RESEARCH ARTICLE

Modular structure in fish co-occurrence networks: A comparison across spatial scales and grouping methodologies

Daniel J. McGarvey[1]*, Joseph A. Veech[2]

1 Center for Environmental Studies, Virginia Commonwealth University, Richmond, Virginia, United States of America, **2** Department of Biology, Texas State University, San Marcos, Texas, United States of America

* djmcgarvey@vcu.edu

Abstract

Network modules are used for diverse purposes, ranging from delineation of biogeographical provinces to the study of biotic interactions. We assess spatial scaling effects on modular structure, using a multi-step process to compare fish co-occurrence networks at three nested scales. We first detect modules with simulated annealing and use spatial clustering tests (interspecific distances among species' range centroids) to determine if modules consist of species with broadly overlapping ranges; strong spatial clustering may reflect environmental filtering, while absence of spatial clustering may reflect positive interspecific relationships (commensalism or mutualism). We then use non-hierarchical, multivariate cluster analysis as an alternative method to identify fish subgroups, we repeat spatial clustering tests for the multivariate clusters, then compare spatial clustering results among modules and clusters. Next, we compare species lists within modules and clusters, and estimate congruence as the proportion of species assigned to the same groups by the two methods. Finally, we use a well-documented nest associate complex (fishes that deposit eggs in the gravel nests of a common host) to assess whether strong within-group associations may, in fact, reflect positive interspecific relationships. At each scale, 2–4 network modules were detected but a consistent relationship between scale and the number of modules was not observed. Significant spatial clustering was detected at all scales for network modules and multivariate clusters but was less prevalent at smaller scales. Congruence between modules and clusters was always < 90% and generally decreased as the number of groups increased. At all scales, the complete nest associate complex was completely preserved within a single network module, but not within a single multivariate cluster. Collectively, our results suggest that network modules are promising tools for studying positive interactions and that smaller scales may be preferable in this research.

Introduction

Efforts to understand the structural [1,2] and functional [3,4] properties of ecological networks are quickly becoming central themes in community ecology and biogeography. This trend has

been driven by a fundamental need to understand how perturbations may propagate through interconnected ecosystems, as well as the enhanced availability of large datasets to represent complex networks [5–9]. In a network graph, distinct entities–usually individuals or species–are represented as nodes or vertices and connections among entities are represented as links or edges. These connections often represent species' co-occurrences [10,11] or food web links between predators and their prey [12,13] when working with unipartite or 'one-mode' networks. Connections within bipartite or 'two-mode' networks may also represent associations between species with distinct functional roles, such as plants and their pollinators [14,15], or species' presences within a matrix of potential habitats [16,17].

One aspect of network structure that is particularly relevant to many ecological questions is modularity. Modularity is the tendency for networks to consist of highly interconnected subgroups of species' nodes that are distinguished from other such groups, or modules, by relatively sparse among-group connections [18]. Depending on one's interest, network modules can serve several purposes. At regional to continental scales, modularity analysis of bipartite networks (species × site data) can be used to detect biogeographical provinces and may be a superior alternative to multivariate ordination and clustering algorithms [19]. Multivariate methods that use species' presence-absence matrices as raw data can be sensitive to the choice of a particular dissimilarity index and to the rules used to combine or agglomerate entities within clusters [20]. But bipartite modularity values are derived solely from observed connections between species and sites, and do not incorporate an abstract measure of dissimilarity [17,18].

At smaller scales (e.g., forest plots and small lakes), modules are used to study the structure and stability of interactive communities. Previous work built upon well-documented, empirical examples of antagonistic and mutualistic interactions within unipartite [21,22] and bipartite networks [14,23]. However, network tools are now being used to infer biotic interactions when direct, observational information on species' interactions is incomplete [11,24,25]. For instance, when working with unipartite co-occurrence networks, it is reasonable to predict that positive interspecific interactions are more likely to exist within modules than among modules [16]. In this way, modularity analysis may be a logical precursor to experimental tests of species' interactions [10].

With network modules now being used for such diverse purposes as detecting biogeographical regions and characterizing species' interactions, it is increasingly necessary to account for the effect of spatial scale [7,24]. For example, strong modular structure within a unipartite co-occurrence network may reflect positive interactions among species within shared modules and/or negative interactions among species in different modules. However, if modularity analyses use species co-occurrence data that are aggregated across large spatial extents that exceed individual species' ranges, frequent co-occurrence within a shared module may be mistaken for a positive interaction when it is in fact the result of environmental filtering [11]. Filtering of species among different habitat types will tend to place species with similar habitat requirements in close proximity, resulting in frequent co-occurrences that may be mistaken for positive interactions [26,27]. Alternatively, if environmental filtering or historical barriers to species' movements create disjunctions in species' ranges, strong modular structure may be misinterpreted as evidence of negative interspecific interactions among species in different modules [28].

In this study, our primary objective is to quantify the effect of spatial scale on perceived modular structure within unipartite networks. Secondary objectives are to determine whether network modules partition species into subgroups in a fundamentally different way than multivariate clusters, and to test if differences between modules and clusters are themselves scale dependent. The later objectives are motivated by recent efforts to assess whether network-

based algorithms for detecting subgroups within ecological communities are more effective than multivariate clustering algorithms [17,29,30].

To accomplish these objectives, we build co-occurrence networks for Mississippi River (USA) fish assemblages at three nested spatial scales. At each scale, we then use a three-step process to examine network modules and to assess scale dependence. First, we delineate network modules and compared the modules with clusters identified by a non-hierarchical clustering algorithm. These module vs. cluster comparisons test for congruence in the numbers of fish species that are assigned to the same groups when modularity and cluster analysis are used. Second, we test for 'spatial clustering' within network modules and multivariate clusters. Our intent is to determine whether modules and/or clusters are comprised of species with broadly overlapping ranges; if so, they may constitute distinct regional species pools or biogeographical units (*sensu* environmental filtering), rather than subsets of species joined by interactions *per se*. Third, we use a nest associate complex–fishes that deposit their eggs within the pebble mound nests of a common host species–as an empirical benchmark to determine whether network modules and/or multivariate clusters consistently assign members of the complex to the same group. This third step is of particular interest because past efforts to infer biotic interactions from co-occurrence data have often been constrained by a lack of corroborating, empirical evidence of real-time interactions [11,31] and by ambiguity in the specific types of interactions that one seeks to identify [24,32,33]. Interactions may be positive or negative, but they may also be symmetric or asymmetric. For instance, mutualism and commensalism are both positive relationships, but the former is symmetric, the later asymmetric. By using the nest associate complex, we focus on a specific interaction that is well-documented and precisely defined as positive symmetric, or mutualistic [34–36]. Finally, we compare results from each of the three steps to determine whether they vary in a consistent manner among spatial scales.

Materials and methods

Fish data within nested spatial units

Fish occurrence data were obtained from a large database of Mississippi River Basin fish assemblage samples, collected by the U.S. Environmental Protection Agency. We combined Mississippi Basin samples from the Environmental Monitoring and Assessment Program [37] and the National Rivers and Streams Assessment [38] databases. These two programs utilized similar, standardized field methods (single-pass backpack electrofishing surveys) that were calibrated with sampling effort curves (i.e., plotting the number of sample units needed to detect 95% of all locally occurring species) to ensure samples would be comparable among different sites [39–41]. Together, these databases provided occurrence records for 300 fish species distributed among 1018 Mississippi River Basin sites. Samples were organized at three spatial scales: the entire Mississippi Basin and two smaller, nested scales. We used 2-digit hydrologic units (HU-2) from the Watershed Boundary Dataset [42] to represent 'medium' sized river basins (mean size within Mississippi Basin = 231,968 km^2) and 4-digit hydrologic units (HU-4) to represent 'small' basins (mean size within Mississippi Basin = 26,190 km^2). The complete Mississippi River Basin was treated as a 0-digit hydrologic unit (HU-0; size = 3,247,552 km^2).

Co-occurrence networks, network modules, and clusters

Fish co-occurrence networks were built for the complete Mississippi Basin and for each of the nested sub-basins that included a sufficient number of sampling sites ($n \geq 70$ for HU-2 basins; $n \geq 35$ for HU-4 basins). Site × species occurrence matrices (species' presence-absence matrices) were first converted to species × species edge lists (two-column lists of species pairs that

co-occur at one or more sites) using the *cooccur* package in R [43]. Edge lists were then used to build unweighted, unipartite networks in R package *igraph* [44]. A workflow diagram of all data conversions and analyses is shown in Fig 1.

Network modules were identified with a simulated annealing algorithm [45]. Specifically, we used the 'cluster_spinglass' function (100 spins, start temperature = 1, stop temperature = 0.01, cooling factor = 0.99) in *igraph*. This function calculated an optimal modularity value for each network and partitioned species among distinct groups or modules, then exported the number of detected modules and lists of species membership within each module for further analysis. Simulated annealing was chosen because it generally outperforms other graph partitioning methods [46,47] and the computational burden in working with the modest sized fish network datasets was acceptable (< 10 minutes processing time for each network).

Next, multivariate cluster analysis was used as a second, alternative method to partition species among groups. We first calculated a Jaccard dissimilarity matrix for each of the original site × species occurrence matrices using the 'vegdist' function in R package *vegan* [48]. The dissimilarity matrices and the robust 'partitioning around medoids' (PAM) alternative to traditional k-means clustering [49], as implemented with the 'pam' function in R package *cluster* [50], were then used to partition species among clusters (see Fig 1). Notably, two strategies were used to specify the number of PAM clusters, k, for each of the fish datasets. First, k was made equal to the number of modules detected within a given network, so that direct tests of congruence in species' groupings among k modules and an equivalent number of PAM clusters could be performed (see next paragraph). Second, we used the gap statistic [51] to infer an optimal number of PAM clusters for each fish dataset and made k equal to the optimum. This latter method avoided potential circularity in the specification of k in cluster analyses, which could generate bias in spatial clustering tests (see 'spatial clustering' below) if the optimal number of PAM clusters within a given fish dataset was smaller or larger than the number of modules within the same dataset. However, the second method did not allow us to perform direct tests of congruence between module and cluster assignments (see next paragraph) because the numbers of modules and clusters tended to differ.

