Acta Societatis Botanicorum Poloniae

ORIGINAL RESEARCH PAPER Acta Soc Bot Pol 83(2):93–102 DOI: 10.5586/asbp.2014.016
Received: 2014-04-19 Accepted: 2014-05-27 Published electronically: 2014-07-04

Application and comparative performance of network modularity algorithms to ecological communities classification

Thorsten Thiergart[1], Ulf Schmitz[1], Giddy Landan[2], William F. Martin[1], Tal Dagan[2]*

[1] Institute of Molecular Evolution, Heinrich-Heine University Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany

[2] Genomic Microbiology Group, Institute of Microbiology, Christian-Albrechts-University of Kiel, Christian-Albrechts-Platz 4, 24118 Kiel, Germany

Abstract

Network modularity is a well-studied large-scale connectivity pattern in networks. The detection of modules in real networks constitutes a crucial step towards a description of the network building blocks and their evolutionary dynamics. The performance of modularity detection algorithms is commonly quantified using simulated networks data. However, a comparison of the modularity algorithms utility for real biological data is scarce. Here we investigate the utility of network modularity algorithms for the classification of ecological plant communities. Plant community classification by the traditional approaches requires prior knowledge about the characteristic and differential species, which are derived from a manual inspection of vegetation tables. Using the raw species abundance data we constructed six different networks that vary in their edge definitions. Four network modularity algorithms were examined for their ability to detect the traditionally recognized plant communities. The use of more restrictive edge definitions significantly increased the accuracy of community detection, that is, the correspondence between network-based and traditional community classification. Random-walk based modularity methods yielded slightly better results than approaches based on the modularity function. For the whole network, the average agreement between the manual classification and the network-based modules is 76% with varying congruence levels for different communities ranging between 11% and 100%. The network-based approach recovered the known ecological gradient from riverside – sand and gravel bank vegetation – to dryer habitats like semidry grassland on dykes. Our results show that networks modularity algorithms offer new avenues of pursuit for the computational analysis of species communities.

Keywords: ecoinformatics; information visualization; vegetation unit; phytosociology; networks; classification

Introduction

Network modularity (community structure) is a well-studied large-scale connectivity pattern in networks [1,2], with several detection algorithms described in the literature (for a review, see [2]). A comparative study of module assignment accuracy by the different algorithms using benchmark networks showed that algorithm performance varies according to network size and the level of inter-module mixing [3]. A comparative evaluation of different community detection algorithms using real ecological data is however lacking so far. Here we test the applicability of network modularity algorithms as a method to classify plant species communities.

Community ecologists seek to understand the processes underlying organism and environment interaction dynamics of diversity, abundance, and composition of species in communities [4]. Vegetation science focuses on the ecology and composition of plant communities [5]. A basic task of the vegetation ecologist is to characterize, identify and distinguish different vegetation units that comprise plant species with similar habitat preferences. A common traditional approach is the making of relevés [6], which comprise a catalog of all plant species that occur in a vegetation plot together with their respective degree of coverage (i.e., frequency). Plant communities are ascertained by sorting the relevés in vegetation tables according to the occurrence of diagnostic species. Large and complex vegetation tables can however become error-prone and do not provide a concise overview of the whole data. A potentially more critical limitation is that the method demands an a priori knowledge about the respective diagnostic species whose identity can be a matter of debate. Diagnostic species include those particular species whose occurrence in the relevés may serve as an important

* Corresponding author. Email: tdagan@ifam.uni-kiel.de

Handling Editor: Andrzej Bodył

telltale for the plant community classification. These include the character species, whose occurrence is typical to specific plant communities, and the differential species, whose occurrence can be used to distinguish related plant communities, but are not limited to a single community. Computer based methods including network applications have been used in the classification of plant communities [7–11]. De Cáceres et al. [7,8] tested the performance of fuzzy clustering for the classification of communities in mesophytic and xerophytic pastures of Spanish highlands. Oliver et al. [9] employed a different approach including construction of dendrogram groups using hierarchical clustering for the classification of a broad diversity of communities from New South Wales. Classification success rates using those methods range between 75 and 80% [7,9].

For the present study, we investigated four different modularity functions that do not require a pre-determined number of expected modules. These four algorithms are based on two main approaches. The modularity maximization (ModMax) function by Girvan and Newman [1] is an application of a binary recursion algorithm that iteratively splits the network into modules. The algorithm seeks to maximize the modularity function defined as the ratio of edge frequency within modules and edge frequency outside modules. In the simulated annealing (SimAnn) algorithm the modularity function is calculated similarly to the ModMax algorithm and optimized by a simulated annealing approach [12]. The Markov cluster algorithm (MCL) [13] applies a flow simulation approach that is equivalent to a random walk along the edges in the network and measuring the probability to pass between different nodes. Edges that link highly connected nodes comprising a module are assumed to be more frequently travelled than edges that connect nodes from different modules. The less travelled edges are gradually omitted and the remaining connected nodes are the resulting modules. The information flow map (InfoMap) algorithm [14] applies a similar strategy to that of MCL, namely that the path of a random walk along the network edges will pass more frequently between nodes in the same module. The protocol of InfoMap incorporates principles from the field of information theory that are used to eliminate uninformative edges from the network. All algorithms except SimAnn include an implementation of weighted networks. All selected algorithms assign the nodes into a single module only.

The utility of the four modularity algorithms for plant communities classification was tested using data surveyed in the Lower Rhine floodplain vegetation plots. We further investigated the effects of several parameters affecting network based classification and determine to what extent this approach, entailing minimal a priori information, can approximate the results of manual classification for the same data.

