OIKOS

Research

Food web aggregation: effects on key positions

Emanuele Giacomuzzo and Ferenc Jordán

E. Giacomuzzo (https://orcid.org/0000-0003-2119-2163), Centre for Ecological Research, Budapest, Hungary. EG also at: Univ. of Zurich, Zurich, Switzerland and Eawag, Swiss Federal Inst. of Aquatic Science and Technology, Dübendorf, Switzerland. – F. Jordán (https://orcid.org/0000-0002-0224-6472) ✉ (jordan.ferenc@gmail.com), Democracy Inst., Central European Univ., Budapest, Hungary and Stazione Zoologica Anton Dohrn, Napoli, Italy.

Oikos
130: 2170–2181, 2021
doi: 10.1111/oik.08541

Subject Editor and
Editor-in-Chief: Dries Bonte
Accepted 8 September 2021

Food webs are often simulated dynamically to explore how trophic interactions influence resource and consumer abundances. As large trophic networks cannot be simulated in their original size – it would be too computationally expansive – they are shrunk by aggregating species together. However, key species may get lumped during this process, masking their unique role in their ecosystem. Therefore, a more systematic understanding of the aggregation effects on key positions is needed. Here, we study how six aggregation methods change 24 importance indices used to find key species in food webs. Our work was carried out on 76 aquatic food webs from the Ecopath with Ecosim database (EcoBase). The aggregation methods we considered were: 1) hierarchical clustering with the Jaccard index; 2) hierarchical clustering with the REGE index; 3) clustering within classic food web modules, which we refer to as 'density-based' modules; 4) clustering within 'predator-based modules' in which species fed on the same preys; 5) clustering within 'prey-based modules' in which species are fed upon by the same predators; and 6) clustering within 'groups' in which species share the same probability to interact with other groups. Hierarchical clustering with the REGE index produced the best results. Therefore, we recommend using it if we were interested in maintaining the identity of key species. The other algorithms could also be used to study specific network processes. However, we need to consider the bias they produce when masking important species.

Keywords: centrality indices, dynamic food web, ecological networks, food web aggregation, trophic role, trophospecies

Introduction

Food web models often include how species abundance (population dynamics) and interaction strength evolve over time (Curtsdotter et al. 2011). Such 'dynamical simulations' allow us to study properties of a community that we would not be able to study otherwise, such as, for example, secondary extinctions (Curtsdotter et al. 2011), standing stock and production at multiple trophic levels (Curtsdotter et al. 2018). As large food webs may include even more than 180 nodes (Martinez 1991), their simulation in their original size would be computationally too expansive. To simulate such networks, nodes are lumped together (Yodzis and Winemiller 1999), a process called 'data aggregation'. For simplicity, throughout the paper, we will refer to nodes as either

species or nodes interchangeably. Therefore, when speaking about species, we refer to a node of a food web, which can represent a species, a life stage, a higher taxon, a functional group or something else.

Since food webs reflect the trophic functioning of the community, the network position of species is informative about their trophic roles. Central positions may be a proxy for functional importance and the community-wide distribution of either centrality values (Bauer et al. 2010) or hypothetical importance values (Mills et al. 1993) provide macroscopic descriptors of ecosystems. Therefore, when shrinking food web size, the network should be simpler, but it must preserve key organisms in their critically important positions, while other species in similar positions (with partly redundant roles) can well be lumped together. For example, a sardine in a wasp-waist food web should not be lumped together with other pelagic fish since its unique position marks its trophic role (Cury et al. 2000). Different aggregates of microscopic and mesoscopic invertebrates, for example, are easier to be lumped without losing essential functional information.

The process of data aggregation assumes that there are nodes in the network that are similar enough that we can consider them functionally equivalent. For example, 'benthic feeders' and 'infauna' can compose the larger functional group 'benthic invertebrates'. Since we can use different aggregation algorithms based on different similarities, our work tried to find which is the best to maintain key species. Similarity can be understood mathematically (equivalent network positions) and biologically (similar trophic habits). As a food web node often represents organisms with different taxonomy and functional traits, lumping nodes according to biological similarities cannot be universally used for shrinking food webs. For example, the taxonomical classification of 'benthic invertebrates' and the body size of 'phytoplankton' would not be useful to find nodes that are similar to them. Therefore, the candidate algorithms we investigated were all based on lumping species with equivalent network positions.

Species can be considered in equivalent network positions if they have the same or similar predators and preys. Species that share a high number of predators and preys are said to be 'structurally equivalent' and can be aggregated, forming 'trophospecies' (a term first appearing in Yodzis 1988). Structural equivalence between nodes can be measured through the Jaccard index (Yodzis and Winemiller 1999). Species that share a high number of similar predators and preys (but not necessarily the same ones) are said to be 'regularly equivalent' and to have the same 'trophic role'. Regular equivalence between nodes can be measured using the REGE index (Luczkovich et al. 2003). Aggregating structurally equivalent and regularly equivalent nodes are the most common ways of performing a mathematical data aggregation in food webs (Sugihara et al. 1989, 1997, Martinez 1991, Johnson et al. 2009, Olivier and Planque 2017).

Species can also be considered in equivalent network positions if they are part of the same module, where a module is defined as a set of species sharing an interaction pattern. Food webs can have different types of modules. The most studied network modules are nodes with different neighbours, but that form a dense subgraph (Guimerà et al. 2010). They are usually just called modules, but here we call them 'density-based modules' as they have been sometimes called (Malliaros and Vazirgiannis 2013); this denomination allows us to distinguish them from the other types of modules. The second and third types of modules are the ones where species interact with the same preys and those where species interact with the same predators. We call these 'predator-based' and 'prey-based' modules, respectively. These modules were first introduced to network science by Guimerà et al. (2007) and then applied to food webs by Guimerà et al. (2010). Another type of module that we can find in food webs is the so-called 'group'. Groups are species with the same probability of interacting with species from other groups; these modules are detected using the group model (Allesina and Pascual 2009).

Here, we study how aggregating species with similar or identical predators and preys (using hierarchical clustering with the Jaccard or the REGE index) or species within the same module (defined as a density-based module, predator-based module, prey-based module or group) preserves the key species of the food web. Key species were found by using 24 of the most used importance indices in food web research. Our investigation was carried out on 76 Ecopath with Ecosim (EwE) food web models available on the EcoBase database (Colléter et al. 2013). Having been constructed with the same methodology (Okey 2004a) provided us with comparable results. See the Supporting information for a list of these networks.

Methods – food web aggregation

Food webs were shrunk using different aggregation methods. Each aggregation had to take the following three steps:

1) Create clusters with similar nodes.
2) Connect the clusters.
3) Assign interaction strength to the connections between clusters.

Because we combined six ways of creating clusters with similar nodes, five ways of connecting the clusters and four ways of assigning interaction strength, we used 120 methods ($6 \times 5 \times 4 = 120$) to aggregate a food web. We applied each of these aggregation methods to each network.

Food web aggregation step 1. Create clusters with similar nodes

The first step in aggregating a food web is clustering its nodes according to a measure of similarity. For each aggregation, we clustered nodes according to one of the following techniques: 1) hierarchical clustering with Jaccard index; 2) hierarchical clustering with REGE index; 3) clustering within density-based modules; 4) clustering within predator-based modules; 5) clustering within prey-based modules; or 6) clustering within groups.

