

LETTER

Food web models: a plea for groups

Stefano Allesina^{1*} and Mercedes Pascual²

¹National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300. Santa Barbara, CA 93101, USA

²Department of Ecology and Evolutionary Biology, and Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI 48109 1048, USA

*Correspondence: E-mail: allesina@nceas.ucsb.edu

Abstract

The concept of a group is ubiquitous in biology. It underlies classifications in evolution and ecology, including those used to describe phylogenetic levels, the habitat and functional roles of organisms in ecosystems. Surprisingly, this concept is not explicitly included in simple models for the structure of food webs, the ecological networks formed by consumer–resource interactions. We present here the simplest possible model based on groups, and show that it performs substantially better than current models at predicting the structure of large food webs. Our group-based model can be applied to different types of biological and non-biological networks, and for the first time merges in the same framework two important notions in network theory: that of compartments (sets of highly interacting nodes) and that of roles (sets of nodes that have similar interaction patterns). This model provides a basis to examine the significance of groups in biological networks and to develop more accurate models for ecological network structure. It is especially relevant at a time when a new generation of empirical data is providing increasingly large food webs.

Keywords

Akaike information criterion, clustering algorithm, compartment, connectance, food web model, group, likelihood, model selection, species richness, trophic role, trophospecies.

Ecology Letters (2009) 12: 652–662

INTRODUCTION

Few concepts in biology are so pervasive as that of a group. Since the work of Linnaeus more than 250 years ago, biologists have tried to classify organisms into species, species into genera, genera into families, and so on. In ecology, species have been grouped according to their habitat, such as benthic vs. pelagic species in marine environments, below-ground and canopy communities in tropical forests, autotrophs, primary consumers and detritivores, based on energy sources, omnivores, specialists, and generalists based on their diet breadth, to name a few examples of grouping species. Surprisingly, however, the concept of group has been largely left out from the construction of simple models for food web structure (Cohen *et al.* 1990; Williams & Martinez 2000; Cattin *et al.* 2004; Allesina *et al.* 2008). Only Cattin *et al.* (2004) implicitly consider that similar predators act in a similar way. To date no model addressed the presence of groups explicitly.

These stochastic models provide a way to construct with a few simple rules ‘realistic’ food webs or networks describing ‘who eats whom’ in an ecosystem. The simplest

models are motivated by a few ecological principles and require only two parameters, the total number of species, or species richness, and the total number of connections, or connectance, in the network. The recent comparisons of these models based on likelihoods have therefore considered the same number of parameters (Allesina *et al.* 2008). However, the use of likelihoods opens the door for more general comparisons among models of varying complexity based on information criteria. We introduce here network models based on the concept of groups and compare for the first time, models with a different number of parameters. We show that dividing species into groups yields critical information for building better models of food web structure, especially for large networks.

In all the simple models of food web structure proposed so far, species are ranked into a one-dimensional hierarchy with this ranking providing the basis to establish species’ interactions (Cohen *et al.* 1990; Williams & Martinez 2000; Cattin *et al.* 2004; Allesina *et al.* 2008). For example, in the cascade model (Cohen *et al.* 1990), a species can prey with a given probability upon any species whose position in the ranking is lower, but cannot prey upon those with higher

ranking. It follows that in all these models, the only characteristic that differentiates species is their 'ranking value'. No other information, on phylogeny, habitat, feeding behaviour etc., is required.

Although the idea of identifying groups in food webs is not new, it has been applied for specific definitions of groups and has not been used in the construction of models. Since the consideration of compartments in the late 70s (Pimm 1979; Pimm & Lawton 1980; Critchlow & Stearns 1982; Yodzis 1982), several papers have dealt with the 'community detection' problem of identifying groups of highly interacting nodes that interact very weakly with the rest of the network. These methods have revealed some underlying structuring of species according to spatial and habitat distributions [e.g. benthic vs. pelagic communities (Newman & Girvan 2004; Krause *et al.* 2003; Melián & Bascompte 2004; Allesina *et al.* 2005; Newman, 2006)]. Another notion of group important to food webs has been that of trophic position, the extension of the familiar idea of trophic levels to include fractional positions (Levine 1980). The concept of trophic similarity was proposed to reduce the inherent complexity of food webs into manageable smaller networks (Yodzis & Winemiller 1999). In this framework, species that share the same predators and prey are considered a trophic species. A generalization of this concept is that of trophic role (Luczkovich *et al.* 2003) in which two species are equivalent if they prey upon equivalent species and are in turn preyed upon by equivalent species. For example, two herbivores are said to be equivalent if they both feed on equivalent plants and are also the prey of equivalent carnivores, even though they do not share necessarily the same plants and carnivores. The recursive definition of trophic roles encompasses that of trophic species as a special case (Luczkovich *et al.* 2003). Interestingly, trophic species and trophic roles are known to sociologists as structural and regular equivalence classes (Luczkovich *et al.* 2003).

These two lines of research, on communities and roles in networks, have typically been separated since they represent different objective functions for the search algorithms employed to detect the groups. Here we present the simplest framework in which these two characteristics of network structure are considered at the same time in the detection of groups: if communities are the main driver of network structure, our algorithm detects communities; if roles provide a better explanation for the interaction patterns, the framework yields groups based on roles; finally, the two features are not mutually exclusive, as the algorithm can detect some groups based on dense within-group interactions (communities) that also have well-defined between-groups interactions (roles). In this sense, our framework provides a generalization in which communities and roles are special cases of the more

general concept of a group. Although our results and examples pertain to food webs, the methods we proposed should apply to any type of network (biological or not, directed or undirected). The general approach is also the simplest one possible for constructing models of food web structure based on groups of species. The resulting models are shown to capture the structure of food webs better than existing ones for sufficiently large networks. Thus the concept of groups become critical for modelling ecological networks at a time when a new generation of large empirical networks is underway (Bascompte *et al.* 2005; Arii *et al.* 2007). We argue that the construction of such models will be impossible without considering the existence of groups in these large networks. We end with a discussion of the open areas related to the biological interpretation of such groups, including the further improvement of the models themselves.