Material and methods

Data collection and syntaxonomy

The ecological data consisted of 282 vegetation plots (relevés) that were recorded between 1996 and 2006 in the floodplains of the lower Rhine area, Germany [15,16]. The sampled plot size ranges between 4 m^2 and 100 m^2. Species communities in the plots were classified according to the Braun-Blanquet method considering characteristic and differential species yielding 13 different communities [15]. The communities span a geographical gradient including sand and gravel bank vegetation, communities of moist meadows, flood swards, reeds, nitrophyte vegetation, riparian forests and communities of semidry to fresh meadows (Fig. 1a). Plant species in each plot where identified and ranked by their coverage following the scale of Braun-Blanquet [6] that ranges between 1 (1–5%) and 5 (75–100%). A total of 232 different species were identified during the survey, species with coverage <1% were excluded from network calculations. The frequency of different species per plot with coverage >1% ranges between 1 and 27 with a median of 6. Syntaxa were named according to Pott [17] and LANUV [18], complemented by Schmitz and Lösch [19].

Vegetation network structure and properties

In the vegetation network, vertices correspond to plots in the vegetation table while the edges designate species composition similarity between the plots that they connect. Edge weights in the network were calculated by two different species similarity measures: the Sørensen similarity index [20] and the weighted similarity index [21].

The Sørensen similarity index (*SSI*) is calculated as:

$$SSI = \frac{2C}{A + B} \quad (1)$$

where A and B are the number of species in plots A and B respectively, and C is the number of species shared by the two plots.

The weighted similarity index (*wSI*) is calculated from the species coverage rankings as:

$$wSi = \frac{{}^1\!/_2\, M_c}{M_A + M_B + {}^1\!/_2\, M_C} \quad (2)$$

where M_A and M_B are the total coverage rankings of species that are present only in plot A and B respectively, and M_C is calculated by the total rankings of all species present in both plots.

To estimate the level of connectivity among plots of different communities we used the mixing parameter μ_t that quantifies the proportion of links connecting a certain node with nodes outside the community [3]:

$$\mu_t = \frac{k_i^{out}}{k_i^{in} + k_i^{out}} \quad (3)$$

The variables k_i^{out} and k_i^{in} designate the number of edges connecting node i with nodes outside and within the community respectively. The weighted mixing parameter μ_w quantifies the inter-community connectivity strength [3]:

$$\mu_w = \frac{w_i^{out}}{w_i^{in} + w_i^{out}} \quad (4)$$

The variables w_i^{out} and w_i^{in} designate the sum of edge weights for edges connecting node i with nodes from different or the same community respectively. It is commonly agreed that in a network where the average proportion