Hierarchical clustering with Jaccard index

As a first clustering method, we clustered structurally equivalent nodes as in Martinez (1991). To do this, we first calculated the similarity between nodes, defined as the fraction of their predators and preys that they shared; this is known as the Jaccard index (Jaccard 1912). After creating a dendrogram based on the Jaccard index, we cut it where the inconsistency between branches was higher than 0.01. See the Supporting information for the hierarchical clustering algorithm.

Hierarchical clustering with REGE index

Our second clustering method consisted of clustering regularly equivalent nodes as in Luczkovich et al. (2003). To do this, we first calculated the degree to which nodes have similar preys and predators using the REGE index (Borgatti and Everett 1993). After creating a dendrogram based on the REGE index, we cut it where the inconsistency between branches was higher than 0.01. See the Supporting information for the hierarchical clustering algorithm.

Clustering within density-based modules

As a third clustering method, we clustered the nodes inside the modules found by maximising the density modularity, as in Guimerà et al. (2010). This type of modularity is expressed as the number of extra links present within the modules compared to those expected by chance. For directed networks, it can be expressed through the following equation of Arenas et al. (2007), which is a generalisation of the Newman–Girvan modularity (Newman 2004)

$$Q = \frac{1}{L}\sum_{ij}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{L}\right]\delta_{m_i m_j} \quad (1)$$

where Q is the network's modularity, L is the number of links in the network, A_{ij} is the element of the adjacency matrix of the directed binary network (links go from j to i), k_i^{in} is the indegree of i, k_j^{out} is the outdegree of j, m_i is the module of i, m_j is the module of j and δ is the Kronecker delta (Kozen and Timme 2007).

The number and composition of the modules were found by using the Leiden algorithm of Traag et al. (2019). This algorithm is an extension of the Louvain algorithm (Blondel et al. 2008). The latter is one of the best performing and fastest for community detection. However, it tends to produce communities that are arbitrarily poorly connected and sometimes even disconnected. The Leiden algorithm not only solves this problem by creating better-connected communities, but it is also faster (Traag et al. 2019). The code used was implemented in the igraph package (Csardi and Nepusz 2006) for R (<www.r-project.org>).

Clustering within prey-based and predator-based modules

As the fourth and fifth clustering methods, we clustered the nodes within every module found by maximising the prey modularity and the predator modularity of the food web, as in Guimerà et al. (2010). In this case, the network's modularity is expressed as to how different nodes connect to the same predators (for prey modularity) or preys (for predator modularity) than expected by chance. Mathematically, it can be expressed by the following equation (Guimerà et al. 2007) for prey modularity

$$Q = \sum_{ij}\left[\frac{c_{ij}^{out}}{\sum_l k_l^{in}\left(k_l^{in}-1\right)} - \frac{k_i^{out}k_j^{out}}{\left(\sum_l k_l^{in}\right)^2}\right]\delta_{\boldsymbol{m}_i m_j} \quad (2)$$

or by the following one for predator modularity

$$Q = \sum_{ij}\left[\frac{c_{ij}^{in}}{\sum_l k_l^{out}\left(k_l^{out}-1\right)} - \frac{k_i^{in}k_j^{in}}{\left(\sum_l k_l^{out}\right)^2}\right]\delta_{\boldsymbol{m}_i m_j} \quad (3)$$

where c_{ij}^{out} is the number of outgoing links that i and j have in common and c_{ij}^{in} is the number of incoming links that i and j have in common. We maximised these two types of modularity using the rnetcarto package (Doulcier and Stouffer 2015) implemented in R. This finds the network's community structure using simulated annealing (Kirkpatrick et al. 1983).

Clustering within groups

As a sixth clustering method, we clustered the nodes inside the modules found by the group model of Allesina and Pascual (2009). This model finds the modules that maximise the probability of randomly retrieving the food web by generating a modular version of an Erdős–Rényi random graph. For an arbitrary number of groups k, the probability of recovering the network is:

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

where $N(S,L)$ is the food web N with S number of nodes and L number of links, $\vec{p}$ is the vector containing the probabilities of a connection between and within clusters, p_{ij} is the probability that a node inside the group i connects to another node inside the group j, L_{ij} is the number of links connecting nodes belonging to the group i to nodes belonging to the group j, S_i is the number of nodes in the cluster i and S_j is the number of nodes in the cluster j.

It is not possible to explore all possible module arrangements because of their high number. To find the best possible solution that our computation power allows us to find, we used the algorithm of Sander et al. (2015). This algorithm relies on a metropolis-coupled Markov chain Monte Carlo (MC^3), also known as 'parallel tempering' (Geyer 1991), with a Gibbs sampler (Yildirim 2012). MC^3 can be considered a Markov chain Monte Carlo (MCMC) with multiple chains running simultaneously (Sander et al. 2015).

Food web aggregation step 2. Connect the clusters

After having created clusters, we had to connect them. The connection of the clusters followed a similar approach to the one described in the seminal work of Martinez (1991). To decide if there was a link from cluster A to cluster B, we used five methods. These methods were all based on the number of possible connections going from A's nodes to B's nodes that were realised. In the first method, these realised connections had to be at least 25%, in the second method at least 50% and in the third at least 75%. In the fourth method, there should have been at least one realised connection (a method known as the 'maximum linkage method', or NMAX), and in the fifth method, all the possible connections had to be realised (a method known as the 'minimum linkage method', or NMIN).

Food web aggregation step 3. Assign interaction strength to the connections between clusters

We used four methods to calculate the interaction strength of the connection between clusters. These methods calculated it from the connections between the nodes of the two clusters as: 1) their minimum interaction strength; 2) their maximum interaction strength; 3) the sum of their interaction strengths; and 4) the mean of their interaction strengths.

Methods – importance indices

For each food web, we calculated the importance indices before and after the aggregation. The importance indices of a node after the aggregation were defined as the ones of its cluster. Let us consider the following example. Before the aggregation, the node 'hake' has a degree centrality of 5. The aggregation process clusters it with other fish nodes, creating a node in the aggregated food web called 'fish'. The degree centrality of 'fish' is 8. In this case, the degree centrality of 'hake' is 5 in the original network and 8 in the aggregated network. The importance indices we used belonged to the following families: degree centrality, closeness centrality, betweenness centrality, status index, keystone index, topological importance and species uniqueness.

Degree centrality

The degree centrality (DC) of a node i is the number of links it has (Wasserman and Faust 1994)

$$\mathrm{DC}_i = \sum_{j=1}^{n} A_{ij} \tag{5}$$

where n is the number of nodes in the food web, and A_{ij} is the element of the adjacency matrix after the network has been transformed into a binary undirected one.

Another type of degree centrality that we considered was the weighted degree centrality (wDC), often referred to as node strength. Its formula is the same as for the non-weighted degree centrality. This time, however, the adjacency matrix is of an undirected weighted network (Fornito et al. 2016)

$$\mathrm{wDC}_i = \sum_{j=1}^{n} A_{ij} \tag{6}$$

Closeness centrality

The closeness centrality (CC) of a node is the average distance of a node from all the others in the network (Wasserman and Faust 1994)

$$\mathrm{CC}_i = \frac{1}{\sum_{j=1}^{n} d(i,j)} \tag{7}$$

where $d(i, j)$ is the shortest path between node i and j.

