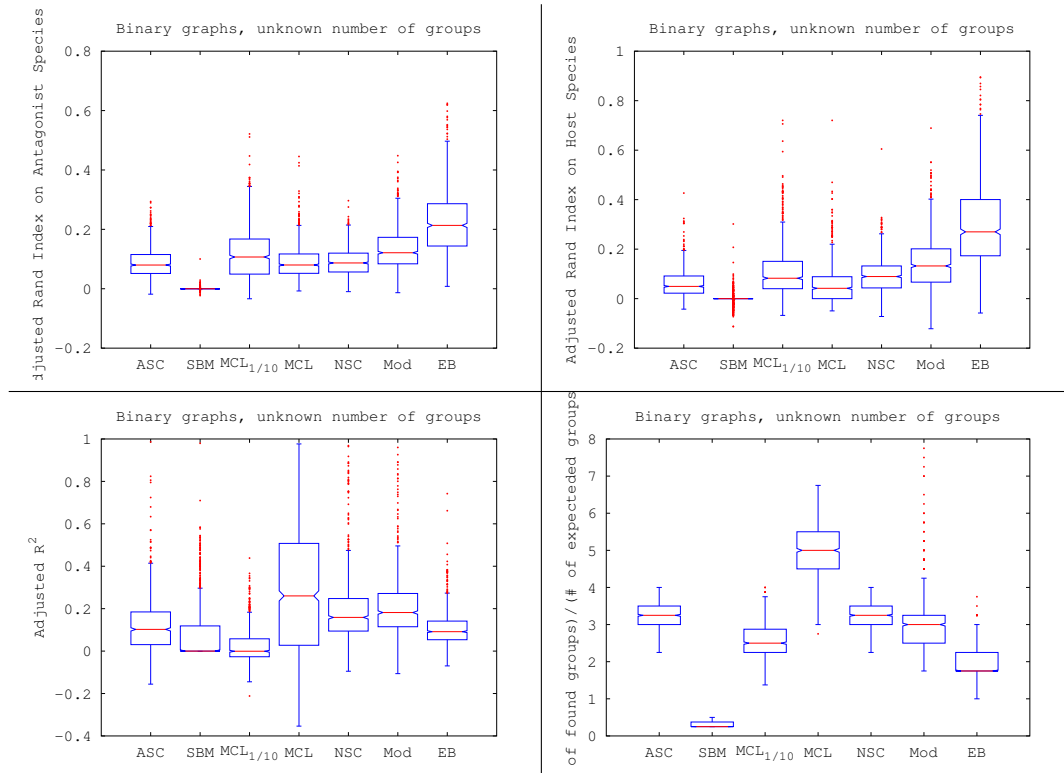


Figure 3.15: Box-plots of the 100 replicates obtained on binary networks for unknown number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges, bottom-right: number of estimated groups divided by the true number (4 for communities methods, 8 for the SHS methods)

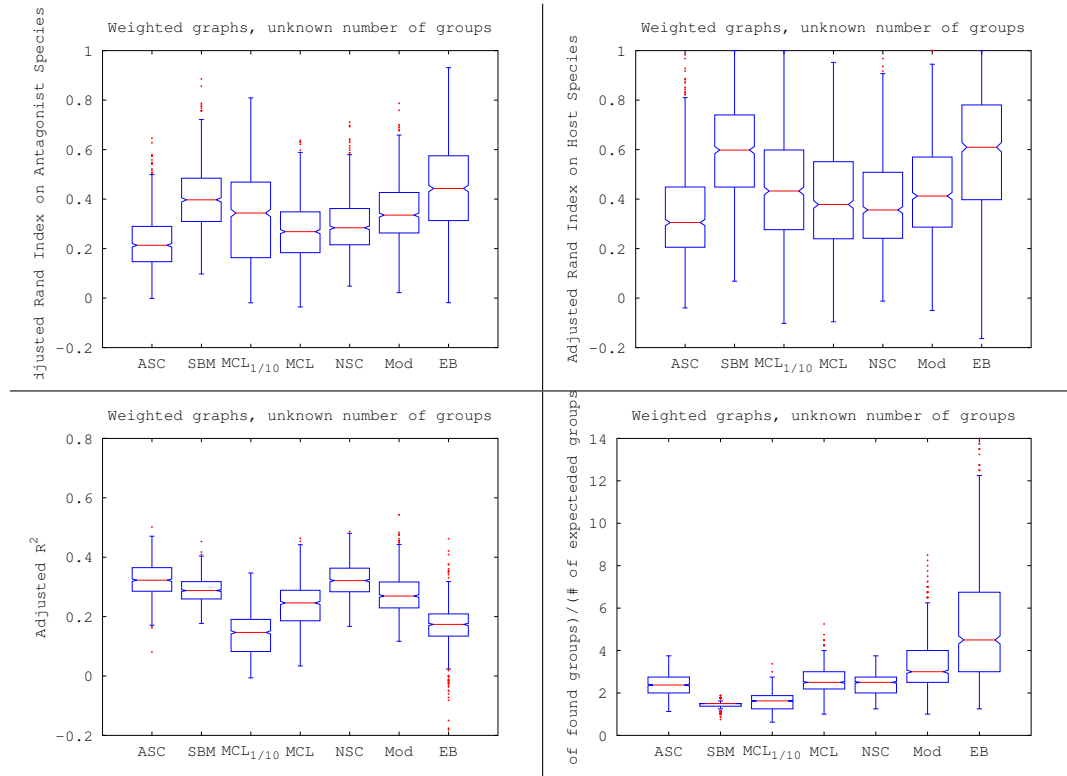
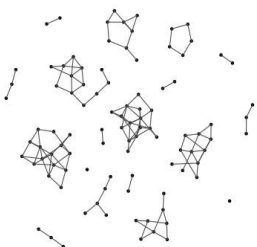


Figure 3.16: Box-plots of the 100 replicates obtained on weighted networks for unknown number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges, bottom-right: number of estimated groups divided by the true number (4 for communities methods, 8 for the SHS methods).



binary (Fig. 3.17) and weighted networks (Fig. 3.18). Knowing the true number of groups improved the performance of SBM on binary networks (Fig. 3.17), but it drastically reduced that of EB for weighted networks (Fig. 3.18).

Execution processor time are presented in Table 3.3. Mod was the fastest method. SBM was the slowest. Real execution time of SBM was lower than processor execution time since SBM used parallelized estimation. However, SBM was also the slowest for real execution time (not shown).

3.3.2 Effect of network properties on the hierarchy between clustering methods

The effect of network properties on the relative performance of the clustering methods was investigated in the case where the number of groups was estimated by the method itself. We first compared the ability of the methods to estimate the true number of groups. In the case of binary networks (Fig. 3.19), SBM does not find any group except the two trophic levels, and all the other methods strongly overestimated the number of groups when the size of the network was high or when the true number of groups was low. EB give the more stable and the best results for all values of the parameters except in the case of high connectivity. Variations in the degree of nestedness and compartmentalization hardly changed the quality of the estimation of the number of groups. In the case weighted networks (Fig. 3.20), all the methods, except SBM and $MCL_{\frac{1}{10}}$, had a tendency to over-estimation of the number of groups when the size of the network was high or when the true number of groups was low. This was particularly true for EB and Mod. EB strongly overestimated the number of groups in most cases, especially when the degree of compartmentalization was low, the degree of nestedness was high, the number of species in the higher network level was high in comparison with the number of species in the basal network level and when the mean weight of links increased.

Variations in the network properties also changed the ability of the methods to retrieve the composition of the groups, as indicated by the variations in the ariA and ariH indexes. In the case of binary networks (Fig. 3.19), all the methods except SBM better retrieved the groups when the degree of compartmentalization was high. SBM did not retrieve the groups at all. The same result was obtained in the case of weighted networks (Fig. 3.20), for all the methods including SBM. In that case, the ability of the methods to retrieve the groups also increased with the number of groups and the connectivity, but decreased with the network size.

Finally, variations in the network properties changed the ability of the methods to predict the associations between groups, as indicated by variations in the R_a^2 index. Variations in the case of weighted networks (Fig. 3.20) were stronger and more consistent than in the case of binary networks (Fig. 3.19). In the case of

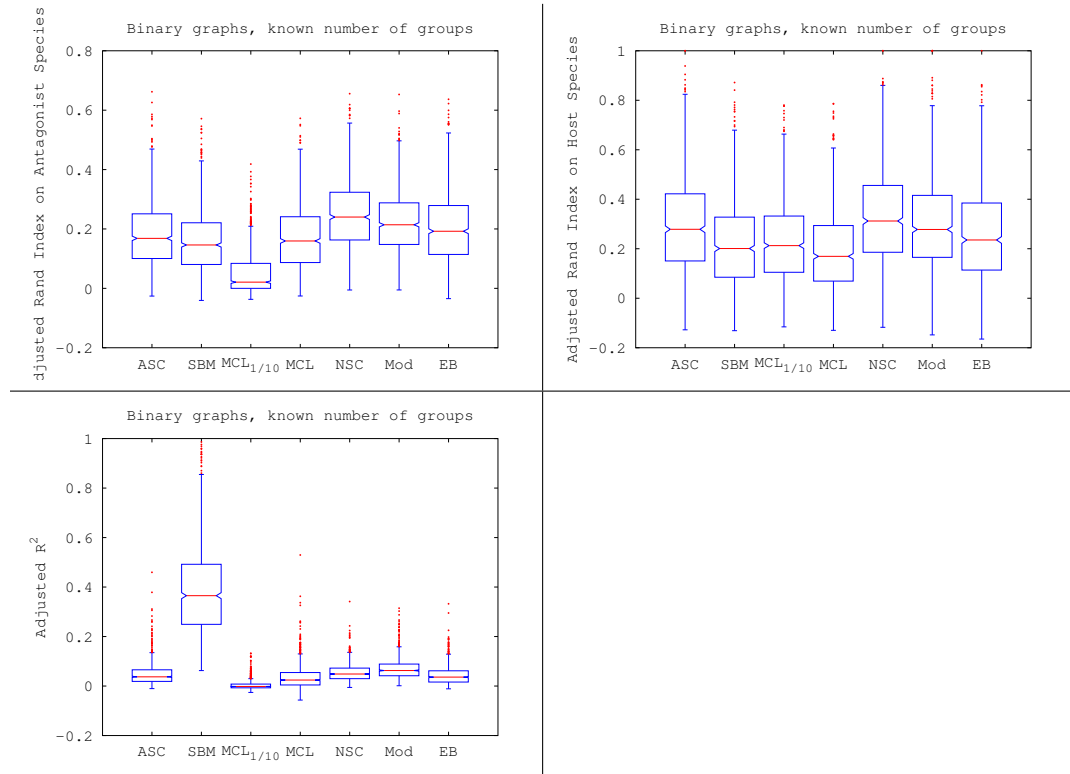
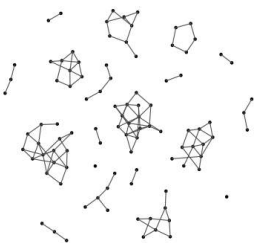


Figure 3.17: Box-plots of the 100 replicates obtained on binary networks for known number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges.



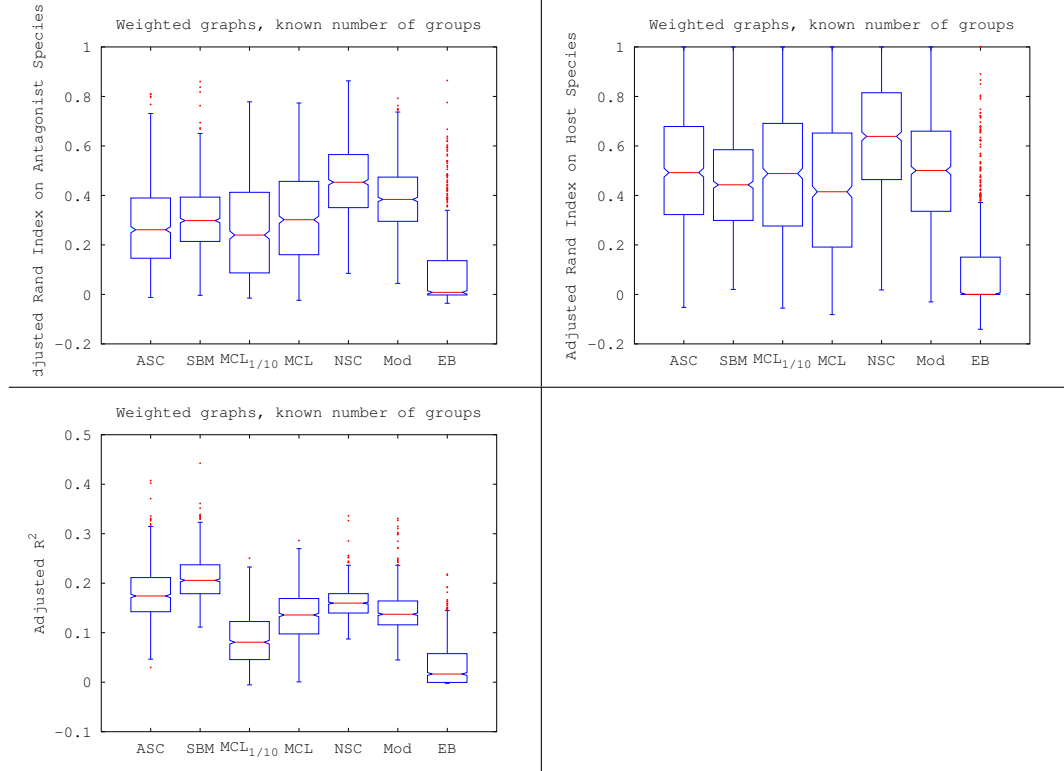
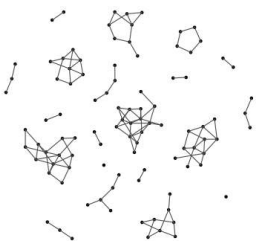


Figure 3.18: Box-plots of the 100 replicates obtained on weighted networks for known number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges.

Method	Unknown number of groups		Known number of groups	
	Binary	Weighted	Binary	Weighted
ASC	1m 35s	1m 30s	1m 27s	1m 21s
SBM	22h 26m	11h 49m	22h 26m	11h 49m
$MCL_{\frac{1}{10}}$	33s	31s	7m 40s	7m 59s
MCL	31s	32s	6m 8s	6m 5s
NSC	1m 25s	1m 23s	1m 24s	1m 17s
Mod	11s	13s	8s	9s
EB	31s	12m 59s	31s	12m 56s

Table 3.3: Execution time of methods for a network on the central point. Given time is processor time, *i.e.* equivalent execution time on a single processor computer which only run one job. Some methods as SBM use parallel computing and processor time must be divided by the number of threads to obtain the real execution time



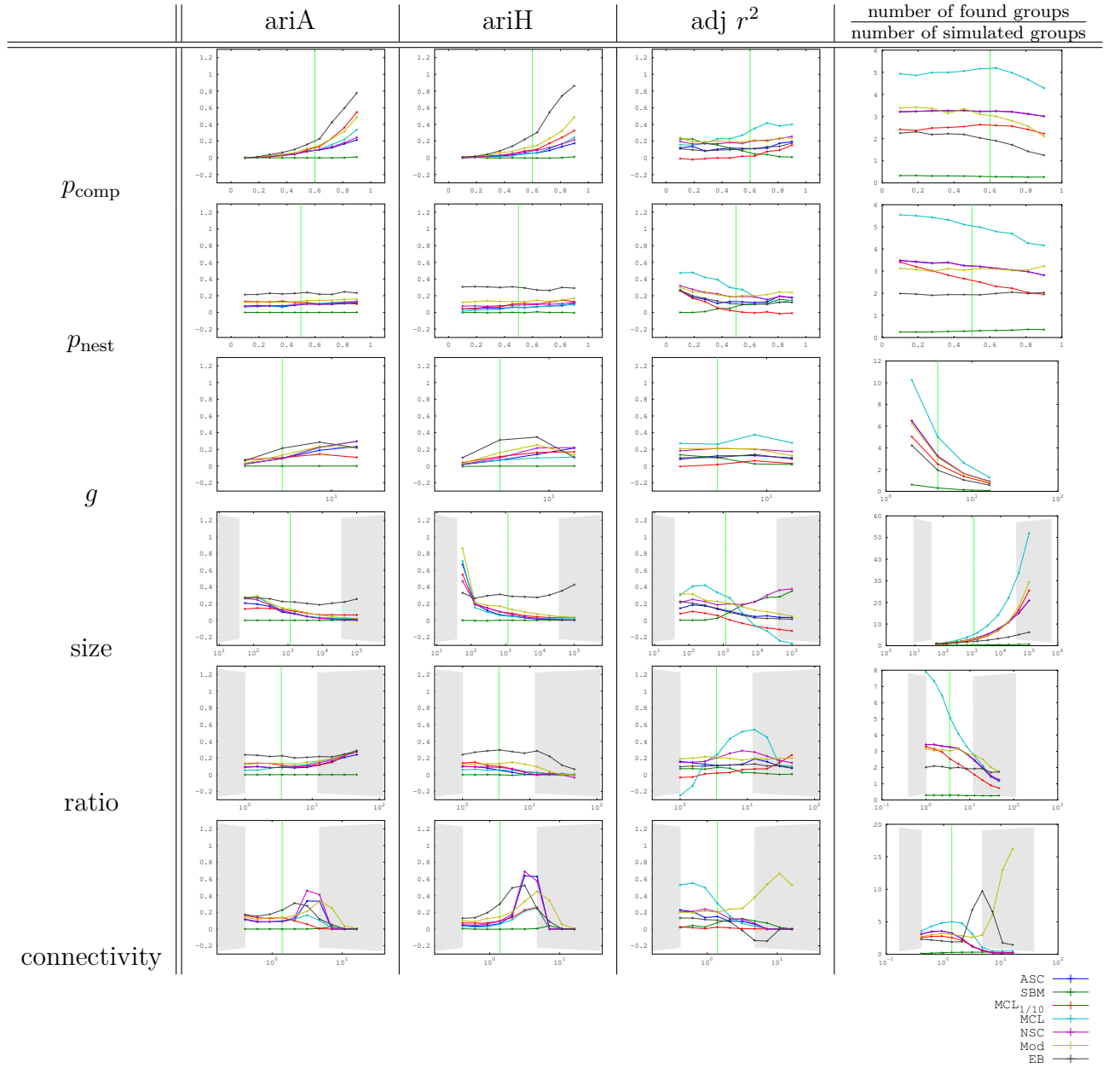


Figure 3.19: Relationships between the performances of the methods and the network parameters for unknown number of groups and binary networks. ariA: adjusted Rand index for antagonist species, ariH: adjusted Rand index for the host species, R_a^2 : adjusted R square for the prediction of edges, last column: number of estimated groups divided by the true number (g for communities methods, $2g$ for the SHS methods), size: $n_1 n_2$, ratio: $\frac{n_1}{n_2}$, connectivity: $\frac{n_1}{(n_1 n_2)^{0.63}}$.