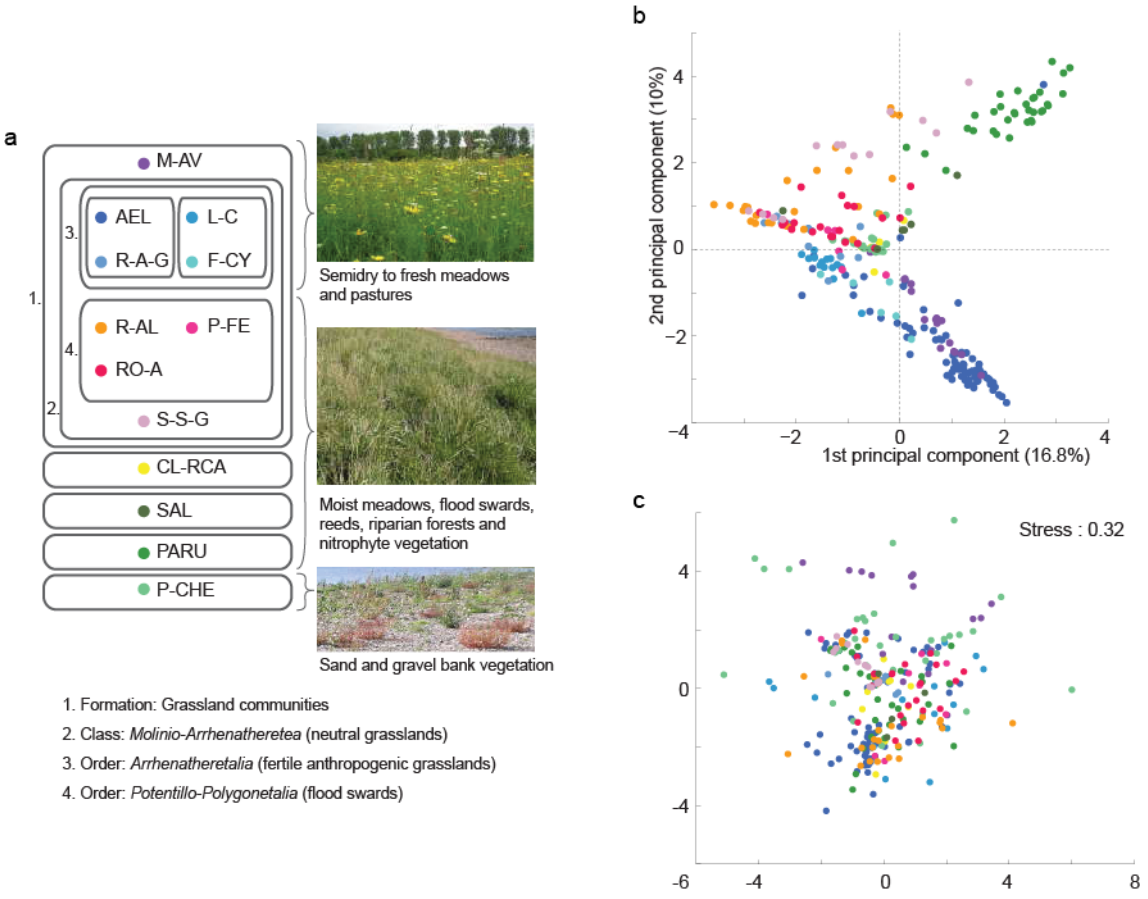

Fig. 1 A hierarchical representation of the ecological communities, results of PCA and MDS. **a** Each colored dot represents one of the 13 plant communities that have been classified in the vegetation data. The boxes describe relationships between plant communities with groups sharing the same syntaxonomy marked by a box. Descriptions next to the boxes provide the threefold gradient present in the data, which is independent from the syntaxonomic community classification. **b** Principal component analysis result. Each dot signifies a plot with colors according to the legend above. **c** Nonmetric multidimensional scaling analysis of the vegetation units based on euclidian distances. AEL – *Arrhenatheretum elatioris*; CL-RCA – *Cuscuto lupuliformis-Rubetum caesii*; F-CY – *Festuco-Cynosuretum/Luzulo-Cynosuretum*; L-C – *Lolio-Cynosuretum*; M-AV – *Medicagini-Avenetum/Mesobrometum alluviale*; P-CHE – *Polygono-Chenopodietum*; P-FE – *Potentillo-Festucetum arundinaceae*; PARU – *Phalaridetum arundinaceae*; R-A-G – *Ranunculus repens-Alopecurus pratensis* community; R-AL – *Ranunculo repentis-Alopecuretum geniculati*; RO-A – *Rorippo-Agrostietum stoloniferae*; SAL – *Salicetum albae*; S-S-G – *Sanguisorba officinalis-Silaum silaus* community.

of edges connecting a given node with nodes outside its community is >50% (i.e., $\mu > 0.5$) the modularity structure (if it exists) would be difficult to detect using computational methods [3]. Network graphs were generated with Cytoscape [22] version 2.8.0 using the "force directed" layout with default parameters [23].

Detection of modules within the network

Modularity applications for SimAn, MCL and InfoMap were downloaded from their dedicated websites. Module detection using ModMax was implemented with a MatLab™ script for the modularity maximization function as described by Newman [24]. The modularity detection accuracy was quantified by the quality of overlap between the modules obtained and the species communities as defined by the expert ecologists. The quality measures were calculated as described by Brohée and van Helden [25]. Standard data analysis and statistical tests were performed using MatLab™ version R2012a (7.14.0.739).

Results

Results of principal component analysis and multidimensional scaling

To test the level of similarity among plots in the different plant communities using traditional methods, we analyzed the data using principal component analysis (PCA) where plant species comprise the variables and the plots are defined as observations. The result does not reveal a clear distinction between plots in the different communities, yet their distribution in the PCA largely conforms to the plots ecological distribution. Plots of the *Phalaridetum arundinaceae* (PARU) and the *Arrhenatheretum elatioris* (AEL) tend to group together but the rest of the communities appear as intermixed (Fig. 1b). The first two components explain in total 26.8% of variability in the data. A further analysis of the plot composition dissimilarity using a multidimensional scaling (MDS) approach results in mixed distribution of the plots and no clear distinction between the communities (Fig. 1c). The results of the multivariate methods reveal some order

in the data but they cannot be used in order to classify the plant communities.

Edge definition and species community mixing

The definition of edge weight is expected to have a strong influence on the network structure and the modularity algorithms performance. To investigate the effect of plot similarity estimation on the network modularity, we constructed networks using six different combinations of edge weight assignment and connectivity rules. Those include networks in which the edge weight is calculated by *SSI* or *wSI* measures as well as different edge exclusion regimes according to the shared dominant species (the species having the highest coverage within the plot; Tab. 1). Network construction based on plant composition similarity among the plots (Tab. 1: network A, Fig. 2a) results in a highly connected network where each plot is, on average, connected to half of the plots in the network (Tab. 1). The average weighted mixing parameter of nodes in network A (Tab. 1) is >0.5, indicating that species communities in this network will be difficult to detect using any modularity function.

Plots of the *Arrhenatherum elatioris* (AEL) and *Phalaridetum arundinaceae* (PARU) species communities (for abbreviations see legend of Fig. 1) are frequently connected with plots in the same community. Consequently the mixing parameter of these communities is low (PARU $\mu_w = 0.52$; AEL: $\mu_w = 0.42$; Fig. 2f), suggesting that these species communities are relatively well defined. This result is in agreement with the distinction level of these two communities in the PCA result.

In network B the edge weight was calculated by a weighted similarity index to include information about the plot coverage of each species (Tab. 1). This results in a slight decrease of the average μ_w in the total network. In addition, the average mixing parameter of nodes in the *Cuscuto lupuliformis-Rubetum caesii* (CL-RCA) species community drops below 0.5 (Fig. 2f). Furthermore, plots in this community are largely indistinguishable in any visible way (Fig. 1b, Fig. 2b).

To reinforce the species community structure within the network we adopted a procedure from the ecological definition of species communities relying on the dominant species for the characterization of plant communities [26]. In network C the plots are connected only if the dominant species is identical. The resulting network is sparser, with 73% fewer edges than networks A or B, and the node connectivity is reduced to 36 nearest neighbors per node on average (Tab. 1). The grouping of plots from several species communities is visible in the network, including M-AV, AEL, L-C, PARU, and CL-RCA (Fig. 2c). Accordingly, the mean weighted mixing parameter of these five communities is <0.5 (Fig. 2f). The overall mean mixing parameter in the network is significantly lower in comparison to networks A and B with an average of 38% inter-community edges per node (Tab. 1). Seven plots do not share their dominant species with any of the plots and are disconnected from the network main component (Fig. 2c).