Congruence (C) between species' assignments within network modules and PAM clusters was quantified for each dataset with the procedure shown in Fig 2, while C significance levels were estimated with a permutation test. In each of 999 permutations, we randomized species' assignments within PAM clusters and re-calculated C between the observed network modules and the permuted clusters. A p-value for C was then estimated as the proportion of permutations in which the randomized C value was equal to or greater than the observed C value. All C tests were performed with a custom Visual Basic for Applications function in Microsoft Excel.

Spatial clustering

Spatial clustering tests focused on the spatial orientations of nodes within and among network modules when node spatial positions were estimated as the centroids of the respective species' ranges (see next paragraph and Fig 1A). We reasoned that if species within the same modules are positioned closely together in space, relative to species in other modules (Fig 1B), then explanations for modular structure that invoke positive species interactions within modules, such as commensalism or mutualism, would be weakened. In this spatial clustering scenario, modular structure could be parsimoniously explained by historical events or environmental filtering processes that lead to overlapping species ranges. Alternatively, if no spatial clustering of species within modules is observed (Fig 1C), then shared module membership may indeed reflect positive or facilitative interactions among species. We also assumed that spatial clustering will be most prevalent at the largest spatial scales, where spatial clusters will tend to

Presence-absence matrix

	Site 1	Site 2	Site 3	Site 4
Sp. A	1	0	0	0
Sp. B	1	1	0	0
Sp. C	0	1	1	1

{cooccur}

{vegan}

Species edge list

Sp. A – Sp. B	----→	0.505
Sp. A – Sp. C	----→	-0.420
Sp. B – Sp. C	----→	0.238

Pairwise effect sizes

Used in nest associate tests

Jaccard dissimilarity matrix

	Sp. A	Sp. B	Sp. C
Sp. A	0.000	0.427	0.981
Sp. B	0.427	0.000	0.546
Sp. C	0.981	0.546	0.000

{igraph}

{cluster}

Unipartite network

PAM clusters
(k = no. modules)

Simulated annealing

Network modules

C; $MD_{within:among}$

$MD_{within:among}$
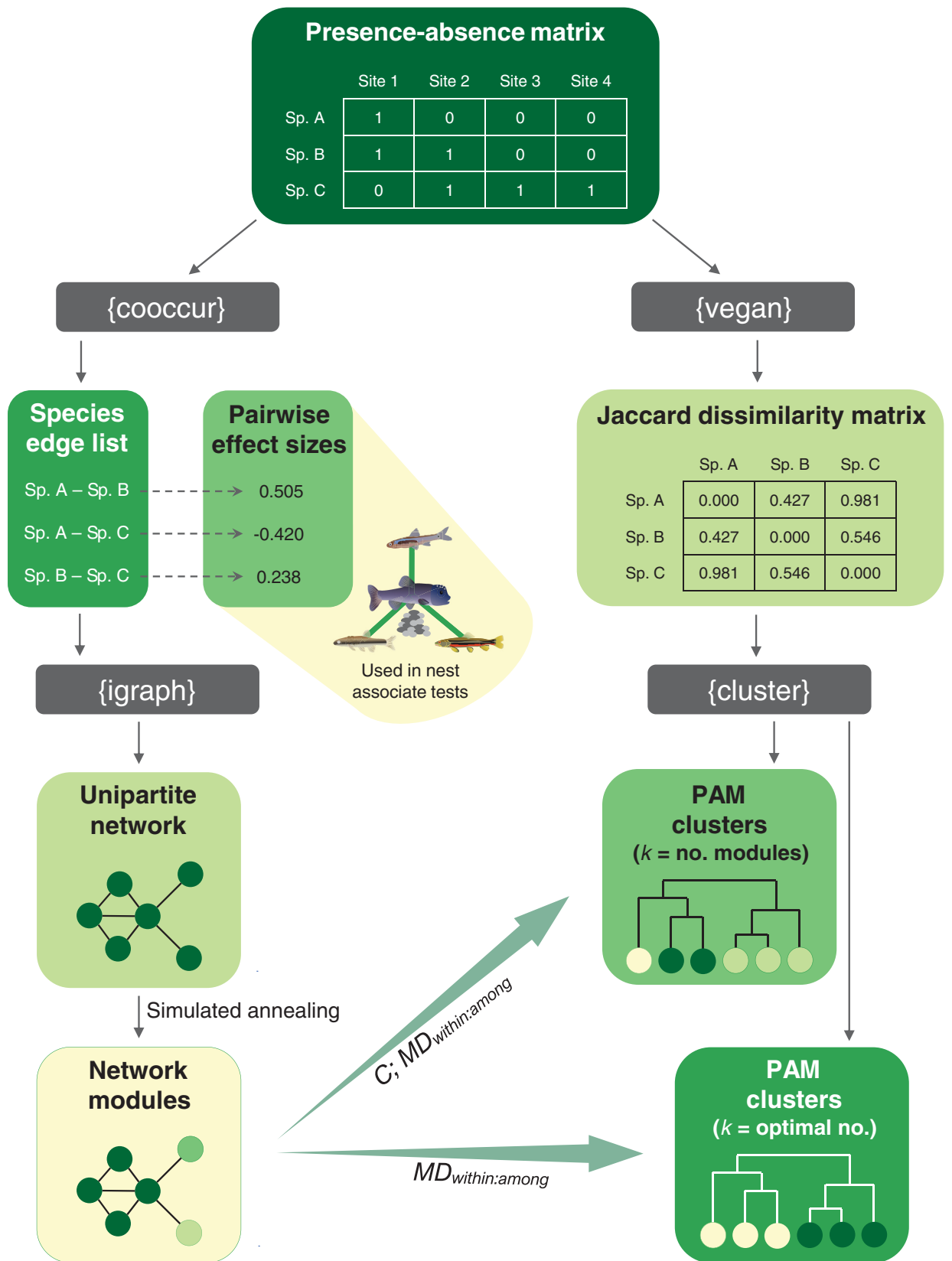
PAM clusters
(k = optimal no.)

Fig 1. Workflow diagram of the network and cluster analyses. For each of the 10 nested river basins included in the study, a species' presence-absence matrix was first compiled then converted to a species' edge list or a species × species Jaccard dissimilarity matrix. Edge lists were used to build unipartite networks, followed by modularity analysis through simulated annealing. (Pairwise effect sizes from cooccurrence analyses were estimated for use in subsequent tests of nest associate species and are not utilized within the workflow diagram.) Dissimilarity matrices were used in PAM cluster analyses. Comparisons of network modules and PAM clusters included congruence (*C*) analysis and ratios of mean distances within and among groups ($MD_{within:among}$) when the number of PAM clusters (*k*) was equal to the number of modules (*k* = no. modules). However, network and cluster comparisons were limited to $MD_{within:among}$ when an optimal number of PAM clusters was independently identified with the gap statistic (*k* = optimal no.). R packages used in each step of the workflow are shown in curly brackets within gray bubbles (e.g., 'igraph').

represent distinct biogeographic provinces, and least prevalent at the smallest scales. We therefore performed spatial clustering tests at each of the three nested scales (HU-0, 2, and 4) and tested for spatial clustering among PAM clusters as well as modules.

To perform spatial clustering tests, we first located the range centroid for each species that was included in one of the co-occurrence networks. We started by importing and analyzing species' distribution data from NatureServe [52] in a geographic information system. The NatureServe fish database combines fish occurrence records (point samples) from published research and state-sponsored natural heritage programs, then uses these records to document species' occurrences within small drainages (mean area within Mississippi River Basin = 4022 km^2), as represented by 8-digit hydrologic units (HU-8; USGS 2013). We used the 'Feature to Point' tool in ArcGIS 10.1 (Environmental Systems Research Institute, Redlands, CA) to interpolate the spatial centroid of every HU-8 within the Mississippi Basin. We then queried all HU-8's from the NatureServe [52] database with known occurrence of a given species and calculated the overall range centroid as the mean *x* coordinate and the mean *y* coordinate among all HU-8 occurrence centroids. This interpolation process is illustrated in Fig 3A and was repeated at each of the three nested scales for every species included in one of the co-occurrence networks. Prior to interpolation, all spatial data were converted to Albers equal-area (NAD83) conic projection.

Next, we used the species' range centroids to calculate mean distances (*MD*) among species within groups (MD_{within}) and mean distances among groups (MD_{among}) for all pairwise combinations of species. For each of the nested fish datasets, MD_{within} and MD_{among} were independently calculated for network modules and for PAM clusters, both when cluster *k* was equal to the number of network modules and when an optimal *k* was determined with the gap statistic. Our process was modeled after the mean similarity study of [53]. However, we calculated MD_{within} and MD_{among} as true Euclidian distances in units of km (see Fig 3), rather than unit-less dissimilarity index (e.g., Jaccard dissimilarity) values. Significance levels were then estimated with the multi-response permutation procedure of Mielke et al. [54]. In each of 999 permutations, we randomly shuffled species among groups (network modules or PAM clusters) then recalculated MD_{within} and MD_{among} for the permuted data. These permutations focused solely on the average spatial proximities of species within and among groups, without invoking a more liberal process of simulating species' ranges (i.e., randomizing species' occurrences within HU-8 units; see Fig 3A) then re-interpolating the range centroids. Finally, a spatial clustering *p*-value was calculated as the proportion of permutations in which $MD_{within} \geq MD_{among}$. This tested the null hypothesis that within-group distances were, on average, smaller than among-group distances (i.e., $MD_{within:among}$ ratios < 1), indicative of spatial clustering. Spatial clustering tests were performed with the 'meandist' and 'mrpp' functions in *vegan*.