Betweenness centrality

The betweenness centrality (BC) of a node is the average number of times it acts as a bridge along the shortest path between two nodes. It can be mathematically expressed as follows (Wasserman and Faust 1994)

$$\mathrm{BC}_i = \sum_{i \neq m \neq n} \frac{\sigma_{mn}(i)}{\sigma_{mn}} \tag{8}$$

where σ_{mn} is the total number of shortest paths going from m to n and $\sigma_{mn}(i)$ is the total number of these paths passing through i.

Status index

The status index of a node is the sum of its distances from all the other nodes inside the network, calculated as their shortest paths following a bottom–up direction (Endrédi et al. 2018)

$$s_i = \sum_{j=1}^{n} d(i,j) \tag{9}$$

where $d(i, j)$ is the shortest path between node i and j. It was first introduced to social networks, followed two years later by its application to food webs by Harary (1959, 1961). Following the same method but in a top–down direction, we obtain the contrastatus (s_i')

$$s_i' = \sum_{j=1}^{n} d(i,j) \tag{10}$$

The difference between the status and the contrastatus is called the net status (Δs_i)

$$\Delta s_i = s_i - s_i' \tag{11}$$

The status index can be meaningful where indirect effects are not becoming weaker but stronger with distance. This can be the case with the bioaccumulation of heavy metals in a food chain. In this case, the second neighbour can be impacted more than the first neighbour. The computation of the status, contrastatus and net status needs to be performed on a network without cycles. See the Supporting information for the algorithm used to create a directed acyclic graph (DAG).

Keystone index

The keystone index was firstly introduced by Jordán et al. (1999) and was inspired by the status index. As in the status index family, the keystone index of a species is calculated by considering the bottom–up and the top–down effects separately (Jordán et al. 2006). Unlike the status index, the keystone index considers how the size of a certain effect gets split between the different neighbours of a node. Every time the effect reaches a certain node connected to multiple nodes, the following nodes receive only a fraction of the total effect. For example, when considering the bottom–up effect, if the prey has two predators, the bottom–up effect received by each predator will be half. The computation of the keystone index also needs to be performed on a network without cycles.

The keystone index of a species i is equal to the sum of its bottom–up and top–down effects

$$K(i) = K_b(i) + K_t(i) \tag{12}$$

where $K(i)$ is the keystone index, $K_b(i)$ is the bottom–up keystone index and $K_t(i)$ is the top–down keystone index. The bottom–up keystone index ($K_b(i)$) is calculated as

$$K_b(i) = \sum_{P=1}^{n} \frac{1}{\text{preys}(i)(P)} + \frac{K_b(P)}{\text{preys}(i)(P)} \tag{13}$$

where P is a predator of i and preys(i)(P) is the number of preys of P. $\frac{1}{\text{preys}(i)(P)}$ is the direct bottom–up effect of i on its predator P. $\frac{K_b(P)}{\text{preys}(i)(P)}$ is the fraction of bottom–up effects of P that are caused by i. The $K_b(P)$ of top predators is set as 0. The top–down keystone index is calculated precisely as the bottom–up keystone index but with the direction of the links inverted.

When calculating the keystone index, we might also be interested in how much the effect of a species depends upon its directed and undirected effects. To do this, we can split the keystone index of a species into its direct and indirect effect components

$$K(i) = K_{\text{dir}}(i) + K_{\text{indir}}(i) \tag{14}$$

where $K_{\text{dir}}(i)$ is its directed component, and $K_{\text{indir}}(i)$ is its indirect component. To calculate them, we need to know that

$$K_{\text{dir}}(i) = K_{b,\text{dir}} + K_{t,\text{dir}} \tag{15}$$

$$K_{\text{indir}}(i) = K_{b,\text{indir}} + K_{t,\text{indir}} \tag{16}$$

where $K_{b,\text{dir}}$ is the directed component of the bottom–up effect, $K_{t,\text{dir}}$ is the directed component of the top–down effect, $K_{b,\text{indir}}$ is the indirect component of the bottom–up effect and $K_{t,\text{indir}}$ is the indirect component of the top–down effect. The directed and indirect components of the bottom–up index are calculated as

$$K_{b,\text{dir}}(i) = \sum_{P=1}^{n} \frac{1}{\text{preys}(i)(P)} \tag{17}$$

$$K_{b,\text{indir}}(i) = \sum_{P=1}^{n} \frac{K_b(P)}{\text{preys}(i)(P)} \tag{18}$$

The direct and indirect components of the top–down effect are calculated in the same way but inverting the direction of the links.

Topological importance

Topological importance (TI) was first introduced to host–parasitoid networks by Müller et al. (1999) and then to food webs by Jordán et al. (2003). The topological importance of a species is the potential it has to create bottom–up effects. It differs from the bottom–up keystone index because instead of considering effects weakening as they spread by being divided by the outdegree of the prey, it considers them weakening by being divided by the indegree of the predator. When calculating topological importance, we need to set a number of steps to which the effect is allowed to spread. For example, let us say that we set this number to three. In this case, topological importance will tell us about the effects a particular species has on species that are maximum three connections away when creating bottom–up effects. If topological importance considers interaction strength, we refer to it as weighted topological importance (WI). See the Supporting information for the algorithm for the computation of topological importance and weighted topological importance.

Species uniqueness

Species uniqueness (STO) represents how redundant the strong interactions of a node are. It was first introduced by Lai et al. (2015) and can be considered as an extension of the trophic field overlap (TO) (Jordán et al. 2009). The trophic field overlap of a node *i* is the number of times that *i* and another node interact strongly with the same predator. Interactions are considered strong when they exceed a certain threshold. To avoid choosing a single threshold over which interactions are considered strong, TO is calculated based on multiple thresholds. The thresholds we used were 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09 and 0.1. We chose 0.1 as the highest threshold because higher thresholds did not modify the value of STO. We chose only ten thresholds because otherwise the analysis would have been too computationally demanding. Each of these TO is then summed to give species uniqueness. We can calculate the topological overlap considering interaction strength by calculating weighted species uniqueness (wSTO). See the Supporting information for the algorithm for the computation of trophic field overlap.

Methods – statistical analysis

To study the effects of aggregation procedures on importance indices, we calculated the correlation between the node ranking before and after the aggregations. This correlation was calculated using Kendall's tau B (τ_B) – a version of Kendall's rank correlation coefficient that makes adjustments for ties (Agresti 2012). For each combination of aggregation method and index of importance, we found the mean τ_B across all food webs and its 95% confidence intervals through bootstrapping (DiCiccio and Efron 1996). Because correlation coefficients are often not normally distributed, we used Fisher's z transformation (Fisher 1915). This technique consists of transforming the values of a distribution into their arctangent, which makes them normally distributed. Therefore, we transformed the τ_B into their arctangent, calculated their mean and confidence intervals, and back-transformed these means and confidence intervals into τ_B by calculating their tangent. τ_B and bootstrapping were implemented in the Statistics and Machine Learning Toolbox for MATLAB (Mathworks Inc. 2019).

Results

Size of the clusters produced

The 76 food webs we used had a median of 25.5 nodes (IQR = 16.0), with a minimum of 14 nodes and a maximum of 55 nodes (Fig. 1). The median size of the aggregated network compared to the original one was 74.5% (IQR = 10.8%) for the Jaccard index, 73.0% (IQR = 7.2%) for the REGE index, 12.8% (IQR = 6.5) for the density-based modules, 35.8% (IQR = 21.3%) for the prey-based modules, 72.1% (IQR = 29.6%) for the predator-based modules and 15.8% (IQR = 6.5%) for the group model (Fig. 2).