In network D the connectivity rule makes use of the dominant species coverage information so that only plots sharing their dominant species in similar levels of coverage are connected. This restriction results in a decrease of 22% of the edges in the network (Tab. 1). The mixing parameter is reduced as well with an average of 30% inter-community links per plot. Nodes in plant community CL-RCA that are characterized by *Cuscuta lupuliformis* (willow dodder) and *Rubus caesius* (European dewberry), form a homogeneous cluster that is disconnected from the main network component (Fig. 2d). Plots of moist meadow and flood sward communities (R-A-G, R-AL, RO-A, P-FE and S-S-G) are frequently inter-connected (Fig. 2d). The high mixing parameter of plots in these communities (Fig. 2e) indicates that their species composition and dominant species are similar, hence the success of the computational approach to distinguish between species communities in this class is low. Plots in the F-CY community have an average $\mu_w = 0.77$ and are mixed accordingly with other communities in the order *Arrhenatheretalia* (fertile anthropogenic meadows and pastures; Fig. 2d). Indeed, network D does not present a complete division between all species communities, yet there is a certain distinction between rarely and often flooded habitats in which the plots were sampled. The top part of the network comprises species communities sampled in semidry to fresh meadows and pastures, the waistline presents plots sampled in moist habitats, while the bottom portraits plots sampled near the Rhine waterside on sand and gravel banks (Fig. 2d).

Using the dominant species for the connectivity rule resulted in a significant overall decrease in the plot inter-community mixing level (Fig. 2g). Further increase of the network modularity may be achieved by using a different edge weight calculation leading to stronger connectivity among plots classified into the same species community. Networks C_w and D_w are constructed using the same connectivity rules as in networks C and D respectively, but with edge weights calculated by the weighted similarity index (*wSI*; Tab. 1). The mixing parameter distribution of nodes in the resulting networks does not differ significantly from the networks where the *SSI* was used for the edge weight calculation ($P_{Cw} = 0.58$, $n_{Cw} = 275$ and $P_{Dw} = 0.72$, $n_{Dw} = 272$, using Kruskal–Wallis test and Tukey posthoc comparisons with $\alpha = 0.05$; Fig. 2g). This counter intuitive result may be explained by the similar distributions of *SSI* and *wSI* in our data. The two similarity measures are significantly linearly correlated ($SSI = 0.89 \times wSI + 0.05$; $P = 0$, $n = 3888$) where changes in *SSI* explain 84% of the variability in *wSI* ($R^2 = 0.84$).

Modules within the network of plant communities

To test the applicability of network modularity algorithms for species community classification we constructed modules in the network using four different algorithms. The results reveal that indeed module detection accuracy improves when the network is constructed using connectivity rules yielding lower average mixing (Fig. 3). The modularity detection in networks A and B using ModMax, MCL, and InfoMap algorithms yielded overall similar results while SimAnn was an exception with 94 modules. The 94 modules included 3 large modules comprising 67% of the plots and 91 single-plot modules (Tab. 2). In what follows, we discuss the resulting modules in network D, which is characterized by the lowest mixing parameter distribution and overall

Tab. 1 Vegetation plot network definitions and properties.

Network	Edge weight	Connectivity rule – plots *i* and *j* are connected if:	No. of edges	Mean connectivity[2]	Mean μ_t	Mean μ_w
A	$a_{ij} = SSI$	$a_{ij} > 0$	18.011	127	0.7	0.61
B	$a_{ij} = wSI$	$a_{ij} > 0$	18.011	127	0.7	0.55
C	$a_{ij} = SSI$	One or more of the most frequent species[1] are identical.	4.988	36	0.38	0.34
D	$a_{ij} = SSI$	One or more of the most frequent species[1] are identical, and the difference in their plot coverage is ±1.	3.888	28	0.31	0.3
C_w	$a_{ij} = wSI$	As in C.	4.988	36	0.38	0.34
D_w	$a_{ij} = wSI$	As in D.	3.888	28	0.31	0.29

[1] The most frequent species are species having the maximum coverage ranking in the plot. [2] Connectivity (C_i) is calculated as the number of nodes connected to node *i* by a single edge.