The framework we propose can be interpreted as a model for food web structure based on groups, but can also be seen as a clustering algorithm or a way to investigate similarity among species in an ecological network. In what follows we concentrate on the stochastic model interpretation, while we will focus on the other aspects in future work. Both the group-based model proposed here and previous stochastic models of food web structure based on a hierarchical order of species can be used to generate families of networks that reproduce aspects of the structure of empirical food webs. Both are motivated by simple principles, in one case the importance of groups in species' interactions, in the other, the similarity of prey and a ranking of species that defines the generality of predators. These principles are not exclusive and future models are likely to combine them in some form. Thus, our model should not be viewed simply as a clustering method. It differs, however, from previous simple models of structure in that the number of parameters is not pre-defined and constant across different empirical webs, since it does not rely on species richness and connectance as the two fundamental parameters. This difference has implications for generating families of networks across ecosystems, to examine for example the dynamical consequences of structure, using the same set of fundamental parameters. That this is not the case with our approach does not mean that comparisons across ecosystems are not possible but that these have to rely on possible differences and specific features of the different communities. We view the two modelling approaches as complementary, as their importance will depend on the size of the networks. Furthermore, future work is likely to reveal regularities in the group-based structure that can serve as principles for general models. In the Discussion, we further explore similarities and differences with previous stochastic models of food web structure.

MATERIALS AND METHODS

We define here a food web N as a network formed by S nodes representing species or a collection of similar species, connected by L directed connections representing consumer–resource interactions and denoted as edges or links. Each network can be associated with a matrix $A[a_{ij}]$ called the adjacency matrix, in which each coefficient a_{ij} is 1 if the row species i is a prey of the column species j and 0 elsewhere.

To reproduce an empirical food web $N(S, L)$ with a simple random process, the simplest model we can think of is one in which there is a fixed probability p of connecting each node with any other node. This is an Erdős–Rényi (Erdős & Rényi 1960) directed random graph. What is then the probability of obtaining exactly the given empirical network $N(S, L)$ using this process? We can see the associated adjacency matrix A as a sequence of zeros and ones. In the random graph, we seek to reproduce such a sequence with the only rule that the probability of obtaining 1 is p and that of obtaining 0 is $(1 - p)$. Therefore the probability of producing exactly N is:

$$\mathcal{P}(N(S, L)|p) = p^L(1 - p)^{S^2-L} \tag{1}$$

where, for consistency, we define $0^0 = 1$. This probability can also be read as the likelihood of the parameter p given the network N :

$$\mathcal{L}(p|N(S, L)) = p^L(1 - p)^{S^2-L} \tag{2}$$

As one can easily show using derivatives, this probability (or likelihood) is maximized by $p = L/S^2$, exactly the quantity known as the directed connectance (or density) of the network.

A way to slightly complicate the above model is to divide the nodes into two groups (‘Green’ and ‘Red’) so that we now need four probabilities, p_{gg} – the probability of a Green node connecting to a Green node; p_{rr} – the probability of a Red node connecting to a Red node; p_{gr} – the probability of a Green node connecting to a Red node; and p_{rg} – the probability of a Red node connecting to a Green node. We can rewrite the equation for the probability of obtaining the graph N when we have two groups of nodes as follows:

$$\begin{aligned} \mathcal{P}(N(S, L)|p_{gg}, p_{rr}, p_{gr}, p_{rg}) = \\ p_{gg}^{L_{gg}}(1 - p_{gg})^{S_g^2-L_{gg}} \cdot p_{rr}^{L_{rr}}(1 - p_{rr})^{S_r^2-L_{rr}} \cdot \\ p_{gr}^{L_{gr}}(1 - p_{gr})^{S_g S_r-L_{gr}} \cdot p_{rg}^{L_{rg}}(1 - p_{rg})^{S_r S_g-L_{rg}} \end{aligned} \tag{3}$$

where L_{xy} is the number of edges connecting nodes belonging to the group X to nodes belonging to the group Y and S_x is the number of nodes in group X . To maximize the probability, it is sufficient to set $p_{xy} = L_{xy}/(S_x S_y)$ for all combinations of groups.

We can generalize the formulation above to an arbitrary number of groups k :

$$\mathcal{P}(N(S, L)|\vec{p}) = \prod_{i=1}^k \prod_{j=1}^k p_{ij}^{L_{ij}}(1 - p_{ij})^{S_i S_j-L_{ij}} \tag{4}$$

where \vec{p} is the vector containing all the probabilities p_{ij} . In the limit, we can assign each node to a different group. In this case, $p_{ij} = L_{ij} = a_{ij}$. Because each probability is set either to 0 or 1, this process will always produce the same network, which is exactly the desired empirical one (N). The probability of recovering the data is therefore 1.

We can compute the probability above for any arrangement of the nodes into groups. The two simplest cases correspond to the two extremes of (1) all the nodes belonging to the same group (a random graph) and (2) each node consisting of a different group by itself ($\mathcal{P} = 1$). Figure 1 shows these two extremes together with two intermediate solutions. Each possible arrangement can also be seen as a model for generating food webs: we divide S nodes according to groups, and then generate a network by connecting two nodes belonging to groups X and Y with probability p_{XY} . The probability of obtaining the original network with this process is that given in eqn 4.

How many possible arrangements are there? Because we seek to partition S nodes into k nonempty groups, we can count the number of arrangements using the Stirling’s numbers of the second kind:

$$\mathcal{S}(S, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k - i)^S \tag{5}$$

The sum of Stirling’s numbers for all k s gives the Bell’s numbers: For example, for 10 nodes, there is only 1 way of arranging them into 1 group, but there are 511 ways to define two groups, 9330 ways to define three and so forth, with a total number of possible arrangements of 115 975. Bell’s number can be written with the following recurrence relation:

$$\mathcal{B}_S = \sum_{k=0}^{S-1} \mathcal{B}_k \binom{S-1}{k} \tag{6}$$

This formula shows how fast the number of possible ways of organizing species into groups grows with the number of nodes. While $\mathcal{B}_3 = 5$, $\mathcal{B}_{30} = 10^{23}$ and $\mathcal{B}_{60} > 10^{59}$. Setting aside the technical difficulties introduced by these huge numbers, we can easily compute a probability for each possible arrangement/model.