Graph clustering methods differ in their ability to detect patterns in bipartite ecological networks

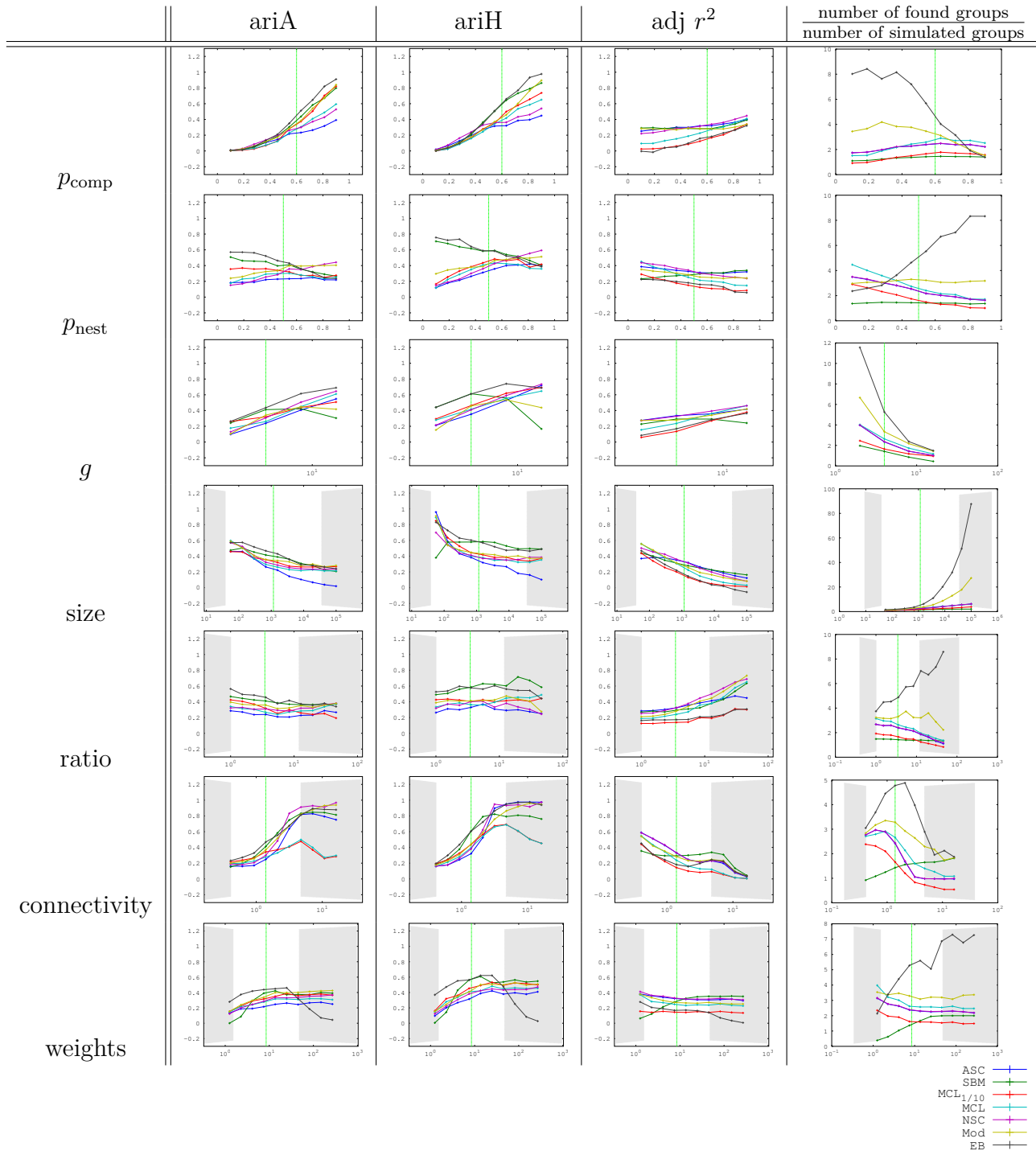
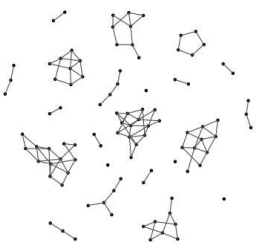


Figure 3.20: Relationships between the performances of the methods and the network parameters for unknown number of groups and weighted networks. ariA: adjusted Rand index for antagonist species, ariH: adjusted Rand index for the host r^2 : adjusted R square for the prediction of edges, last column: number of groups divided by the true number (g for communities methods, $2g$ for SBM methods), size: $n_1 n_2$, ratio: $\frac{n_1}{n_2}$, connectivity: $\frac{n_1}{(n_1 n_2)^{0.63}}$.



weighted networks (Fig. 3.20) , all the methods gave a better prediction when the degree of compartmentalization was high, the true number of groups was high, the network size was small and the network connectivity was low.

In the case where the number of groups was known, the effect of network properties on the relative performance of the clustering methods was investigated. For binary networks results are shown figure 3.21 and for weighted networks figure 3.22.

3.4 Discussion

Our results showed that some methods of graph clustering performed better than others at retrieving sub-groups of highly interacting nodes. Among the clustering methods included in this study, and in the case of binary bipartite networks simulated with ecologically relevant parameters, the edge-betweenness algorithm consistently gave the best results. It retrieved the composition of the sub-groups and kept a moderate over-estimation of the number of sub-groups. The edge-betweenness algorithm was also very efficient at retrieving the composition of the sub-groups in the case of weighted bipartite networks, but it strongly over-estimated the number of sub-groups in that case. The stochastic block model was another very good method for weighted networks, but it was the slowest method in terms of computation time. Contrary to the edge-betweenness algorithm, it estimated precisely the number of sub-groups in weighted networks. It is noteworthy that the stochastic block model was not able to retrieve the sub-groups in binary networks, despite its good performance on weighted networks.

Modularity maximization, which is the most popular clustering method in ecology, gave satisfactorily results for both weighted and binary networks, although it had a tendency to over-estimate the number of sub-groups (by a factor of between 2 and 3). It was among the three best methods for a wide range of ecological networks. However this clustering method suffers a major drawback: it can be very time consuming depending on the method used to maximize the modularity. For instance, the simulated annealing optimization approach (Guimera and Amaral, 2005) implemented in the software NETCARTO, which has often been used in ecological studies (Martos et al., 2012; Gómez et al., 2010; Guimera et al., 2010; Meskens et al., 2011; Traveset et al., 2013; Danieli-Silva et al., 2012; Krasnov et al., 2012; Genini et al., 2012), is particularly slow. We could not use it in our study to cluster several thousands of networks. Therefore, we used another approach to maximize the modularity (Newman, 2006). This approach is implemented in the new software MODULAR (Marquitti et al., 2013), which has been developed for analyzing the structure of thousands of networks in a reasonable amount of time. This software offers the choice between several maximization approaches and also between several modularity metrics. The modularity metrics used in our study is

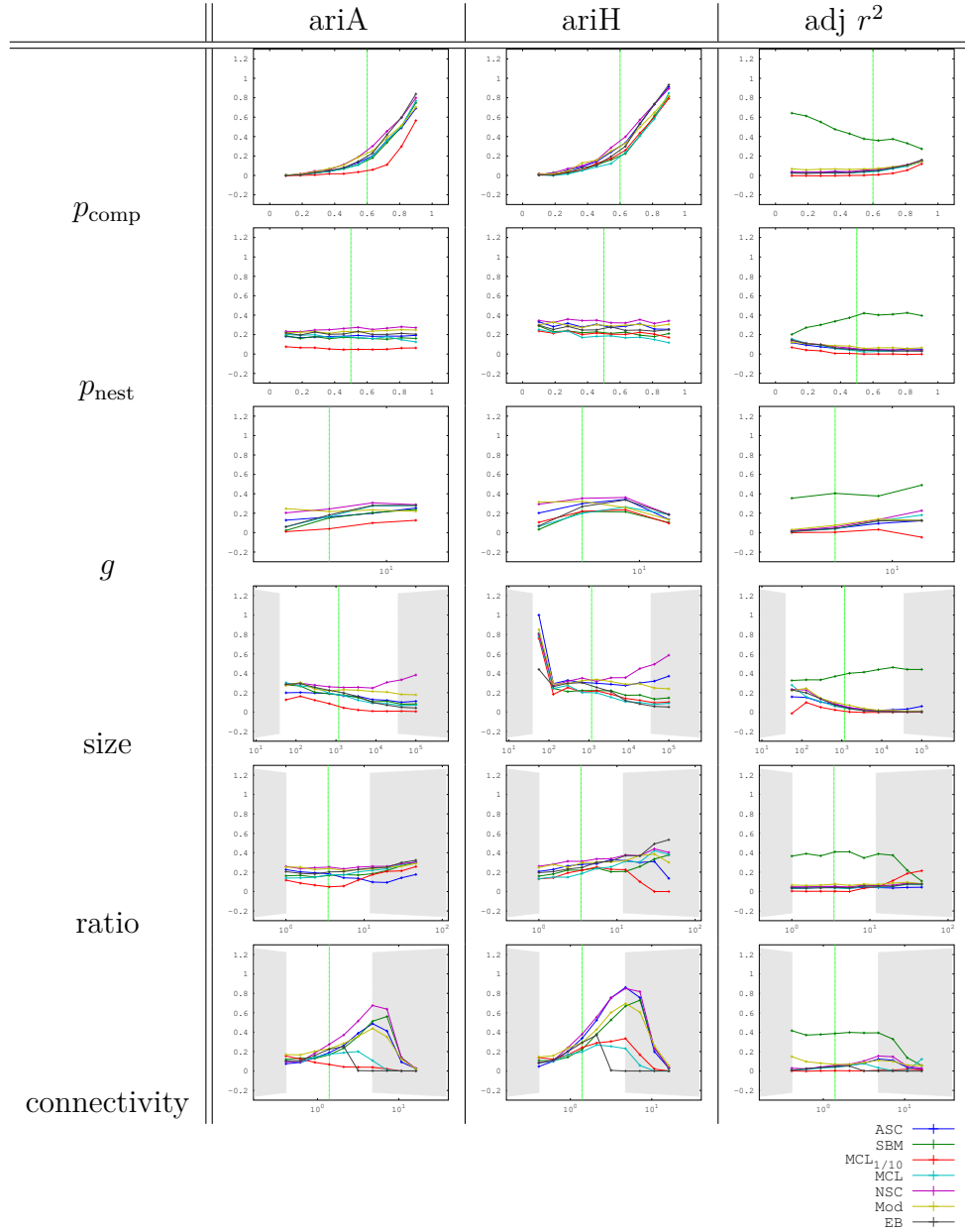
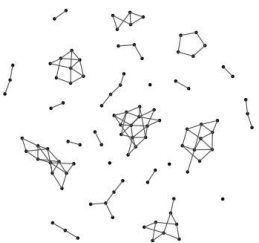


Figure 3.21: Relationships between the performances of the methods and the network parameters for known number of groups and binary networks. ariA: adjusted Rand index for antagonist species, ariH: adjusted Rand index for the host species, R_a^2 : adjusted R square for the prediction of edges, size: $n_1 n_2$, ratio: $\frac{n_1}{n_2}$, connectivity: $\frac{n_l}{(n_1 n_2)^{0.63}}$.



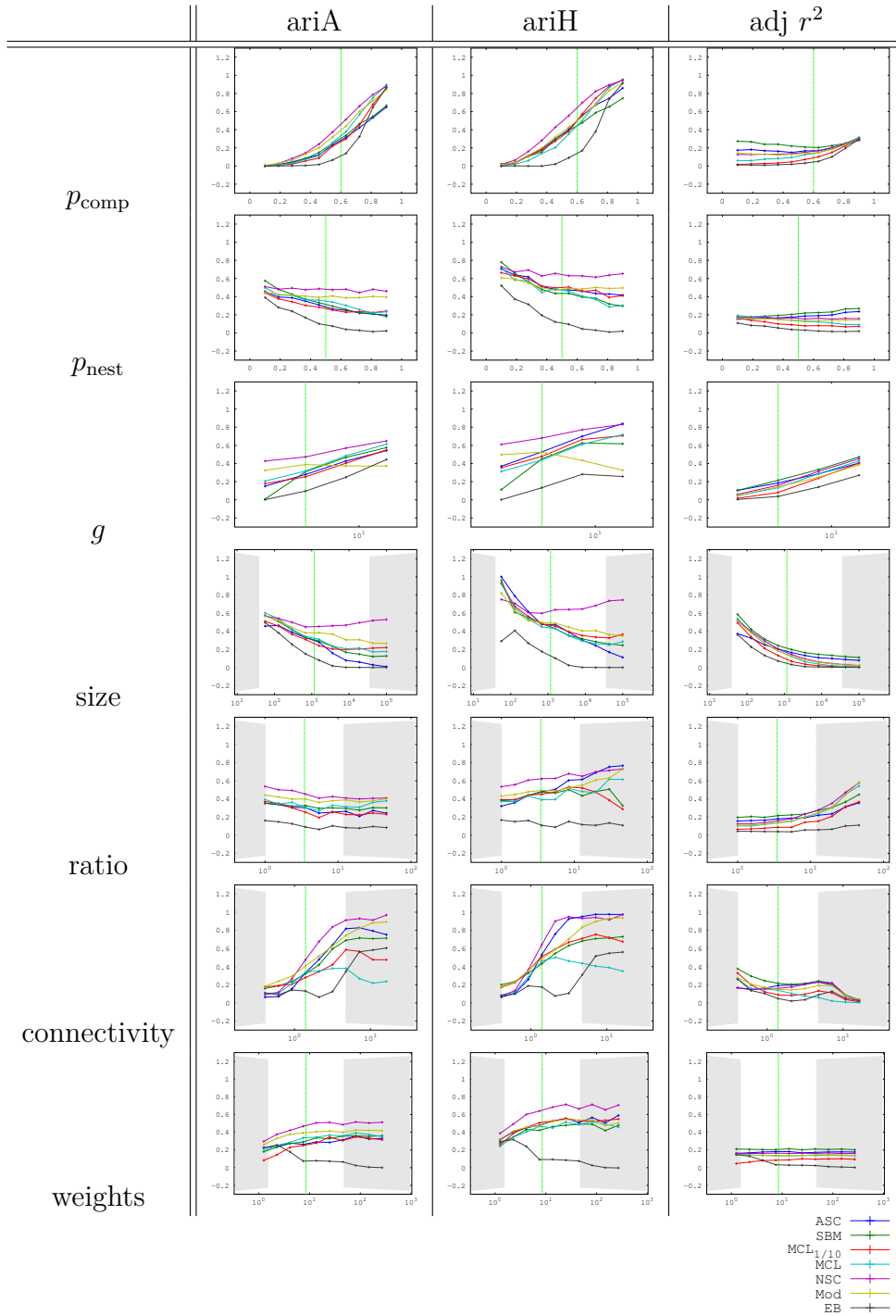


Figure 3.22: Relationships between the performances of the methods and the network parameters for known number of groups and weighted networks. ariA : adjusted Rand index for antagonist species, ariH : adjusted Rand index for the host species, R_a^2 : adjusted R square for the prediction of edges, size : $n_1 n_2$, ratio : $\frac{n_1}{n_2}$, connectivity : $\frac{n_1}{(n_1 n_2)^{0.63}}$.

the one proposed by Newman (2006). We chose it because it is the most common modularity metrics and because it has been applied to bipartite ecological networks, although it was developed for unipartite networks. Other modularity metrics, developed specifically for bipartite networks (Barber, 2007; Guimerà et al., 2007), may have been used.

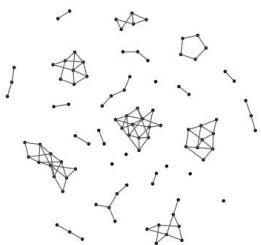
Our results also showed that the accurate estimation of the number of sub-groups, which is an input parameter in several clustering methods, is sometimes as important as the clustering step itself. Among the clustering methods included in this study, the number of sub-groups is an input parameter for spectral clustering methods, the stochastic block model and the edge betweenness algorithm. The number of sub-groups must be selected according to a criterion before the clustering step. For the other methods the number of sub-groups is automatically chosen by the algorithm.

All the methods, except the stochastic block model, over-estimate the true number by a factor greater or equal to 2, leading to the building of chimeric groups. However the corrected Rand index is not necessarily bad with too many sub-groups, if these estimated sub-groups are just divisions of the true sub-groups.

The criterion ICL associated with the stochastic block model has been built in order to avoid over-estimation, and this is checked in the simulation results. However in the case of binary networks, the number of subgroups was consistently underestimated by ICL. The clustering obtained with the stochastic block model was just able to separate the two trophic levels of the bipartite network. In contrast, when the accurate number of subgroups was selected, the stochastic block model performed much better. The criterion used to estimate the number of sub-groups should thus be improved, in order to increase the applicability of the stochastic block model to binary, bipartite ecological networks with similar structure of network obtained by the simulation algorithm.

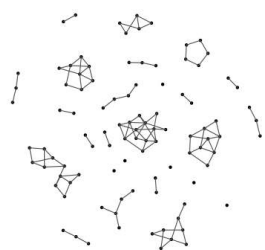
Another important result of our study is that the retrieved sub-groups were generally more accurate for host species than for antagonist species. This result may be accounted for by the lower number of host species in ecological networks. The amount of information per host species contained by the network is thus higher than the amount of information per antagonist species, leading to a better classification of host species. Interestingly, a relationship between subgroups and phylogeny has often been found for host species but not for antagonist species (Vacher et al., 2008a; Krasnov et al., 2012). Evolutionary processes have been proposed to account for this asymmetric pattern. Our results suggest that the clustering method is an alternative explanation.

Finally, our results showed that the clustering methods which are the best at retrieving the subgroups are not those which best retrieve the links between the subgroups. As already mentioned, detecting sub-groups of highly interacting



species is important for ecologists because it provides a simplified picture of the network (Allesina and Pascual, 2009). Estimating the number of sub-groups and their composition allows drawing the first part of the picture. The picture would be very incomplete without the links between the sub-groups. Our results showed that spectral clustering methods were the most efficient at predicting these links in the case of weighted networks, although they were less efficient than other methods at retrieving the composition of the sub-groups. The stochastic block model and modularity maximization also gave good predictions, contrary to the edge betweenness algorithm.

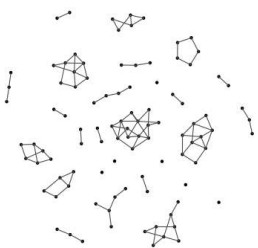
In conclusion, according to our results, researchers in ecology should favor the edge-betweenness algorithm with the modularity criterion for group number selection in order to retrieve sub-groups of highly interacting species in binary bipartite networks. In the case of weighted bipartite networks, they should rather favor the stochastic block model, since it accurately estimates the number of sub-groups and it retrieves well their composition and the links between them. Unfortunately, this method is very time consuming. Modularity maximization is a good alternative, knowing that recent developments allow analyzing thousands of networks in a reasonable amount of time.



Bibliography

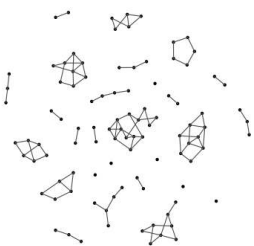
- S. Allesina and M. Pascual. Food web models: a plea for groups. *Ecology letters*, 12(7):652–662, 2009.
- M. T. K. Arroyo, R. Primack, and J. Armesto. Community studies in pollination ecology in the high temperate Andes of central Chile. I. Pollination mechanisms and altitudinal variation. *American journal of botany*, pages 82–97, 1982.
- M. J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007.
- S. C. Barrett and K. Helenurm. The reproductive biology of boreal forest herbs. I. Breeding systems and pollination. *Canadian Journal of Botany*, 65(10):2036–2046, 1987.
- J. Bascompte, P. Jordano, and J. M. Olesen. Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science*, 312(5772):431–433, 2006.
- N. Blüthgen and K. Fiedler. Preferences for sugars and amino acids and their conditionality in a diverse nectar-feeding ant community. *Journal of Animal Ecology*, 73(1):155–166, 2004.
- A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- F. E. Clements and F. L. Long. *Experimental pollination: an outline of the ecology of flowers and insects*. Number 336. Carnegie Institution of Washington, 1923.
- A. Danieli-Silva, J. M. T. de Souza, A. J. Donatti, R. P. Campos, J. Vicente-Silva, L. Freitas, and I. G. Varassin. Do pollination syndromes cause modularity and predict interactions in a pollination network in tropical high-altitude grasslands? *Oikos*, 121(1):35–43, 2012.
- J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008. ISSN 0960-3174.