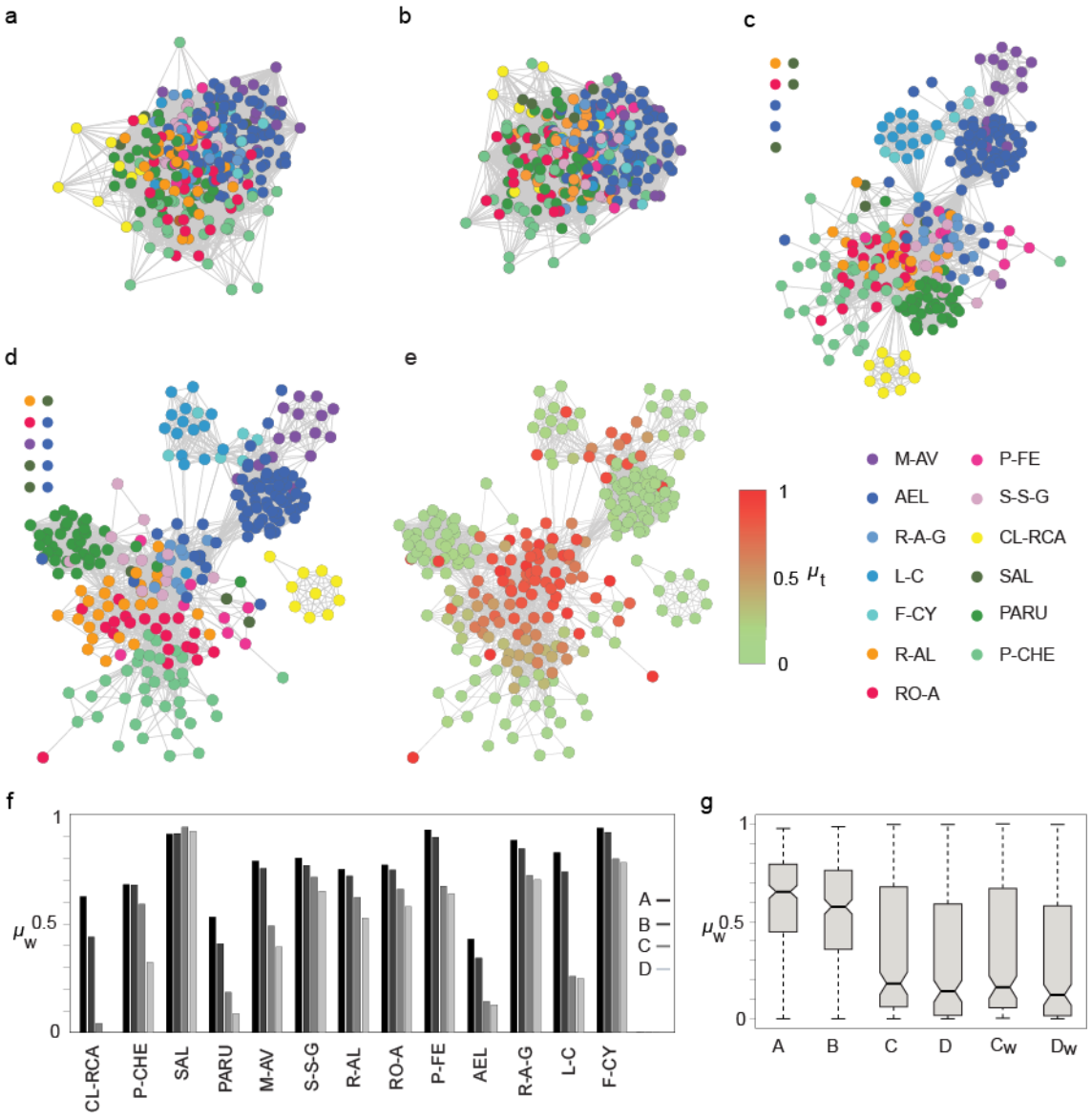

Fig. 2 Plant community networks. **a–d** Networks A–D calculated according to their definitions in Tab. 1. **e** Distribution of the node mixing parameter, μ_t, in network D. The nodes are painted by their μ_t, ranging between 0 and 1 according to the color scale-bar on the right. **f** The distribution of community μ_w in networks A–D. **g** Distribution of node μ_w in the six tested networks (Tab. 1).

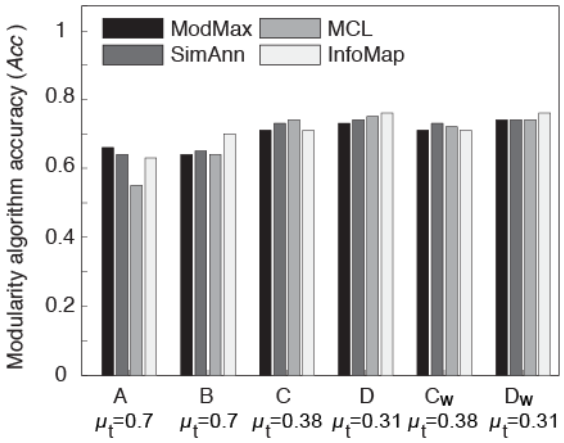

Fig. 3 Module detection accuracy.

higher module detection accuracy. The modularity maximization based methods ModMax and SimAnn (with default starting resolution) resulted in nine and seven modules accordingly, while the information flow algorithms MCL (with default Inflation parameter) and InfoMap resulted in fourteen modules each. The latter two methods resulted in a higher positive predictive value (*PPV*) in comparison to the modularity maximization methods, indicating that modules by the MCL and InfoMap result in a better prediction of the species communities (Tab. 2).

The sensitivity (*Sn*) measure used to quantify the success rate of assigning plots from the same plant community into the same module is higher using the ModMax and SimAnn functions (Tab. 2). These algorithms yield fewer and larger modules that contain whole communities, leading to higher sensitivity values. Weighting the *PPV* and *Sn* values to assess the algorithm accuracy (*Acc*) reveals somewhat better precision using the flow information methods with a slight advantage to InfoMap (Tab. 2). The community separation (Sep_{com}) reveals that MCL yields the best overlap between species communities and modules, while the module separation (Sep_{mod}) shows that modules in the SimAnn result in higher community mixing and somewhat better overall separation in comparison to the other methods (Tab. 2). Testing the impact of the SimAnn starting resolution parameter and the MCL inflation parameter on the performance of these two algorithms showed that the best accuracy was achieved using their default parameters (Tab. 3).

All modularity methods except SimAnn yield modules that overlap well with species communities of the fresh meadows and pastures (Fig. 4). From among the modules assigned to these communities, the module that specifies the A-EL community has the best overlap, followed closely by plots of the M-AV community (Tab. 4). Plots in the L-C and F-CY species communities are grouped together in all methods (Fig. 4). This result supports the view expressed by several authors that F-CY should be included in the L-C community as a nutrient-poor variant of the latter [27,28]. The ModMax algorithm resulted in a slightly higher separation value of L-C community than the other algorithms because it succeeds in identifying a plot that is densely connected with 39 plots from other communities (the plot is marked by an

Tab. 2 Comparison of community detection algorithms.