Increasing the number of groups used to partition the network will never decrease the likelihood $\mathcal{L}(\vec{p}, k|N(S, L))$ (in fact, if k groups do not produce better probabilities than $k - 1$, we can always leave one of the groups empty, recovering the better likelihood), but it will greatly increase

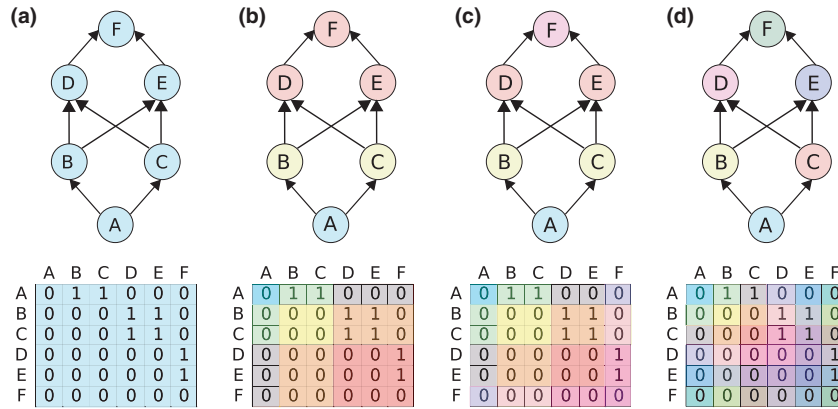


Figure 1 Arrangement into groups of a hypothetical food web. From left to right: (a) all species are included in the same group (random graph with $p = 0.22$); the log-likelihood is -8.28 and the Akaike Information Criterion (Akaike 1974) is $AIC = 30.56$. (b) The species are divided into three groups, which yields nine probabilities (e.g. $p_{3,3} = 2/9$). The log-likelihood is -3.73 and the $AIC = 37.46$. (c) and (d) divide the matrix into sub-matrices containing either all 1s or all 0s. This means that the log-likelihood is 0 and therefore the AIC is given by $2S + 2k^2$, and equals 44 in the case (c) and 84 in the case (d).

the number of parameters. This consideration leads naturally to issues of model selection. We can use Akaike’s AIC (Akaike 1974) to select the best model and arrangement into groups. The parameters for each model will be (a) a vector specifying the assignment of each species to a group and (b) k^2 probabilities, where k is the number of groups. We can write the AIC for any arrangement as:

$$AIC(N(S, L) | \vec{p}, k) = 2k^2 + 2S - 2 \log \left(\prod_{i=1}^k \prod_{j=1}^k p_{ij}^{L_{ij}} (1 - p_{ij})^{S_i S_j - L_{ij}} \right) \quad (7)$$

where k is the number of groups. If we define $0 \log 0 = 0$, this can be reduced to a sum:

$$AIC(N(S, L) | \vec{p}, k) = 2k^2 + 2S - 2 \left(\sum_{i=1}^k \sum_{j=1}^k L_{ij} \log(p_{ij}) + (S_i S_j - L_{ij}) \log(1 - p_{ij}) \right) \quad (8)$$

Minimizing the AIC yields the group arrangement that maximizes the likelihood of reproducing the empirical data while keeping the number of groups, and therefore the number of free parameters, low. The use of AIC for model selection is rooted in information theory, as AIC is clearly connected with the Kullback–Leibler information (Burnham & Anderson 2002). We report in Fig. 1 the AIC values for four possible arrangements of a simple food web.

The next section describes an algorithm to find the grouping of species that minimizes AIC , i.e. the best simple model based on groups for a given network.

A modified genetic algorithm

To search for the optimal solution (the lowest AIC) for a given network $N(S, L)$ and k groups we implemented a simple modification of a genetic algorithm. Specifically, a hill-climbing procedure is added to the algorithm to speed up the search. In a genetic algorithm, chromosomes encode possible solutions for the problem. In our case, each chromosome is simply a vector of length S whose elements are numbers between 1 and k , defining a way of grouping the species. A population of chromosomes is built and initialized at random. At each generation, each chromosome locally searches for a minimum for the AIC as follows: the algorithm considers all the possible groupings that can be

obtained by changing a single element in the vector (effectively computing the AIC of all its neighbours). If any of the neighbouring solutions is better than the one currently encoded in the chromosome, the chromosome will ‘mutate’ to the better solution. This procedure is repeated until no neighbour yields a lower AIC . In this way, each chromosome is at a local minimum. Then, the proper genetic algorithm takes place. Each chromosome is assigned a fitness, equal to $1/AIC$, and reproduces (i.e. is copied in the next generation) with a probability proportional to its fitness. While copying chromosomes mutations can occur with small probability. This procedure is repeated until a

new population of chromosomes is created. The whole process is repeated for a given number of generations or until all chromosomes are identical.

We applied this algorithm with the following parameters: size of the chromosomes' population = 1000; number of generations = 100; number of mutations in the offspring = 3. (The code is available upon request). To test whether the algorithm is sufficiently powerful we contrasted the results with the ones obtained using adaptive simulated annealing and a classical genetic algorithm. In all cases, our algorithm produced better or equal solutions, and found the local optima much faster than the other algorithms. For each food web, we searched for the best solution for all the $k \in [1, 2, \dots, 20]$.

RESULTS

We analysed 10 published empirical food webs (Table 1) and found the best way of dividing each network into groups. These networks are medium sized, and were chosen because they were previously used to compute the likelihood of existing models for food web structure. This allows us to compare our model with previous simple models based on the AIC values. For each network, we searched for the best configuration using $k = 1, 2, \dots, 20$ groups. In all cases, we found that the k minimizing the AIC is small (in all cases but one we found that the minimum AIC is found for $k \leq 10$, Table 1). Also, the best AIC varies smoothly as a function of k (Fig. 2).

Given a network, we can set an upper boundary for the number of groups that can possibly lead to the minimum AIC . Let us start from the simplest arrangement: all species in one group $k = 1$. There is only one such an arrangement, yielding an AIC value we denote with AIC_1 . Now, consider any arrangement making use of x groups: if

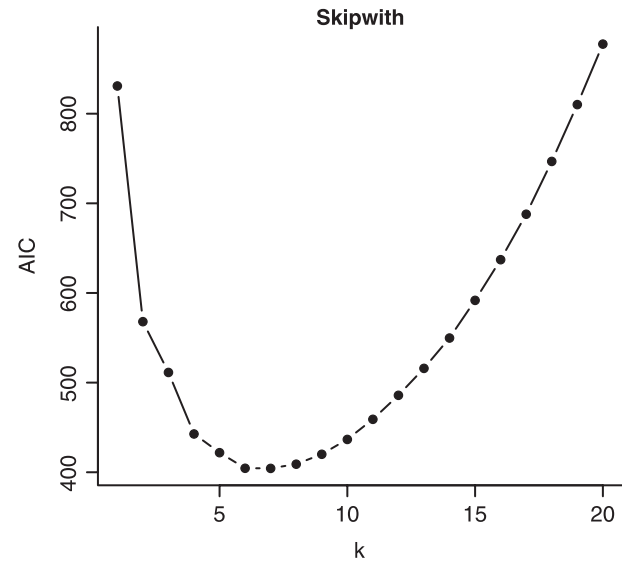


Figure 2 Profile of the best AIC value found by the algorithm for the Skipwith Pond food web (Warren 1989) for a varying number of groups (k). The curve has the same shape for all the 10 empirical networks analysed. In this case, the minimum AIC is obtained for $k = 7$ groups.