- D. Davidson and B. Fisher. Symbiosis of ants with *Cecropia* as a function of light regime. *Huxley, C, R., Cutler, D, F ed (s). Ant-plant interactions. Oxford Univ. Press: Oxford, etc*, pages 289–309, 1991.
- D. W. Davidson, R. R. Snelling, and J. T. Longino. Competition among ants for myrmecophytes and the significance of plant trichomes. *Biotropica*, pages 64–73, 1989.
- W. E. Donath and A. J. Hoffman. Lower Bounds for the Partitioning of Graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.
- C. F. Dormann and R. Strauss. Detecting modules in quantitative bipartite networks: the QuaBiMo algorithm. *arXiv preprint arXiv:1304.3218*, 2013.
- Y. L. Dupont, D. M. Hansen, and J. M. Olesen. Structure of a plant–flower–visitor network in the high-altitude sub-alpine desert of Tenerife, Canary Islands. *Ecography*, 26(3):301–310, 2003.
- H. Elberling and J. M. Olesen. The structure of a high latitude plant–flower visitor system: the dominance of flies. *Ecography*, 22(3):314–323, 1999.
- C. R. Fonseca and G. Ganade. Asymmetries, compartments and null interactions in an Amazonian ant–plant community. *Journal of Animal Ecology*, pages 339–347, 1996.
- C. Fontaine, P. R. Guimarães, S. Kéfi, N. Loeuille, J. Memmott, W. H. Van Der Putten, F. J. Van Veen, and E. Thébault. The ecological and evolutionary implications of merging different types of networks. *Ecology letters*, 14(11): 1170–1181, 2011.
- S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010. ISSN 0370-1573. doi: 10.1016/j.physrep.2009.11.002. URL <http://www.sciencedirect.com/science/article/pii/S0370157309002841>.
- J. Genini, M. C. Côrtes, P. R. Guimaraes Jr, and M. Galetti. Mistletoes Play Different Roles in a Modular Host–Parasite Network. *Biotropica*, 44(2):171–178, 2012.
- M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821, 2002.
- J. M. Gómez, M. Verdú, and F. Perfectti. Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature*, 465(7300):918–921, 2010.



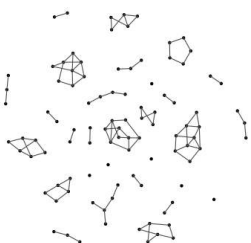
- R. Guimera and L. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005. ISSN 0028-0836.
- R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3):036102, 2007.
- R. Guimera, D. Stouffer, M. Sales-Pardo, E. Leicht, M. Newman, and L. Amaral. Origin of compartmentalization in food webs. *Ecology*, 91(10):2941–2951, 2010.
- B. Hocking. Insect-flower associations in the high Arctic with special reference to nectar. *Oikos*, pages 359–387, 1968.
- D. W. INOUE and G. H. PYKE. Pollination biology in the Snowy Mountains of Australia: comparisons with montane Colorado, USA. *Australian Journal of Ecology*, 13(2):191–205, 1988.
- A. Joern. Feeding patterns in grasshoppers (Orthoptera: Acrididae): factors influencing diet specialization. *Oecologia*, 38(3):325–347, 1979.
- P. Jordano, J. Bascompte, and J. M. Olesen. Invariant properties in coevolutionary networks of plant–animal interactions. *Ecology letters*, 6(1):69–81, 2003.
- B. R. Krasnov, M. A. Fortuna, D. Mouillot, I. S. Khokhlova, G. I. Shenbrot, and R. Poulin. Phylogenetic signal in module composition and species connectivity in compartmentalized host-parasite networks. *The American Naturalist*, 179(4):501–511, 2012.
- S. R. Leather. Feeding specialisation and host distribution of British and Finnish *Prunus* feeding macrolepidoptera. *Oikos*, pages 40–48, 1991.
- J.-B. Leger, C. Vacher, and J.-J. Daudin. Detection of structurally homogeneous subsets in graphs. *Statistics and Computing*, pages 1–18, 2013.
- M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *Ann Appl Stat*, 4:715–742, 2010.
- F. M. D. Marquitti, P. R. Guimaraes Jr, M. M. Pires, and L. F. Bittencourt. MODULAR: Software for the Autonomous Computation of Modularity in Large Network Sets. *arXiv preprint arXiv:1304.2917*, 2013.
- F. Martos, F. Munoz, T. Pailler, I. Kottke, C. Gonneau, and M.-A. SELOSSE. The role of epiphytism in architecture and evolutionary constraint within mycorrhizal networks of tropical orchids. *Molecular Ecology*, 21(20):5098–5109, 2012.

- C. McMullen. Flower-visiting insects of the Galapagos Islands. *Pan-Pacific Entomologist*, 69(1):95–106, 1993.
- D. Medan, N. H. Montaldo, M. Devoto, A. Mantese, V. Vasellati, G. G. Roitman, and N. H. Bartoloni. Plant-pollinator relationships at two altitudes in the Andes of Mendoza, Argentina. *Arctic, Antarctic, and Alpine Research*, pages 233–241, 2002.
- C. Meskens, D. Mckenna, T. Hance, and D. Windsor. Host plant taxonomy and phenotype influence the structure of a neotropical host plant–hispine beetle food web. *Ecological Entomology*, 36(4):480–489, 2011.
- T. Mosquin and J. Martin. Observations on the pollination biology of plants on Melville Island, NWT, Canada. *Canadian Field Naturalist*, 81:201–205, 1967.
- A. F. Motten. Pollination ecology of the spring wildflower community of a temperate deciduous forest. *Ecological Monographs*, pages 21–42, 1986.
- M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- M. E. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.
- J. M. Olesen, L. I. Eskildsen, and S. Venkatasamy. Invasion of pollination networks on oceanic islands: importance of invader complexes and endemic super generalists. *Diversity and Distributions*, 8(3):181–192, 2002.
- J. M. Olesen, J. Bascompte, Y. L. Dupont, and P. Jordano. The modularity of pollination networks. *Proceedings of the National Academy of Sciences*, 104(50):19891–19896, 2007.
- J. Ollerton, S. D. Johnson, L. CRANMER, and S. Kellie. The pollination ecology of an assemblage of grassland asclepiads in South Africa. *Annals of Botany*, 92(6):807–834, 2003.
- J. Ollerton, D. McCollin, D. G. Fautin, and G. R. Allen. Finding NEMO: nestedness engendered by mutualistic organization in anemonefish and their hosts. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):591–598, 2007.
- N. RAMIREZ and Y. BRITO. Pollination biology in a palm swamp community in the Venezuelan Central Plains. *Botanical Journal of the Linnean Society*, 110(4):277–302, 1992.



- W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- E. L. Rezende, E. M. Albert, M. A. Fortuna, and J. Bascompte. Compartments in a marine food web associated with phylogeny, body mass, and habitat structure. *Ecology Letters*, 12(8):779–788, 2009.
- C. Robertson. Flowers and insects lists of visitors of four hundred and fifty three flowers. 1929.
- K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional Stochastic Block Model. *The Annals of Statistics*, 39(4):1878–1915, 2011.
- D. W. Schemske, M. F. Willson, M. N. Melampy, L. J. Miller, L. Verner, K. M. Schemske, and L. B. Best. Flowering ecology of some spring woodland herbs. *Ecology*, pages 351–366, 1978.
- E. Small. Insect pollinators of the Mer Bleue peat bog of Ottawa. *Canadian Field Naturalist*, 90:22–28, 1976.
- D. B. Stouffer and J. Bascompte. Compartmentalization increases food-web persistence. *Proceedings of the National Academy of Sciences*, 108(9):3648–3652, 2011.
- E. Thébault and C. Fontaine. Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science*, 329(5993):853–856, 2010.
- A. Traveset, R. Heleno, S. Chamorro, P. Vargas, C. K. McMullen, R. Castro-Urgal, M. Nogales, H. W. Herrera, and J. M. Olesen. Invaders of pollination networks in the Galápagos Islands: emergence of novel communities. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758), 2013.
- C. Vacher, D. Piou, and M. Desprez-Loustau. Architecture of an antagonistic tree/fungus network: the asymmetric influence of past evolutionary history. *PLoS One*, 3(3):1740, 2008a. ISSN 1932-6203.
- C. Vacher, D. Piou, and M.-L. Desprez-Loustau. Architecture of an antagonistic tree/fungus network: the asymmetric influence of past evolutionary history. *PLoS One*, 3(3):e1740, 2008b.
- S. Van Dongen. Graph clustering by flow simulation. *University of Utrecht*, 275, 2000.

- D. P. Vázquez. *Interactions among introduced ungulates, plants, and pollinators: a field study in the temperate forest of the southern Andes*. PhD thesis, University of Tennessee, Knoxville, 2002.
- U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. ISSN 0960-3174.
- H. C. White, S. A. Boorman, and R. L. Breiger. Social Structure From Multiple Networks. *American Journal of Sociology*, 81:730–780, 1976.



Chapter 4

Wmixnet: Software for Clustering the Nodes of Binary and Valued Graphs using the Stochastic Block Model

This article was submitted to *Journal of Statistical Software*.

Abstract

Clustering the nodes of a graph allows the analysis of the topology of a network.

The *stochastic block model* is a clustering method based on a probabilistic model. Initially developed for binary networks it has recently been extended to valued networks possibly with covariates on the edges.

We present an implementation of a variational EM algorithm. It is written using C++, parallelized, available under a GNU General Public License (version 3), and can select the optimal number of clusters using the ICL criteria. It allows us to analyze networks with ten thousand nodes in a reasonable time.

4.1 Introduction

Complex networks are being more and more studied in different domains such as social sciences and biology. The network representation of the data is graphically attractive, but there is clearly a need for a synthetic model, giving an enlightening representation of complex networks. Statistical methods have been developed for analyzing complex data such as networks in a way that could reveal underlying data patterns through some form of classification.

Unsupervised classification of the vertices of networks is a rapidly developing area with many applications in social and biological sciences. The underlying idea is that common connectivity behavior shared by several vertices leads to their grouping in one *meta-vertex*, without losing too much information. Thus, the initial complex network can be reduced to a simpler *meta-network*, with few *meta-vertices* connected by few *meta-edges*. Picard et al. (2009) show applications of this idea to biological networks and Nowicki and Snijders (2001) and Handcock et al. (2007) to social networks.

Model-based clustering methods model the heterogeneity between nodes by grouping the nodes into classes. The model used in this paper is an extension of the *stochastic block model* (SBM) (Nowicki and Snijders, 2001). This model assumes that the nodes are distributed into groups, and connectivity between nodes is driven by node group memberships.

SBM for non-binary graphs, with or without covariates has been introduced in Mariadassou et al. (2010). In this paper, a *variational Expectation–Maximization* algorithm has been used to estimate parameters and to predict groups.

This article introduces **wmixnet**, an implementation of the *variational expectation–maximization* algorithm for this extension of the *stochastic block model* with or without covariates for three families of laws of probability: Bernoulli, Poisson, Gaussian.

This implementation allows us to estimate parameters and to predict node groups and covariate effects for graphs which are valued or binary, directed or not, and with or without covariates.

4.2 SBM model with covariates

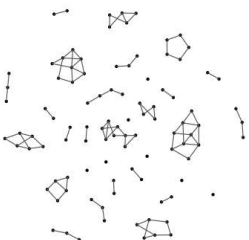
We introduce here the *stochastic block model* with covariates and three probability distributions.

4.2.1 Notations

Graph. Consider a graph $G = (V, E, w)$, where

- V is the set of nodes, labelled in $\{1, \dots, n\}$,
- E is the set of edges, which is a subset of V^2 ,
- $w: E \rightarrow \mathbf{R}$, is the function which gives edge weights.
- $Y: V^2 \rightarrow \mathbf{R}^p$, is the function which gives the covariate vector associated to each couple of nodes.

We assume *without loss of generality* that $E = V^2$, with the convention $w(i, j) = 0$ if there is no edge from vertex i to vertex j .



Groups. Consider Q classes of nodes. For a given partition $(\mathcal{C}_1, \dots, \mathcal{C}_Q)$ of V , for a node i and a group q , let Z be defined as $Z_{iq} = 1 \Leftrightarrow i \in \mathcal{C}_q$. And let $Z_i = (Z_{i1}, \dots, Z_{iQ})$.

4.2.2 The model

Nodes. The class memberships of the nodes are driven by independent identically distributed multinomial distributions:

$$\forall i \in V \quad Z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{M}(1, \alpha)$$

where $\alpha = (\alpha_1, \dots, \alpha_Q)$ and $\sum_q \alpha_q = 1$.

Edges. For each couple of nodes (i, j) the probability law of the link is driven by their class memberships and the (i, j) covariate $Y(i, j)$:

$$(w(i, j) | (i, j) \in \mathcal{C}_q \times \mathcal{C}_l) \sim \mathcal{F}_{ql}(Y(i, j)).$$

4.2.3 Probability laws

Generally, various probability laws can be used. The probability distributions which are implemented in `wmixnet` are the following:

— *Bernoulli*:

without covariates: $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{B}(\pi_{ql})$. This model does not use covariates and can model only binary networks. This is the classical *stochastic block model* model.

with covariates (with homogeneous effects): $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{B}(\pi_{ql} \frac{1}{1 + \exp(-\beta^T Y_{ij})})$. This model uses covariates and can model only binary networks. The effect of covariates is the same for all pairs of classes.

with covariates (with heterogeneous effects): $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{B}(\pi_{ql} \frac{1}{1 + \exp(-\beta_{ql}^T Y_{ij})})$. This model uses covariates and can model only binary networks. The effect of covariates is *not* the same for all pairs of classes.

— *Poisson*:

without covariates: $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{P}(\lambda_{ql})$. This model does not use covariates and can model networks with non negative integer weights.

with covariates (with homogeneous effects): $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{P}(\lambda_{ql}(Y(i, j)))$ where $\lambda_{ql}(Y(i, j)) = \lambda_{ql} \exp(\beta^T Y(i, j))$. This model uses covariates and can model networks with non negative integer weight. The effect of covariates is the same for all pairs of classes.

with covariates (with heterogeneous effects): $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{P}(\lambda_{ql}(Y(i, j)))$ where $\lambda_{ql}(Y(i, j)) = \lambda_{ql} \exp(\beta_{ql}^T Y(i, j))$. This model uses covariates and can

model networks with non negative integer weight. The effect of covariates is *not* the same for all pairs of classes.

— *Gaussian*:

without covariates: $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{N}(\mu_{ql}, \sigma^2)$. This model does not use covariates and can model networks with real weight.

with covariates (with homogeneous effects): $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{N}(\mu_{ql}(Y(i, j)), \sigma^2)$ where $\mu_{ql}(Y(i, j)) = \mu_{ql} + \beta^T Y(i, j)$. This model uses covariates and can model networks with real weight. The effect of covariates is the same for all pair of classes.

with covariates (with heterogeneous effects): $\mathcal{F}_{ql}(Y(i, j)) = \mathcal{N}(\mu_{ql}(Y(i, j)), \sigma^2)$ where $\mu_{ql}(Y(i, j)) = \mu_{ql} + \beta_{ql}^T Y(i, j)$. This model uses covariates and can model networks with real weight. The effect of covariates is *not* the same for all pair of classes.

4.2.4 Analysis of groups when covariates are used

Without covariates, groups are sets of nodes which have the same connectivity behavior (in probability), and groups can be easily interpretable using the connectivity matrix ($[\pi_{ql}]$, $[\lambda_{ql}]$ or $[\mu_{ql}]$).

With covariates, groups are sets of nodes which have the same connectivity behavior (in probability) *conditionally to covariates*. Two nodes of the same group can have different connectivity behavior due to different values of covariates.

For a model with covariates, groups are covariate-residual groups. There are two points of view:

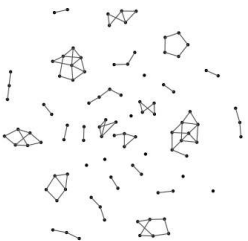
- the focus is on the effects of the covariates and groups model the (residual) connectivity which is *not* explained by covariates,
- the focus is on the groups which helps in suggesting some sources of heterogeneity after correcting the artefact due to covariates.

One can test the effect of covariates using a likelihood ratio test between models with and without covariates.

4.3 Estimation method

The estimation method is described in Mariadassou et al. (2010). The likelihood is not computable in a reasonable time, and a variational approximation is done and a *variational expectation-maximization* is used. The ICL criterion is used for choosing the number of groups, see Mariadassou et al. (2010).

Some estimation implementation details which differ from the framework introduced in Mariadassou et al. (2010) are explained here.



4.3.1 Initialization

As in the general case on *expectation-maximization* algorithm, the initialization plays a major role in the quality of the local maximum found.

In Mariadassou et al. (2010), the authors propose to use a hierarchical clustering to initialize the algorithm. In a real case of network analysis this initialization is often an extremal one (most of the initialized groups contain only one node) and the *expectation-maximization* algorithm converges to a local maximum which may be far from the global maximum.

The Absolute Value Spectral Clustering algorithm is consistent for finding groups in SBM (with Bernoulli probability law without covariates), see Rohe et al. (2011). We use the absolute spectral clustering to initialize the *expectation-maximization* algorithm.

When there are covariates, the spectral clustering is done on the residual graph, after eliminating the effect of covariates by regression.

4.3.2 Smoothing

To determine if an estimation for Q groups has reached a bad local maximum, we use two findings:

- With an ascending number of groups, models are nested. A model with Q groups can be interpreted as a model with $Q + 1$ groups, so the likelihood must increase with Q .
- Empirical findings make us say that the ICL criterion is convex.

A reinitialization of the *expectation-maximization* can be done. The new initialization is obtained in two ways:

- merging two groups of the $Q + 1$ result (descend mode)
- splitting one group into two groups of the $Q - 1$ result (ascend mode), this split is done by a spectral clustering of the residual graph on $Q - 1$ groups.

There are two modes of reinitialization:

- the **minimal** one, reinitializations are done each time one of the two findings (see above) is not respected,
- the **exhaustive** one, all reinitializations are done; while it improves likelihood, this option is very time-consuming and cannot be used with non small graphs.

4.3.3 Parallelism

Many steps of the estimation can be done independently:

- The *expectation-maximization* algorithm for various Q
- Reinitialization in ascend and descend mode

Considering that computers and computing units have more than one logical processor, this implementation uses threads to parallelize the implementation as much as possible.

4.4 wmixnet program

This section introduces the `wmixnet` program and the program usage.

4.4.1 Sources availability and installation

`wmixnet` is provided on the GNU General Public Licence version 3, and C++ sources are available on the `wmixnet` page:

- <http://www.agroparistech.fr/mia/productions:logiciels>
- <http://www.agroparistech.fr/mia/productions:logiciel:wmixnet>

`wmixnet` should be installable from sources on any Linux distribution, when dependencies are provided:

- IT++ library, used for matrix calculation. This library uses `blas` and `lapack`, well-known algebra libraries.
- `boost` library, for many aspects including parallelism.

4.4.2 Input format

The input format is a plain text with the following specifications:

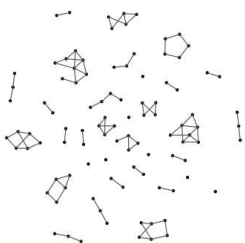
- each line describes a node
- for each line the first two columns describe the indexes of starting and ending nodes
- for each line the third column describes the weight of the edge
- for each line the fourth to end columns (if present) describe the covariates associated to the edge.

There are some constraints:

- node indexes must start from 1 to the number of nodes
- each edge must have the same number of covariates.

If an edge is not present, and if no covariates are used, the corresponding lines can be omitted; otherwise the line must be present with a weight of zero.

Functions are provided to write a file following these specifications, with adjacency matrices, and covariate matrices, for GNU R, and MATLAB or GNU Octave.



4.4.3 Output format

The output format contains the model parameters for all explored numbers of groups.

Model parameters are:

- α , the parameters of the multinomial distribution
- θ , the parameters of the probability law of the edge weight conditionally to groups of nodes,
 - for the Bernoulli model, $\theta = (\pi)$,
 - for the Poisson model, $\theta = (\lambda)$,
 - for the Poisson model with covariates $\theta = (\lambda, \beta)$,
 - for the Gaussian model $\theta = (\mu, \sigma^2)$,
 - for the Gaussian model with covariates $\theta = (\mu, \sigma^2, \beta)$.

The output contains variational parameter estimates (τ) which give the nodes membership in groups.

The output also contains values of criteria such as pseudo-likelihood and the ICL criterion.

There are three output formats provided:

- Plain text output format (named `text`), which is a human readable file.
- GNU R file output format, which is an GNU R loadable file. Nevertheless this file can be easily read by a human.
- MATLAB or GNU Octave file output format, which is a MATLAB and GNU Octave loadable file. Nevertheless this file can be easily read by a human.