	ModMax	SimAnn	MCL	InfoMap
Network A				
No. modules	3	94	3	3
PPV	0.51	0.73	0.31	0.43
Sn	0.86	0.57	0.96	0.92
Acc	0.66	0.64	0.55	0.63
Sep_{com}	0.18	0.57	0.15	0.20
Sep_{mod}	0.80	0.08	0.66	0.86
Sep	0.38	0.21	0.32	0.41
Network B				
No. modules	5	94	5	5
PPV	0.51	0.73	0.45	0.56
Sn	0.79	0.57	0.90	0.89
Acc	0.64	0.65	0.64	0.70
Sep_{com}	0.23	0.57	0.22	0.26
Sep_{mod}	0.59	0.08	0.57	0.68
Sep	0.37	0.21	0.36	0.42
Network C				
No. modules	8	7	11	13
PPV	0.65	0.65	0.67	0.71
Sn	0.76	0.82	0.82	0.70
Acc	0.71	0.73	0.74	0.71
Sep_{com}	0.42	0.42	0.50	0.51
Sep_{mod}	0.68	0.78	0.59	0.51
Sep	0.54	0.57	0.54	0.51
Network C_w				
No. modules	8	7	14	14
PPV	0.62	0.65	0.68	0.72
Sn	0.80	0.83	0.77	0.69
Acc	0.71	0.73	0.72	0.71
Sep_{com}	0.39	0.43	0.51	0.52
Sep_{mod}	0.63	0.80	0.47	0.48
Sep	0.49	0.58	0.49	0.50
Network D				
No. modules	9	7	14	14
PPV	0.65	0.67	0.75	0.75
Sn	0.82	0.82	0.75	0.76
Acc	0.73	0.74	0.75	0.76
Sep_{com}	0.43	0.42	0.57	0.55
Sep_{mod}	0.61	0.78	0.53	0.51
Sep	0.51	0.57	0.55	0.53
Network D_w				
No. modules	9	7	14	14
PPV	0.65	0.67	0.71	0.75
Sn	0.83	0.82	0.76	0.76
Acc	0.74	0.74	0.74	0.76
Sep_{com}	0.43	0.42	0.55	0.55
Sep_{mod}	0.63	0.78	0.51	0.51
Sep	0.52	0.57	0.53	0.53

Abbreviations are explained in the text.

Tab. 3 Comparison of community detection accuracy in network D by MCL and SimAnn algorithms using different parameters.

MCL	Inflation parameter	Accuracy
	5.00	0.693
	4.00	0.689
	3.00	0.749
	2.50	0.745
	2.20	0.741
	2.00	0.752
	1.80	0.741
	1.50	0.738
	1.20	0.660
SimAnn	**Resolution**	**Accuracy**
	0.30	0.680
	0.35	0.700
	0.40	0.717
	0.45	0.738
	0.50	0.739
	0.55	0.739
	0.60	0.737
	0.65	0.735

arrow in Fig. 3a; Tab. 4). The multitude of inter-community connections of this plot is due to the presence of six equally coverage-ranked species in the plot that are all defined as dominant using our connectivity rule. These species are dominant in multiple species communities and hence the high connectivity of this plot.

All modularity functions have 100% success rate in specifying the CL-RCA community. Using the dominant species and coverage for the connectivity rule disconnected the CL-RCA plots from the network and enabled the accurate detection of this plant community (Tab. 4, Fig. 4). The next best-identified community is PARU with 94% of its plots classified into one module containing a marginal frequency (5%) of plots from other communities (Tab. 4). Plots of the orders *Arrhenatheretalia* (fertile anthropogenic grasslands) and *Potentillo-Polygonetalia* (flood swards) are classified into two modules or more, but better distinction between the communities was unsuccessful. The P-CHE community comprising sand and gravel bank vegetation is well detected by all algorithms except ModMax (Tab. 4). ModMax yielded a low-resolution modularity structure where the P-CHE plots are grouped together with R-AL and RO-A plots into one module.

Discussion

Various statistical methods such as PCA and MDS have been used in the past to analyze and identify species communities. They provide a general overview of the similarity distribution across the sampled plots, yet these methods are not suited for de novo classification of species communities. In the present work, we found that networks can attain an accuracy of 76% relative to manual classification using existing modularity algorithms. The network modularity structure, even in cases where it is not highly accurate relative to manual methods, is helpful in providing a general overview over the global distribution of shared plot composition in a given dataset. This includes gradients in ecological properties of different habitats, such as moisture or temperature, having an impact on observed species composition.

None of modularity functions were able to distinguish between plots in the L-C (*Lolio-Cynosuretum*) and F-CY (*Festuco-Cynosuretum*) species communities. These two communities were distinguished based on species composition by many authors [29–33]. However, smooth transitions in plant community assembly between both communities occur frequently [34]. Consequently, several studies have considered the F-CY community as a nutrient-poor form of L-C community and included it into the latter [27,28]. The modularity structure of our network supports that view. Such community continuums pose a challenge to the network-based community detection methods and would require the addition of either additional ecological/biological information or further analysis of network higher-order structure. On the other hand, the distinctiveness of the CL-RCA (*Cuscuto lupuliformis-Rubetum caesii*), which was first described by Schmitz and Lösch [19], is strongly recovered in the networks.

Using the a priori information of species communities in this study enabled the test of different connectivity rules while studying their impact on the network modularity. For example, the use of dominant species as the hallmark of species communities in the connectivity rule increased the network modularity and improved the performance of the modularity functions considerably. However, the rigorous use of the dominant species may also bias the modularity results, especially in plots with more than one dominant species of the same coverage. Furthermore, the dominant species is not necessarily the character species, which was used for the manual ecological classification of species communities in our data. Our computational approach is hence more similar to the "Uppsala school" (dominant species) rather than the "Zurich-Montpelier school" (character species) in vegetation studies. This difference between the manual and computational classification approaches here stems from the difficulty in finding a rigorous definition of character species based on their coverage ranking alone. From the present study, a critical parameter for the utility of networks for classifying species communities are the connectivity rule and the function used to identify characteristic species.