$2S + x^2 \geq AIC_1$, this means that the AIC of any arrangement encompassing x groups will have a value greater than that of the random graph (AIC_1): the best solution must have less than x groups. Similarly, the boundary can be refined. Let us denote with AIC_g the best solution making use of g groups, and say that $AIC_g < AIC_1$. We can use AIC_g to refine the boundary we set using AIC_1 . Take a number of groups $y < x$. If $2S + y^2 \geq AIC_g$, we know that the best solution must have fewer than y groups. In our search for the best AIC , each time we come across a better AIC value we can refine the boundary, making the search

Table 1 Results of the analysis for 10 empirical food webs

Food Web	S	L	k	AIC_{Groups}	AIC_{MinPot}	Reference
Chesapeake	31	68	7	386.090	356.220	Baird & Ulanowicz (1989)
Grass	61	97	7	738.765	715.880	Martinez <i>et al.</i> (1999)
Bridgebrook	25	107	6	297.559	238.360	Havens (1992)
Skipwith	25	197	7	404.372	393.340	Warren (1989)
Benguela	29	203	7	481.477	489.040	Yodzis (1998)
Stmartin	42	205	10	829.810	864.120	Goldwasser & Roughgarden (1993)
Stmarks	48	221	9	1077.601	1108.980	Christian & Luczkovich (1999)
Broom	85	223	9	1081.523	1427.080	Memmott <i>et al.</i> (2000)
Coachella	29	262	9	524.282	654.200	Polis (1991)
Reef	50	556	13	1636.190	1973.420	Optiz (1996)

We report the AIC values for the group-based model and the Minimal Potential Niche (Allesina *et al.* 2008).

S , number of species or groups of species in the food web (number of nodes); L , number of connections; k , number of groups yielding the minimum AIC for the group-based model described in the main text.

more efficient. For example, we can see that for the *bridgebrooke* (Warren 1989) system AIC_6 is 238.36. Given that, for this network, $AIC_{10} = 2 \cdot 25 + 2 \cdot 10^2 + L_{10} \geq 250$, we already see that the best configuration cannot include more than nine groups.

How do the AIC for the group model compare with those of previous models for food web structure? In Table 1 we report the AIC for the Minimal Potential Niche Model [the best simple model – in terms of likelihoods – proposed so far (Allesina *et al.* 2008)] and the minimum AIC we found by using the simple model based on groups. The Minimal Potential Niche Model assumes that species can be ordered and that predators tend to prey among similarly-ranked species. This model makes use of $S + 3$ parameters. The AIC is simply twice the number of parameters minus twice the log-likelihood, and can therefore readily be computed using values published by Allesina *et al.* When we order the food webs according to their number of links, a clear pattern emerges. The group-based models have a better likelihood when the networks contain more than 200 connections (regardless of the number of species), and worse values otherwise. We explore this relation further, together with the effect of the number of groups in Fig. 3, where we show the correlation among the values in Table 1.

The number of species S is weakly correlated with the number of links L and the number of groups k in the best configuration. It exhibits however a strong correlation with the AIC of the models. The number of links and more strikingly, the number of groups k , are strongly correlated with the AIC . In fact L can be used to predict k : a linear regression $k = 5.5 + 0.0136 L$ gives an $r^2 = 0.78$ and a highly significant p value. It follows that more complex networks should produce configurations with more groups than simpler ones, and that this relation is basically linear. Also there is a strong correlation between the AIC for the two models: a linear regression, $AIC_{Groups} = 120.65 + 0.76 AIC_{MinPots}$ shows that for more complex networks the group based solution should be much better than the one obtained with the Minimal Potential Niche Model ($r^2 = 0.97$, $P < 2 \times 10^{-7}$). This result is consistent with the findings of Table 1, and implies that for the large data sets that are currently appearing in the literature, with tens of thousands of links, the groups-based models should outperform the Minimal Potential Niche Model.

The difference in the performance increases significantly as the networks become more complex. For example, the difference in our model selection criterion, AIC , is of 345.56 and 337.23 for the *broom* (Memmott *et al.* 2000) and the *reef* (Optiz 1996) food webs respectively. One can use Akaike

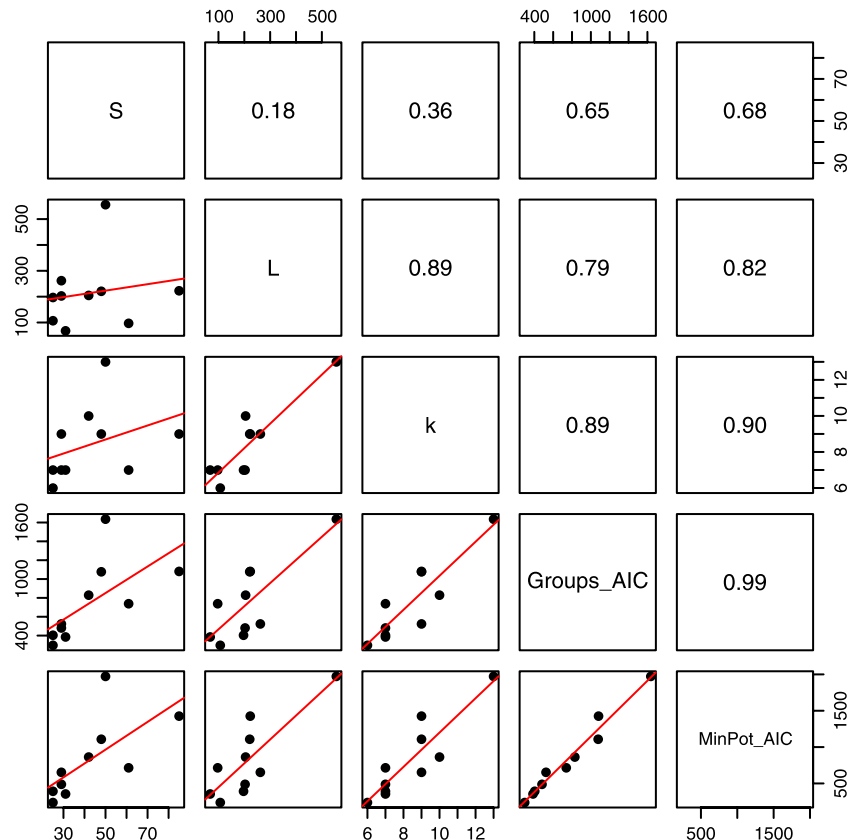


Figure 3 Correlations among the values in Table 1. In the upper part of the matrix we report the Pearson's coefficients. The red lines represent the best fit linear regression to the data.