Correlation of importance indices before and after the aggregation

By focusing only on the clustering method and ignoring the linkage and interaction strength methods, we can select the best clustering for each combination of centrality indices and clustering methods (see the Supporting information for a heatmap with the correlation between the ranking before and after the aggregation including linkage and interaction strength methods). This provides us with a clearer heat map

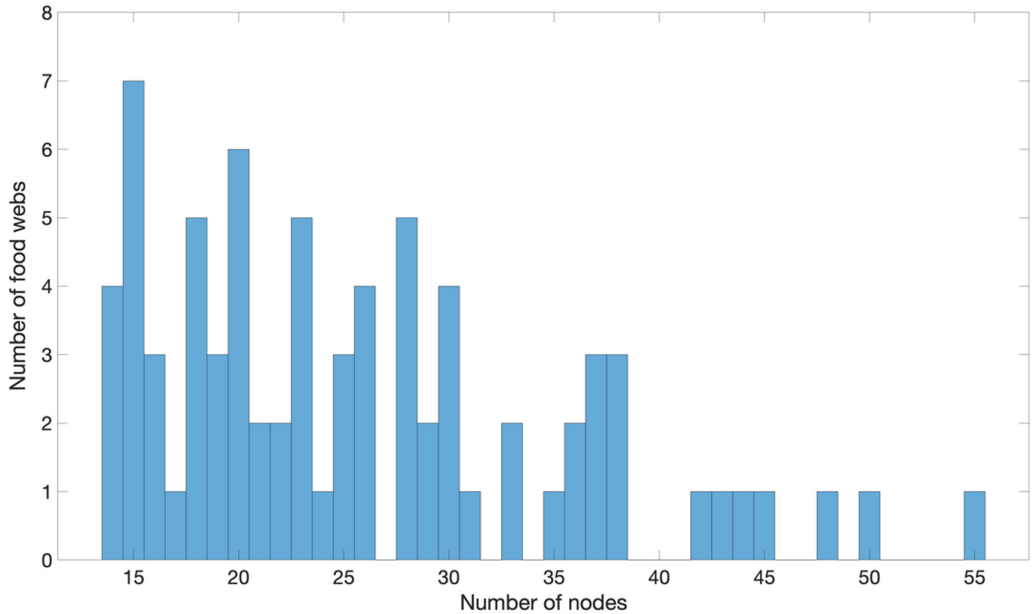

Figure 1. Size of the food webs we used in our study.

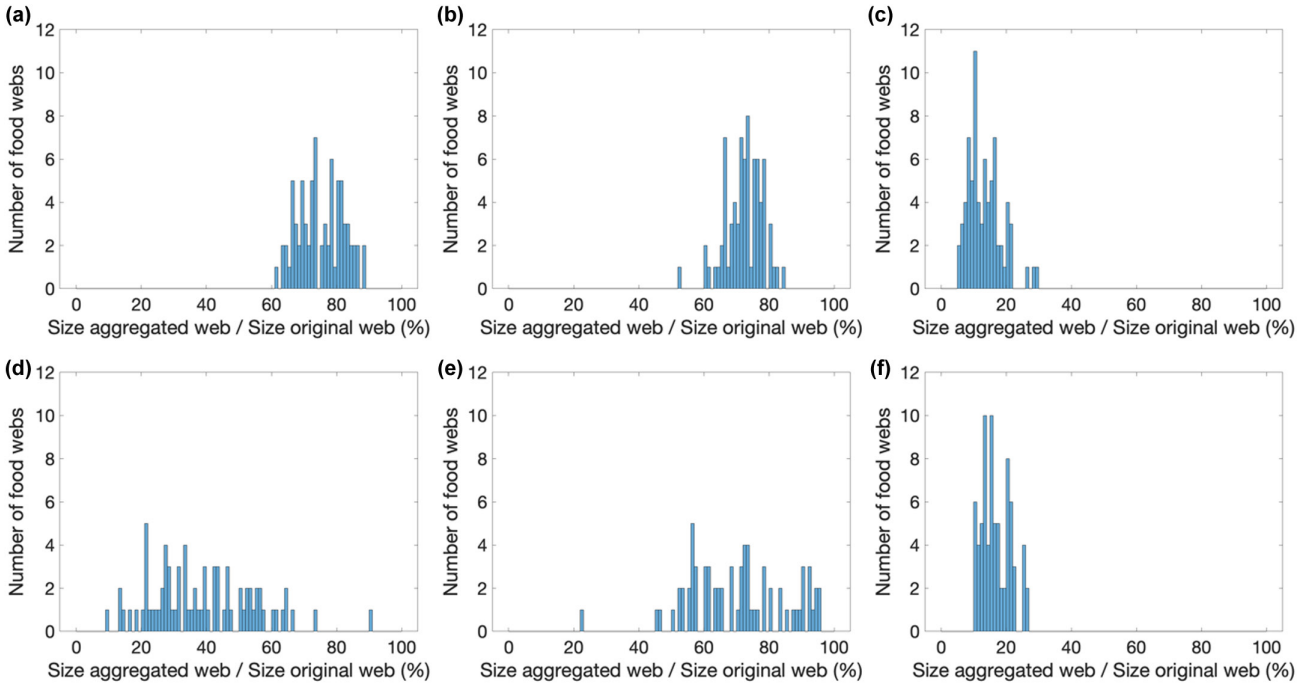

Figure 2. Size of the food webs produced by the different clustering methods. (a) Hierarchical clustering with Jaccard index, (b) hierarchical clustering with REGE index, (c) clustering within density-based modules, (d) clustering within prey-based modules, (e) clustering within predator-based modules, (f) clustering within groups.

(Fig. 3). Ranking the clustering algorithms in Fig. 3 produces Table 1. Density modularity was consistently ranked as the worst clustering algorithm. Prey-based modules and group model were consistently ranked as either fourth or fifth. Except for BC and s', the clustering of predator modules ranked consistently as third. Excluding the results of contrastatus, the hierarchical clustering based on the Jaccard index and the hierarchical clustering based on the REGE