4.4.4 Command line usage

`wmixnet` is usable with command line, and the following arguments must be provided:

- `--input` to specify the input file,
- `--symmetric` to indicate if the graph is an undirected graph if applicable,
- `--model` to specify the model in
 - `bernoulli` for Bernoulli without covariate
 - `BH` for Bernoulli with covariates (homogeneous effects)
 - `BI` for Bernoulli with covariates (heterogeneous effects)
 - `poisson` for Poisson without covariate
 - `PRMH` for Poisson with covariates (homogeneous effects)
 - `PRMI` for Poisson with covariates (heterogeneous effects)
 - `gaussian` for Gaussian without covariate
 - `GRMH` for Gaussian with covariates (homogeneous effects)
 - `GRMI` for Gaussian with covariates (heterogeneous effects)

- `--Qmax` to specify the maximum number of groups, or `--Qauto` to let the program choose the maximum number of groups,
- `--smoothing` to specify the smoothing mode
 - `none` no reinitialization is done (by default)
 - `minimal` reinitializations are done for detected problems
 - `exhaustive` all reinitialization are done (time-consuming option, only for small graphs)
- `--output` to specify the output file,
- `--output-format` to specify the output format
 - `text` (by default)
 - `R` for GNU R loadable file
 - `matlab` or `octave` which are synonymous for MATLAB and GNU Octave loadable file.

4.4.5 Empirical complexity

Some simulations suggest the following estimation of complexity:

$$t = C_{\text{model}} n^{2.46} g^{2.1} 1.03^p$$

with

- t the total processor time (equivalent time on a mono-core computer, without parallelization, which executes only this job)
- C_{model} a constant which depends on the model. Since absolute values are not pertinent, ratios are given:

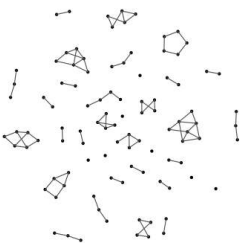
$$\begin{aligned} \frac{C_{\text{poisson}}}{C_{\text{bernoulli}}} &= 3.9 \\ \frac{C_{\text{PRMH}}}{C_{\text{bernoulli}}} &= 21 \\ \frac{C_{\text{gaussian}}}{C_{\text{bernoulli}}} &= 840 \\ \frac{C_{\text{GRMH}}}{C_{\text{bernoulli}}} &= 1350 \end{aligned}$$

This ratio is dependent on the way each model is implemented. Some models allow us to vectorize some steps, have explicit maxima, and thus are significantly faster

- n the number of nodes
- g the number of groups found
- p the number of covariates (the size of the covariate vector)

4.4.6 Capacity of extension

In the `wmixnet` program, the estimation procedure and other model-common parts are implemented once. Only model-specific functions are present for each model. Therefore it is relatively easy to add other models in the `wmixnet` program.



4.5 Example

Here we introduce the analysis of two ecological networks, obtained from one host–parasite network from the Département Français des Forêts database. These networks are the same as the ones used in Mariadassou et al. (2010) to illustrate the estimation method, refer to Mariadassou et al. (2010) for details.

4.5.1 Projected networks

We considered two undirected, valued networks having parasitic fungal species ($n = 154$) and tree species ($n = 51$) as nodes, respectively. Edge strengths was defined as the number of shared host species and the number of shared parasitic species, respectively.

4.5.2 Covariate data

For each network we have covariates for each pair of nodes:

- Taxonomic distance

For trees, based on the NCBI Taxonomy Browser, species are evenly distributed over two taxonomic classes (Magnoliophyta and Coniferophyta) and further subdivided in 8 orders, 13 families and 26 genera. Following Poulin (2005), we considered that the taxonomic distance is equal to 0 if species are the same, 1 if they belong to the same genus, 2 to the same family, 3 to the same order, 4 to the same taxonomic class and 5 if their only point in common lies in belonging to the phylum Streptophyta.

For fungi, with the same considerations as trees, the taxonomic distance between species is computed and has values in the range from 0 to 6.

- For the tree species network, a geographical distance can be introduced between each pair of tree species. The geographical distance is the Jaccard distance computed on the profiles of presence/absence in 309 geographical units covering the entire French territory.

4.5.3 Example of command line

For the analysis of the tree species network, for the Poisson model without covariates, the command line is:

```
wmixnet --input Trees.spm --symmetric \  
  --model poisson \  
  --Qauto --smoothing exhaustive \  
  --output Trees.m --output-format octave
```

4.5.4 Results

On the tree species network

In Figure 4.1, we plot the ICL criterion for Poisson model without and with covariates (taxonomic distance, geographical distance or both). For the model without covariates the maximum is reached with 7 groups. With the geographical covariates, the maximum is reached for 6 groups, with a little improvement of the ICL criterion. For the model with the taxonomic covariates, the maximum is reached for 4 groups, with a larger improvement of the ICL criterion. Adding the geographical covariates to the taxonomic covariates does not improve the criterion.

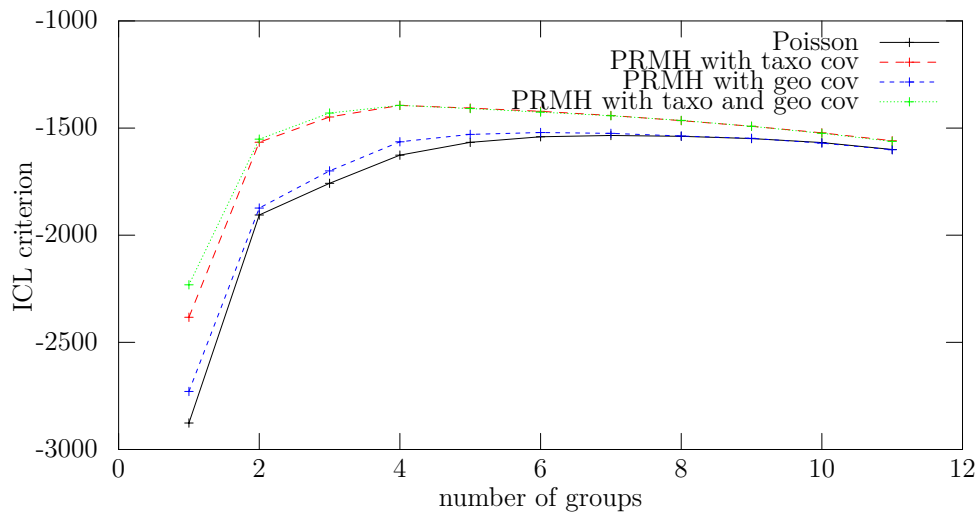
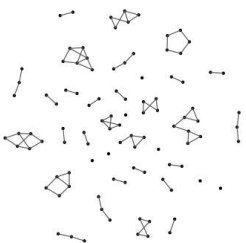


Figure 4.1: ICL criterion values obtained for Poisson and Poisson with covariates on the trees network

Tree species are divided into two taxonomic classes *Coniferophyta* and *Magnoliophyta*, and we can observe the repartition of each class into groups for each analysis. In Table 4.1, we present proportions obtained with the best (for the ICL criterion) number of groups for the Poisson model without covariates. We can see groups which contain in most cases only trees from one class, so the information extracted by groups contains the taxonomic information. In Table 4.2, we present proportions obtained with the best number of groups (for the ICL criterion) for the Poisson model with the taxonomy covariate. There is less information extracted (4 groups instead of 7) but group information is not redundant with taxonomy information.



Group	1	2	3	4	5	6	7
Group proportion (%)	11.8	17.6	17.8	7.9	13.7	11.6	19.5
Coniferophyta proportion (%)	100	22.2	22	0.1	0	84.5	100
Magnoliophyta proportion (%)	0	77.8	78	99.9	100	15.5	0

Table 4.1: Group proportions and group composition for the analysis with the Poisson model without covariate on the tree species network.

Group	1	2	3	4
Group proportion (%)	15.7	21.6	37.2	25.5
Coniferophyta proportion (%)	25	63.6	47.4	53.9
Magnoliophyta proportion (%)	75	36.4	52.6	46.1

Table 4.2: Group proportions and group composition for the analysis with the Poisson model (PRMH) with taxonomic covariate on the tree species network.

On the fungi network

In Figure 4.2, we plot the ICL criterion for the Poisson model without and with covariates (taxonomic distance). For the model with and without taxonomic covariates the maximum is reached with 15 groups in both cases. There is no real improvement by adding the taxonomic covariates to the model, and one conclusion may be that taxonomic covariates are not able to explain a part of the fungus species graph structure.

This conclusion can be made using the difference in the ICL criterion, which contains the difference in the log-likelihood and a penalty for the number of parameters, or a *test* can be done on the log-likelihood difference.

Acknowledgments

We thank the Département Santé des Forêts (DSF) of the French Ministère de l'Agriculture et de la Pêche for allowing us to use their database.

We thank Pierre Barbillon for the Bernoulli model with covariates (BH and BI) and for helping find some bugs in the estimation procedure.

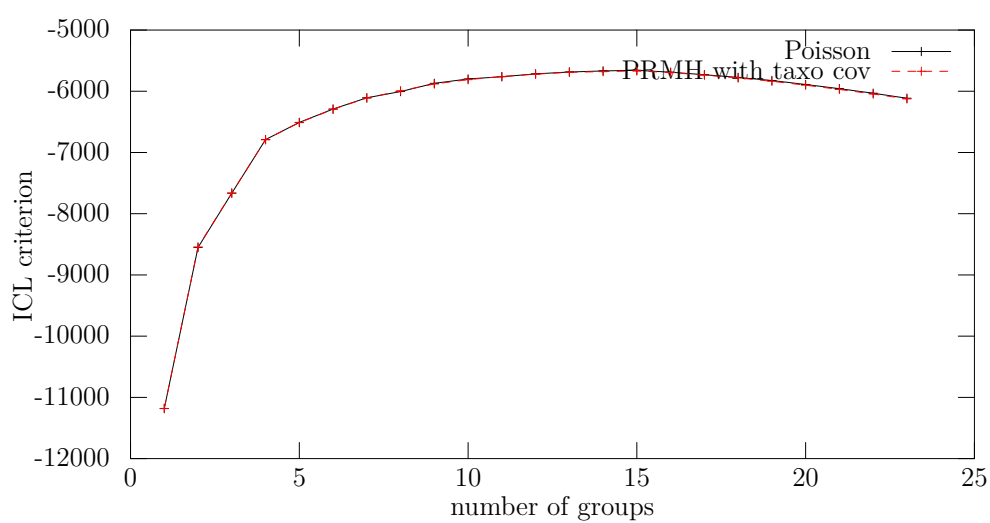
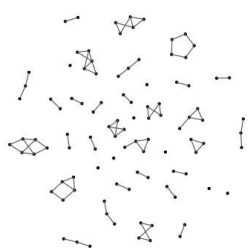
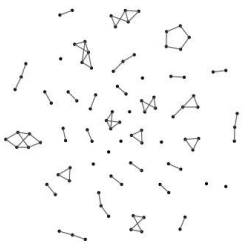


Figure 4.2: ICL criterion values obtained for the Poisson model and the Poisson with covariates model on the fungus species network.



Bibliography

- M. Handcock, A. Raftery, and J. Tantrum. Model-based clustering for social networks. *JRSSA*, 54:301–354, 2007.
- M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *Ann Appl Stat*, 4:715–742, 2010.
- K. Nowicki and T. Snijders. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, 96:1077–1087, 2001.
- F. Picard, V. Miele, J.-J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics*, 10, 2009.
- R. Poulin. Relative infection levels and taxonomic distances among the host species used by a parasite: insights into parasite specialization. *Parasitology*, 130(01):109–115, 2005.
- K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.



Chapter 5

Deciphering the mechanisms shaping ecological networks: a framework and a method

Abstract

Most methods for identifying communities in ecological networks cannot integrate additional information or covariates such as sampling effort. In this study we present a statistical model, the Stochastic Block Model with covariates, applicable to weighted bipartite networks, that is able to retrieve the structure of the network and integrate covariates. It can be used to evaluate the relative contribution of several factors potentially shaping the interactions between species. We will apply this model to two antagonistic networks, one host-parasite network and one plant-insect network. ^a

5.1 Introduction

Understanding whom interacts with whom and why is a salient question in ecology, as evidenced by the proliferation of studies on ecological networks in the

^a. This paper is in preparation and will be completed on three points:

- Two networks are considered. Only the results on the tree-fungal network are presented here. Some more work is needed for the tree-insect network.
- One of the sampling covariates (Sampling Antagonist, SA) is not yet integrated in the analysis. However the complete model, including the covariate SA, is presented in the method section.
- The result and discussion sections have to be completed.

past 20 years (Ings and Chittka, 2008). In the case of food webs, several conceptual models (reviewed by Capitán et al. 2013), often based on species body size, have been proposed to explain the complex structure of the observed interactions. In the case of host-parasite interactions, which are usually represented by bipartite networks, species body size is not a relevant trait for predicting interactions. Other conceptual models have thus been proposed.

For instance, Combes (2001) proposed the concept of filters to summarize the mechanisms shaping host-parasite interactions. Two species may interact with each other if they pass over the encounter filter and then the compatibility filter. The mechanisms acting on the encounter filter include the degree of overlap in the species geographical range, the degree of similarity in their habitat requirements and the degree of temporal synchronization between the interacting phases of the two species. Species behavior, when it exists, may also prevent or favor the encounter. The mechanisms acting on the compatibility filter include the quality of the resource and the degree of complementarity between the interaction traits of the two species.

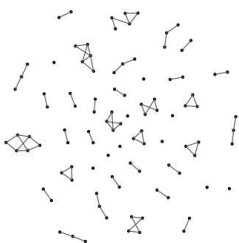
Vázquez et al. (2009) proposed a similar framework in the case of plant-mutualist interactions, with some mechanisms depending on the geographical distribution of the species and some mechanisms depending on the traits of the interacting species. He also added the effect of sampling on the observed network structure. In the present study, we considered a similar but simplified framework, with three main mechanisms acting on the observed network structure (Fig. 5.1) the probability of species encounter, the degree of species compatibility and the sampling effort.

Our aim is to evaluate the relative contribution of each of these three forces on the structure of weighted, bipartite ecological networks. For that, we developed a statistical model, applicable to weighted bipartite networks, that can be used to evaluate the relative contribution of several factors potentially shaping the interactions between species. We apply it to two antagonistic networks, one host-parasite network and one plant-insect network.

5.2 Materials and methods

5.2.1 The network data

The network data were collected by the French Department of Forest Health (DSF) between 1986 and 2006, over whole France. By using the 10744 records of symptoms caused by parasitic fungal species to forest trees, we reconstructed a quantitative network of interactions between 46 tree species and 140 fungal species. By using the 28547 records of damage caused by herbivorous insects to forest trees,



we reconstructed a quantitative network of interactions between tree species and insect species. All the records were georeferenced on a grid with a resolution of 16km, by using the GPS point of the record (when available) or the name of the town in which the record was made. Values of interaction strength, defined as the total number of records for a species pair, ranged from 0 to 895 for the tree-fungus network and 0 to 3611 for the tree-insect network.

5.2.2 The statistical model

Let $X_{i,j}$ be the observed interaction strength between tree species i and antagonist species j . The probability distribution of $X_{i,j}$ is assumed to be a Poisson distribution with parameter $\lambda_{i,j}$. The model takes into account the three main mechanisms acting on the observed network structure: compatibility, encounter and sampling Figure 5.1.

Sampling We assumed that the sampling effort of tree species i depends on its total abundance and on its economic value. Indeed, DSF observers have a higher probability of observing abundant tree species and thus a higher probability of observing an interaction with an antagonist species. Similarly, higher economic value of a tree species induces more intense health monitoring for this species. We assumed that the abundance of tree species and their economic value are well synthesized by the total number of records ST_i for tree species i in the whole DSF database, including all types of tree health problems (Piou, pers. comm). The probability that foresters working for the DSF record an interaction between tree species i and antagonist species j may also depend on the total abundance of antagonist species j , on the visibility of the symptoms or damage it causes and on the knowledge of foresters about this species. We assumed that these variables are well synthesized by the total number of records SA_j for antagonist species j in the database. The sampling effort on the interaction between species i and j is assumed to be the result of the product of the two sampling efforts, denoted by $S_{i,j}$ and defined by

$$S_{i,j} = ST_i^{\beta_{ST}} SA_j^{\beta_{SA}}$$

with β_{ST} and β_{SA} parameters of the model. These parameters are unknown and were estimated using the network data.

Encounter We assumed that the frequency of encounter $E_{i,j}$ between tree species i and antagonist species j depends on their frequency of encounter in each of the 2083 quadrats of the grid in the DSF database. We assumed that the frequency of encounter between tree species i and antagonist species

j in quadrat k is proportional to the products of their abundances. The abundance $ET_{k,i}$ of tree species i in quadrat k was estimated by using the surface given by the Inventaire Forestier National (IFN).

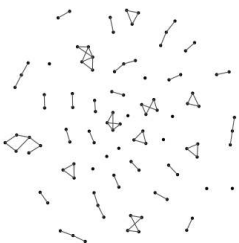
However, we noticed that some tree species were recorded as having a health problem in a given quadrat in the DSF database, although they were absent from this same quadrat according to the IFN database. This was particularly the case for tree species having a sparsed distribution or present in sub-urban areas, such as plane trees (*Platanus hybrida*). In order to correct the inconsistencies between the two databases, we added a grove of 10ha in every quadrat where a tree species was present according to the DSF database, but absent according to the IFN database. Overall, this addition hardly increased the total forest area (+0,47%).

The abundance $EA_{k,j}$ of antagonist species j in quadrat k was estimated by using the DSF database itself, since we did not have any other data source about the geographical distribution of antagonist species. We assumed that $EA_{k,j}$ was proportional to the total number of records of symptoms or damage caused by antagonist species j in quadrat k . Finally, we defined $E_{i,j}$ as:

$$E_{i,j} = \left(\sum_k ET_{k,i} EA_{k,j} \right)^{\beta_E}$$

where β_E is a parameter of the model. This parameter is unknown and was estimated using the network data.

Compatibility The degree of compatibility between tree species i and antagonist species j should theoretically be influenced by the degree of matching between their interaction traits. However, in practice, we do not have a list of the traits involved in the interactions, with their values, for the hundreds of species included in the networks. Thus no covariate is available for measuring the degree of compatibility between tree species and antagonist species. In such case it is possible to use latent (not observed but inferred using the network data) variables. Tree species and antagonist species were thus classified in Q_T and Q_A groups of species respectively. The probability distribution of the interaction strength between two species only depends on the groups to which these species belong. Hence, species belonging to the same group tend to have similar connectivity behaviors. This assumption is well suited to our data, since previous analyses have shown the existence of such groups in the studied networks. Specifically, closely related tree species tend to have similar parasitic fungal species (see Vacher et al. 2008; Vacher et al. 2010; Mariadassou et al. 2010 and Daudin et al., 2010). The discrete latent variable model is



$$\lambda_{i,j} \parallel (i \in G_q, j \in G_l) = \lambda_{q,l}$$

with $Q_T Q_A$ parameters $\lambda_{q,l}$. These parameters are unknown and were estimated using the network data.

However, this part of the model contains but cannot be reduced to the compatibility mechanism. It contains all the mechanisms shaping the network structure, except the sampling and encounter mechanisms.

Finally the statistical model including the three main mechanisms acting on the observed network structure (Fig. 5.1) is the following:

$$\begin{aligned} \lambda_{i,j} \parallel (i \in G_q, j \in G_l) &= \lambda_{q,l} E_{i,j} S_{i,j} \\ &= \lambda_{q,l} \left(\sum_k ET_{k,i} EA_{k,j} \right)^{\beta_E} ST_i^{\beta_{ST}} SA_j^{\beta_{SA}} \end{aligned}$$

This model is called Stochastic Block Model (SBM) with covariates.

5.2.3 Application of the model to the data

Missing data

The statistical treatment of the model does not allow missing data. Hence the covariates have to be known for each tree-antagonist species pair included in the analysis. Moreover the estimation step of the statistical model needs the network to be connected. Given these constraints, we applied the model to a subset of the data. The subset of the tree-fungus network was a connected component composed 41 tree species belonging to 4 families and 140 fungal species belonging to 25 families, with each family composed of at least three species. The subset of the tree-insect network was a connected component composed 48 tree species belonging to 10 families and 278 insect species belonging to 51 families, with each family composed of at least three species.