weights (Burnham & Anderson 2002) to determine the strength of evidence for each model. The Akaike weight for a model x is defined as:

$$w_x = \frac{e^{-\frac{AIC_x - AIC^*}{2}}}{\sum_{\forall j} e^{-\frac{AIC_j - AIC^*}{2}}} \quad (9)$$

where AIC^* is the minimum AIC among the models. The weight w_x expresses the probability that x is the best model for the data (Burnham & Anderson 2002). For all the networks with more than 200 links, the corresponding weight w_{Groups} is almost exactly 1, i.e. the probability that this is in fact the best model is close to 1. The network with the worst performance of the group-based model is *bridgebrook* (Warren 1989), for which the difference is -59.2 (in this case computing the Akaike weights shows that the Minimal Potential model is almost certainly the best model).

DISCUSSION

We have presented a formulation and approach to model the structure of food webs based on the concept of groups. These models perform substantially better than current simple models when the networks are large, containing more than 200 links. From our results, we expect the difference in AIC to grow linearly with the number of links. Given that a third generation of food web data is currently underway (if we consider the data in Cohen *et al.* as the first generation and the data collected in the 90s as the second generation – J. Dunne, personal communication), with a much larger number of links than ever before of the order of several thousand (Bascompte *et al.* 2005; Arii *et al.* 2007), the difference in performance between the group-based models and their predecessors based on a one-dimensional hierarchy is likely to be considerable.

By using likelihoods and a model selection criterion, we were able to directly compare models of different complexity. Until recently, food web models were compared using ‘summary statistics’ that measured, for a set of networks generated with a given model, a set of network characteristics such as the number of top predators, the number of cannibal species, the average trophic level, etc. (Williams & Martinez 2000; Cattin *et al.* 2004; Allesina *et al.* 2008). Thus, the ability of the model to generate the network was evaluated through its ability to reproduce these values. This approach is clearly not feasible for models with a different number of parameters. How much closer to the observed number of top predators must the model predictions be to justify a larger number of parameters? AIC solves this problem in a simple and elegant way. The results of the models based on groups show that complex networks contain sufficient information to justify models with more

than 100 parameters. There is clearly room for considering models with a much larger number of parameters than the ones currently in use. In fact, the group-based models presented here do not contain any *a priori* ecological information and are a simple collection of random sub-graphs. We foresee the development of better group models that do incorporate ecological information. The formulation proposed here provides the simplest null-model based on groups and as such, a baseline for further improvements. The fact that the null-model yields better results than previous simple formulations motivated by ecological considerations suggests that the latter are too simple to show a good fit to the data. The use of complex models with several parameters has been proved useful to understand different problems in biology (Huelsenbeck *et al.* 2001; King *et al.* 2008).

One possible disadvantage of a variable number of parameters is the lack of a fixed set of ecological quantities that can be used universally to construct food webs with structures similar to those of nature. A special place has been so far given to two ecological quantities, connectance and species richness, which are used in all the simple food web models as the fundamental parameters. This has allowed for example the evaluation of dynamical consequences of structure as a function of these quantities (Williams & Martinez 2000). The group-based models do not possess such fundamental parameters; they can provide however better models for specific systems, and as the analyses of a large number of networks develop, give rise to regularities that emerge from the data. Dynamical consequences of structure can still be addressed, for example on the specific role of groups, beyond the much debated role of only one type of group, that of compartments (Yodzis 1982; Pimm *et al.* 1991).

The search for the best group configuration can be interpreted by examining at the adjacency matrix obtained through the ordering of the species into groups. Figure 4 illustrates such an adjacency matrix for the *broom* system (Memmott *et al.* 2000) and Fig. 5 shows the corresponding network partitioned into groups. One can easily see from eqn 4 that the likelihood is maximized when inside each ‘sub-matrix’ defined by two groups the density of links is either maximized or minimized. For example, in Fig. 4 we can see that in the given configuration the single node in the pink group does not receive any link, but sends links to all the nodes of the green, violet and cyan groups. In the same way, all nodes in the cyan group are connected to all the nodes in the violet group. Several other examples of very high (or low) density of connections are represented in the figure. Thus, groups are defined by the relation they have with each other (e.g. the violet group is the one that receives links from the pink, cyan and violet nodes and sends most of its links to violet and orange nodes). This