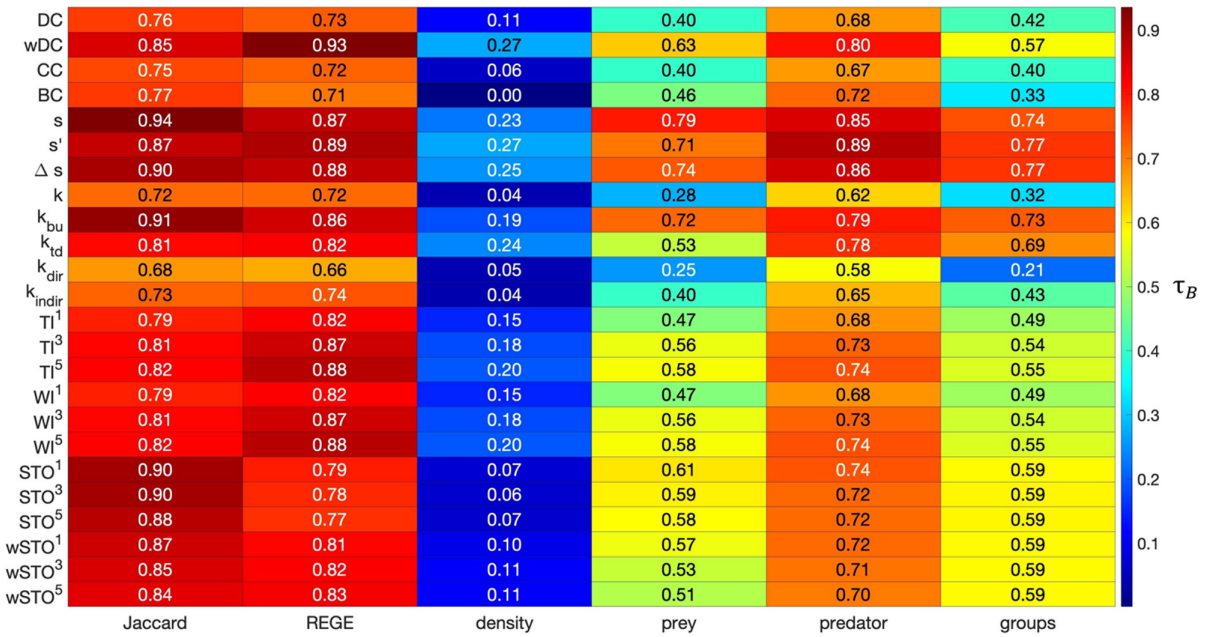

Figure 3. Heat map of the best Kendall's rank correlation coefficient for each combination of clustering methods and importance indices. The best correlation is selected across linkage methods and methods of determining interaction strength. Jaccard = hierarchical clustering using Jaccard index, REGE = hierarchical clustering using REGE index, density = clustering within density-based modules, prey = clustering within prey-based modules, predator = clustering within predator-based modules, groups = clustering within groups. See the Supporting information for the confidence intervals associated with these values and the linkage method and interaction strength method used.

Table 1. Best Kendall's correlation coefficients, as in Fig. 3. They are ranked from the clustering that produced the best correlation to the clustering that produced the worst correlation. Green = hierarchical clustering with Jaccard index, red = hierarchical clustering with REGE index, grey = clustering within density-based modules, blue = clustering within predator-based modules, purple = clustering within groups produced by the group model, yellow = clustering within prey-based modules. Index = importance index, C1 = best clustering, C2 = second-best clustering, C3 = third-best clustering, C4 = fourth-best clustering, C5 = fifth-best clustering, C6 = sixth-best clustering.

Index	C1	C2	C3	C4	C5	C6
DC	0.76	0.73	0.68	0.42	0.40	0.11
wDC	0.93	0.85	0.80	0.63	0.57	0.27
CC	0.75	0.72	0.67	0.40	0.40	0.06
BC	0.77	0.72	0.71	0.46	0.33	0.00
s	0.94	0.87	0.85	0.79	0.74	0.23
s'	0.89	0.89	0.87	0.77	0.71	0.27
Δs	0.90	0.88	0.86	0.77	0.74	0.25
k	0.72	0.72	0.62	0.32	0.28	0.04
k_{bu}	0.91	0.86	0.79	0.73	0.72	0.19
k_{td}	0.82	0.81	0.78	0.69	0.53	0.24
k_{dir}	0.68	0.66	0.58	0.25	0.21	0.05
k_{indir}	0.74	0.73	0.65	0.43	0.40	0.04
TI^1	0.82	0.79	0.68	0.49	0.47	0.15
TI^3	0.87	0.81	0.73	0.56	0.54	0.18
TI^5	0.88	0.82	0.74	0.58	0.55	0.20
WI^1	0.82	0.79	0.68	0.49	0.47	0.15
WI^3	0.87	0.81	0.73	0.56	0.54	0.18
WI^5	0.88	0.82	0.74	0.58	0.55	0.20
STO^1	0.90	0.79	0.74	0.61	0.59	0.07
STO^3	0.90	0.78	0.72	0.59	0.59	0.06
STO^5	0.88	0.77	0.72	0.59	0.58	0.07
wSTO^1	0.87	0.81	0.72	0.59	0.57	0.10
wSTO^3	0.85	0.82	0.71	0.59	0.53	0.11
wSTO^5	0.84	0.83	0.70	0.59	0.51	0.11

index were consistently ranked as the best clustering methods. Jaccard index was better than REGE for weighted and unweighted species uniqueness, unweighted topological importance, degree centrality, closeness centrality and betweenness centrality. REGE was better for weighted topological importance and weighted degree centrality. According to which index of those two families we considered, the status index and keystone index were maintained better either by Jaccard or REGE. We can qualitatively say that the correlation between the ranking before and after the aggregation seems to increase with the size of the aggregated food web.

Discussion

Effects of data aggregation on key species

Food webs are often simulated dynamically to study how the abundance of species, strength of interactions and community-wide properties (e.g. standing stock and production at multiple trophic levels) evolve over time (Curtsdotter et al. 2018). However, large food webs cannot be dynamically simulated, as such a task would be too computationally expansive. To overcome this problem, the size of the network is reduced by lumping nodes together, producing a smaller version that can be then simulated (Yodzis and Winemiller 1999). This practice, however, has the potential to aggregate key species. These species produce large effects on their community by spreading perturbations across the food web; for example, by acting as a bridge between other nodes, spreading effects in a bottom–up direction, or spreading effects in a top–down direction (Jordán et al. 2007).

Our study is the first to systematically compare the effects of different aggregation algorithms on key species, where key species are not restricted to any taxa. Multiple studies investigated the effects of data aggregation on food web properties such as connectance, food chain length, and the ratio between the bottom, intermediate and top species (Angelini and Agostinho 2005, Johnson et al. 2009, Gauzens et al. 2013, Olivier and Planque 2017). However, no study has looked at the effects of data aggregation on lumping important species, except for Essington and Plagányi (2014) and Plagányi and Essington (2014). These two papers, however, focused only on forage fish.

Here, we studied the effects of aggregation on several indices used to find key species in food web ecology (Estrada 2007, Olmo Gilabert et al. 2019). A species is considered key if one of these indices is disproportionally high compared to the other species. The aggregation methods we compared were the two most used for mathematical aggregation (hierarchical clustering using the Jaccard index and hierarchical clustering using the REGE index) and four used for finding modules (we clustered species within density-based modules, prey-based modules, predator-based modules and groups). The latter four are typically used for community detection, but we explored using them for data aggregation.

Hierarchical clustering with Jaccard or REGE can shrink food webs to the desired size, as we can choose where to cut the dendrogram. However, each community detection algorithm can shrink a particular food web to only one size. How much they aggregate will determine whether the size of the produced network is small enough to be dynamically simulated. Our results show that different community detection algorithms have different potentials to shrink large food webs: aggregating within density-based modules and groups produced a substantial aggregation of the network (original size was shrunk to a median of 12.8% and 15.8%, respectively), within predator-based modules produced a weak aggregation (original size was shrunk to a median of 72.1%), and within prey-based modules produced an intermediate aggregation (original size was shrunk to a median of 35.8%) (Fig. 5). The maximum size that each of these four algorithms will shrink will depend upon the largest size of the network that can be simulated, depending on different factors (e.g. network connectance, the machine where the simulation is taking place, and the software used).

Our results also show that different aggregation methods maintain the relative importance of species in different ways. Therefore, they have different potential to change the food web's key species (Fig. 4). Except for the contrastatus index (s') and betweenness centrality (BC), hierarchical clustering with Jaccard and REGE outperformed the other methods.