Contribution of each mechanism

The fit of a given model is given by the loglikelihood. However increasing the number of covariates and/or latent variables increases the loglikelihood. To correct this artifact, we used a penalized approximate loglikelihood criterion, called ICL (Mariadassou et al., 2010). Higher values of ICL correspond to a better fit. To assess the relative contribution of a mechanism to the network structure, we computed the difference between the ICL of the full model (including the three mechanisms) with the ICL of the model without this mechanisms.

Choice of the number of groups

The numbers of groups Q_T and Q_A were chosen using the ICL criteria for the full model.

Modification of the estimation algorithm for bipartite graph

The algorithm for SBM was modified to take into account the bipartite structure of the network.

Computing the association between latent groups and taxonomy

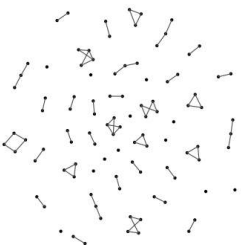
Cross-tabulation tables between latent groups and families were computed for each trophic level. P-values chi-square were computed to assess the statistical significance of the association. To account for small frequency values in the cross-tabulation, the P-values were computed using simulations (100000 replicates) in place of the usual chi-square asymptotic P-value, using the function *chisq.test* of the R-package.

5.3 Results

The higher ICL value was obtained for 10 groups (Figure 5.2). Host species and antagonist species were separated, giving 10 groups of tree species and 9 groups of antagonist species described (Table 5.1 and 5.2). A very significant association between tree species groups and genus was found (P-value = $4 \cdot 10^{-05}$). On the opposite the association between antagonist species groups and genus was not statistically significant (P-value = 0.268). The same findings were obtained for the associations between species groups and classes (P-values respectively equal to $1.8 \cdot 10^{-4}$ and $5.9 \cdot 10^{-2}$).

The model was able to predict the observed values of the network, as shown by the correlation between the predicted and observed values (Figure 5.3) and the comparison of the heatmaps of the observed and predicted adjacency matrices (Figure 5.4). The estimates of the parameters of the covariates are $\hat{\beta}_{ST} = 0.238$ and $\hat{\beta}_E = 1.02$, β_{SA} has not been estimated. The estimates of the parameters λ_{ql} are given in Figure 5.5.

The most important value de $\lambda_{q,l}$ was between *Castanea sativa* and *Fomitopsis cytisina* (between T_7 and A_8). This is an outlier, because interaction between these two species was observed only once. Given the low values for the sampling and encounter covariates, the observation of an interaction between these two species is highly improbable. This explains the high $\lambda_{7,8}$ estimated value. More interestingly, the strongest links (after correction by the encounter and sampling) were between



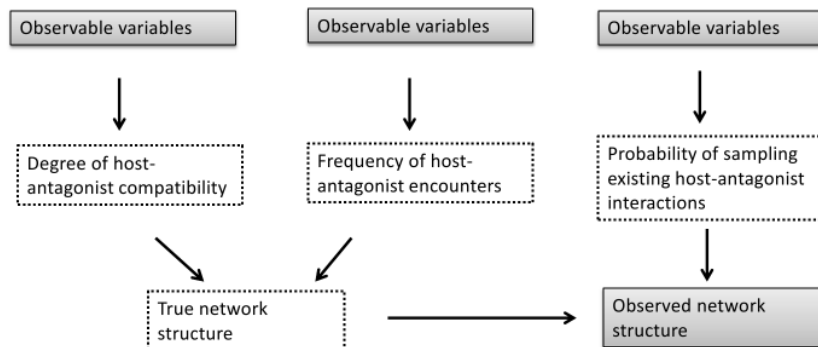


Figure 5.1: Variables potentially accounting for the probability that foresters working for the French Department of Forest Health (DSF) observe host-antagonist interactions. Observable variables are shown in grey boxes. Those which are included in the model are in bold. Latent variables, which cannot be observed nor directly measured, are shown in dotted boxes

Group	genus	Class
T1	Abies Carpinus Fraxinus Pinus(3) Tilia	Coniferopsida (3), Magnoliophyta(3)
T2	Abies Cedrus Picea	Coniferopsida (3)
T3	Picea Pinus Pseudotsuga Sorbus	Coniferopsida (3), Magnoliophyta
T4	Alnus Fagus Robinia	Magnoliophyta(3)
T5	Pinus Platanus Populus(6) Prunus	Coniferopsida, Magnoliophyta(8)
T6	Quercus(5)	Magnoliophyta(5)
T7	Castanea	Magnoliophyta
T8	Abies Larix Pinus(3)	Coniferopsida (5)
T9	Acer(2) Larix Quercus	Coniferopsida, Magnoliophyta(3)
T10	Acer Quercus	Magnoliophyta(2)

Table 5.1: Tree species composition of the groups obtained by the 10-groups model with covariate ST and SA .

Group	Class
A1	Pucciniomycetes(5), Leotiomycetes(4), Agaricomycetes(3), Sordariomycetes(2), Dothideomycetes
A2	Sordariomycetes(6), Dothideomycetes(5), Leotiomycetes(4), Agaricomycetes(4), Pucciniomycetes
A3	Agaricomycetes(7), Sordariomycetes(6), Leotiomycetes, Dothideomycetes
A4	Sordariomycetes(3), Leotiomycetes(3), Dothideomycetes(3), Dacrymycetes
A5	Sordariomycetes(11), Leotiomycetes(4), Pucciniomycetes(3), Taphrinomycetes(2), Agaricomycetes, Dothideomycetes
A6	Sordariomycetes(5), Dothideomycetes(2)
A8	Agaricomycetes
A9	Agaricomycetes(17), Sordariomycetes(12), Leotiomycetes(4), Dothideomycetes(3), Eurotiomycetes, Pucciniomycetes, Pezizomycetes, Taphrinomycetes
A10	Leotiomycetes(2), Dothideomycetes(2), Agaricomycetes

Table 5.2: Fungus species composition of the groups obtained by the 10-groups model with covariate ST and SA .

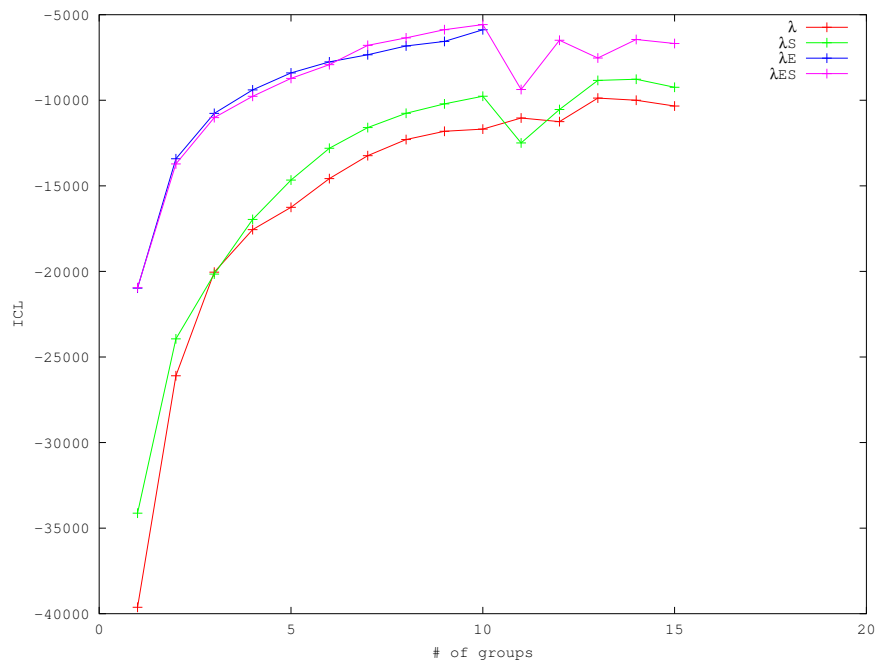
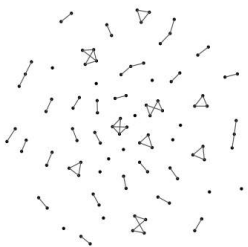


Figure 5.2: ICL Values obtained by the inference in function of the number of groups. Shown model are λ , for the model only with group effect, without covariates; λS , for the model with groups effect and sampling effect; λE , for the model with groups effect and encounter effect; λES , for the model with groups effect, sampling and encounter effect.



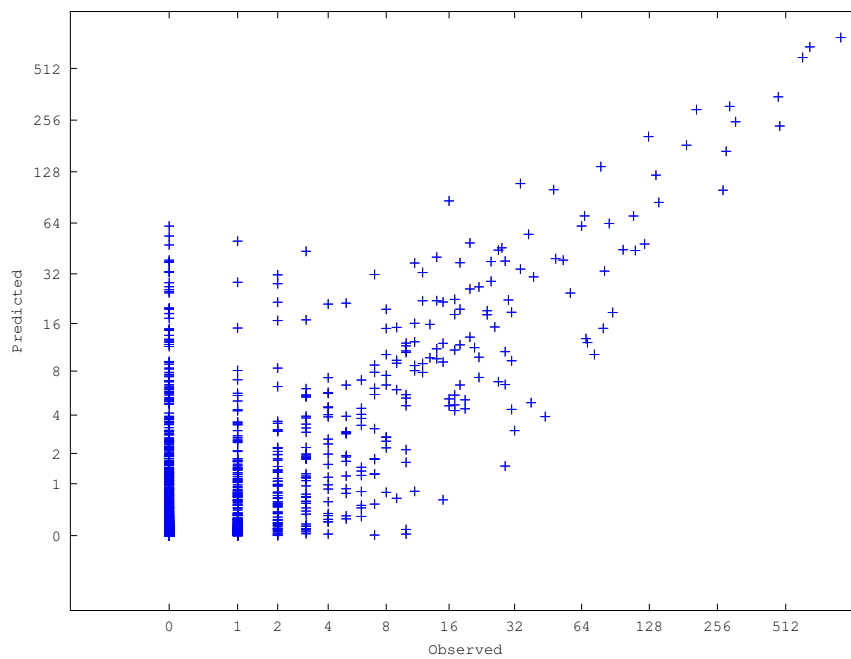


Figure 5.3: Observed interaction strenghts in the tree-fungus network against the predicted ones by the 10-groups model with covariates ST and SA .

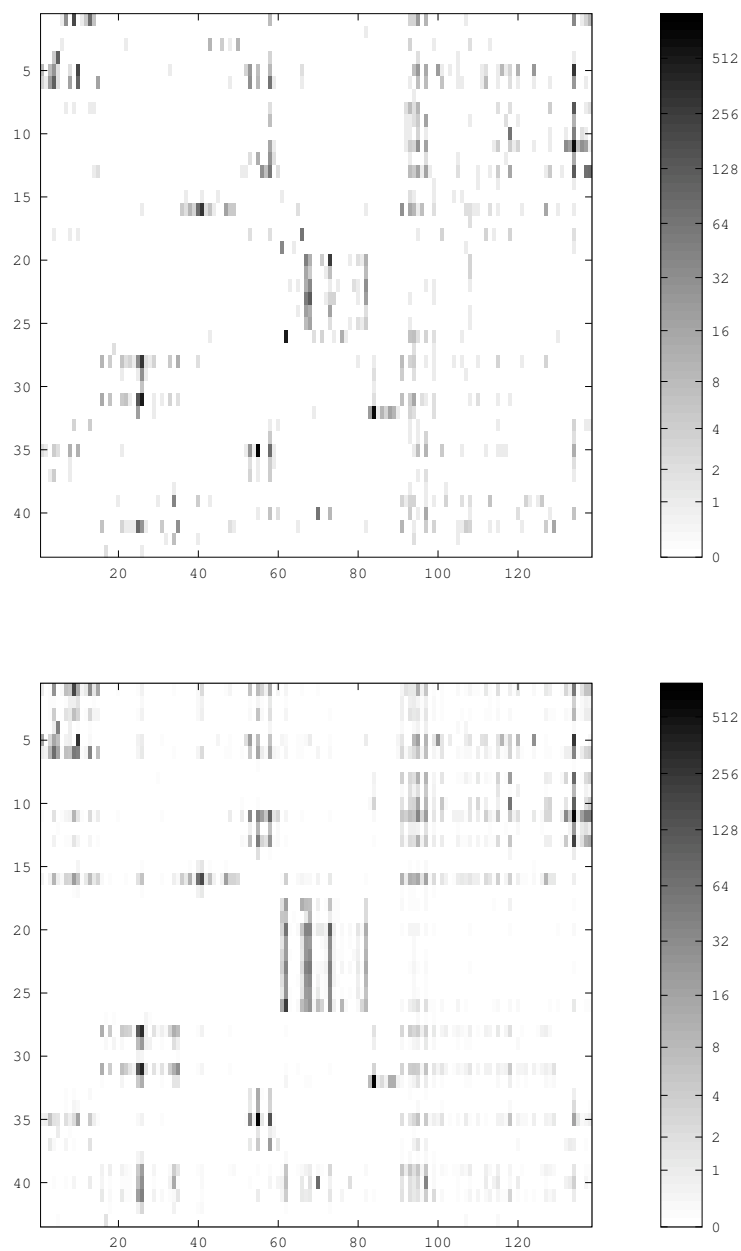
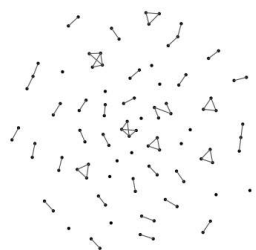


Figure 5.4: Observed interaction strenghts in the tree-fungus network (*top*). Predicted interaction strenghts in the tree-fungus network (*bottom*) by the 10-groups model with covariates ST and SA . Species are ordered by the group membership obtained by the inference.



groups T5 and group A5. One may notice also the strong links between groups 4 and 8 and 9 and 2. Low values for $\lambda_{q,l}$ underline the low level of interaction or the absence of interaction between some groups of tree species and some groups of antagonist species even when probability of encounter between these species is high. This may be accounted for by a lack of compatibility between these groups of species. For example few or no interactions are present between (T1, T4, T10) and A6, T4 and A4.

5.3.1 Relative importance of the three mechanisms

The table gives the ICL for different models. The quantitative importance of each mechanism is summarized by the differences between ICLs: 302.0 for the sampling, 3610.6 for the encounter and 15372.38 for the other mechanisms including compatibility.

5.4 Ongoing work and discussion

The results for the tree-insect network have not been obtained yet, due to the very long computation time. Results obtained for the tree-fungus network suggest that the main mechanism is the compatibility followed by the encounter and that the sampling has a marginal effect on the structure of network. This preliminary interpretation of the results will be confirmed later, because the interpretation of $\lambda_{q,l}$ values requires further discussion. estimated well enough. Applying the method to simpler networks, with better known covariates, would be useful.

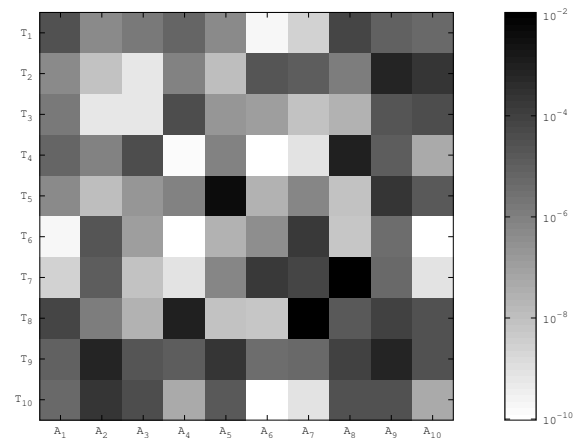
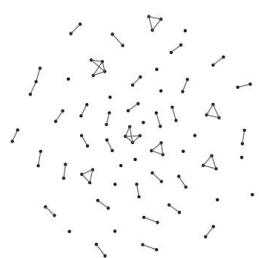


Figure 5.5: The estimates of the parameters λ_{qi} for the 10-groups model with covariate ST and SA .

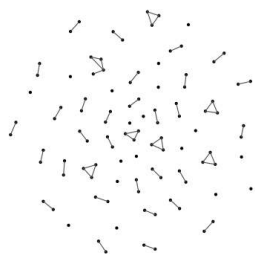
Model	ICL
compatibility,encounter, sampling	-5578.72
compatibility,encounter	-5880.75
compatibility,sampling	-9491.36
encounter, sampling	-20951.1
none	-39627.7

Table 5.3: ICL values obtained for the different models with 10 groups.



Bibliography

- J. A. Capitán, A. Arenas, and R. Guimerà. Degree of intervality of food webs: from body-size data to models. *Journal of Theoretical Biology*, 2013.
- C. Combes. *Parasitism*. University of Chicago Press, 2001.
- J.-J. Daudin, L. Pierre, and C. Vacher. Model for Heterogeneous Random Networks Using Continuous Latent Variables and an Application to a Tree-Fungus Network. *Biometrics*, 66(4):1043–51, 2010.
- T. C. Ings and L. Chittka. Speed-accuracy tradeoffs and false alarms in bee responses to cryptic predators. *Current Biology*, 18(19):1520–1524, 2008.
- M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, 4(2):715–742, 2010.
- C. Vacher, D. Piou, and M.-L. Desprez-Loustau. Architecture of an antagonistic tree/fungus network: the asymmetric influence of past evolutionary history. *PLoS One*, 3(3):e1740, 2008.
- C. Vacher, J.-J. Daudin, D. Piou, and M.-L. Desprez-Loustau. Ecological integration of alien species into a tree-parasitic fungus network. *Biological Invasions*, 12(9):3249–3259, 2010.
- D. P. Vázquez, N. Blüthgen, L. Cagnolo, and N. P. Chacoff. Uniting pattern and process in plant–animal mutualistic networks: a review. *Annals of Botany*, 103(9):1445–1457, 2009.



Appendix A

Putting the Biological Species Concept to the Test: Using Mating Networks to Delimit Species

Putting the Biological Species Concept to the Test: Using Mating Networks to Delimit Species

Lélia Lagache^{1,2}, Jean-Benoist Leger^{3,4}, Jean-Jacques Daudin^{3,4}, Rémy J. Petit^{1,2}, Corinne Vacher^{1,2*}

1 INRA, UMR1202 BIOGECO, Cestas, France, **2** Univ. Bordeaux, BIOGECO, UMR 1202, Talence, France, **3** INRA, UMR 518 MIA, Paris, France, **4** AgroParisTech, UMR 518 MIA, Paris, France

Abstract

Although interfertility is the key criterion upon which Mayr's biological species concept is based, it has never been applied directly to delimit species under natural conditions. Our study fills this gap. We used the interfertility criterion to delimit two closely related oak species in a forest stand by analyzing the network of natural mating events between individuals. The results reveal two groups of fertile individuals connected by only few mating events. These two groups were largely congruent with those determined using other criteria (morphological similarity, genotypic similarity and individual relatedness). Our study, therefore, shows that the analysis of mating networks is an effective method to delimit species based on the interfertility criterion, provided that adequate network data can be assembled. Our study also shows that although species boundaries are highly congruent across methods of species delimitation, they are not exactly the same. Most of the differences stem from assignment of individuals to an intermediate category. The discrepancies between methods may reflect a biological reality. Indeed, the interfertility criterion is an environment-dependant criterion as species abundances typically affect rates of hybridization under natural conditions. Thus, the methods of species delimitation based on the interfertility criterion are expected to give results slightly different from those based on environment-independent criteria (such as the genotypic similarity criteria). However, whatever the criterion chosen, the challenge we face when delimiting species is to summarize continuous but non-uniform variations in biological diversity. The grade of membership model that we use in this study appears as an appropriate tool.