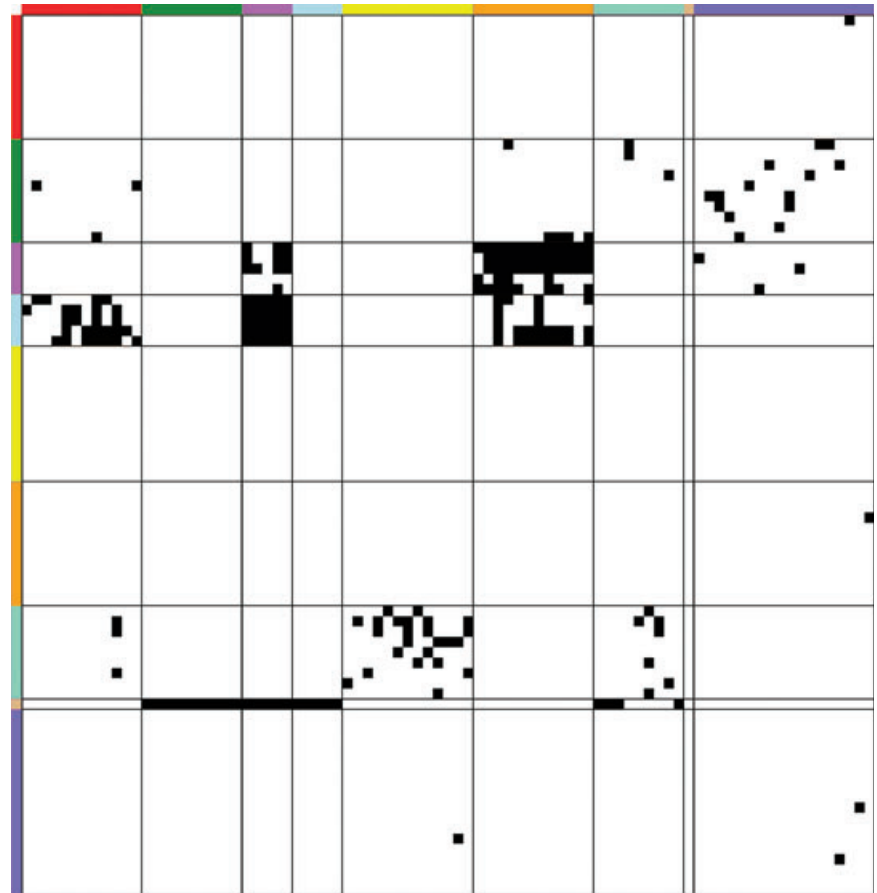


Figure 4 Adjacency matrix for the *broom* food web in which the nodes are ordered according to their group. As explained in the text, the algorithm seeks a partition of the nodes into groups that makes the density of the connections within each sub-matrix either maximal or minimal.

interpretation of groups leads naturally to the definition of roles in networks, and therefore to a clear relationship of our algorithm with ‘block-modelling’ in social networks (White *et al.* 1976) (even though the two approaches differ). Sociologists have focused on the image-graph obtained by condensing all the nodes belonging to a group into a single node (Luczkovich *et al.* 2003). The image-graph is a simplified picture of the role of nodes in networks. Although ‘block-modelling’ has a long tradition in sociology, to our knowledge information criteria have not been used before to find the best way of dividing networks into roles.

The definition of roles is evident also by examining the network itself (Fig. 5). For example, the *broom* system is a web obtained by sampling a single plant (Scotch broom, *Cytisus scoparius*), its herbivores and their predators. This is one of the first food webs in which parasites, parasitoids and pathogens were included. It is sensible then that the algorithm singled-out the plant (in pink), and divided herbivores (dark green, cyan) from the omnivores (light green, violet). Although a full examination of the biological significance of the groups found by our models is beyond the scope of this paper, we note that similar patterns are

present in all the networks examined. Figure 6 shows for example that the only primary producer is in a group by itself for the Skipwith pond food web.

We drew a parallel between our framework and that of block-modelling in social networks. Another very similar model was recently proposed by Newman & Leicht (2007). In their model, they group together all the nodes that have similar relations to other nodes in the network, that is they consider the probabilities φ_{xy} that a node belonging to group x is connected to node j . Our model deals with the probabilities p_{xy} that a node belonging to group x is connected to a node in group y . The similarity between the contributions is carried forward by the fact that Newman and Leicht used likelihoods to evaluate possible solutions (Newman & Leicht 2007). The mathematical definition of groups and the model selection techniques differ however, defining two separate frameworks. Another approach that uses likelihoods to identify communities is found in the preprint ‘Identifying Community Structures from Network Data’, by Čopič, J., Jackson, M.O. and Kirman, A. (en.scientificcommons.org/40597570). In this case, a network is produced according to two probabilities: the probability of interactions among nodes belonging to the same group

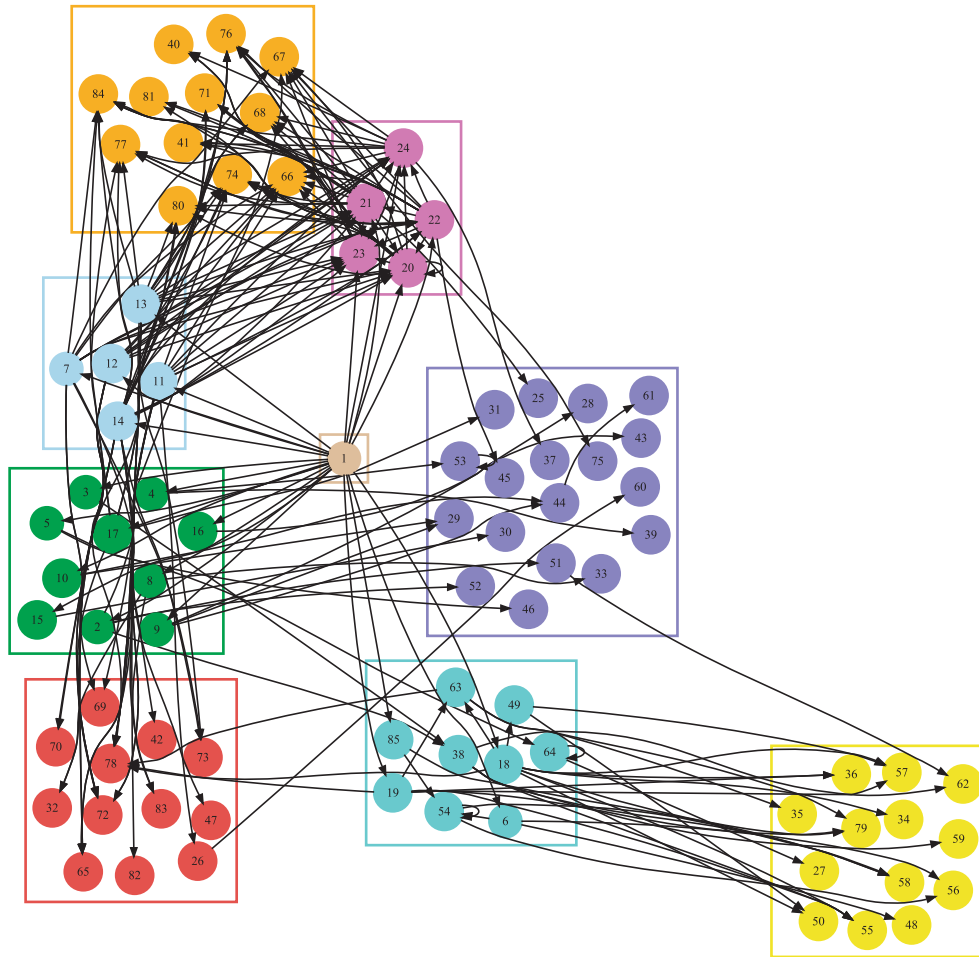


Figure 5 Best configuration for the *broom* food web. The groups are the same than those in Fig. 4.