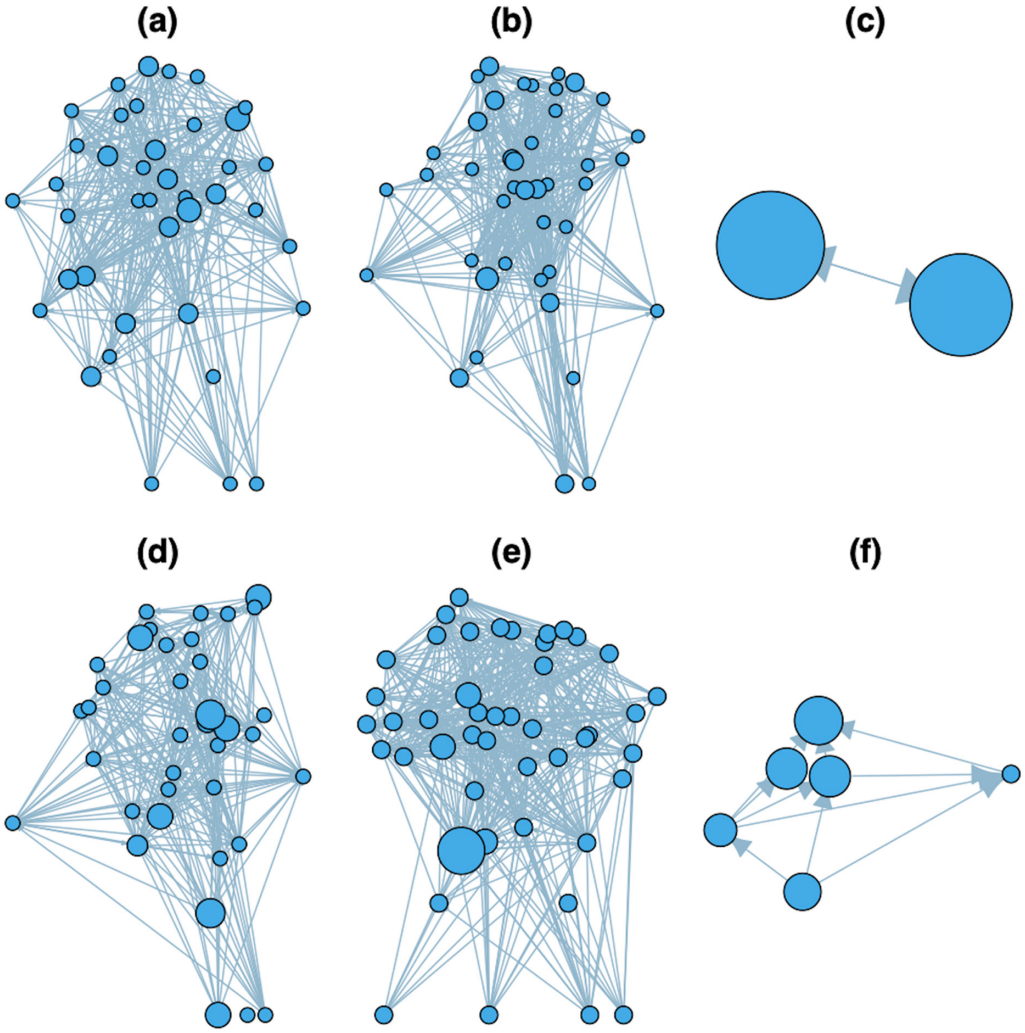

Figure 4. Aggregation of the food web of Fig. 4 according to different clustering algorithms. (a) Hierarchical clustering using Jaccard index, (b) hierarchical clustering using REGE index, (c) clustering within density-based modules, (d) clustering within prey-based modules, (e) clustering within predator-based modules, (f) clustering within groups. The linkage method and the interaction strength method used for each clustering were the ones that produced the highest Kendall's rank correlation between the ranking before and after the aggregation. The size of the nodes is proportional to the number of nodes that have been aggregated into them. Also in this case, we omitted self-loops.

Jaccard and REGE maintained different indices differently – some were better preserved by Jaccard and others by REGE. When deciding which among these two methods preserved key species the best, we need to consider that not all indices have the same power to predict key species. Gouveia et al. (2021) looked at topological important indices and how their findings correlated with the results of the dynamical important index keystoneness (Libralato et al. 2006). They found that the most reliable topological importance index was the weighted degree centrality (wDC). wDC could predict the most important species for dynamic processes in 70.1% of the cases. It was followed by a combination of wDC and the 5-step weighted topological importance (WI^5), which increased this percentage to 78.4%. In light of these findings, REGE might be considered the best clustering algorithm to maintain key species, as it maintains wDC and WI^5 the best.

When choosing among these methods, we recommend using the algorithm that better fits the research question but keeping in mind its effects on key positions. Hierarchical clustering with the REGE index can be used to maintain the key species of the system but might miss some other information. As hierarchical clustering with the Jaccard index maintained unaggregated species that were key according to other definitions (e.g. species uniqueness), we might want to use it if we were interested in such species. Also, Jaccard and REGE reveal two different things: Jaccard should be used to model guilds of competitors, and REGE should be used if we want to get rid of functional redundancy. The group model could be used to reveal structures such as, for example, trophic guild and habitat patterns (Baskerville et al. 2011, Eklöf et al. 2012, Michalska-Smith et al. 2018, Sander et al. 2015). Density-based modules could be used, for instance, to study