Citation: Lagache L, Leger J-B, Daudin J-J, Petit RJ, Vacher C (2013) Putting the Biological Species Concept to the Test: Using Mating Networks to Delimit Species. PLoS ONE 8(6): e68267. doi:10.1371/journal.pone.0068267

Editor: Enrico Scalas, Università del Piemonte Orientale, Italy

Received: January 17, 2013; **Accepted:** May 28, 2013; **Published:** June 20, 2013

Copyright: © 2013 Lagache et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funding was provided by the LinkTree project (ANR BIODIVERSA), by the EU Network of Excellence EvoTree, by ANR-10-EQPX-16 XYLOFOREST and by the INRA (CJS grant). The genotyping was funded by grants from the Conseil Régional d'Aquitaine n°20030304002FA, n°20040305003FA and from the European Union, FEDER n°2003227. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: corinne.vacher@pierroton.inra.fr

Introduction

According to the biological species concept, the ability to interbreed (i.e. interfertility) is a defining property of species [1]. Yet, to our knowledge, the interfertility criterion has never been used to delimit species on the basis of mating events observed under natural conditions. Only artificial crosses have been used for this purpose, including in fungi (e.g. [2]), plants [3], or insects [4]. However, this approach has been criticized (e.g. [5,6]) because artificial crosses bypass some pre-mating barriers to hybridization: mating events observed under artificial conditions might not reflect what would naturally occur. Hence, to date, there is no satisfactory example of the use of the interfertility criterion to delimit species. In fact, the methods used most frequently for species delimitation are not derived from the well-known biological species concept but are derived from other concepts such as the phylogenetic species concept,

the genotypic species concept and the morphological species concept. Species definitions according to these concepts and possible associated criteria for species delimitation are listed in Table 1.

One potential method of species delimitation based on the interfertility criterion is the analysis of mating networks. Mating networks represent mating events between individuals [7]. Nodes of the network represent the individuals and links connect the individuals between whom mating events have occurred. Applying methods of network clustering [8–10] to mating networks should allow the identification of subsets of strongly interconnected nodes that correspond to species. If the biological species concept is strictly interpreted, then a species should correspond to a connected component of the mating network (Figure 1A). A connected component is a subset of nodes within the network that are directly or indirectly connected but are not connected to nodes not contained in the

Table 1. Major species concepts with associated possible criterion for species delimitation.

Species concept	Species definition according to this concept	Possible criterion for species delimitation derived from this definition	Possible method of species delimitation using this criterion	First application of this method at the study site
Biological species concept	Species are "groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups" [1]. According to Hausdorf [17], "natural populations" can be replaced by "individuals" in this statement without change of meaning.	Higher natural interfertility between individuals within than among species	Clustering of the network of natural mating events between individuals using Continuous Stochastic Block Model (C-SBM) [24].	this study
Phylogenetic species concept	A species is "a diagnosable cluster of individuals within which there is a parental pattern of ancestry and descent, beyond which there is not, and which exhibits a pattern of phylogenetic ancestry and descent among units of like kind" [29].	Higher genetic relatedness between individuals within than among species	Clustering of the network of relatedness relationships between individuals using C-SBM [24].	this study
Genotypic species concept	A species is a "genotypic cluster of individuals that can overlap without fusing with its siblings" [17,52]	Higher genotypic similarity between individuals within a species	Clustering of the individuals based on their multilocus genotype with STRUCTURE [50]	Guichoux <i>et al.</i> 2012 [23]
Morphological species concept	Species are "the smallest detected samples of self-perpetuating organisms that have unique sets of characters" [53,54].	Higher morphological similarity between individuals within than among species	Clustering of the individuals based on several morphological traits with a factorial discriminant analysis [55].	Bacilieri <i>et al.</i> 1996 [26]

subset. According to a relaxed biological species concept, which allows for some level of hybridization between species [11–13], a species should correspond to a community in the mating network (Figure 1B). Communities are subsets of nodes with a high density of links within the group and a lower density of links between different groups [8]. It is in this latter case, when species hybridize, that species delimitation based on the interfertility criterion is particularly challenging and network analysis may be particularly useful.

The idea of analyzing mating networks to delimit species according to the biological species concept was proposed more than 40 years ago by Sokal and Crovello [14] but it does not appear to have been put into practice. Building a mating network is indeed a difficult task as it requires a very large data set of mating events collected under natural conditions. The species should be sympatric and have semi-permeable reproductive barriers so that the issue of species delimitation is relevant. Furthermore, the species should be not only outcrossing (with a low selfing rate) but also highly polygamous and have multiple offspring per generation so that actual mating events are representative of potential mating events between individuals at a given time [15–17]. If such data were available, would the analysis of mating networks be an effective method to delimit species based on the interfertility criterion? Would the boundaries between species be the same as those obtained using other species delimitation criteria?

To answer these questions, we investigate the congruence between four methods of species delimitation, derived from the biological, morphological, genotypic and phylogenetic species concepts (Table 1), by applying them to two hybridizing tree species living in sympatry. The study site is a 5 ha mixed stand of *Quercus robur* and *Q. petraea* comprising 298 adult trees originating from natural regeneration [18]. As many other closely related plant species [19], these two oak species

hybridize under natural conditions [20], including in the studied stand [21–23]. To delimit species according to the interfertility criterion, we analyze the network of observed natural mating events between pairs of adult trees by using a method of network clustering. Each node of the mating network corresponds to an adult tree and each link corresponds to at least one mating event between two trees. To cluster individuals, we selected among available methods of network clustering [8–10] the Continuous Stochastic Block Model (C-SBM) recently introduced by Daudin *et al.* [24]. C-SBM synthesizes the heterogeneity of a real network by producing a simplified version of the network composed of a few virtual nodes, called extremal hypothetical nodes (EHNs). Unlike many methods of network clustering, which assume that each node belongs to only one group, C-SBM allows nodes to exhibit mixed connectivity behavior by assuming that each node of the real network is a mixture of the EHNs. This method is thus particularly suited to our study. Indeed, because the two previously identified oak species [23,25] are known to hybridize [22], we expected to find some individuals with a mixed reproductive behavior, i.e. breeding with both species. The same method was used to delimit species based on genetic relatedness between individuals. In that case, each node of the network corresponds to an adult tree and links connect the individuals that are considered to be related based on their genotype. Finally, we compare individual assignments obtained by analyzing the mating network and the relatedness network with those previously obtained in the same study site using criteria of morphological and genotypic similarities [23,25]. We then discuss how to summarize continuous but non-uniform variations in biological diversity.

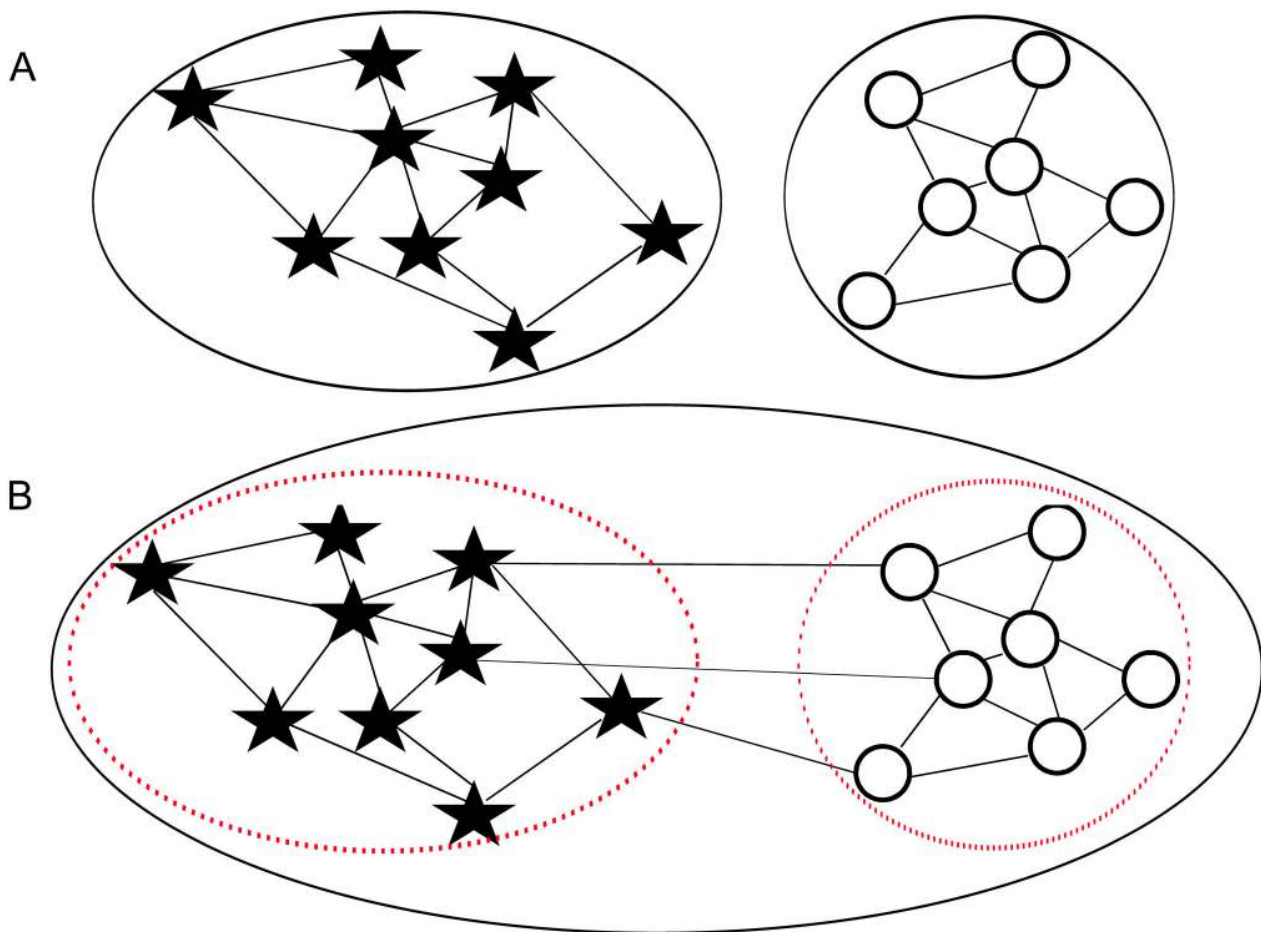


Figure 1. Example of mating networks with species boundaries. Each node of the network, represented by a black star or a white circle, is an individual. Each link of the network, represented by a thin black line, corresponds to a mating event between two individuals. In A, there is no mating event between the two groups of individuals whereas in B, a few mating events occur between groups. Species boundaries according to a strict application of the biological species concept are indicated by a continuous thick black line. Species boundaries according to a relaxed interpretation of the biological species concept, allowing interspecific hybridization, are indicated by a broken red line. In network theory, the continuous black line delimits the connected components of the network whereas the broken red line delimits communities.

doi: 10.1371/journal.pone.0068267.g001

Results

Species Delimitation based on Interfertility

According to the AIC criterion, the best model for the mating network was the one with four EHNs, followed by the models with five and three EHNs (Figure S1 in File S1). We selected the model with three EHNs because the two other models highlighted the structure of the sampling design (Text S1 in File S1). According to the connectivity matrix for the EHNs (Figure 2A), EHN_0 corresponds to a virtual node not connected to the whole network. This EHN, which is systematically present in the network models produced by C-SBM [24], makes it possible to take into account the variation in the number of links attached to the nodes of the real network. The two other

EHNs, called EHN_{B1} and EHN_{B2} , were strongly connected within themselves and were not connected to the other EHNs.

The nodes of the mating network (each corresponding to an individual) were then represented in a triangle, with one EHN at each point (Figure 2A). The higher the proportion of a given EHN in the mixture of a node, the closer the node was to this EHN in the triangle. According to the connectivity matrix for the EHNs (Figure 2A), the nodes that had a high proportion of EHN_0 in their mixture were weakly connected to the mating network. The nodes that had a high proportion of EHN_{B1} in their mixture belonged to a group of nodes strongly connected to each other and weakly connected to nodes with a high proportion of EHN_{B2} . Conversely, the nodes that had a high proportion of EHN_{B2} in their mixture belonged to a group of nodes strongly connected to each other and weakly connected

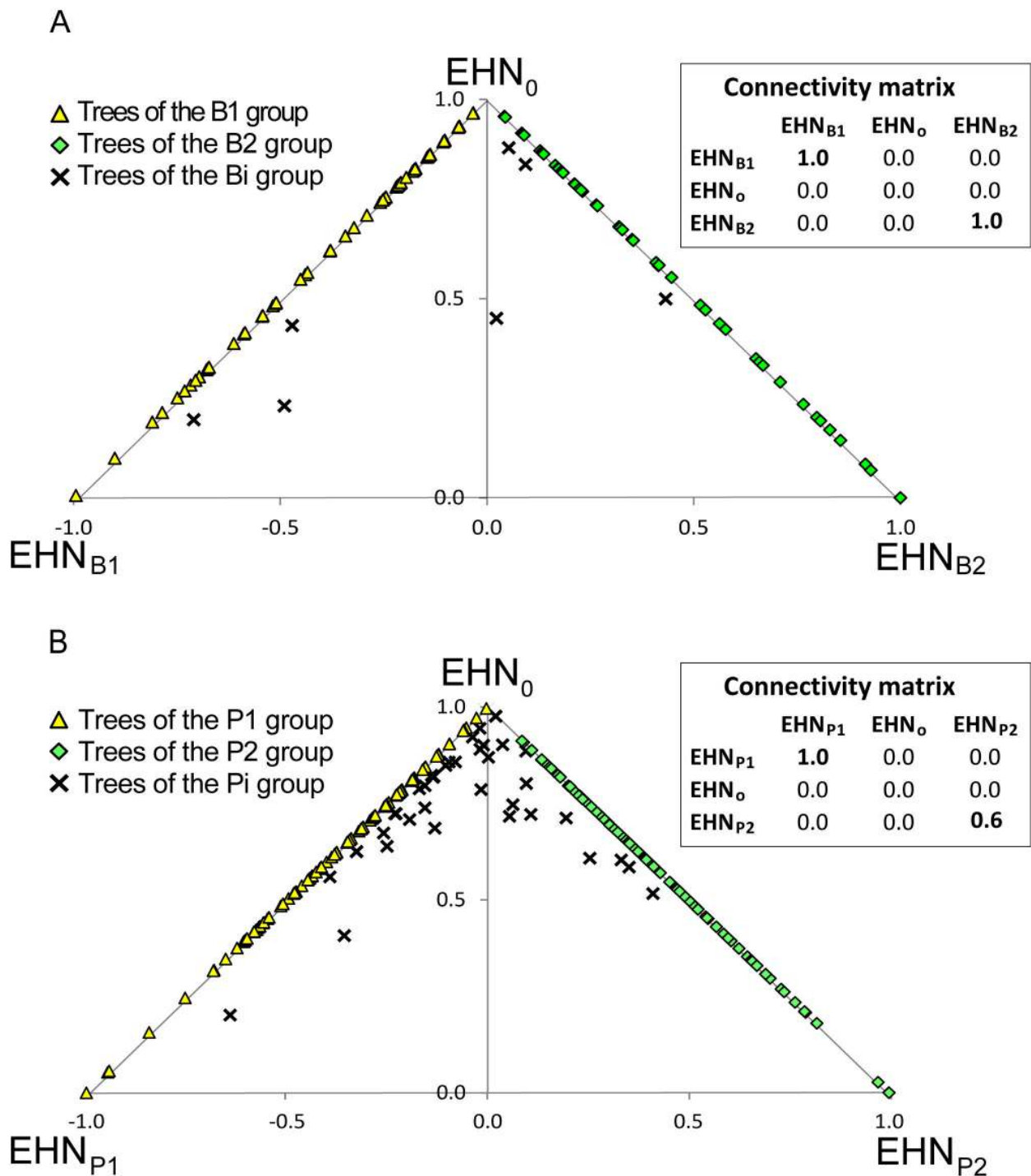


Figure 2. Triangular representation of the nodes of (A) the mating network and (B) the relatedness network, indicating the mixture of EHNs (i.e. Extremal Hypothetical Nodes) for each node according to C-SBM. In A, nodes that are on the edge between EHN_o and EHN_{B1} are classified in group B1 whilst nodes on the edge between EHN_o and EHN_{B2} are classified in group B2. Other individuals are classified as intermediates (group Bi). In B, nodes that are on the edge between EHN_o and EHN_{P1} are classified in group P1 whilst nodes on the edge between EHN_o and EHN_{P2} are classified in group P2. Other individuals are classified as intermediates (group Pi). Connectivity matrices for the EHNs are presented next to each triangular representation. Non-zero values are given in bold.

doi: 10.1371/journal.pone.0068267.g002

to nodes with a high proportion of EHN_{B1} . There were, therefore, two groups of adult trees in the mating network within which mating events were frequent and between which mating events were rare. The graphical representation of the network confirmed this result (Figure 3A). According to the relaxed interpretation of the biological species concept, these two groups of individuals should correspond to two biological species (Figure 1B).

In order to assign the individuals to the two species, we classified the nodes of the mating network according to their relative proportions of EHN_{B1} and EHN_{B2} . We assumed that an individual belonged to species B1 if the corresponding node was a mixture of EHN_0 and EHN_{B1} and only of these two nodes. Conversely, we assumed that an individual belonged to species B2 if the corresponding node was a mixture of EHN_0 and EHN_{B2} . Other individuals were classified as being reproductively intermediate (group Bi). In the triangular representation (Figure 2A), individuals assigned to species B1 were on the edge between EHN_0 and EHN_{B1} ($n=78$ individuals) whilst individuals assigned to species B2 were on the edge between EHN_0 and EHN_{B2} ($n=121$ individuals). Intermediate individuals were within the triangle ($n=7$ individuals). The three groups are shown in different colors in the network representation (Figure 3A).

Species Delimitation based on Relatedness

According to the AIC criterion, the optimal number of EHNs in the relatedness network was six. Models with three, four, five and seven EHNs were also good models (Figure S2 in File S1). As we did not find any satisfactory way to identify the best model (Text S2 in File S1), we selected the model with three EHNs to facilitate a comparison between the relatedness network structure and the mating network structure. According to the connectivity matrix for the EHNs (Figure 2B), EHN_0 corresponded to a virtual node not connected to the whole network. The two other EHNs, called EHN_{P1} and EHN_{P2} , were strongly connected within themselves and were not connected to the other EHNs. Like the mating network, the individuals were, therefore, classified into three groups called P1, P2 and Pi. Group P1 ($n=70$ individuals located on the edge between EHN_0 and EHN_{P1} in the triangular representation; Figure 2B) and group P2 ($n=108$ individuals located on the edge between EHN_0 and EHN_{P2} ; Figure 2B) comprised individuals with high within-group and low between-group degrees of relatedness. The third group Pi ($n=28$ individuals located within the triangle; Figure 2B) included trees related to both P1 and P2 individuals and trees with few relatives. The three groups are shown in different colors in the network representation (Figure 3B).

Species Delimitation based on Morphology and Multilocus Genotypes

The morphological similarity criterion has previously been used by Bacilieri *et al.* [26] to identify all trees from the study site. Based on their results, we assigned the individuals to two pure morphological groups (called M1 and M2 in this study and corresponding to *Q. robur* and *Q. petraea*, respectively) and to a morphologically intermediate class (called Mi). Guichoux *et al.* [23] used genotypic similarity as a criterion to assign the

trees of the study site to species. Based on their results, we classified the adult trees in two purebred groups (hereafter called G1 and G2) and one genetically intermediate class (Gi).

Congruence between the Four Methods of Species Delimitation

In order to assess the congruence between the four methods of species delimitation, we compared the spatial distribution of the three groups of individuals identified with each method. The species boundaries are very similar (Figure 4). Among the 206 adult trees included in the mating network and in the relatedness network, there were 97 trees classified consistently in the B1, P1, G1 and M1 groups and 63 trees classified consistently in the B2, P2, G2 and M2 groups. We therefore re-named groups B1, P1, G1 and M1 *Q. robur* and groups B2, P2, G2 and M2 *Q. petraea*. Based on this classification, there were only four species inversions associated with the delimitation methods (Table S1 in File S1). Among the 206 adult trees, 42 were classified as intermediates according to at least one method. Surprisingly, no individual was classified as intermediate according to all four methods. Therefore, 91% of the discrepancies between the four methods were caused by assignments to the intermediate class (Figure S3 and Table S1 in File S1).