and the probability of interaction for two nodes belonging to different groups. The idea is similar to the one developed here, although in the Čopič *et al.* approach there are only two probabilities (instead of k^2) and *AIC* cannot be used to discriminate among arrangements (all partitions have the same number of parameters). Finally, Rosvall & Bergstrom (2007) used entropy and information to directly partition networks into communities. If communities were the main driver behind the patterns we obtained in our analyses, the submatrices on the diagonal would contain very high densities of links, while the off-diagonal sub-matrices would be sparsely connected. The community definition becomes in this context an extreme case of the division of nodes into roles. In all the food webs analysed, we found very few highly connected submatrices on the diagonal: the influence of roles seems to be stronger than that of communities. To test this further, we took advantage of the fact that the *AIC* can be computed for any partition of the network. We therefore partitioned the two networks shown in Figs 5 and 6 using two recent community-detection algorithms (Gui-

merà *et al.* 2007; Rosvall & Bergstrom 2007). We then contrasted the *AIC* for the partitions obtained using community algorithms with those in Table 1. For the Skipwith pond, the *AIC* of the group-based model is 404.372, while the 'best' partition into communities yields 792.629 using Guimerà *et al.* or 776.505 using Rosvall & Bergstrom algorithm. We conclude that in this case the best partition obtained using communities yields much worse values than those derived above. The same pattern is found for the Scotch Broom system (the *AIC* values are 1081.523, 2125.602 and 2364.942 respectively).

The proposed framework opens several avenues for future research. We have begun to address the biological signature behind the groups identified by the algorithm. Are the groups related to trophic levels, body size or phylogenetic relationships? A match between biological characteristics and group structure will support to develop better models for food webs. A second possible extension of this work is indeed the development of such models by considering together group structure and some of the basic

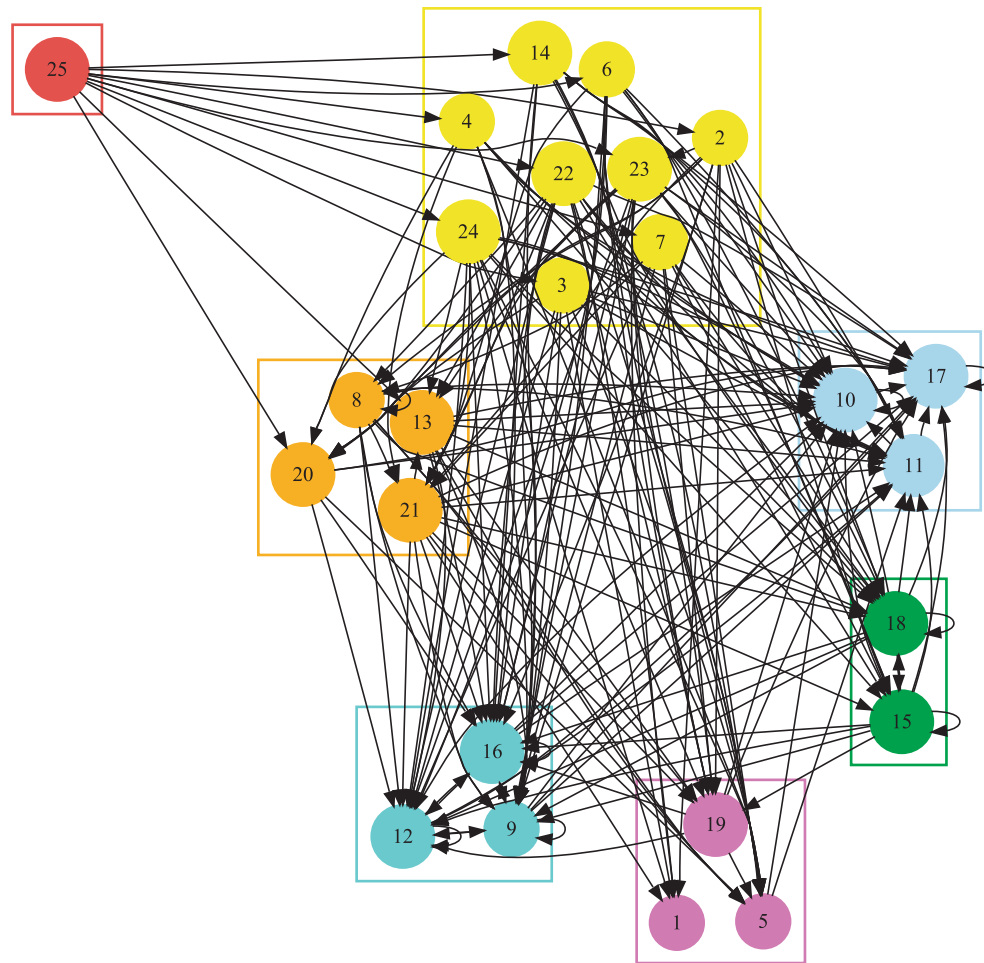


Figure 6 Best configuration for the Skipwith Pond food web. Also in this network, as in the one depicted in Fig. 5, the algorithm singled out the only primary producer in the network.

principles of existing models. A third future direction will address the aggregation problem in food webs. The algorithm can be used to produce networks of reduced complexity that retain the basic information about the roles of nodes in the network (like the image-graph). This might provide an alternative to the concept of trophic species in the aggregation and simplification of networks. Ways to produce such reduced networks should provide valuable tools for comparing systems of different sizes and levels of resolution, by reducing the networks to their essential features.

Consideration of different levels of resolution and a systematic way to approach the problem of aggregation would also be useful to explore how network properties vary as a result of coarser node identity. This may reveal network properties that are scale invariant and do not depend on the level of aggregation. Alternatively, properties that depend strongly on aggregation can be identified. Aggregation further relates to the important concept of spatial (or

temporal) boundaries in the definition of nodes. For example, how far should we aggregate the same species in two different locations in an ecosystem? Network structure may implicitly take into account spatial information and the analysis of groups can tell us whether this information is relevant or should instead be aggregated into a single node. This aspect relates to the current interest in the role of space in food web dynamics (Rooney *et al.* 2008).