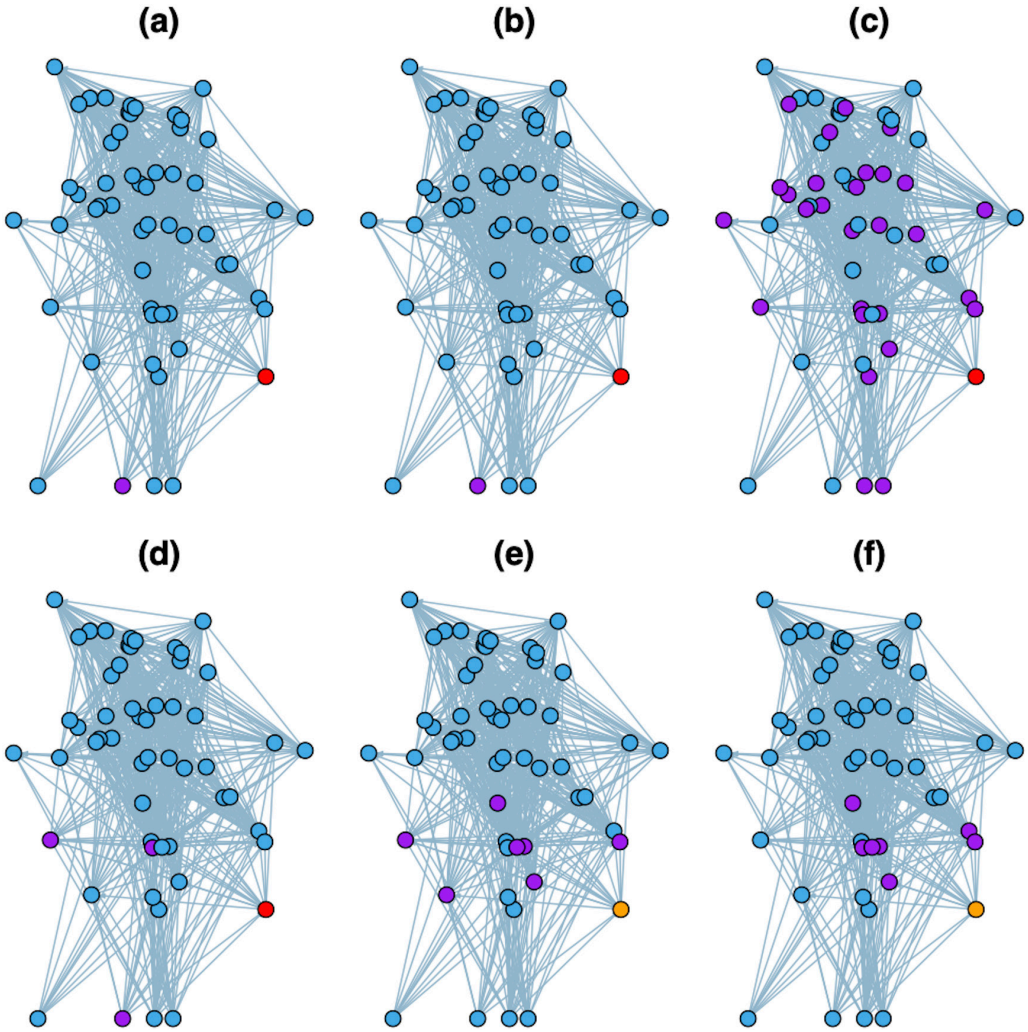

Figure 5. Effect of aggregation on most the most important nodes in a food web. Red: most important before and central after. Orange: most important before but not after. Purple: not the most important before and most important after. Blue: not the most important before and after. The self-loops are not included in the figure for clarity. (a) Hierarchical clustering using Jaccard index, (b) hierarchical clustering using REGE index, (c) clustering within density-based modules, (d) clustering within prey-based modules, (e) clustering within predator-based modules, (f) clustering within groups. The food web here depicted is the one of the West Florida Shelf (Okey 2004b). It is the largest network used in this study (55 nodes). To make the figure clearer, we omitted the self-loops.

the connection among different trophic levels across different habitats (Kortsch et al. 2015). Predator-based and prey-based modules, as they have not been widely studied, have not been found to reveal any ecological process yet. However, we do not exclude that they might do in the future. Each method reveals biological similarities between nodes (although these four module types are not always present in food webs). The key point is to know the effects of these aggregation procedures and keep them in mind when applying them and evaluating the properties of the aggregated network.

This paper focuses on a well-defined set of food webs that are methodologically strictly comparable (all created by the EwE methodology, Heymans et al. 2014). Studying aggregation would be more exciting for larger networks, but these are rarely comparable (lacking standards for their description). When data will be available, interaction networks representing various interaction types will be very important to analyse from our present perspective: different interaction types may be differently sensitive to aggregation. A significant advancement would be the analysis of aggregation effects on dynamical food web models. To see how dynamical properties can be altered or changed as an effect of aggregation algorithms would be a significant step towards predictive food web modelling. If the effects of aggregation algorithms are clearly understood, network aggregation can produce smaller networks with trustable properties, and we can better understand and interpret the difference between smaller and larger versions of food webs. This knowledge about the difference between small and big networks is crucial when the dynamical analyses of smaller food webs must be interpreted.

In the future, it will also be essential to understand how these algorithms affect other food web properties (e.g.

frequency of motif occupancy, Cirtwill et al. 2018). Olivier and Planque (2017) showed that aggregating using hierarchical clustering with REGE biases properties such as connectance, the ratio between predators and preys, percentage of cannibal species, and dispersion of in-degree and out-degree distributions more than when using Jaccard. However, the size of this bias depended upon the threshold chosen for REGE and Jaccard to cut the dendrogram. For certain thresholds, Jaccard and REGE produced a similar bias. However, no study has compared the effects of the other four algorithms. Therefore, such a study would go complementing the one of Olivier and Planque (2017). Furthermore, another study could investigate whether Jaccard and REGE have a threshold to preserve food web properties and key positions simultaneously.

Conclusion

In conclusion, we have shown that different aggregation methods maintain key species in a food web to different degrees. Hierarchical clustering with the REGE index maintained the key species unaggregated the most, followed by hierarchical clustering with the Jaccard index. Therefore, we recommend using hierarchical clustering with the REGE index to maintain key species unaggregated. The choice of the aggregation algorithm, however, should be driven by our research question. If we were interested in the key species maintained better by the Jaccard index, we should use it instead of REGE. The community detection algorithms could be used if we were interested in the connection among the structures they reveal (e.g. spatial guilds for the group model, Baskerville et al. 2011). However, we should keep in mind that they are likely to be aggregating key species. Future research should be carried out on larger food webs, on networks constructed with different methodology, on dynamical food webs and aimed to study the effects of aggregation on other network properties.

Acknowledgements – We would like to thank Wei-Chung Liu for providing the code for computing some importance indices, Stefano Allesina and Elizabeth Sander for providing the code for the computation of the group model, and Anett Endrédi for providing us support with data management. Finally, we would like to thank Timothée Poisot whose excellent comments improved our manuscript significantly.
Funding – FJ was supported by AtlantECO (H2020 BG-08-2018-2019, grant no. SEP-210591007).

Author contributions

Emanuele Giacomuzzo: Conceptualization (equal); Formal analysis (lead); Investigation (lead); Methodology (lead); Project administration (lead); Software (lead); Validation (equal); Visualization (lead); Writing – original draft (lead). **Ferenc Jordán**: Conceptualization (equal); Data curation (lead); Software (supporting); Supervision (lead); Validation (equal); Writing – review and editing (lead).

Data availability statement

Data are available from Figshare <https://figshare.com/articles/dataset/Ecopath_with_Ecosim_EwE_adjacency_matrices/14439242> (Giacomuzzo and Jordán 2021).

References

Agresti, A. 2012. Analysis of ordinal categorical data: second edition. – Wiley. doi: 10.1002/9780470594001
Allesina, S. and Pascual, M. 2009. Food web models: a plea for groups. – Ecol. Lett. 12: 652–662.
Angelini, R. and Agostinho, A. A. 2005. Food web model of the Upper Paraná River Floodplain: description and aggregation effects. – Ecol. Model. 181: 109–121.
Arenas, A. et al. 2007. Size reduction of complex networks preserving modularity. – New J. Phys. 9: 176.
Baskerville, E. B. et al. 2011. Spatial guilds in the Serengeti food web revealed by a bayesian group model. – PLoS Comput. Biol. 7: e1002321.
Bauer, B. et al. 2010. Node centrality indices in food webs: rank orders versus distributions. – Ecol. Complex. 7: 471–477.
Blondel, V. D. et al. 2008. Fast unfolding of communities in large networks. – J. Stat. Mech. Theory Exp. 10: P10008.
Borgatti, S. P. and Everett, M. G. 1993. Two algorithms for computing regular equivalence. – Soc. Netw. 15: 361–376.
Cirtwill, A. R. et al. 2018. A review of species role concepts in food webs. – Food Webs 16: e00093.
Colléter, M. et al. 2013. EcoBase: a repository solution to gather and communicate information from EwE models. – Report no. 21. Fisheries Centre, Univ. of British Columbia, Canada. doi: 10.14288/1.0354309.
Csardi, G. and Nepusz, T. 2006. The igraph software package for complex network research. – InterJ. Complex Syst. 1695(5): 1–9.
Curtsdotter, A. et al. 2011. Robustness to secondary extinctions: comparing trait-based sequential deletions in static and dynamic food webs. – Basic Appl. Ecol. 12: 571–580.
Curtsdotter, A. et al. 2018. Ecosystem function in predator–prey food webs – confronting dynamic models with empirical data. – J. Anim. Ecol. 88: 196–210.
Cury, P. et al. 2000. Small pelagics in upwelling systems: patterns of interaction and structural changes in 'wasp-waist' ecosystems. – ICES J. Mar. Sci. 57: 603–618.
DiCiccio, T. J. and Efron, B. 1996. Bootstrap confidence intervals. – Stat. Sci. 11: 189–228.
Doulcier, G. and Stouffer, D. B. 2015. Rnetcarto: fast network modularity and roles computation by simulated annealing. – R package ver. 0.2, 4. <https://cran.r-project.org/web/packages/rnetcarto/index.html>.
Eklöf, A. et al. 2012. Relevance of evolutionary history for food web structure. – Proc. R. Soc. B 279: 1588–1596.
Endrédi, A. et al. 2018. Food web dynamics in trophic hierarchies. – Ecol. Model. 368: 94–103.
Essington, T. E. and Plagányi, É. E. 2014. Pitfalls and guidelines for 'recycling' models for ecosystem-based fisheries management: evaluating model suitability for forage fish fisheries. – ICES J. Mar. Sci. 71: 118–127.
Estrada, E. 2007. Characterization of topological keystone species. Local, global and 'meso-scale' centralities in food webs. – Ecol. Complex. 4: 48–57.