Nest associates

Nest association within freshwater fish assemblages is a specialized (but not uncommon) reproductive strategy in which associate species seek out and deposit their eggs in pebble mound nests built by a host species [55]. At a minimum, this relationship constitutes
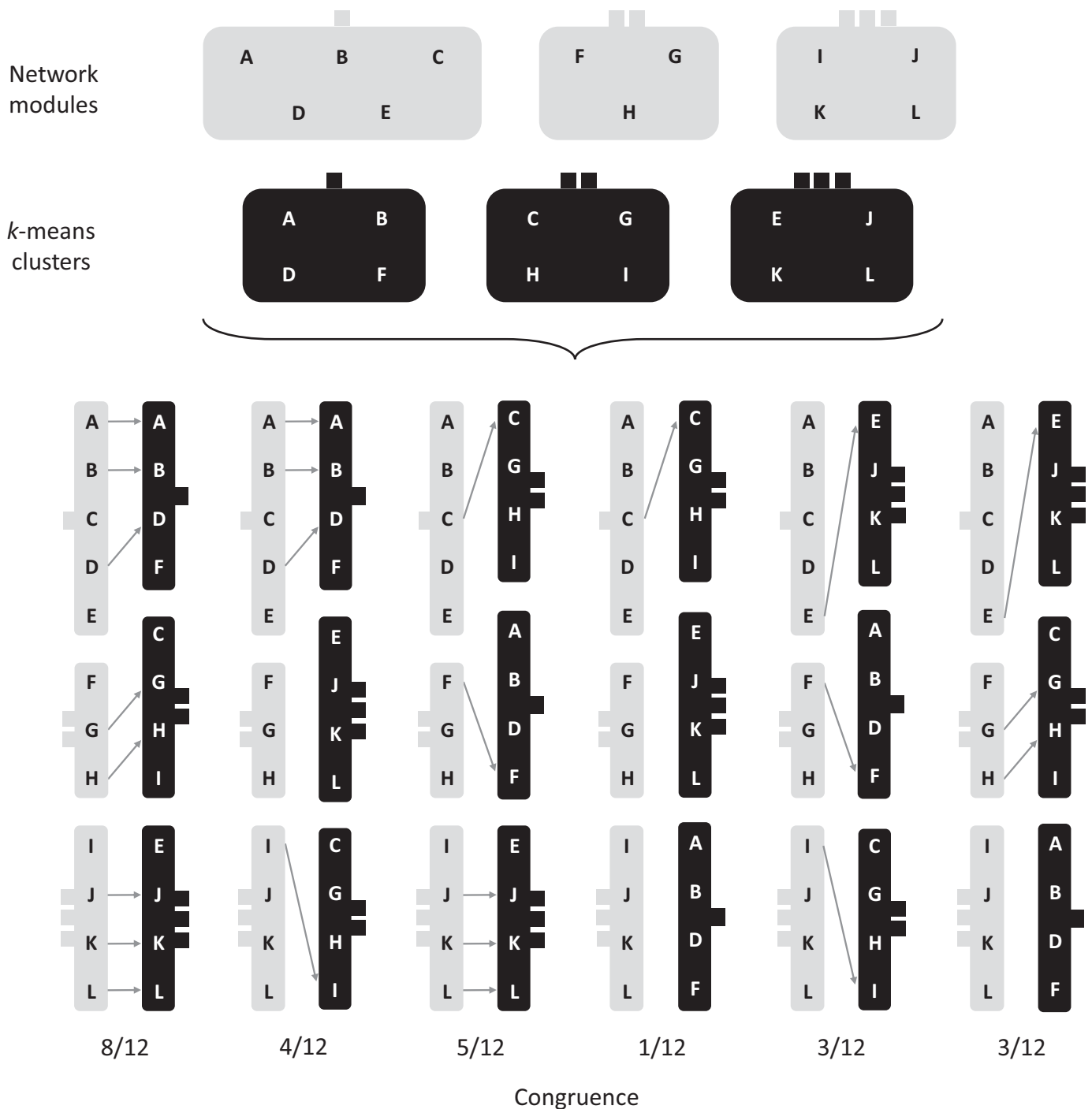
Fig 2. Illustration of the process used to quantify congruence (*C*) among network modules and PAM clusters. Hypothetical results are shown at the top of the diagram for network modules and PAM clusters: the same 12 species (A—L) were first partitioned among three network modules (grey boxes), then among three PAM clusters (black boxes). Partitioning of species among network modules and PAM clusters was conducted independently, though the number of PAM clusters was determined by (i.e., equivalent to) the number of network modules identified by the simulated annealing algorithm. Note that the number of species assigned to each module and cluster may vary and is determined by the annealing and clustering algorithms respectively. In this hypothetical example, species numbers are variable among modules, with five, three, and four species assigned to the first, second, and third modules respectively, but four species are assigned to each of the three clusters. Congruence between network modules and PAM clusters is based on the number of instances in which species are grouped together in the same module and cluster. For example, in the first pair of columns shown at the lower-left side of the diagram, a total of eight species are assigned to the same module and cluster groups, as indicated by gray arrows. Congruence in this instance is calculated as eight matches divided by the total number of species (i.e., $C = 8 \div 12 = 0.67$). To aid in interpretation, modules and clusters are identified by the number of 'tabs' assigned to each; one, two, or three tabs per module or cluster are shown and are consistent

among the upper and lower parts of the diagram (e.g., the black cluster with two tabs consistently contains species C, G, H, and I). When assessing *C*, it is critical to recognize that the labels used to identify network modules and PAM clusters (one, two, or three tabs in this illustration) are arbitrary. Only the shared identities of the species' lists within each module or cluster are important. Therefore, a complete test of *C* cannot be performed by simply comparing module 1 vs. cluster 1, module 2 vs. cluster 2, etc. Rather, the level of congruence between modules and clusters must account for multiple module vs. cluster combinations. This is achieved by 'rotating' the clusters in a combinatorial manner, as shown in the 2nd through 6th pairs of columns in the lower part of the diagram. The goal is to investigate all possible combinations of modules and clusters while searching for the *highest possible* level of *C*, given the constraint of the observed species' assignments within modules and clusters (shown at top of diagram). Note, however, that the network modules do not need to be rotated during the combinatorial comparisons with the PAM clusters, as the objective is to assess the degree to which species' assignments within clusters match species' assignments within modules. Thus, with a system of three modules and three clusters, six cluster rotations are needed to explore all possible module vs. cluster combinations. The observed *C* value for each of the six rotations is shown at the bottom of the diagram. Hence, the first combination of modules and clusters (i.e., first two columns at lower-left) in this illustration leads to the highest possible *C* value which is then taken as the overall amount of congruence between the modules and clusters.

asymmetric commensalism; associate species incubate their eggs in well-sorted substrates with superior aeration and parental guarding, at no cost to the nest host [56,57]. But in many instances, this relationship may constitute symmetric mutualism, as egg survival is maximized for nest-building species through a predatory dilution effect [36]. One specific, well-documented example of nest association in streams of the eastern Mississippi River Basin see [34,35,58]) is the connection between the relatively large-bodied, nest-building Bluehead Chub (*Nocomis leptocephalus*) and six species of smaller-bodied minnows (family Cyprinidae), including Rosyside Dace (*Clinostomus funduloides*), Mountain Redbelly Dace (*Chrosomus oreas*), Central Stoneroller (*Campostoma anomalum*), White Shiner (*Luxilus albeolus*), Rosefin Shiner (*Lythrurus ardens*), and Crescent Shiner (*Luxilus cerasinus*).

Focusing exclusively on the river basin in which Bluehead Chub nest associates have been most carefully studied–the Kanawha River Basin (HU-4 scale), nested within the Ohio River Basin (HU-2 scale) and Mississippi River Basin (HU-0 scale)–we used the nest associate complex to evaluate the network analysis results in two ways. First, we assessed the degree to which the nest associate complex was preserved within network modules and PAM clusters (*k* = number of modules and the optimal number of clusters), at each of the three nested spatial scales. The underlying logic was that the majority or entirety of the nest associate complex should be preserved within a single group if modules or clusters are comprised of species linked by positive associations.

Second, we calculated an effect size for every pairwise link within a given network, then compared effect sizes that were exclusive to the nest associate complex with average effect sizes from the complete networks. This tested the hypothesis that links between nest associates would be among the strongest in the network. Pairwise effect sizes were calculated directly from the raw fish occurrence data, using the probabilistic co-occurrence model of Veech [59,60]. Briefly, this model uses a combinatorics approach to calculate the probability that two species will co-occur at *j* sites, then to compare this expected value with an observed co-occurrence value. Observed co-occurrence values that exceed the expected value by a significant margin provide evidence of positive associations and vice-versa. Following Veech [60], effect sizes were calculated as the differences between observed and expected co-occurrences for each pairwise association, using R package *cooccur*.