The ranking of modularity function performance for classification of vegetation plots in our study differs from their ranking based on the analysis of benchmark networks. A recent comparison of community detection algorithms using a simulated (artificial) weighted network in which the modules are known revealed that the InfoMap function performs much better than MCL or SimAnn algorithms [3]. In our networks comprised of real data, MCL and InfoMap performed similarly well. Indeed, the present analyses embrace the notion that lower mixing parameter leads to better community detection performance [3]. Variation

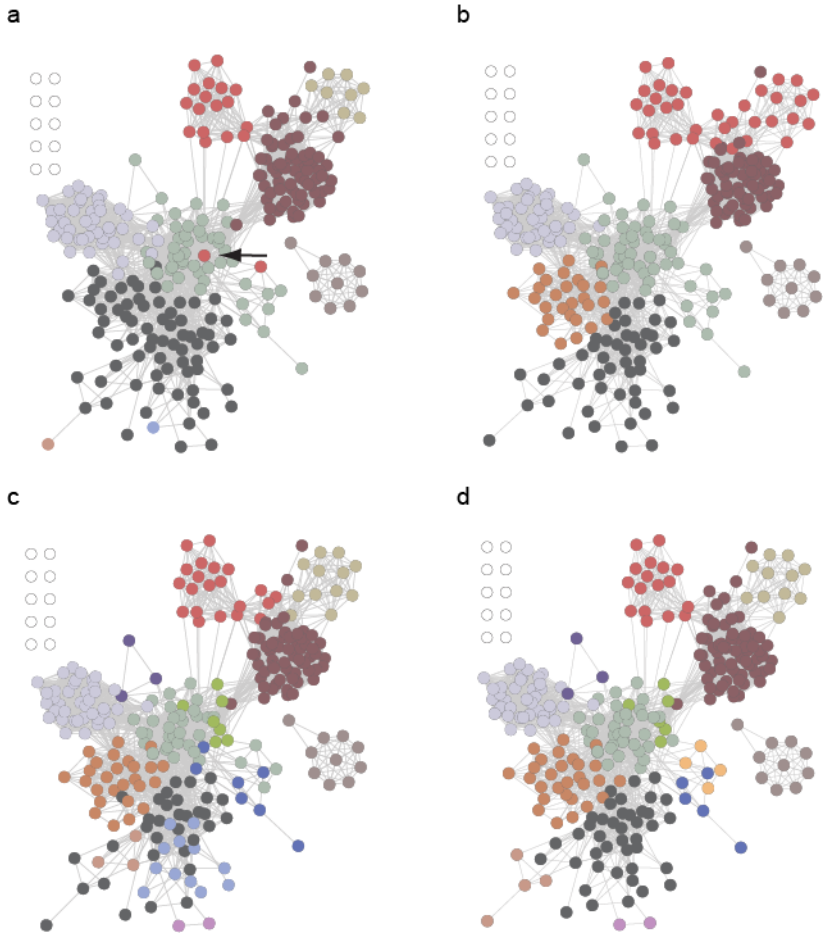

Fig. 4 Modules detected in network D by the four modularity functions. Nodes having the same color were grouped into the same module. **a** ModMax. **b** MCL. **c** Infomap. **d** SimAnn.

Tab. 4 Quality of community separation using the different algorithms for network D.

Species community	No. of plots	No. of classified plots	μ_w (mean ±std)	Sepcom			
				ModMax	SimAnn	MCL	Infomap
M-AV	16	15	0.26 ±0.35	0.58	0.32	0.8	0.74
AEL	76	72	0.19 ±0.28	0.74	0.79	0.82	0.79
R-A-G	10	10	0.67 ±0.17	0.2	0.18	0.23	0.24
L-C	15	15	0.21 ±0.2	0.78	0.37	0.62	0.77
F-CY	6	6	0.76 ±0.06	0.1	0.17	0.28	0.11
R-AL	26	25	0.48 ±0.24	0.19	0.45	0.53	0.45
RO-A	21	20	0.54 ±0.19	0.3	0.26	0.35	0.27
P-fe	8	8	0.45 ±0.47	0.12	0.11	0.65	0.43
S-S-G	13	13	0.54 ±0.26	0.15	0.18	0.38	0.4
Cl-RCA	10	10	0	1	1	1	1
SAL	6	3	0.82 ±0.15	0.06	0.05	0.08	0.34
PARU	39	39	0.12 ±0.21	0.78	0.87	0.9	0.9
P-Che	36	36	0.18 ±0.25	0.48	0.67	0.74	0.69

in connectivity rules had a strong impact on community detection accuracy. This result stresses the importance of edge definition in networks constructed from real data. Furthermore, our findings reveal a more heterogeneous view of mixing parameter distribution over the network nodes in comparison to benchmark networks. Consequently, different communities may vary in their detection accuracy, in that several communities detected in high accuracy (e.g., PARU) and some that are not identified at all (e.g., SAL). Indeed, an estimation of network modularity (community mixing) for real data is impossible. Yet, our results show that for a network based species communities' analysis it is recommended to eliminate edges connecting plots that do not share any species, and in addition to employ information regarding diagnostic species in the connectivity rule.

Our results show that network methods can readily be used to visualize and analyze vegetation tables for the identification and study of plant communities. Modularity detection algorithms could also be applied to other biological systems, for example the study of microbial species community structure, which entails even larger datasets [35].

Acknowledgments

Work in the authors' laboratories is supported by the European Research Council (grant No. 232975 to WFM; grant No. 281357 to TD).

Authors' contributions

The following declarations about authors' contributions to the research have been made: carried out the computational analysis and drafted the manuscript: TT; collected the underlying ecological data and drafted parts of the manuscript: US; contributed to the computational analysis: GD; conceived the study and drafted the manuscript: WFM, TD.