Fisher, R. A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. – Biometrika 10: 507–521.
Fornito, A. et al. 2016. Fundamentals of brain network analysis. – Academic Press.
Gauzens, B. et al. 2013. Food-web aggregation, methodological and functional issues. – Oikos 122: 1606–1615.
Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. – In: Computing science and statistics. Proc. 23rd Symp. on the Interface, pp. 156–163.
Giacomuzzo, E. and Jordán, F. 2021. Data from: Food web aggregation: effects on key positions. – Figshare Digital Repository, <https://figshare.com/articles/dataset/Ecopath_with_Ecosim_EwE_adjacency_matrices/14439242>.
Gouveia, C. et al. 2021. Combining centrality indices: maximizing the predictability of keystone species in food webs. – Ecol. Indic. 126: 107617.
Guimerà, R. et al. 2007. Module identification in bipartite and directed networks. – Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 76: 1–8.
Guimerà, R. et al. 2010. Origin of compartmentalization in food webs. – Ecology 91: 2941–2951.
Harary, F. 1959. Status and contrastatus. – Sociometry 22: 23.
Harary, F. 1961. Who eats whom? – General Syst. 6: 41–44.
Heymans, J. J. et al. 2014. Global patterns in ecological indicators of marine food webs: a modelling approach. – PLoS One 9: e95845.
Jaccard, P. 1912. The distribution of the flora in the alpine zone. – New Phytol. 11: 37–50.
Johnson, G. A. et al. 2009. The effects of aggregation on the performance of the inverse method and indicators of network analysis. – Ecol. Model. 220: 3448–3464.
Jordán, F. et al. 1999. A reliability theoretical quest for keystones. – Oikos 86: 453–462.
Jordán, F. et al. 2003. Quantifying the importance of species and their interactions in a host–parasitoid community. – Commun. Ecol. 4: 79–88.
Jordán, F. et al. 2006. Topological keystone species: measures of positional importance in food webs. – Oikos 112: 535–546.
Jordán, F. et al. 2007. Quantifying positional importance in food webs: a comparison of centrality indices. – Ecol. Model. 205: 270–275.
Jordán, F. et al. 2009. Trophic field overlap: a new approach to quantify keystone species. – Ecol. Model. 220: 2899–2907.
Kirkpatrick, S. et al. 1983. Optimization by simulated annealing. – Science 220: 671–680.
Kortsch, S. et al. 2015. Climate change alters the structure of arctic marine food webs due to poleward shifts of boreal generalists. – Proc. R. Soc. B 282: 20151546.
Kozen, D. and Timme, M. 2007. Indefinite summation and the Kronecker delta. – Ecommons.Cornell.Edu.
Lai, S. M. et al. 2015. A trophic overlap-based measure for species uniqueness in ecological networks. – Ecol. Model. 299: 95–101.
Libralato, S. et al. 2006. A method for identifying keystone species in food web models. – Ecol. Model. 195: 153–171.
Luczkovich, J. J. et al. 2003. Defining and measuring trophic role similarity in food webs using regular equivalence. – J. Theor. Biol. 220: 303–321.
Malliaros, F. D. and Vazirgiannis, M. 2013. Clustering and community detection in directed networks: a survey. – Phys. Rep. 533: 95–142.
Martinez, N. D. 1991. Artifacts or attributes? Effects of resolution on the little rock lake food web. – Ecol. Monogr. 61: 367–392.
Mathworks Inc. 2019. Matlab statistics and machine learning toolbox. – <www.mathworks.com/products/statistics.html>.
Michalska-Smith, M. J. et al. 2018. Understanding the role of parasites in food webs using the group model. – J. Anim. Ecol. 87: 790–800.
Mills, L. S. et al. 1993. The keystone-species concept in ecology and conservation. – BioScience 43: 219–224.
Müller, C. B. et al. 1999. The structure of an aphid–parasitoid community. – J. Anim. Ecol. 68: 346–370.
Newman, M. E. J. 2004. Fast algorithm for detecting community structure in networks. – Phys. Rev. E Stat. Phys. Plasmas Fluids Related Interdiscip. Top. 69: 066133.
Okey, T. A. 2004a. Shifted community states in four marine ecosystems: some potential mechanisms. – Univ. of British Columbia, Canada.
Okey, T. A. 2004b. Simulating community effects of sea floor shading by plankton blooms over the West Florida Shelf – Ecol. Model. 172: 339–359.
Olivier, P. and Planque, B. 2017. Complexity and structural properties of food webs in the Barents Sea. – Oikos 126: 1339–1346.
Olmo Gilabert, R. et al. 2019. Body size and mobility explain species centralities in the Gulf of California food web. – Commun. Ecol. 20: 149–160.
Plagányi, É. E. and Essington, T. E. 2014. When the SURFs up, forage fish are key. – Fish. Res. 159: 68–74.
Sander, E. L. et al. 2015. What can interaction webs tell us about species roles? – PLoS Comput. Biol. 11(7): e1004330.
Sugihara, G. et al. 1989. Scale invariance in food web properties. – Science 245: 48–52.
Sugihara, G. et al. 1997. Effects of taxonomic and trophic aggregation on food web properties. – Oecologia 112: 272–284.
Traag, V. A. et al. 2019. From Louvain to Leiden: guaranteeing well-connected communities. – Sci. Rep. 9: 1–12.
Wasserman, S. and Faust, K. 1994. Social network analysis: methods and applications. – Cambridge Univ. Press.
Yildirim, I. 2012. Bayesian inference: gibbs sampling. – Technical Note, Univ. of Rochester. Retrieved from <www.mit.edu/~ilkery/papers/GibbsSampling.pdf>.
Yodzis, P. 1988. The indeterminacy of ecological interactions as perceived through perturbation experiments. – Ecology. 69(2): 508–515.
Yodzis, P. and Winemiller, K. O. 1999. In search of operational trophospecies in a tropical aquatic food web. – Oikos 87: 327–340.