Data and code

Complete fish occurrence data, species' range centroids, R code to reproduce all network and cluster analyses, and the Visual Basic for Applications function used in congruence tests are available on Figshare at https://doi.org/10.6084/m9.figshare.c.4151780.v1

Results

Sample sizes within basins were adequate to build four fish co-occurrence networks at the HU-2 scale and five at the HU-4 scale, in addition to the HU-0 scale network for the entire
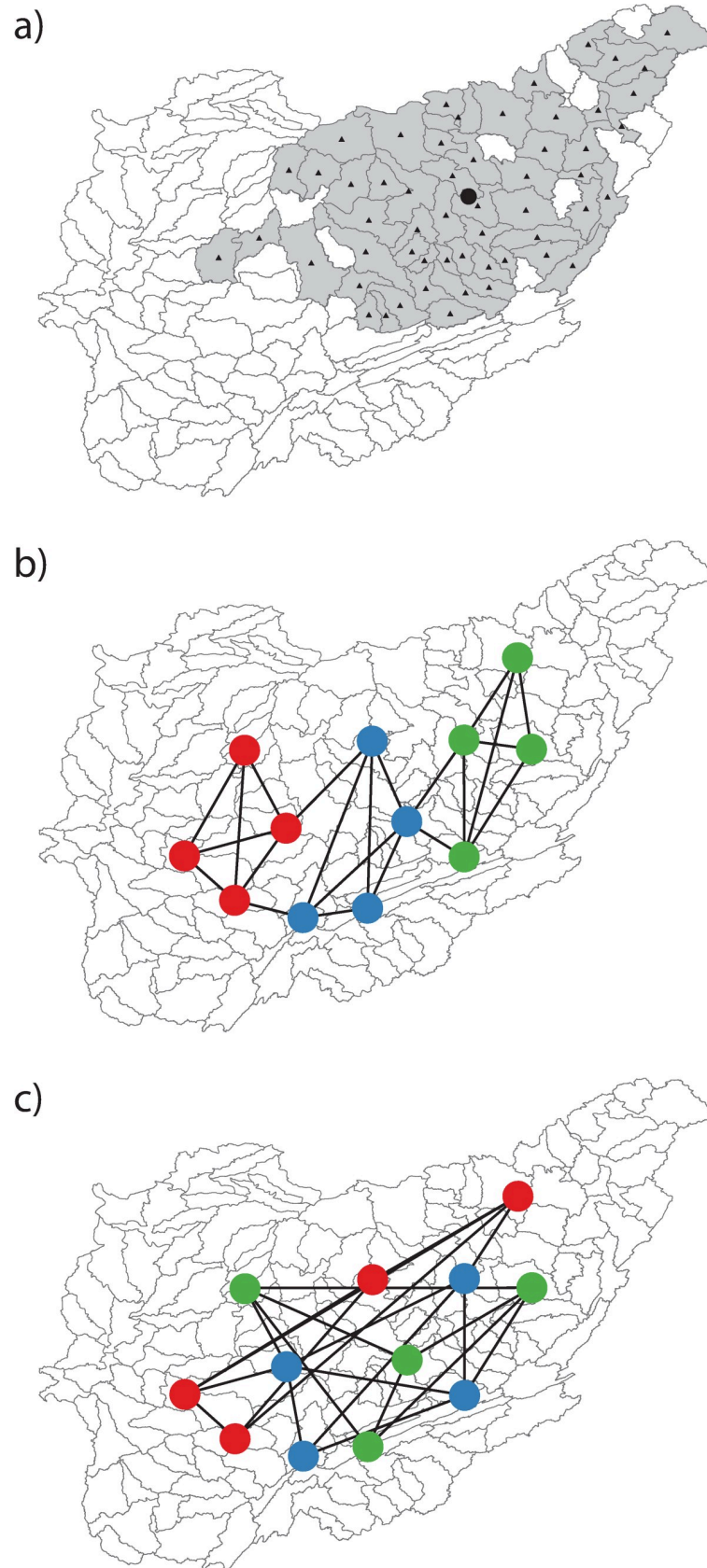
Fig 3. Hypothetical maps of fish co-occurrence networks. These maps demonstrate spatial clustering and the absence of spatial clustering. Each map is centered on the Ohio River Basin with 8-digit hydrologic units delineated by grey lines. Panel **a** illustrates the process used to interpolate range centroids for individual species. In this example, the native range of the Variegate Darter (Etheostoma variatum) is indicated by shaded grey polygons (see main text for source data). The center or 'centroid' of each range polygon, interpolated in a two-dimensional Cartesian plane, is indicated by a black triangle. The master centroid of the species' native range, calculated as the grand mean of the x and y coordinates for individual range polygon centroids, is shown as a black circle. Panel **b** illustrates a hypothetical network of 12 fish species, partitioned into three distinct network modules (red, blue, and green circles). In this instance, strong spatial clustering is evident. Panel **c** illustrates a similar fish network, but one that is characterized by a lack of spatial clustering.

Mississippi River Basin (Table 1). The number of modules detected in each network ranged from 2–4 and was not a clear function of spatial scale, as the minimum and maximum number of modules were both associated with the smallest (HU-4 scale) networks.

Congruence in species' membership among network modules and an equivalent number of PAM clusters (k = no. modules) was variable, ranging from 0.48–0.88, and generally decreased as the number of modules increased (Table 1). This inverse trend between C and the number of modules was intuitive because placement of species within the same groups is, on average, less likely when the number of groups is larger. For all networks, the C permutation test results were highly significant. In no instance did the randomized C value exceed the observed C value ($p < 0.001$ in all permutation tests; not shown in Table 1). Thus, we found no evidence to support the hypothesis that the observed C values can be attributed to random partitioning of species among an equivalent number of network modules and PAM clusters.

Spatial clustering tests revealed a high level of clustering at each of the three spatial scales. For network modules, MD_{within} was significantly ($p < 0.05$) smaller than MD_{among} (i.e., $MD_{within:among}$ ratios < 1) in 9 of 10 basins (Table 1). For PAM clusters, $MD_{within:among}$ ratios were significantly < 1 in 7 of 10 basins when k was equal to the number modules (k = no. modules; Table 1), and in 6 of 7 basins when an optimal k value was determined with the gap statistic (k = optimal no.; Table 1). Notably, the prevalence and magnitude of spatial clustering did vary among scales. At the smallest HU-4 scale, highly significant spatial clustering ($p \leq 0.005$) was detected in only 2 of 5 river basins (Upper Tennessee and Missouri). At the larger HU-0 and HU-2 scales, highly significant ($p \leq 0.005$) spatial clustering was detected in all river basins and MD_{within}:MD_{among} ratios were consistently smaller than at the HU-4 scale. In all cases, spatial clustering test results were generally similar among network modules and PAM clusters; MD_{within}:MD_{among} ratios and p-values did not deviate strongly among modules and clusters (inclusive of both k = no. modules and k = optimal no. results) when comparing results for a given river basin (Table 1).

Analysis of the nest associate data suggested that network modules may outperform PAM clusters when the objective is to detect groups of species that are linked by positive interactions. At each of the three spatial scales, the complete nest associate complex (i.e., the Bluehead Chub and its six associates) was preserved in a single network module (Table 2). However, when species were partitioned among PAM clusters, results were more variable. At the HU-0 scale, all members of the nest associate complex were assigned to the same cluster. But at the smaller HU-2 and HU-4 scales, only 2 of 6 and 3 of 6 nest associates were included in the same cluster as the Bluehead Chub, respectively, when the number of PAM clusters was equal to the number of network modules. When an optimal number of PAM clusters was used (two clusters rather than four; see Table 1), all of the nest associate species were again included in the same cluster. In Fig 4, we illustrate the complete HU-4 scale fish network for the Kanawha River (panel a), with distinct modules indicated by node colors. Links between the Bluehead

Table 1. Congruence (C) and spatial clustering results for fish network modules and partitioning around medoids (PAM) clusters.

				Spatial clustering						
				Network modules		PAM clusters				
						k = no. modules		k = optimal no.		
Scale	River basin	Modules	C	$MD_{within:among}$	p	$MD_{within:among}$	p	GS no.	$MD_{within:among}$	p
HU-0	Mississippi	4	0.50	0.75	0.001	0.75	0.001	3	0.83	0.001
HU-2	Ohio	4	0.51	0.85	0.001	0.76	0.001	4	[a]	
HU-2	Upper Mississippi	4	0.69	0.80	0.001	0.82	0.001	3	0.90	0.005
HU-2	Missouri	3	0.65	0.54	0.001	0.67	0.001	2	0.60	0.001
HU-2	Arkansas-White-Red	3	0.73	0.79	0.001	0.74	0.001	2	0.76	0.001
HU-4	Allegheny	3	0.80	1.02	0.372	1.07	0.185	1	[b]	
HU-4	Upper Ohio	2	0.88	0.85	0.001	0.86	0.001	4	0.79	0.001
HU-4	Kanawha	4	0.48	0.88	0.001	0.75	0.001	2	0.82	0.001
HU-4	Upper Tennessee	4	0.56	0.97	0.030	0.97	0.078	2	0.98	0.139
HU-4	Missouri	3	0.61	0.89	0.043	1.07	0.933	3	[a]	

Results are shown at three spatial scales, corresponding to hydrologic units (HU-0, -2, -4). *Modules* is the number of distinct network modules detected with simulated annealing for each fish network. *C* is the maximum proportion of fish species that were assigned to the same groups by modularity and PAM cluster analyses, when the number of PAM clusters (k) was equivalent to the number of network modules. $MD_{within:among}$ is the ratio of the mean Euclidian distance (between species' centroids; see main text) within and among groups. $MD_{within:among}$ values are shown with permutation test (999 iterations) p-values for network modules, for PAM clusters when k was equivalent to the number of modules, and for PAM clusters when k was equal to the optimal number of clusters identified with the gap statistic. *GS no.* is the number of PAM clusters detected when the gap statistic was used.

[a] Values not reported because the optimal number of PAM clusters was equal to the number of network modules and spatial clustering test results were therefore identical.