The methods presented here are applicable to several other types of networks since they are valid for both directed and undirected graphs. We envision future applications to different social and biological systems including metabolic networks.

ACKNOWLEDGEMENTS

The authors thank D. Alonso and C. Melian, three anonymous referees and F. Jordan for useful comments and discussion. This work was funded by the NSF

EF-0827493 and by the James S. McDonnell Foundation. SA is a postdoctoral fellow at the National Center for Ecological Analysis and Synthesis, funded by NSF DEB-0553768, the University of California, Santa Barbara, and the State of California. M. Pascual is a Howard Hughes Medical Investigator. This work is dedicated to Luca E. Allesina, who was born while the algorithm was running.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19, 716–723.
- Alllesina, S., Bodini, A. & Bondavalli, C. (2005). Ecological subsystems via graph theory: the role of strongly connected components. *Oikos*, 110, 164–176.
- Alllesina, S., Alonso, D. & Pascual, M. (2008). A general model for food web structure. *Science*, 320, 658–661.
- Arii, K., Derome, R. & Parrott, L. (2007). Examining the potential effects of species aggregation on the network structure of food webs. *Bull. Math. Biol.*, 69, 119–133.
- Baird, D. & Ulanowicz, R.E. (1989). The seasonal dynamics of the Chesapeake Bay ecosystem. *Ecol. Monog.*, 59, 329–364.
- Bascompte, J., Melian, C. & Sala, E. (2005). Interaction strength combinations and the overfishing of a marine food web. *Proc. Natl Acad. Sci. USA*, 102, 5443–5447.
- Burnham, K. & Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*. Springer, New York.
- Cattin, M.F., Bersier, L.F., Banasek-Richter, C., Baltensperger, R. & Gabriel, J.P. (2004). Phylogenetic constraints and adaptation explain food-web structure. *Nature*, 427, 835–839.
- Christian, R.R. & Luczkovich, J. (1999). Organizing and understanding a winter's seagrass foodweb network through effective trophic levels. *Ecol. Model.*, 117, 99–124.
- Cohen, J., Briand, F. & Newman, C. (1990). *Community Food Webs: Data and Theory*. Springer-Verlag, Berlin.
- Critchlow, R. Jr & Stearns, S. (1982). The structure of food webs. *Am. Nat.*, 120, 478–499.
- Erdős, P. & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5, 17–61.
- Goldwasser, L. & Roughgarden, J. (1993). Construction of a large caribbean food web. *Ecology*, 74, 1216–1233.
- Guimerà, R., Sales-Pardo, M. & Amaral, L. (2007). Module identification in bipartite and directed networks. *Phys. Rev. E*, 76, 036102.
- Havens, K. (1992). Scale and structure in natural food webs. *Science*, 257, 1107–1109.
- Huelsenbeck, J., Ronquist, F., Nielsen, R. & Bollback, J. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294, 2310–2314.
- King, A., Ionides, E., Pascual, M. & Bouma, M. (2008). Inapparent infections and cholera dynamics. *Nature*, 454, 877–880.
- Krause, A.E., Frank, K.A., Mason, D.M., Ulanowicz, R.E. & Taylor, W.W. (2003). Compartments revealed in food-web structure. *Nature*, 426, 282–285.
- Levine, S. (1980). Several measures of trophic structure applicable to complex food webs. *J. Theor. Biol.*, 83, 195–207.
- Luczkovich, J.J., Borgatti, S.P., Johnson, J.C. & Everett, M.G. (2003). Defining and measuring trophic role similarity in food webs using regular equivalence. *J. Theor. Biol.*, 220, 303–321.
- Martinez, N.D., Hawkins, B.A., Dawah, H.A. & Feifarek, B.P. (1999). Effects of sampling effort on characterization of food-web structure. *Ecology*, 80, 1044–1055.
- Melián, C. & Bascompte, J. (2004). Food web cohesion. *Ecology*, 85, 352–358.
- Memmott, J., Martinez, N.D. & Cohen, J.E. (2000). Predators, parasitoids and pathogens: species richness, trophic generality and body sizes in a natural food web. *J. Anim. Ecol.*, 69, 1–15.
- Newman, M. (2006). Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA*, 103, 8577–8582.
- Newman, M.E. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69, 026113.
- Newman, M.E.J.E. & Leicht, E.A.A. (2007). Mixture models and exploratory analysis in networks. *Proc. Natl Acad. Sci. USA*, 104, 9564–9569.
- Optiz, S. (1996). *Trophic Interactions in Caribbean Coral Reefs*. Technical Report 43, ICLARM, Manila.
- Pimm, S. (1979). The structure of food webs. *Theor. Popul. Biol.*, 16, 144–158.
- Pimm, S.L. & Lawton, J.H. (1980). Are food webs divided into compartments? *J. Anim. Ecol.*, 49, 879–898.
- Pimm, S., Lawton, J. & Cohen, J. (1991). Food web patterns and their consequences. *Nature*, 350, 669–674.
- Polis, G. (1991). Complex trophic interactions in deserts: an empirical critique of food-web theory. *Am. Nat.*, 138, 123–155.
- Rooney, N., McCann, K. & Moore, J. (2008). A landscape theory for food web architecture. *Ecol. Lett.*, 11, 867–881.
- Rosvall, M. & Bergstrom, C. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl Acad. Sci. USA*, 104, 7327–7331.
- Warren, P. (1989). Spatial and temporal variation in the structure of a freshwater food web. *Oikos*, 55, 299–311.
- White, H., Boorman, S. & Breiger, R. (1976). Social structure from multiple networks. i: blockmodels of roles and positions. *Am. J. Sociol.*, 81, 730–780.
- Williams, R.J. & Martinez, N.D. (2000). Simple rules yield complex food webs. *Nature*, 404, 180–183.
- Yodzis, P. (1982). The compartmentation of real and assembled ecosystems. *Am. Nat.*, 120, 551–570.
- Yodzis, P. (1998). Local trophodynamics and the interaction of marine mammals and fisheries in the benguela ecosystem. *J. Anim. Ecol.*, 67, 635–658.
- Yodzis, P. & Winemiller, K. (1999). In search of operational trophospecies in a tropical aquatic food web. *Oikos*, 87, 327–340.

Editor, Ferenc Jordan

Manuscript received 6 February 2009

First decision made 7 March 2009

Manuscript accepted 31 March 2009