There were nine discrepancies between the individual assignments according to the genotypic and morphological similarity criteria on the one hand and the interfertility criterion on the other hand. We investigated whether the biotic environment of the individuals might account for these discrepancies. Our hypothesis is that the neighborhood of each tree influences its mating system and might thus influence its assignment to species based on the interfertility criterion, whereas it would hardly affect its assignment to species based on the genotypic and morphological criteria. We therefore examined the neighborhood of each tree for which the assignment to species based on genotypic and morphological similarity criteria were congruent ($N=192$). For each tree, we calculated the proportion of allospecific neighbors within a radius of 69m (corresponding to the average distance of pollen dispersal within stand for *Q. petraea*, the species with the smallest dispersal ability [22]). We found, by performing a logistic regression, that the proportion of allospecific neighbors had a significant effect on the congruence between the individual assignments according to the genotypic and morphological similarity criteria on the one hand and the interfertility criterion on the other hand ($\chi^2=6.5$, $df=1$, $p\text{-value}=0.01$). The individuals with congruent assignments had fewer allospecific neighbors on average (29%, versus 51% for individuals with incongruent assignments). Hence, individual species assignments based on the interfertility criterion were environment-dependent.

Discussion

To our knowledge, this is the first time that the interfertility criterion is used successfully to delimit species under natural conditions. The analysis of a network of mating events between pairs of adult trees, constructed on the basis of a powerful

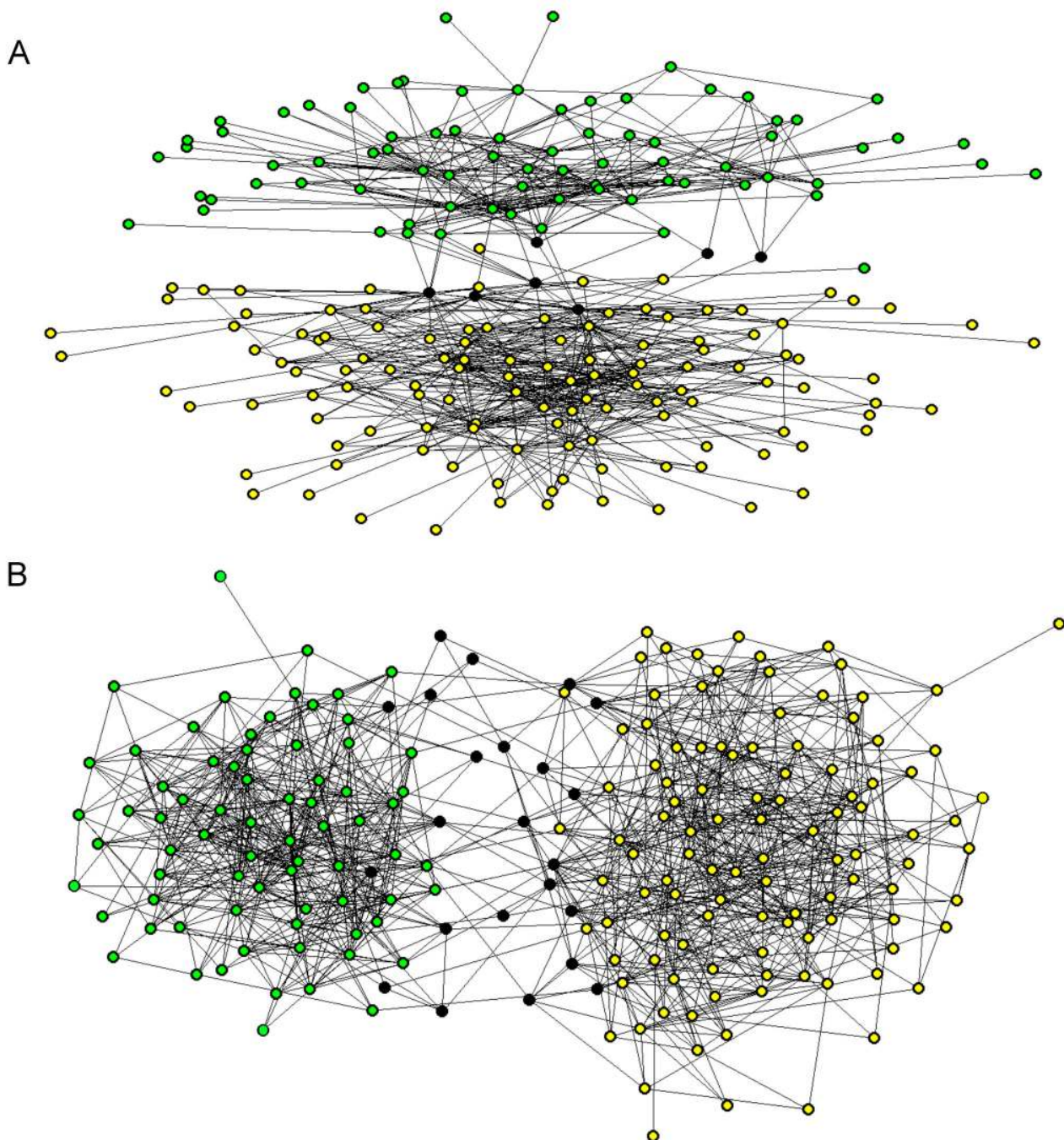


Figure 3. Graphical representation of (A) the mating network and (B) the relatedness network, using the software PAJECK with the following parameters: Draw/Layout/Energy/Kamada-Kawai/Separate Components. Individuals classified into the B1 group (in A) or the P1 group (in B) are shown in green, individuals belonging to the B2 group (in A) or the P2 group (in B) are shown in yellow, and intermediate individuals are shown in black.

doi: 10.1371/journal.pone.0068267.g003

paternity analysis of a large number of seedlings produced under natural conditions, allowed us to identify two groups of interfertile individuals with only a few mating events between

groups. The two groups that were delimited, corresponding to two species according to a relaxed interpretation of the biological species concept (Figure 1B), were closely congruent

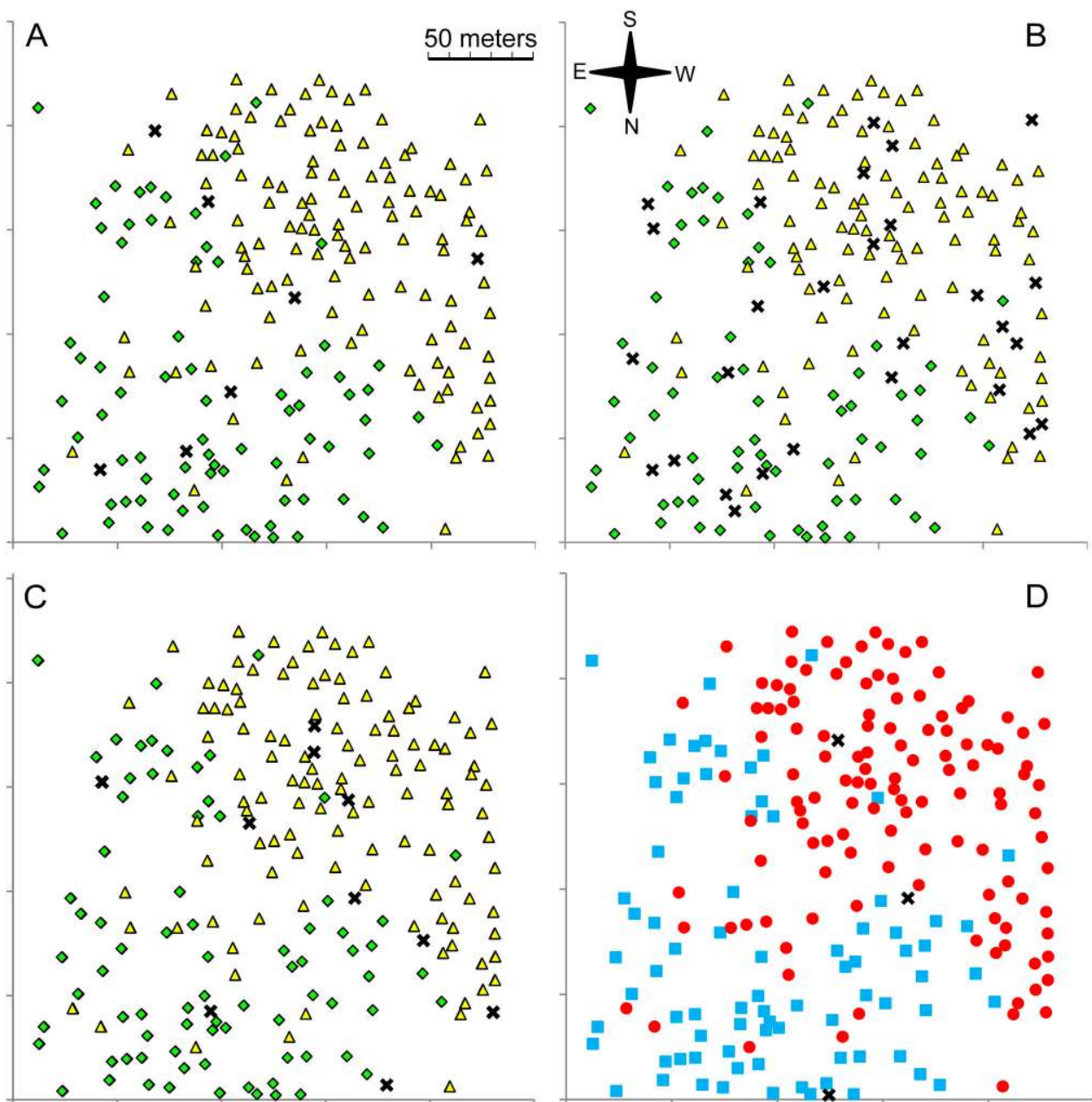


Figure 4. Species boundaries based on interfertility (A), relatedness (B), genotypic similarity (C) and morphological similarity (D) criteria, represented on the map of the stand. In A, B and C, individuals classified into the B1, P1 or G1 species, respectively, are represented by yellow triangles. Individuals classified into the B2, P2 or G2 species are represented by green diamonds. Intermediate individuals are represented by black crosses. In D, individuals classified into M1 are shown in red, individuals classified into M2 in blue and morphologically intermediate individuals are indicated by black crosses. Individuals of the M1 group are assigned to *Q. robur* and individuals of the M2 group to *Q. petraea* on the basis of current taxonomical practices.

doi: 10.1371/journal.pone.0068267.g004

with those obtained previously using morphological and genotypic similarity as criteria for species delimitation [23,26]. Indeed, 88% of the individuals were classified consistently according to the interfertility, morphological similarity and

genotypic similarity criteria. Our results do not support earlier claims that the interfertility criterion cannot be applied in the field (e.g. [14,15]), particularly in the genus *Quercus* [27]. They show instead that the analysis of mating networks can be used

for delimiting species according to the biological species concept, as first suggested by Sokal and Crovello [14].

However this method of species delimitation has two main drawbacks. First, adequate network data are difficult to assemble. In our study we performed a paternity analysis on as many as 3046 offspring produced by 51 mothers in order to construct the mating network for adult trees. Despite the very large number of offspring, our network data did not allow us to assign all the individuals in the forest stand to species. Not all individuals sired offspring and some sired too few offspring to be reliably connected to the network. For example, three of the individuals whose assignment based on the interfertility criterion differed from that based on the three other criteria were represented by a single offspring in the progeny test. They were thus connected to the mating network through just a single link. Second, the sampling design may generate some heterogeneity in the network structure that blurs the biological heterogeneity caused by the existence of different species. This happened in our network data because we harvested the offspring of only 20% of the trees in the stand. The harvested trees (i.e. mother-trees), therefore, had more links in the mating network than the other trees. To solve both problems, one would have to harvest seeds from all the individuals in the stand, assuming that all of them produced seeds. In principle, this goal could be achieved with our biological system by extending sampling over multiple years, because oak species are perennial and monoecious. However this would be impossible for annual or dioecious species. Another possibility to reduce the noise caused by sampling would be to introduce the sampling structure as a covariate in the statistical model (e.g. [28]). Unfortunately, the Continuous Stochastic Block Model [24], which was selected for this study because it allows modeling continuous variations in the connectivity properties of the nodes, does not currently allow the incorporation of covariates.

Our results further show that the analysis of the network of contemporary relatedness relationships is a relevant method for delimiting species. The two groups found in our study might be interpreted as corresponding to two different 'phylogenetic species' [29], if phylogenetic relationships are considered in a broad sense so as to include contemporary pedigree relationships. Methods of species delimitation derived from the phylogenetic species concept have almost exclusively focused on deep ancestry using tree-based phylogenetic methods (reviewed in 30, but see 31). These methods are not well-suited for delimiting hybridizing species because horizontal gene transfers between species, caused by hybridization and subsequent backcrossing events, produce conflicts between gene trees and species trees [32,33]. Compared to data on mating events, data on relatedness were easier to acquire and there was no sampling issue. The analysis of the relatedness network revealed two groups of individuals with high within-group and low between-group degrees of relatedness. These two groups were highly congruent with those obtained using interfertility, morphological similarity and genotypic similarity as criteria, indicating that the analysis of relatedness networks may have potential for species delimitation. However, this method also has some drawbacks: the best model had five

groups of related individuals and we did not find any hypothesis accounting for their origin; the number of species should thus be known in advance in order to apply this method.

By comparing the results obtained with the four criteria used for species delimitation (i.e. interfertility, relatedness relationships, morphological or genotypic similarities), we showed that the species boundaries were largely congruent across methods of species delimitation. Our analyses confirmed the existence of two groups of individuals that were both morphologically and genetically differentiated. We also showed that the individuals of each group preferentially mated and were more related with each other than with individuals from the other group. Therefore, there were two 'evolutionary lineages' in the studied stand. The Lineage Species Concept introduced by Simpson [34,35], then taken up by Wiley [36] and de Queiroz [16,37,38], focuses on the question of congruence among methods of species delimitation. For these authors, modern species concepts (e.g. morphological, phylogenetic, genotypic and biological) assimilate, explicitly or implicitly, species 'to separately evolving (segments of) metapopulation lineages' and are thus all by-products of the lineage species concept [16,17]. This should account for the high degree of congruence among species delimitation methods.

Another important result of this comparison is that, irrespective of the criterion used for delimiting species, we found intermediate individuals that had features of both species. Interestingly, the individuals classified as intermediates often differed across methods. In particular, no individual was consistently classified as intermediate according to all four methods. These discrepancies might be explained by the thresholds that were chosen empirically to delimit purebred species and by data quality problems. As mentioned above, examining more offspring per parent tree may improve species delimitation based on the interfertility criterion. Similarly, a greater number of molecular markers [39] may improve methods of species delimitation based on the genotypic and relatedness criteria. Likewise, a larger number of morphological markers [26] may improve morphological species delimitation. However, we believe that these discrepancies may also reflect a biological reality. Indeed, as shown in other studies [40–42], including in oaks [22,43], species relative abundance affects hybridization rate. An individual tends to reproduce with its neighbours. If it is surrounded by numerous *allospecifics* and few *conspecifics* (e.g. [22,42]), this can result in much hybridization. Such an individual will tend to be assigned to another species or to a reproductively intermediate class, according to methods of delimitation based on interfertility. Therefore, we expect some discrepancies in species assignments between methods based on environment-dependent criteria (such as that based on the interfertility criterion) and methods based on environment-independent criteria (such as that based on the genotypic similarity criterion). Because of these fundamental differences among methods, it is impossible to compute a reference dataset that would give the correct assignment of each individual. Our results thus cannot be used to identify one method of species delimitation that would produce more reliable assignments than the others. Instead, our results show that different methods of

species delimitation produce slightly different results when applied to real biological data.

Conclusion

Our results confirmed the existence of two differentiated groups of individuals at the study site, corresponding to two species: *Quercus robur* and *Q. petraea*. However, depending on the criterion used for assigning individuals to species (i.e. interfertility, relatedness, morphological or genotypic similarities), the boundary between species was not exactly the same. Most of the differences stem from assignment of individuals to an intermediate category. This finding illustrates the continuous nature of variation between species. The model we used, which belongs to a category called 'grade of membership models' (reviewed in [10]) is appropriate for synthesizing continuous (but not uniform) variations in biological diversity. However, to get closer to the species concepts, which generally define species as groups of individuals, we finally classified the individuals into non-overlapping groups. Our approach, therefore, illustrates the influence of concepts on our (mis)representation of species and on our understanding of biological diversity. Frost and Hillis [44], as well as Mayr [45], proposed defining species as 'a whole' instead of as a group of individuals. According to our study, species could also be defined as an 'extreme point' to which individuals are more or less close, thus allowing the possibility of an individual being a mixture of two different species.

Materials and Methods

Species Delimitation based on Interfertility

To construct the mating network for the adult trees, we made use of a progeny test involving 3046 offspring resulting from open pollination, harvested from 51 mother-trees distributed across the entire stand (Figure S4 in File S1). A paternity analysis was conducted [22] by genotyping all the offspring from the test and all the adults trees for which DNA was available, using 12 multiplexed microsatellite (SSR) markers developed by Guichoux *et al.* [46]. According to the paternity analysis, 1575 offspring had only one possible father in the stand, 54 offspring had several potential fathers in the stand and 1417 offspring had no father in the stand [22]. Based on the offspring for which only a single father was found, we identified 198 father-trees in the stand. These trees included 43 trees that were also mothers, because oak trees are monoecious. Based on these results, we reconstructed 1629 mating events between 206 adult trees within the stand. These mating events allowed us to identify 751 couples of trees that mated at least once, indicating that they were interfertile under natural conditions. These data were represented by an undirected and unweighted network in which each of the 206 nodes corresponded to an adult tree and each of the 751 links corresponded to at least one mating event between two trees.

Then, the network was modeled with C-SBM [24]. The parameters of the model are the connectivity coefficients between the EHNs and the coefficients of the mixture of EHNs

for each node of the real network. For each possible number of EHNs, the parameters of the model were inferred by the maximum likelihood method, derived using the MATLAB program C-Mixnet (available at <http://www.agroparistech.fr/mia/doku.php?id=productions:logiciels/>). Then, the optimal number of EHNs in the network was determined by using the AIC criterion [24]. The results were visualized with the software PAJEK [47].

Species Delimitation based on Relatedness

In order to build the relatedness network, we estimated the relatedness of the 206 adult trees included in the mating network. The estimation was performed with the software COANCESTRY [48], which offers seven different estimators of relatedness. As recommended by Wang [48], we used the 1629 offspring for which both parents were known to determine the most suitable estimator. The triadic likelihood estimator (denoted TrioML in COANCESTRY [49]) was selected because it produced relatedness values closest to zero for unrelated offspring, closest to 0.25 for half-sibs and closest to 0.5 for full-sibs. With this estimator, the highest relatedness value between two unrelated offspring was 0.22. We therefore treated 0.22 as a threshold: trees whose relatedness value was higher than this were considered to be related individuals and the other trees were considered to be unrelated. The relatedness relationships were then represented by an unweighted and undirected network with 206 nodes, each corresponding to an adult tree, and 1078 links connecting the individuals considered to be related. As in the case of the mating network, we modeled the network structure using C-SBM [24] and we visualized the results with the software PAJEK [47].

Species Delimitation based on Morphology

The morphological similarity criterion has previously been used by Bacilieri *et al.* [26] to identify all trees from the study site. These authors performed a factorial discriminant analysis (FDA) based on 31 leaf morphological traits to delimit the species. Their study revealed the presence of two groups of individuals differing in their morphology. The first axis of the FDA accounted for 33% of the total variance and was highly correlated to the morphological markers traditionally used by taxonomists to distinguish *Q. robur* from *Q. petraea*. The distribution of the individuals along this axis was used to assign, graphically, the individuals to two pure morphological groups (called M1 and M2 in this study and corresponding to *Q. robur* and *Q. petraea* respectively) and to a morphologically intermediate class (called Mi). Among the 206 adult trees included in the mating and relatedness networks, 123 trees were assigned to M1, 80 to M2 and 3 to Mi (Figure S5 in File S1).

Species Delimitation based on Multilocus Genotypes

Guichoux *et al.* [23] used genotypic similarity as a criterion to assign the trees of the study site to species. These authors genotyped the adult trees with the multiplex of 12 SSRs developed by Guichoux *et al.* [46] and with a chip of 262 single-nucleotide polymorphisms (SNP) enriched with markers highly

differentiated between species [23]. They used the software STRUCTURE [50] to group the individuals into genotypic clusters but did not formally determine the optimal number of genotypic clusters in the stand before performing the clustering. Here we used the ΔK statistic [51] to identify the number of genetically different groups. The optimal number of clusters was two (Figure S6 in File S1), as previously assumed by Guichoux *et al.* [23]. The adult trees were therefore classified in two purebred groups and one genetically intermediate class. Among the 206 adult trees included in the mating and relatedness networks, 78 trees were assigned to the first purebred group (hereafter called G1), 118 to the second purebred group (G2) and 10 to the genetically intermediate class (Gi) (Figure S7 in File S1).