[b] Values not reported because the optimal number of PAM clusters was one (i.e., among group comparisons were not possible).

Chub and its nest associates are highlighted in Fig 4B, where we isolate and magnify the module that contains the nest associate complex.

Comparisons of co-occurrence effect sizes indicated that the 'signal' of the nest association is strongest at the smallest scale. The mean effect size between the Bluehead Chub and each of its six nest associates decreased by a large margin with successive increases in scale (Table 2). But at each scale, the mean effect size of the nest associate complex was much larger than the mean effect size when calculated for all species pairs within the network. For instance, in the Ohio River Basin (HU-2 scale), the mean effect size within the nest associate complex (0.027) was approximately one order of magnitude greater than the mean effect size among all species pairs within the network (0.002; see Table 2).

Table 2. Representation of the Bluehead Chub (*Nocomis leptocephalus*) nest associate complex at three spatial scales.

| | | Completeness | | | Effect size | |
| | | | PAM cluster | | | |
Scale	River basin	Network module	k = no. modules	k = optimal no.	Nest associates only (1 s.d.)	All species in network (1 s.d.)
HU-0	Mississippi	1.00	1.00	1.00	0.007 (0.004)	0.001 (0.005)
HU-2	Ohio	1.00	0.33	0.33	0.027 (0.015)	0.002 (0.010)
HU-4	Kanawha	1.00	0.50	1.00	0.072 (0.036)	0.016 (0.021)

At each of the three scales (HU-0, -2, -4), the completeness of the nest associate network (i.e., the proportion of the six known Bluehead Chub nest associates that was included in the same module or cluster as the Bluehead Chub) is shown for network modules and for PAM clusters. Separate completeness results are shown for PAM clusters when the number of clusters was equal to the number of network modules and when an optimal number of clusters was determined with the Gap Statistic (see main text). Mean effect sizes and standard deviations (s.d.) from the probabilistic co-occurrence model are also shown at each scale for species-pairs that were limited to the Bluehead Chub nest associates and for all species-pairs within a given network.
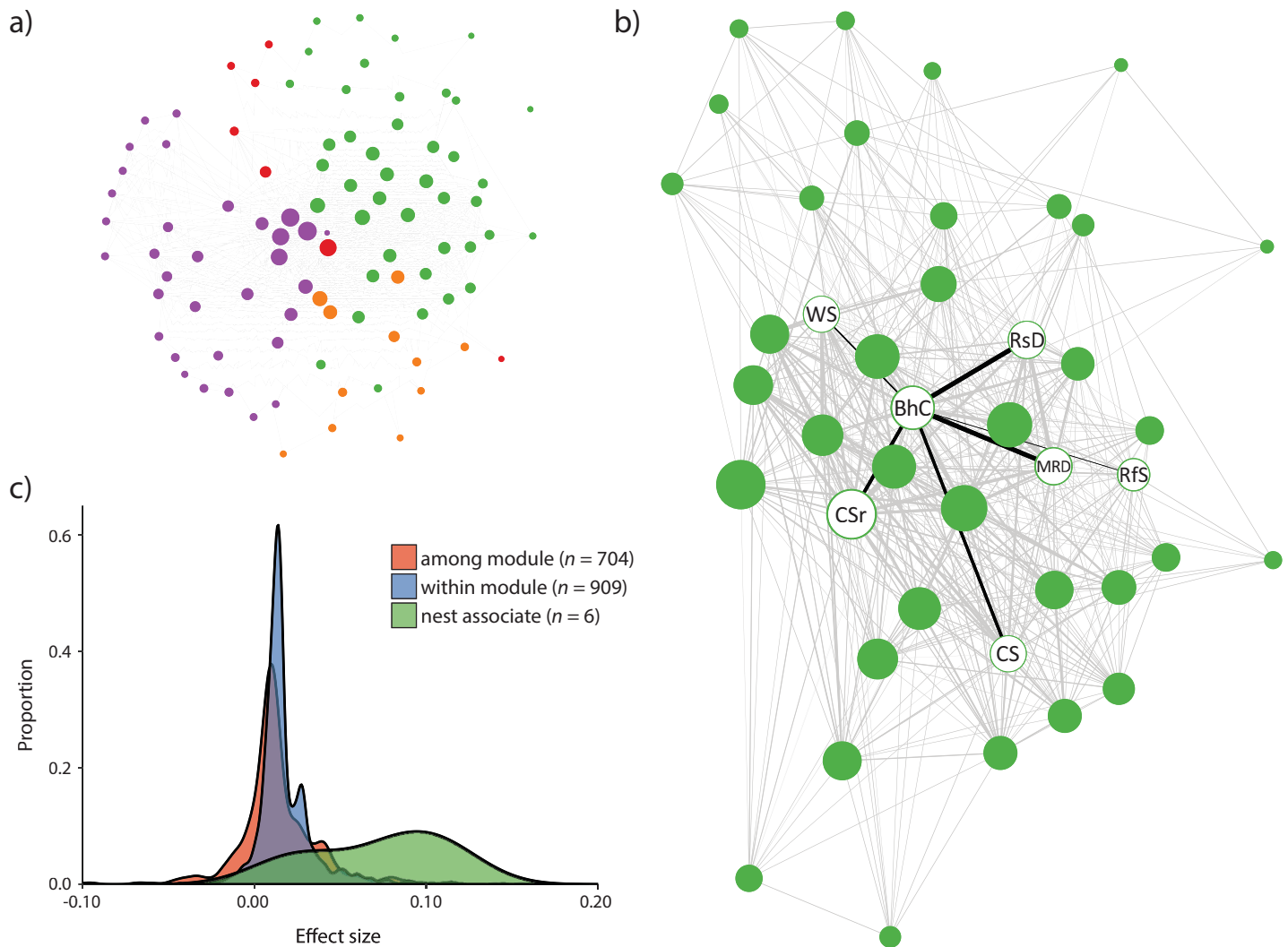
Fig 4. The Kanawha River Basin (HU-4 scale) fish network. Panel **a** shows the complete network of 94 fish species, with co-occurrences among species indicated by light grey edges and the four distinct modules indicated by node colors. The network was plotted with the Kamada-Kawai force-directed layout, which positions the most highly connected nodes near the center and weakly connected nodes along the periphery. (Note that this network does not incorporate species' centroids in the layout.) Panel **b** magnifies the Kanawha River network module (green nodes) that includes the Bluehead Chub ('BhC'; *Nocomis leptocephalus*) and its six known nest associates: White Shiner ('WS'; *Luxilus albeolus*), Rosyside Dace ('RsD'; *Clinostomus funduloides*), Mountain Redbelly Dace ('MRD'; *Chrosomus oreas*), Rosefin Shiner ('RfS'; *Lythrurus ardens*), Crescent Shiner ('CS'; *Luxilus cerasinus*), and Central Stoneroller ('CSr'; *Campostoma anomalum*). Edge widths within the green module are proportional to the effect sizes estimated with the probabilistic model of co-occurrence (see main text) and nest associate pairs are highlighted by black edges. Panel **c** illustrates density functions (kernel estimates) for three groups of co-occurrence effect sizes within the Kanawha River fish network: all edges between species in different modules ('among module'), all edges between species in the same modules ('within module'), and a group that is exclusive to the six edges between Bluehead Chub and each of its nest associates.

Discussion

Spatial scale and network modularity

Network analyses are now common in ecological and biogeographical research, yet many basic questions on methodology and scale dependence remain unanswered [24,28]. We examined the effect of spatial scale on network modularity, using freshwater fishes within the Mississippi River Basin as a case study. Specifically, we tested for spatial clustering within network modules at three spatial scales to determine whether modular patterns reflect a combination of

overlapping species' ranges within modules and disjunct species' ranges among modules. In this way, we used spatial clustering as a null model for tests of biotic interactions within modules. It is perhaps logical to expect that the dense associations *within* network modules reflect positive interactions such as commensalism or mutualism, while the sparse associations *among* modules reflect negative interactions, such as amensalism or competition. But if spatial clustering within modules is strong, as indicated by $MD_{within:among}$ ratios < 1, then explanations for modular structure that invoke biotic interactions may be unjustified. In such scenarios, positive co-occurrence may simply be due to environmental filtering (e.g., shared habitat preferences) rather than direct positive interactions.

Significant spatial clustering ($p < 0.01$) was detected at all three spatial scales (Table 1). However, the strength of the clustering was variable and declined at smaller scales, as indicated by larger $MD_{within:among}$ ratios for HU-4 scale networks. Significant clustering was also least common for HU-4 scale networks; only 2 of 5 HU-4 scale networks exhibited significant clustering at $p < 0.01$. These observations matched our initial prediction that spatial clustering would be least pronounced at the smallest scale. By extension, they were consistent with the hypothesis that inferences regarding species' interactions are less likely to be confounded with non-interactive processes, such as environmental filtering, when network analyses are conducted at smaller spatial scales.

Interestingly, the number of modules detected within each river basin was not a clear function of spatial scale. At each scale, a comparable number of modules, usually 3–4, were detected (Table 1). This was surprising because the river basins were spatially nested; HU-4 scale rivers were nested within the HU-2 rivers and each of the HU-2 rivers was nested within the HU-0 scale Mississippi River Basin. If modules within the HU-4 basins are truly discrete subcommunities that share few links with each other, they might well be preserved at larger spatial scales, leading to an additive increase in the number of modules at larger scales. For instance, three, two, and four modules were detected in the HU-4 scale Allegheny, Upper Ohio, and Kanawha River Basins, respectively, and each was nested within the HU-2 scale Ohio River Basin. Thus, a reasonable expectation would have been that the Ohio River fish network would include more than four modules. The fact that it did not warrants a cautious approach when using modules to infer species' interactions, but it may also be an artifact of species' occurrences in more than one of the HU-4 scale rivers. For example, 81 fish species occurred in both the Allegheny River and the Kanawha River. Accordingly, each of these species was featured more than once in the HU-4 scale analyses (once per network × multiple networks), but only once in the HU-2 scale Ohio River network. Because each species within a network must be assigned to a single module, it is difficult to anticipate how transitions between nested scales will affect the number or composition of network modules. We therefore suggest that a hierarchical algorithm capable of detecting submodules within modules, such as map equation [61,62], would be a useful next step in the analysis of the fish co-occurrence networks.