References

1. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci USA. 2002;99(12):7821–7826. http://dx.doi.org/10.1073/pnas.122653799
2. Fortunato S. Community detection in graphs. Phys Rep. 2010;486(3–5):75–174. http://dx.doi.org/10.1016/j.physrep.2009.11.002
3. Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;80(5). http://dx.doi.org/10.1103/PhysRevE.80.056117
4. Vellend M. Conceptual synthesis in community ecology. Q Rev Biol. 2010;85(2):183–206.
5. Mucina L. Classification of vegetation: past, present and future. J Veg Sci. 1997;8(6):751–760. http://dx.doi.org/10.2307/3237019
6. Braun-Blanquet J, Fuller GD, Conrad HS. Plant sociology: the study of plant communities. New York, NY: Hafner Press; 1965.
7. de Cáceres M, Font X, Vicente P, Oliva F. Numerical reproduction of traditional classifications and automatic vegetation identification. J Veg Sci. 2009;20(4):620–628. http://dx.doi.org/10.1111/j.1654-1103.2009.01081.x
8. de Cáceres M, Font X, Oliva F. The management of vegetation classifications with fuzzy clustering: fuzzy clustering in vegetation classifications. J Veg Sci. 2010;21(6):1138–1151. http://dx.doi.org/10.1111/j.1654-1103.2010.01211.x
9. Oliver I, Broese EA, Dillon ML, Sivertsen D, McNellie MJ. Semi-automated assignment of vegetation survey plots within an classification of vegetation types. Methods Ecol Evol. 2013;4(1):73–81. http://dx.doi.org/10.1111/j.2041-210x.2012.00258.x
10. Roleček J, Tichý L, Zelený D, Chytrý M. Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. J Veg Sci. 2009;20(4):596–602. http://dx.doi.org/10.1111/j.1654-1103.2009.01062.x
11. Tichý L, Chytrý M, Hájek M, Talbot SS, Botta-Dukát Z. OptimClass: using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. J Veg Sci. 2010;21(2):287–299. http://dx.doi.org/10.1111/j.1654-1103.2009.01143.x
12. Guimerà R, Sales-Pardo M, Amaral L. Modularity from fluctuations in random graphs and complex networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;70(2). http://dx.doi.org/10.1103/PhysRevE.70.025101
13. van Dongen S. Graph clustering by flow simulations [PhD thesis]. Utrecht: University of Utrecht; 2000.
14. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA. 2008;105(4):1118–1123. http://dx.doi.org/10.1073/pnas.0706851105
15. Schmitz U, Lösch R. Neophyten und C4-Pflanzen in der Auenvegetation des Niederrheins. Decheniana. 2005;158:55–77.
16. Schmitz U. Increase of alien and C4 plant species in annual river bank vegetation of the River Rhine. Phytocoenologia. 2006;36(3):393–402. http://dx.doi.org/10.1127/0340-269X/2006/0036-0393
17. Pott R. Die Pflanzengesellschaften Deutschlands. 2nd ed. Stuttgart: E. Ulmer Verlag; 1995.
18. LANUV. Vegetationstypenliste (list of vegetation types and their abbreviation) [Internet]. 2014; Available from: http://www.naturschutzinformationen-nrw.de/methoden/web/babel/media/vegetationstypen.xlsx
19. Schmitz U, Lösch R. Vorkommen und Soziologie der Cuscuta-Arten in der Ufervegetation des Niederrheins. Tuexenia. 1995;15:373–385.
20. Sørensen TJ. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Copenhagen: I kommission hos E. Munksgaard; 1948. (Biologiske Skrifter; vol 5).
21. Ellenberg H, Walter H. Einführung in die Phytologie. Stuttgart: Ulmer; 1956.
22. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–2504. http://dx.doi.org/10.1101/gr.1239303
23. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Softw Pr Exp. 1991;21(11):1129–1164. http://dx.doi.org/10.1002/spe.4380211102
24. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA. 2006;103(23):8577–8582. http://dx.doi.org/10.1073/pnas.0601602103
25. Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics. 2006;7(1):488. http://dx.doi.org/10.1186/1471-2105-7-488
26. Whittaker RH. Ordination and classification of communities. The Hague: Junk; 1973.
27. Dierschke H. *Molinio-Arrhenathertea* (E1). Kulturgrasland und verwandte Vegetationstypen. Teil 1: *Arrhenatheretalia* Wiesen und Weiden frischer standorte. Synopsis der Pflanzengesellschaften Deutschlands.

Göttingen: Selbstverlag der Floristisch-soziologischen Arbeitsgemeinschaft; 1997.
28. Rennwald E. Rote Liste der Pflanzengesellschaften Deutschlands mit Anmerkungen zur Gefährdung. Schriftenreihe Für Veg. 2000;35:393–592.
29. Runge F. Die Pflanzengesellschaften Mitteleuropas. Münster: Aschendorff; 1990.
30. Wilmanns O. Ökologische Pflanzensoziologie. Heidelberg: Quelle & Meyer; 1998.
31. Schubert R, Hilbig W, Klotz S. Bestimmungsbuch der Pflanzengesellschaften Mittel- und Nordostdeutschlands. Wiesbaden: Spektrum Akademischer Verlag; 2001.
32. Foerster E. Pflanzengesellschaften des Grünlandes in Nordrhein-Westfalen. Münster: Landwirtschaftsverlag; 1983.
33. Verbücheln G, Hinterlang D, Pardey A, Pott R, Raabe U, van de Weyer K. Rote Liste der Pflanzengesellschaften in Nordrhein-Westfalen. Recklinghausen: LÖBF-Schriftenreihe; 1995.
34. Ellenberg H. Vegetation Mitteleuropas mit den Alpen in ökologischer, dynamischer und historischer Sicht. Stuttgart: Eugen Ulmer; 1996.
35. Gonzalez A, Clemente JC, Shade A, Metcalf JL, Song S, Prithiviraj B, et al. Our microbial selves: what ecology can teach us. EMBO Rep. 2011;12(8):775–784. http://dx.doi.org/10.1038/embor.2011.137