Supporting Information

File S1. (PDF)

Acknowledgements

We are grateful to Alexis Ducousso who established the Petite Charnie progeny test and shared information about the stand,

and for his help, together with that of Stefanie Wagner, during sampling. We thank ONF for the management of the stand. Patrick Léger greatly helped with microsatellite genotyping. The genotyping was performed at the Genome-Transcriptome facility of Bordeaux. We are particularly grateful to François Hubert for helpful discussions about phylogeny and to Bastien Castagnérol, Virgil Fievet, Cyril Dutech, Antoine Kremer and two anonymous reviewers for their constructive comments on the manuscript.

Author Contributions

Conceived and designed the experiments: LL RJP. Performed the experiments: LL. Analyzed the data: LL JBL JJD CV. Contributed reagents/materials/analysis tools: JBL JJD. Wrote the manuscript: LL JBL JJD RJP CV.

References

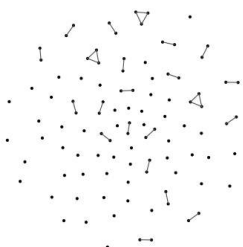
- Mayr E (1942) Systematics and the origin of species. New York: Columbia University Press.
- Dettman JR, Jacobson DJ, Turner E, Pringle A, Taylor JW (2003) Reproductive isolation and phylogenetic divergence in *Neurospora*: comparing methods of species recognition in a model eukaryote. *Evolution* 57: 2721–2741. doi:10.1554/03-074. PubMed: 14761052.
- Marcussen T, Borgen L (2011) Species delimitation in the Ponto-Caucasian *Viola sieheana* complex, based on evidence from allozymes, morphology, ploidy levels, and crossing experiments. *Plant Syst Evol* 291: 183–196. doi:10.1007/s00606-010-0377-z.
- Marin J, Crouau-Roy B, Hemptinne J-L, Lecompte E, Magro A (2010) *Coccinella septempunctata* (Coleoptera, Coccinellidae): a species complex? *Zool Scripta* 39: 591–602. doi:10.1111/j.1463-6409.2010.00450.x.
- Hibbett DS, Fukumasa-Nakai Y, Tsuneda A, Donoghue MJ (1995) Phylogenetic diversity in shiitake inferred from nuclear ribosomal DNA sequences. *Mycologia* 87: 618–638. doi:10.2307/3760806.
- Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM *et al.* (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genet Biol* 31: 21–32. doi:10.1006/fgbi.2000.1228. PubMed: 11118132.
- Fortuna MA, García C, Guimarães PR Jr, Bascompte J (2008) Spatial mating networks in insect-pollinated plants. *Ecol Lett* 11: 490–498. doi:10.1111/j.1461-0248.2008.01167.x. PubMed: 18318718.
- Newman M (2003) The structure and function of complex networks. *Siam Rev* 45: 167–256. doi:10.1137/S003614450342480.
- Fortunato S (2010) Community detection in graphs. *Phys Rep* 486: 75–174. doi:10.1016/j.physrep.2009.11.002.
- Léger J-B, Vacher C, Daudin J-J (2013) Detection of structurally homogeneous subsets in graphs. *Statist Comput* (in press).
- Orr HA (2001) Some doubts about (yet another) view of species. *J Evol Biol* 14: 870–871. doi:10.1046/j.1420-9101.2001.00340.x.
- Noor MAF (2002) Is the Biological Species Concept showing its age? *Trends Ecol Evol* 17: 153–154. doi:10.1016/S0169-5347(02)02452-7.
- Coyne JA, Orr HA (2004) Speciation. Sunderland, Mass., USA: Sinauer Associates.
- Sokal RR, Crovello TJ (1970) The Biological Species Concept: a critical evaluation. *Am Nat* 104: 127–153. doi:10.1086/282646.
- de Meeûs T, Durand P, Renaud F (2003) Species concepts: what for? *Trends Parasitol* 19: 425–427. doi:10.1016/S1471-4922(03)00195-8. PubMed: 14519574.
- de Queiroz (2005) Ernst Mayr and the modern concept of species. *Proc Natl Acad Sci U S A* 102: 6600–6607. doi:10.1073/pnas.0502030102. PubMed: 15851674.
- Hausdorf B (2011) Progress toward a general species concept. *Evolution* 65: 923–931. doi:10.1111/j.1558-5646.2011.01231.x. PubMed: 21463293.
- Streff R, Ducousso A, Kremer A (1998) Spatial genetic structure and pollen gene flow in a mixed oak stand. *Genet Sel Evol* 30: S137–S152. doi:10.1186/1297-9686-30-S1-S137.
- Rieseberg LH, Carney SE (1998) Plant hybridization. *New Phytol* 140: 599–624. doi:10.1046/j.1469-8137.1998.00315.x.
- Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2003) Hybridization as a mechanism of invasion in oaks. *New Phytol* 161: 151–164. doi:10.1046/j.1469-8137.2003.00944.x.
- Lepais O, Gerber S (2011) Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution* 65: 156–170. doi:10.1111/j.1558-5646.2010.01101.x. PubMed: 20722727.
- Lagache L, Klein EK, Guichoux E, Petit RJ (2013) Fine-scale environmental control of hybridization in oaks. *Mol Ecol* 22: 423–436. doi:10.1111/mec.12121. PubMed: 23173566.
- Guichoux E, Garnier-Géré P, Lagache L, Lang T, Bourry C *et al.* (2013) Outlier loci highlight the direction of introgression in oaks. *Mol Ecol* 22: 450–462. doi:10.1111/mec.12125. PubMed: 23190431.
- Daudin J-J, Pierre L, Vacher C (2010) Model for heterogeneous random networks using continuous latent variables and an application to a tree–fungus network. *Biometrics* 66: 1043–1051. doi:10.1111/j.1541-0420.2009.01378.x. PubMed: 20105159.
- Bacilieri R, Ducousso A, Petit RJ, Kremer A (1996) Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution* 50: 900–908. doi:10.2307/2410861.
- Bacilieri R, Ducousso A, Kremer A (1996) Comparison of morphological characters and molecular markers for the analysis of hybridization in sessile and pedunculate oak. *Ann Sci Forestières* 53: 79–91. doi:10.1051/forest:19960106.
- Donoghue MJ (1985) A critique of the Biological Species Concept and recommendations for a phylogenetic alternative. *Bryologist* 88: 172–181. doi:10.2307/3243026.
- Mariadassou M, Robin S, Vacher C (2010) Uncovering latent structure in valued graphs: a variational approach. *Annals Appl Statistics* 4: 715–742. doi:10.1214/10-AOAS361.
- Eldredge N, Cracraft J (1980) Phylogenetic patterns and the evolutionary process. New York: Columbia University Press.
- Sites JWW, Jonathon CM (2004) Operational criteria for delimiting species. *Annu Rev Ecol Syst* 35: 199–227. doi:10.1146/annurev.ecolsys.35.112202.130128.

31. Moalic Y, Arnaud-Haond S, Perrin C, Pearson GA, Serrao EA (2011) Travelling in time with networks: revealing present day hybridization versus ancestral polymorphism between two species of brown algae, *Fucus vesiculosus* and *F. spiralis*. BMC Evol Biol 11: 33. doi:10.1186/1471-2148-11-33. PubMed: 21281515.
32. Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63: 1-19. doi:10.1111/j.1558-5646.2008.00549.x. PubMed: 19146594.
33. Maddison WP (1997) Gene trees in species trees. Syst Biol 46: 523-536. doi:10.1093/sysbio/46.3.523.
34. Simpson GG (1951) The species concept. Evolution 5: 285-298. doi:10.2307/2405675.
35. Simpson GG (1961) Principles of animal taxonomy. New York: Columbia University Press.
36. Wiley EO (1978) The Evolutionary Species Concept reconsidered. Syst Biol 27: 17-26.
37. de Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56: 879-886. doi:10.1080/10635150701701083. PubMed: 18027281.
38. de Queiroz K (1998) The General Lineage Concept of species, species criteria, and the process of speciation. Oxford University Press.
39. Vähä J-P, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Mol Ecol 15: 63-72. PubMed: 16367830.
40. Field DL, Ayre DJ, Whelan RJ, Young AG (2008) Relative frequency of sympatric species influences rates of interspecific hybridization, seed production and seedling performance in the uncommon *Eucalyptus aggregata*. J Ecol 96: 1198-1210. doi:10.1111/j.1365-2745.2008.01434.x.
41. Focke WO (1881) Die Pflanzenmischlinge. Berlin: Bornträger.
42. Hubbs CL (1955) Hybridization between fish species in nature. Systematics Zoology 4: 1-20.
43. Lepais O, Petit RJ, Guichoux E, Lavabre JE, Alberto F et al. (2009) Species relative abundance and direction of introgression in oaks. Mol Ecol 18: 2228-2242. doi:10.1111/j.1365-294X.2009.04137.x. PubMed: 19302359.
44. Frost DR, Hillis DM (1990) Species in concept and practice: herpetological applications. Herpetologica 46: 86-104.
45. Mayr E (1992) A local flora and the Biological Species Concept. Am J Bot 79: 222-238. doi:10.2307/2445111.
46. Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). Mol Ecol Resour 11: 578-585. doi:10.1111/j.1755-0998.2011.02983.x. PubMed: 21481218.
47. Batagelj V, Mrvar A (1998) PAJEK - Program for large network analysis.
48. Wang JL (2010) Coancestry: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. Mol Ecol Resour 11: 141-145. PubMed: 21429111.
49. Wang JL (2007) Triadic IBD coefficients and applications to estimating pairwise relatedness. Genet Res 89: 135-153. PubMed: 17894908.
50. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959. PubMed: 10835412.
51. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol 14: 2611-2620. doi:10.1111/j.1365-294X.2005.02553.x. PubMed: 15969739.
52. Mallet J (1995) A species definition for the modern synthesis. Trends Ecol Evol 10: 294-299. doi:10.1016/0169-5347(95)90031-4. PubMed: 21237047.
53. Nelson G, Plantick N (1981) Systematics and biogeography : cladistics and vicariance. New York: Columbia University Press.
54. Mishler BD (1985) The morphological, developmental, and phylogenetic basis of species concepts in bryophytes. Bryologist 88: 207-214. doi:10.2307/3243030.
55. Legendre G, Legendre P (1984) Ecologie numérique. Paris, France: Masson.

List of Figures

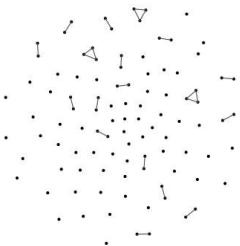
2.1	Difference between communities and structural homogeneous subsets in a hub structure network	25
2.2	Bipartite toy-example	27
2.3	Various clusters obtained for the toy-example	28
2.4	The toy example graph transformed with Jaccard's measure of similarity	29
2.5	MCL applied to the toy example with 4 combinations of tuning parameters	32
2.6	Pons-Latapy distance matrices for $k = 4$ corresponding to the toy example (Figure 2.2) with unitary (M_1) and $\frac{1}{10}$ -weighted self-loops ($M_{\frac{1}{10}}$). Vertices are ordered as $N_1 \cdots N_5, N_6 \cdots N_{10}$	33
2.7	Clusters obtained with Ng-normalized and Absolute Eigenvalue Spectral Clustering, with $k \in \{2, 4\}$	35
2.8	Clusters obtained with Edge-Betweenness with $k \in \{2, 4\}$	38
2.9	Clusters obtained by maximizing the Modularity (each of them have the same modularity)	39
2.10	Clusters obtained with Cut cost with $k \in \{2, 4\}$	40
2.11	Zachary's Karate Club network. Colors show real fission of the club	48
2.12	Edge-betweenness for 2 and 4 groups method applied to the Zachary's Karate Club network	48
2.13	Hierarchical clustering method to Pons-Latapy distance for 2 and 4 groups applied to the Zachary's Karate Club network	49
2.14	Modularity method without choice of number of groups (<i>left</i>) and MCL method without choice of number of groups (<i>right</i>) applied to the Zachary's Karate Club network	49
2.15	Spectral clustering (unweighted variant) method for 2 and 4 groups applied to the Zachary's Karate Club network	50
2.16	Spectral clustering (Absolute Eigenvalues variant) method for 2 and 4 groups applied to the Zachary's Karate Club network	50
2.17	Stochastic Block Model method applied for 2 and 4 groups applied to the Zachary's Karate Club network	51

2.18	Continuous Stochastic Block Model method for 2 and 4 groups applied to the Zachary's Karate Club network	51
3.1	Complementary cumulative distribution of the degrees of the antagonist species for the simulated networks at the central point, on a log-scale. The linear trend is typical of the scale free distribution observed by Jordano et al. (2003) for real ecological networks. . . .	62
3.2	Complementary cumulative distribution of the degrees of the host species for the simulated networks at the central point, on a log-scale. The linear trend is typical of the scale free distribution observed by Jordano et al. (2003) for real ecological networks. . . .	62
3.3	Frequency distribution of dependence values as defined by Bascompte et al. (2006), for antagonist species for simulated networks at the central point. The distribution is similar to those observed by Bascompte et al. (2006) for real ecological networks.	63
3.4	Frequency distribution of dependence values as defined by Bascompte et al. (2006), for host species for simulated networks at the central point. The distribution is similar to those observed by Bascompte et al. (2006) for real ecological networks.	63
3.5	Frequency distribution of symmetry values as defined by Bascompte et al. (2006) for simulated networks at the central point. The distribution is similar to those observed by Bascompte et al. (2006) for real ecological networks.	65
3.6	Relationship between the number of hosts species n_1 , and the number of antagonist species n_2 , on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_1 and n_2 are clearly linked.	65
3.7	Relationship between the ratio of nodes types $\frac{n_1}{n_2}$, and the product $n_1 n_2$, on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_1/n_2 and $n_1 n_2$ are not linked	66
3.8	Empirical distribution of the product of number of species of both types, $n_1 n_2$, for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line. . . .	66

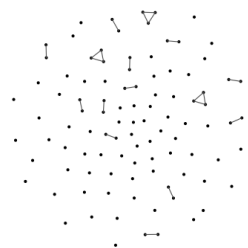


- 3.9 Empirical distribution of the ratio of the numbers of species of both types, n_1/n_2 , for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line. 68
- 3.10 Relationship between the number of edges n_l , and the product of the numbers of species of both types n_1n_2 , on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_l and n_1n_2 are clearly linked. The red line is the regression line. 68
- 3.11 Relationship between the ratio $\frac{n_l}{(n_1n_2)^{0.63}}$, and the product of the numbers of species of both types n_1n_2 , on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. There is no relation between them. 69
- 3.12 Empirical distribution of the ratio $\frac{n_l}{(n_1n_2)^{0.63}}$, for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line. 69
- 3.13 Relationship between the total weight of edges n_w , and the number of edges, n_l , on a log-scale. Each cross represents an ecological network among the 47 networks taken from the Interaction Web Database. n_w and n_l are clearly related. 70
- 3.14 Empirical distribution of the mean weight by present edge, $\frac{n_w}{n_l}$, for 47 networks taken from the Interaction Web Database. The white zone corresponds to the ecological range, the studied range includes the light grey zone. The dark grey zone is not studied. The central point is represented by the green vertical line. 70
- 3.15 Box-plots of the 100 replicates obtained on binary networks for unknown number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges, bottom-right: number of estimated groups divided by the true number (4 for communities methods, 8 for the SHS methods) 73

3.16	Box-plots of the 100 replicates obtained on weighted networks for unknown number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges, bottom-right: number of estimated groups divided by the true number (4 for communities methods, 8 for the SHS methods).	74
3.17	Box-plots of the 100 replicates obtained on binary networks for known number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges.	76
3.18	Box-plots of the 100 replicates obtained on weighted networks for known number of groups at the central point. Up-left: adjusted Rand index for antagonist species, up-right: adjusted Rand index for host species, bottom-left: R_a^2 for prediction of edges.	77
3.19	Relationships between the performances of the methods and the network parameters for unknown number of groups and binary networks. ariA: adjusted Rand index for antagonist species, ariH: adjusted Rand index for the host species, R_a^2 : adjusted R square for the prediction of edges, last column: number of estimated groups divided by the true number (g for communities methods, $2g$ for the SHS methods), size: n_1n_2 , ratio: $\frac{n_1}{n_2}$, connectivity: $\frac{n_l}{(n_1n_2)^{0.63}}$	79
3.20	Relationships between the performances of the methods and the network parameters for unknown number of groups and weighted networks. ariA: adjusted Rand index for antagonist species, ariH: adjusted Rand index for the host species, R_a^2 : adjusted R square for the prediction of edges, last column: number of estimated groups divided by the true number (g for communities methods, $2g$ for the SHS methods), size: n_1n_2 , ratio: $\frac{n_1}{n_2}$, connectivity: $\frac{n_l}{(n_1n_2)^{0.63}}$	80
3.21	Relationships between the performances of the methods and the network parameters for known number of groups and binary networks. ariA: adjusted Rand index for antagonist species, ariH: adjusted Rand index for the host species, R_a^2 : adjusted R square for the prediction of edges, size: n_1n_2 , ratio: $\frac{n_1}{n_2}$, connectivity: $\frac{n_l}{(n_1n_2)^{0.63}}$	82
3.22	Relationships between the performances of the methods and the network parameters for known number of groups and weighted networks. ariA: adjusted Rand index for antagonist species, ariH: adjusted Rand index for the host species, R_a^2 : adjusted R square for the prediction of edges, size: n_1n_2 , ratio: $\frac{n_1}{n_2}$, connectivity: $\frac{n_l}{(n_1n_2)^{0.63}}$	83
4.1	ICL criterion values obtained for Poisson and Poisson with covariates on the trees network	102



4.2	ICL criterion values obtained for the Poisson model and the Poisson with covariates model on the fungus species network.	104
5.1	Variables potentially accounting for the probability that foresters working for the French Department of Forest Health (DSF) observe host-antagonist interactions. Observable variables are shown in grey boxes. Those which are included in the model are in bold. Latent variables, which cannot be observed nor directly measured, are shown in dotted boxes	113
5.2	ICL Values obtained by the inference in function of the number of groups. Shown model are λ , for the model only with group effect, without covariates; λS , for the model with groups effect and sampling effect; λE , for the model with groups effect and encounter effect; λES , for the model with groups effect, sampling and encounter effect.	114
5.3	Observed interaction strenghts in the tree-fungus network against the predicted ones by the 10-groups model with covariates ST and SA	115
5.4	Observed interaction strenghts in the tree-fungus network (<i>top</i>). Predicted interaction strenghts in the tree-fungus network (<i>bottom</i>) by the 10-groups model with covariates ST and SA . Species are ordered by the group membership obtained by the inference.	116
5.5	The estimates of the parameters λ_{ql} for the 10-groups model with covariate ST and SA	118



List of Tables

2.1	Examples of SBM	44
2.2	Summary of the clustering methods	52
3.1	Table 3.1: List of the real antagonistic and mutualistic bipartite networks used in this study. Networks were extracted from the interaction web database (http://www.nceas.ucsb.edu/interactionweb/).	61
3.2	Names, expression, ecological and studied range of the seven parameters of the simulations. The central point corresponds to the central values for each parameter.	73
3.3	Execution time of methods for a network on the central point. Given time is processor time, <i>i.e.</i> equivalent execution time on a single processor computer which only run one job. Some methods as SBM use parallel computing and processor time must be divided by the number of threads to obtain the real execution time	78
4.1	Group proportions and group composition for the analysis with the Poisson model without covariate on the tree species network.	103
4.2	Group proportions and group composition for the analysis with the Poisson model (PRMH) with taxonomic covariate on the tree species network.	103
5.1	Tree species composition of the groups obtained by the 10-groups model with covariate <i>ST</i> and <i>SA</i>	113
5.2	Fungus species composition of the groups obtained by the 10-groups model with covariate <i>ST</i> and <i>SA</i>	113
5.3	ICL values obtained for the different models with 10 groups.	118