Network modules vs. multivariate clusters

Several authors have compared network modules with multivariate clusters, assessing whether one method provides unique insight or otherwise outperforms the other. For instance, in a continental scale study of Australian plants, Bloomfield et al. [30] concluded that network modules were more effective than multivariate clusters in detecting biogeographical regions, as well as fine-scale transition zones among regions. Similarly, Vilhena and Antonelli [29] reported that network modules were superior to multivariate clusters when searching for biogeographical regions at continental to global scales. Carstensen et al. [17] refrained from

labeling one method as superior to the other but did offer a fundamental distinction: 'Whereas the distance-based clustering methods group [species] according to calculated distances between pairs of [species], the network approach seeks to account for the entire link structure of the network by minimizing links between modules.'

We built upon previous comparisons of network modules and multivariate clusters in two novel ways. First, we repeated the spatial clustering tests from the network modules for the PAM clusters. $MD_{within:among}$ ratios were compared between modules and clusters (in a given river basin) to determine whether one method is more prone to detect groups of species with spatially aggregated ranges than the other. When the number of PAM clusters k was constrained to match the number of modules detected within a given fish co-occurrence dataset, we observed no major differences in spatial clustering results. $MD_{within:among}$ ratios and spatial clustering p-values were similar for modules and clusters when compared among the 10 nested river basins (Table 1). Furthermore, these similarities were preserved when the gap statistic was used to identify an optimal, unconstrained number of PAM clusters. These results suggest that network modules are neither more nor less likely to identify groups of spatially clustered species than multivariate clusters.

Second, we measured congruence C in the numbers of species that were assigned to the same groups by modularity and clustering algorithms. To our knowledge, this is the first study to focus explicitly on species' lists within modules and clusters, and to compare lists between methods (see Fig 2). We found that species are not partitioned among network modules and PAM clusters in the same way. Across all spatial scales and river basins, median C in individual species' assignments among modules and clusters was 0.64 and C never exceeded 0.80 when at least three groups were being compared (Table 1). Results from the C tests should be interpreted with caution, as they were only performed for PAM clusters when k was equal to the number of modules. We did not calculate C when the numbers of modules and clusters (based on the gap statistic optimal k) differed because in these situations $C = 1$ was, by definition, an impossible outcome. In this way, we have addressed the question 'how closely does species composition of PAM modules match species composition in network modules' without evaluating the opposite question (i.e,. calculating C when the number of modules is constrained to equal k from an optimal PAM clustering solution). Nevertheless, we have shown that when species are partitioned among network modules, species' lists within modules will tend to differ, sometimes by a large margin, from group lists that would be obtained with a multivariate clustering algorithm.

Our C test results are particularly relevant in a regional species pool context. The regional species pool is defined as the set of regionally distributed species that could potentially colonize a given locality within that region, and it is a fundamental unit in biogeography and community ecology [63–66]. For instance, when species' occurrences are sampled from a common species pool, the sign (+ vs. -) of an interspecific relationship can sometimes be inferred from co-occurrence data [67–70]. Species that rarely or never co-occur at the same sites may provide evidence of competitive exclusion, while frequent co-occurrence may provide evidence of positive interactions such as nest associations. But these co-occurrence patterns can easily be conflated with environmental filtering if the spatial scale of a study is large enough to effectively combine two or more regional pools. It is therefore critical to define the regional species pool in an objective and explicit manner. Network modules and multivariate clusters are both reasonable approximations of regional species pools [17] but as the C results show, they are not equivalent and will likely lead to different outcomes in studies of community assembly.

Biotic interactions

Discrepancy in the *C* results between modules and PAM clusters begs an obvious question: which method is preferable? We suggest the nest associate results provide a meaningful context to address this question. Network modules were clearly superior to PAM clusters in preserving pairwise associations between the nest building Bluehead Chub and its nest associates. At all three spatial scales, the six documented nest associates were assigned to the same network module as the nest host (Table 2). However, when fishes from the same river basins were partitioned among an equivalent number of PAM clusters (*k* = no. modules) at the HU-2 and HU-4 scales, no more than one-half of the nest associates were placed in the same cluster as the host. The complete nest associate complex was only maintained in a single cluster at the HU-0 scale and this was likely due to spatial clustering within the large Mississippi River Basin, in which the Ohio River Basin and its fish fauna comprised a distinct biogeographic region. When the number of PAM clusters was not constrained to equal the number of modules (optimal *k* determined from the gap statistic) in the HU-4 scale Kanawha River Basin, the entire nest associate complex was preserved in a single cluster (Table 2). However, this was potentially an artifact of the smaller number of groups included in the optimal PAM solution; the odds of assigning the complete 7-species nest associate complex to a single group were logically greater when the Kanawha River fishes were partitioned among two PAM clusters, rather than four network modules (Table 1). Thus, we suggest that network modules may be superior to non-hierarchical, multivariate clustering tools, such as the PAM clusters used here, when the objective is to identify species linked by positive interactions.

Co-occurrence effect sizes further emphasize the importance of the nest associate results. At each of the three spatial scales, effect sizes within the nest associate complex were conspicuously larger than average effect sizes throughout the entire network (see Table 2 and Fig 4C). This result is intriguing because it shows that strong positive interactions among species can potentially be detected at very coarse scales. Even at the complete Mississippi River Basin (HU-0) scale, the mean effect size among nest associates was seven-fold larger than the mean effect size for the entire network. Our results now join a growing number of studies that have shown, on both theoretical [24] and empirical grounds [11,71,72], that co-occurrence data can be used to detect positive interactions across a range of spatial scales.

Moving forward, a significant challenge will be to determine which of the many remaining network links reflect biotic interactions and which reflect species that occur in the same habitats but are not interactive *per se*. For example, the complete HU-4 scale Kanawha River fish network includes 909 within-module links (summed among the four modules) and 704 among-module links (Fig 4A). Focusing exclusively on the module that includes the Bluehead Chub (green module in Fig 4A) reveals that the nest associate complex is central to the structure of the module; the force-directed layout used to build this network graph placed the most highly connected nodes near the center (Fig 4B). But the same graph also shows that the six nest associate links account for a tiny fraction of the 503 links within the module. Clearly, many other factors must have an influence on fish coexistence in the Kanawha River Basin.

We suggest that several non-exclusive strategies may be helpful in sorting through the large number of unexamined network links. First, effect sizes for all species pairs could be ranked and a significance threshold used to remove relatively weak links, prior to building and analyzing the networks [11]. This process would be straightforward because the probabilistic co-occurrence model used here to calculate effect sizes also includes a significance testing algorithm (see [59,60]). In a similar way, replacing the species' presence-absence data used in each of our analyses (see Fig 1) with species' abundance data (density estimates or relative abundances) might enhance our ability to detect and characterize species' interactions. Presence-

absence data, which are more readily available than abundance data, have been used extensively in traditional community ecology research (e.g., [73–75]) as well as more recent network-based studies (e.g., [16,72,76]). But turnover in species' composition is rarely a binary or punctuated event. Rather, the addition or loss of a species from a local community is most often the result of a gradual increase or decrease in population size. Abundance data may therefore offer greater power to characterize interspecific relationships or to detect subtle changes in them [77–80].

Second, functional traits could be appended to species within the networks and used to search for traits-based patterns within or among modules [3,81]. For instance, the Bluehead Chub nest associate complex is particularly well-documented, but other stream fishes exhibit similar nest building and/or association behaviors [36,82]. It is therefore likely that analogous functional trait patterns or 'motifs' (*sensu* [83]) may occur within the network modules. In the Kanawha River Basin example, a comparable nest association motif might exist in the 'purple' module of Fig 4A, where another *Nocomis* nest builder, the River Chub (*Nocomis micropogon*), is prevalent. Recognition of repetitive functional motifs among modules might even be an effective way to guide the formulation and testing of hypotheses on other instances of positive interactions. The most likely interactions, as inferred from network links and species' traits, could be identified, then tested through direct observational studies.

Finally, we note that the addition of phylogenetic information could resolve some of the remaining uncertainties regarding modular structure and species' interactions. For example, if a recurring, convergent pattern of functional motifs is detected among network modules, a logical next step would be to test for a parallel pattern of phylogenetic overdispersion among species within the modules. This would test the hypothesis that contemporary functional motifs are attributable to historical partitioning (among modules) of closely related species that play similar ecological roles [84,85]. Alternatively, if species' functional roles within modules are clustered, such that each module has a unique (rather than repeated) functional profile, then a similar pattern of phylogenetic conservatism could serve as a simple, historical null model of network structure [86].

In future work, we hope to pursue each of these ideas. But for the moment, we submit that careful consideration of spatial scale will be necessary to advance ecological network analysis. Our tests of spatial scaling provided modest support for the hypothesis that at larger scales, network modules are more likely to reflect biogeographical provinces than subgroups of interactive species. However, examination of the Bluehead Chub nest associates complex suggested that network modules may more useful than multivariate clusters for characterizing positive interspecific relationships. Furthermore, modules may be capable of detecting positive interactions across a wide range of spatial scales.

Acknowledgments

Author Contributions

Conceptualization: Daniel J. McGarvey, Joseph A. Veech.

Data curation: Daniel J. McGarvey.

Formal analysis: Daniel J. McGarvey.

References

1. Bascompte J, Jordano P, Melián CJ, Olesen JM. The nested assembly of plant–animal mutualistic networks. Proc Natl Acad Sci U S A. 2003: 100; 9383–9387. https://doi.org/10.1073/pnas.1633576100 PMID: 12881488

2. Strona G, Veech JA. A new measure of ecological network structure based on node overlap and segregation. Methods Ecol Evol. 2015: 6; 907–915. https://doi.org/10.1111/2041-210X.12395

3. Eklöf A, Jacob U, Kopp J, Bosch J, Castro-Urgal R, Chacoff NP, et al. The dimensionality of ecological networks. Ecol Lett. 2013: 16; 577–583. https://doi.org/10.1111/ele.12081 PMID: 23438174

4. Zanata TB, Dalsgaard B, Passos FC, Cotton PA, Roper JJ, Maruyama PK, et al. Global patterns of interaction specialization in bird–flower networks. J Biogeogr. 2017: 44; 1891–1910. https://doi.org/10.1111/jbi.13045

5. Proulx SR, Promislow DEL, Phillips PC. Network thinking in ecology and evolution. Trends Ecol Evol. 2005: 20; 345–353. https://doi.org/10.1016/j.tree.2005.04.004 PMID: 16701391

6. Strogatz SH. Exploring complex networks. Nature. 2001: 410; 268–276. https://doi.org/10.1038/35065725 PMID: 11258382

7. Olesen JM, Dupont YL, O'Gorman E, Ings TC, Layer K, Melián CJ, et al. From broadstone to Zackenberg: space, time and hierarchies in ecological networks. Adv Ecol Res. 2010: 42; 1–69. https://doi.org/10.1016/B978-0-12-381363-3.00001-0

8. Tylianakis JM, Morris RJ. Ecological networks across environmental gradients. Annu Rev Ecol Evol Syst. 2017: 48; 25–48. https://doi.org/10.1146/annurev-ecolsys-110316-022821

9. Woodward G, Benstead JP, Beveridge OS, Blanchard J, Brey T, Brown LE, et al. Ecological networks in a changing climate. Adv Ecol Res. 2010: 42; 71–138. https://doi.org/10.1016/B978-0-12-381363-3.00002-2

10. Borthagaray AI, Arim M, Marquet PA. Inferring species roles in metacommunity structure from species co-occurrence networks. Proc Biol Sci. 2014: 281; 20141425. http://dx.doi.org/10.1098/rspb.2014.1425 PMID: 25143039

11. Morueta-Holme N, Blonder B, Sandel B, McGill BJ, Peet RK, Ott JE, et al. A network approach for inferring species associations from co-occurrence data. Ecography. 2016: 39; 1139–1150. https://doi.org/10.1111/ecog.01892

12. Dunne JA, Williams RJ, Martinez ND. Food-web structure and network theory: the role of connectance and size. Proc Natl Acad Sci U S A. 2002: 99; 12917–12922. https://doi.org/10.1073/pnas.192407699 PMID: 12235364

13. Tylianakis JM, Tscharntke T, Lewis OT. Habitat modification alters the structure of tropical host-parasitoid food webs. Nature. 2007: 445; 202–205. http://dx.doi.org/10.1038/nature05429 PMID: 17215842

14. Olesen JM, Bascompte J, Dupont YL, Jordano P. The modularity of pollination networks. Proc Natl Acad Sci U S A. 2007: 104; 19891–19896. https://doi.org/10.1073/pnas.0706375104 PMID: 18056808

15. Dalsgaard B, Martín González AM, Olesen JM, Timmermann A, Andersen LH, Ollerton J. Pollination networks and functional specialization: a test using Lesser Antillean plant–hummingbird assemblages. Oikos. 2008: 117; 789–793. https://doi.org/10.1111/j.0030-1299.2008.16537.x

16. Thébault E. Identifying compartments in presence–absence matrices and bipartite networks: insights into modularity measures. J Biogeogr. 2013: 40; 759–768. https://doi.org/10.1111/jbi.12015

17. Carstensen DW, Lessard JP, Holt BG, Krabbe BM, Rahbek C. Introducing the biogeographic species pool. Ecography. 2013: 36; 1310–1318. https://doi.org/10.1111/j.1600-0587.2013.00329.x

18. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A. 2006: 103; 8577–8582. https://doi.org/10.1073/pnas.0601602103 PMID: 16723398

19. Kreft H, Jetz W. A framework for delineating biogeographical regions based on species distributions. J Biogeogr. 2010: 37; 2029–2053. https://doi.org/10.1111/j.1365-2699.2010.02375.x

20. Gotelli NJ, Ellison AM. A primer of ecological statistics. Sunderland, Massachusetts: Sinauer Associates, Inc.; 2004.

21. Stouffer DB, Bascompte J. Compartmentalization increases food-web persistence. Proc Natl Acad Sci U S A. 2011: 108; 3648–3652. https://doi.org/10.1073/pnas.1014353108 PMID: 21307311

22. Guimerà R, Stouffer DB, Sales-Pardo M, Leicht EA, Newman MEJ, Amaral LAN. Origin of compartmentalization in food webs. Ecology. 2010: 91; 2941–2951. https://doi.org/10.1890/09-1175.1 PMID: 21058554

23. Prado PI, Lewinsohn TM. Compartments in insect-plant associations and their consequences for community structure. J Anim Ecol. 2004: 73; 1168–1178. https://doi.org/10.1111/j.0021-8790.2004.00891.x

24. Araújo MB, Rozenfeld A. The geographic scaling of biotic interactions. Ecography. 2014: 37; 406–415. https://doi.org/10.1111/j.1600-0587.2013.00643.x

25. Kéfi S, Berlow EL, Wieters EA, Joppa LN, Wood SA, Brose U, et al. Network structure beyond food webs: mapping non-trophic and trophic interactions on Chilean rocky shores. Ecology. 2015: 96; 291–303. https://doi.org/10.1890/13-1424.1 PMID: 26236914

26. Kraft NJB, Adler PB, Godoy O, James EC, Fuller S, Levine JM. Community assembly, coexistence and the environmental filtering metaphor. Funct Ecol. 2015: 29; 592–599. https://doi.org/10.1111/1365-2435.12345

27. McManamay RA, Frimpong EA. Hydrologic filtering of fish life history strategies across the United States: implications for stream flow alteration. Ecol Appl. 2015: 25; 243–263. https://doi.org/10.1890/14-0247.1 PMID: 26255371

28. Galiana N, Lurgi M, Claramunt-López B, Fortin M-J, Leroux S, Cazelles K, et al. The spatial scaling of species interaction networks. Nat Ecol Evol. 2018: 2; 782–790. https://doi.org/10.1038/s41559-018-0517-3 PMID: 29662224

29. Vilhena DA, Antonelli A. A network approach for identifying and delimiting biogeographical regions. Nat Commun. 2015: 6: 6848. https://dx.doi.org/10.1038/ncomms7848 PMID: 25907961

30. Bloomfield NJ, Knerr N, Encinas-Viso F. A comparison of network and clustering methods to detect biogeographical regions. Ecography. 2018: 41; 1–10. https://doi.org/10.1111/ecog.02596

31. Schluter D. A variance test for detecting species associations, with some example applications. Ecology. 1984: 65; 998–1005. https://doi.org/10.2307/1938071

32. García-Callejas D, Molowny-Horas R, Araújo MB. Multiple interactions networks: towards more realistic descriptions of the web of life. Oikos. 2018: 127; 5–22. https://doi.org/10.1111/oik.04428

33. Saiz H, Gómez-Gardeñes J, Nuche P, Girón A, Pueyo Y, Alados CL. Evidence of structural balance in spatial ecological networks. Ecography. 2017: 40; 733–741. https://doi.org/10.1111/ecog.02561

34. Pendleton RM, Pritt JJ, Peoples BK, Frimpong EA. The strength of *Nocomis* nest association contributes to patterns of rarity and commonness among New River, Virginia Cyprinids. Am Midl Nat. 2012: 168; 202–217. https://doi.org/10.1674/0003-0031-168.1.202

35. Frimpong EA. A case for conserving common species. PLoS Biol. 2018: 16; e2004261. https://doi.org/10.1371/journal.pbio.2004261 PMID: 29415000

36. Johnston CE. Nest association in fishes: evidence for mutualism. Behav Ecol Sociobiol. 1994: 35; 379–383. https://doi.org/10.1007/BF00165839

37. USEPA. Research strategy: Environmental Monitoring and Assessment Program. EPA 620/R-02/002. Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Research and Development; 2002. Available from: http://www.epa.gov/emap/html/pubs/docs/resdocs/EMAP_Research_Strategy.pdf.

38. USEPA. National Rivers and Streams Assessment 2008–2009: a collaborative survey. EPA/841/D-13/001. Washington, D.C.: U.S. Environmental Protection Agency; 2016. Available from: http://water.epa.gov/type/rsl/monitoring/riverssurvey/upload/NRSA0809_Report_Final_508Compliant_130228.pdf.

39. Lazorchak JM, Klemm DJ, Peck DV. Environmental Monitoring and Assessment Program—surface waters: field operations and methods for measuring the ecological condition of wadeable streams. EPA/620/R-94/004F. Washington, D.C.: U.S. Environmental Protection Agency; 1998. Available from: http://www.epa.gov/emap/html/pubs/docs/groupdocs/surfwatr/field/MAHAWadeableStreams.pdf.

40. USEPA. National Rivers and Streams Assessment: field operations manual. EPA-841-B-07-009. Washington, D.C.: U.S. Environmental Protection Agency; 2007. Available from: http://water.epa.gov/type/rsl/monitoring/riverssurvey/upload/NRSA_Field_Manual_4_21_09.pdf.

41. Cao Y, Larsen DP, Hughes RM. Evaluating sampling sufficiency in fish assemblage surveys: a similarity-based approach. Can J Fish Aquat Sci. 2001: 58; 1782–1793. https://doi.org/10.1139/f01-120

42. USGS. Federal standards and procedures for the National Watershed Boundary Dataset (WBD): techniques and methods. Reston, VA: U.S. Geological Survey and U.S. Department of Agriculture, Natural Resources Conservation Service; 2013. Available from: http://pubs.usgs.gov/tm/tm11a3/.

43. Griffith DM, Veech JA, Marsh CJ. cooccur: probabilistic species co-occurrence analysis in R. J Stat Softw. 2016: 69. http://dx.doi.org/10.18637/jss.v069.c02

44. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006: 1695. Available from: http://www.necsi.edu/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf.

45. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983: 220; 671–680. https://doi:10.1126/science.220.4598.671 PMID: 17813860

46. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004: 69; 026113. https://doi.org/10.1103/PhysRevE.69.026113

47. Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. J Stat Mech. 2005: 2005; P09008. https://doi:10.1088/1742-5468/2005/09/P09008

48. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. Vegan: community ecology package (version 2.5–2). 2018. Available from: https://cran.r-project.org/web/packages/vegan/vegan.pdf.

49. Kaufman L, Rousseeuw PJ. Partitioning around medoids (program PAM). In: Kaufman L, Rousseeuw PJ, editors. Finding groups in data: an introduction to cluster analysis. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2008. pp. 68–125.

50. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster analysis basics and extensions. 2018. Available from: https://cran.r-project.org/web/packages/cluster/cluster.pdf.

51. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Roy Stat Soc Ser B. 2001: 63; 411–423. https://doi.org/10.1111/1467-9868.00293

52. NatureServe. Digital distribution maps of the freshwater fishes in the conterminous United States. Version 3.0. Arlington, Virginia: NatureServe; 2010. Available from: http://www.natureserve.org/conservation-tools/data-maps-tools/digital-distribution-native-us-fishes-watershed.

53. Van Sickle J. Using mean similarity dendrograms to evaluate classifications. J Agric Biol Environ Stat. 1997: 2; 370–88. https://doi:10.2307/1400509

54. Mielke PW, Berry KJ, Johnson ES. Multi-response permutation procedures for *a priori* classifications. Commun Stat Theory Methods. 1976: 5; 1409–1424. https://doi.org/10.1080/03610927608827451

55. Johnston CE, Page LM. The evolution of complex reproductive strategies in North American minnows (Cyprinidae). In: Mayden RL, editor. Systematics, historical ecology, and North American freshwater fishes. Stanford, California: Stanford University Press; 1992. pp. 600–621.

56. Cooper JE. Egg, larval and juvenile development of Longnose Dace, *Rhinichthys cataractae*, and River Chub, *Nocomis micropogon*, with notes on their hybridization. Copeia. 1980: 1980; 469–478.

57. Johnston CE. The benefit to some minnows of spawning in the nests of other species. Environ Biol Fishes. 1994: 40; 213–218. https://doi.org/10.1007/BF00002547

58. Peoples BK, Frimpong EA. Biotic interactions and habitat drive positive co-occurrence between facilitating and beneficiary stream fishes. J Biogeogr. 2016: 43; 923–931. https://doi.org/10.1111/jbi.12699

59. Veech JA. A probability-based analysis of temporal and spatial co-occurrence in grassland birds. J Biogeogr. 2006: 33; 2145–2153. https://doi.org/10.1111/j.1365-2699.2006.01571.x

60. Veech JA. A probabilistic model for analyzing species co-occurrence. Global Ecol Biogeogr. 2013: 22; 252–260. https://doi.org/10.1111/j.1466-8238.2012.00789.x

61. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A. 2008: 105; 1118–1123. https://doi.org/10.1073/pnas.0706851105 PMID: 18216267

62. Edler D, Guedes T, Zizka A, Rosvall M, Antonelli A. Infomap bioregions: interactive mapping of biogeographical regions from species distributions. Syst Biol. 2017: 66; 197–204. https://doi.org/10.1093/sysbio/syw087 PMID: 27694311

63. Cornell HV, Harrison SP. What are species pools and when are they important? Annu Rev Ecol Evol Syst. 2014: 45; 45–67. https://doi.org/10.1146/annurev-ecolsys-120213-091759

64. Zobel M, Otto R, Laanisto L, Naranjo-Cigala A, Pärtel M, Fernández-Palacios JM. The formation of species pools: historical habitat abundance affects current local diversity. Global Ecol Biogeogr. 2011: 20; 251–259. https://doi.org/10.1111/j.1466-8238.2010.00593.x

65. Ricklefs RE. Disintegration of the ecological community. Am Nat. 2008: 172; 741–750. https://doi.org/10.1086/593002 PMID: 18954264

66. Srivastava DS. Using local-regional richness plots to test for species saturation: pitfalls and potentials. J Anim Ecol. 1999: 68; 1–16. https://doi.org/10.1046/j.1365-2656.1999.00266.x

67. Stone L, Roberts A. The checkerboard score and species distributions. Oecologia. 1990: 85; 74–79. https://doi.org/10.1007/BF00317345 PMID: 28310957

68. Veech JA. The pairwise approach to analysing species co-occurrence. J Biogeogr. 2014: 41; 1029–1035. https://doi.org/10.1111/jbi.12318

69. Gotelli NJ. Null model analysis of species co-occurrence patterns. Ecology. 2000: 81; 2606–2621. https://doi.org/10.1890/0012-9658(2000)081[2606:NMAOSC]2.0.CO;2

70. Diamond JM. Assembly of species communities. In: Cody ML, Diamond JM, editors. Ecology and evolution of communities. Cambridge, Massachusetts: Harvard University Press; 1975. pp. 342–444.

71. Belmaker J, Zarnetske P, Tuanmu M-N, Zonneveld S, Record S, Strecker A, et al. Empirical evidence for the scale dependence of biotic interactions. Global Ecol Biogeogr. 2015: 24; 750–761. https://doi.org/10.1111/geb.12311

72. Freilich MA, Wieters E, Broitman BR, Marquet PA, Navarrete SA. Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in ecological communities? Ecology. 2018: 99; 690–699. https://doi.org/10.1002/ecy.2142 PMID: 29336480

73. Gilpin ME, Diamond JM. Are species co-occurrences on islands non-random, and are null hypotheses useful in community ecology? In: Strong DRJ, Simberloff D, Abele LG, Thistle AB, editors. Ecological communities: conceptual issues and the evidence. Princeton, New Jersey: Princeton University Press; 1984. pp. 297–315.

74. Connor EF, Simberloff D. Neutral models of species' co-occurrence patterns. In: Strong DRJ, Simberloff D, Abele LG, Thistle AB, editors. Ecological communities: conceptual issues and the evidence. Princeton, New Jersey: Princeton University Press; 1984. pp. 316–331.

75. Atmar W, Patterson BD. The measure of order and disorder in the distribution of species in fragmented habitat. Oecologia. 1993: 96; 373–382. https://doi.org/10.1007/BF00317508 PMID: 28313653

76. Araújo MB, Rozenfeld A, Rahbek C, Marquet PA. Using species co-occurrence networks to assess the impacts of climate change. Ecography. 2011: 34; 897–908. https://doi.org/10.1111/j.1600-0587.2011.06919.x

77. Ulrich W, Gotelli NJ. Null model analysis of species associations using abundance data. Ecology. 2010: 91; 3384–3397. https://doi.org/10.1890/09-2157.1 PMID: 21141199

78. Bell JR, King RA, Bohan DA, Symondson WOC. Spatial co-occurrence networks predict the feeding histories of polyphagous arthropod predators at field scales. Ecography. 2010: 33; 64–72. https://doi.org/10.1111/j.1600-0587.2009.06046.x

79. Poisot T, Stouffer DB, Gravel D. Beyond species: why ecological interaction networks vary through space and time. Oikos. 2015: 124; 243–251. https://doi.org/10.1111/oik.01719

80. Morris RJ, Gripenberg S, Lewis OT, Roslin T. Antagonistic interaction networks are structured independently of latitude and host guild. Ecol Lett. 2014: 17; 340–349. https://doi.org/10.1111/ele.12235 PMID: 24354432

81. Kéfi S, Miele V, Wieters EA, Navarrete SA, Berlow EL. How structured is the entangled bank? The surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience. PLoS Biol. 2016: 14; e1002527. https://doi.org/10.1371/journal.pbio.1002527 PMID: 27487303

82. Maurakis EG, Woolcott WS, Sabaj MH. Reproductive-behavioral phylogenetics of *Nocomis* species-groups. Am Midl Nat. 1991: 126; 103–110. https://doi:10.2307/2426154

83. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet. 2002: 31; 64–68. https://doi.org/10.1038/ng881 PMID: 11967538

84. Cavender-Bares J, Ackerly DD, Baum DA, Bazzaz FA. Phylogenetic overdispersion in Floridian oak communities. Am Nat. 2004: 163; 823–43. https://doi.org/10.1086/386375 PMID: 15266381

85. Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. Phylogenies and community ecology. Annu Rev Ecol Syst. 2002: 33; 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448

86. Vamosi SM, Heard SB, Vamosi JC, Webb CO. Emerging patterns in the comparative analysis of phylogenetic community structure. Mol Ecol. 2009: 18; 572–92. https://doi.org/10.1111/j.1365-294X.2008.04001.x PMID: 19037898