Food web aggregation: effects on key positions

Emanuele Giacomuzzo[1], Ferenc Jordán[2,3]

[1]Centre for Ecological Research, Budapest, Hungary

[2]Democracy Institute, Central European University, Budapest, Hungary

[3]Stazione Zoologica Anton Dohrn, Napoli, Italy

Abstract

Providing standard definitions of what should be considered as a node in food webs is still an unsolved problem. Especially for comparative and predictive food web modelling, a more systematic understanding is needed for the effects of trophic aggregation procedures. Aggregation is unavoidable during data management. Therefore, it is crucial to know whether food web properties are conserved during this process.

Here, we study how different aggregation methods change the positional importance of species in food webs. In particular, we investigated the effects of various aggregation algorithms on 24 indices of importance. Our work was carried out on 76 aquatic food webs coming from the Ecopath with Ecosim database (EcoBase). We considered six main types of aggregation, according to the way that the nodes were clustered. These were (i) hierarchical clustering based on the Jaccard index, (ii) hierarchical clustering based on regular equivalence (REGE), (iii) maximisation of directed modularity, (iv) maximisation of modularity according to modules in which species fed on the same preys, (v) maximisation of modularity according to modules in which species are fed upon by the same predators, and (vi) clustering through the group model.

Hierarchical clustering based on the Jaccard index and REGE index outperformed the other four methods on maintaining the relative importance of species for all the indices of importance (except for the contrastatus index, *s'*). The choice between these two methods

should follow our research question and the importance index we are interested in studying. The other four aggregation methods change more the centrality of species, especially the one based on maximising directed modularity. When using these aggregation algorithms, one has to keep in mind that the network will not only be smaller but also provides different information.

Keywords: keystone species, centrality indices, trophic role, ecological networks, data aggregation, trophospecies

Introduction

Trophic data management is something that ecologists always must deal with when working with food webs. Trophic interactions can be described among individuals, life stages, species, higher taxa, functional groups, and several other, appropriately defined nodes of food webs. Some kind of aggregation is unavoidable, and even the most highly resolved food webs contain big aggregates (e.g., “cyanobacteria", see D’Alelio, Libralato, Wyatt, & Ribera D’Alcalà (2016)). At the same time, even the least resolved food webs may contain species (e.g., “hake”, see Yodzis (1998)). Data aggregation can happen also at later stages, during data analysis, especially in large networks, where the study of hundreds of nodes would be unfeasible (Yodzis & Winemiller, 1999).

Data aggregation methods are problem dependent. Not considering this can bias the way by which we interpret the results of food web models (Hall & Raffaelli, 1993; Paine, 1988). For instance, various levels of aggregation at different trophic levels might bias our interpretation if we are trying to characterise the structure of a network (Yodzis & Winemiller, 1999). Both low- and high-resolution networks can be useful or useless, the key challenge is to properly match the problem, the data management, and the model construction. Even if this seems like a ubiquitous problem in food web ecology, standards for whether and how to aggregate data in a meaningful way does not exist yet.

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, two fishes from the same genus might be aggregated into a node of the genus (e.g., *Poecilia sphenops* and *Poecilia reticulata* could be aggregated into *Poecilia*). Similarity can be understood mathematically (equivalent network positions) and biologically (similar trophic habits). Yodzis & Winemiller (1999) and Luczkovich et al. (2003) tried to find nodes in equivalent

network positions by borrowing two definitions from social networks. Yodzis & Winemiller (1999) borrowed the concept of structural equivalence – where two nodes are similar when sharing a high number of neighbours – and called the aggregation of structurally equivalent species" tropho-species". Luczkovich et al. (2003) borrowed the concept of regular equivalence – where two nodes are similar when sharing a high number of similar but not necessarily the same neighbours. Nodes belonging to the same equivalence class are said to be sharing the same trophic role.

Groups of nodes that have different neighbours but form dense subgraphs are called modules. Species in food web modules can play different roles (e.g., predator and prey), but they maintain well-defined multispecies processes (e.g., connecting benthic and pelagic organisms). Aggregating the modules of a food web has been suggested already by Allesina & Pascual (2009). The two most reliable ways of finding modules in food webs are through the group model and modularity maximisation. The group model was firstly developed by Allesina & Pascual (2009) and then made more computationally efficient by Sander, Wootton, & Allesina (2015). Modularity maximisation was applied to food webs by Guimerà et al. (2010) following three definitions of modularity. The first one, which we will refer to as density-based modularity, is the degree to which nodes inside modules interact more among themselves than with nodes of other modules. The second one, which we call prey-based modularity, is the degree to which nodes inside modules tend to interact with the same predators. The third one, which we gave the name of predator-based modularity, is the degree to which nodes inside modules tend to interact with the same preys.

The positional importance of species differs in both highly aggregated and highly resolved networks. Central positions may be a proxy for functional importance and the community-wide distribution of either centrality values (Bauer, Jordán, & Podani, 2010) or hypothetical importance values (Mills, Doak, & Soulé, 1993) provide macroscopic

descriptors of ecosystems. Here, we investigate how different aggregation methods maintain the relative importance of species according to 24 of the most used indices of importance 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). By having been constructed with the same methodology (see Okey (2004)), they provided us with comparable results. See supporting information for a list of these food webs.

Material and methods

Clustering techniques

To cluster similar nodes, we used the following six clustering techniques.

Hierarchical clustering with Jaccard index

As a first clustering method, we clustered structurally equivalent nodes as in Yodzis & Winemiller (1999). We used the Jaccard similarity index (Jaccard, 1912) as a measure of structural equivalence. See supporting information for the clustering algorithm.

Hierarchical clustering with REGE index

Our second clustering method consisted of clustering regularly equivalent nodes as in Luczkovich et al., (2003). The measure of regular equivalence we used was the REGE index (Borgatti & Everett, 1993). See supporting information for the clustering algorithm.

Clustering of 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 the ones expected by chance. For directed networks, it can be expressed through the following equation of Arenas, Duch, Fernández, & Gómez, (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 modularity of the network, 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 & Timme, 2007).

The number and composition of the modules were found by using the Leiden algorithm of Traag, Waltman, & van Eck (2019). This algorithm is an extension of the Louvain algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). The latter is one of the best performing and fastest for community detection (Traag et al., 2019). However, it tends to produce communities that are arbitrarily poorly connected to each other and sometimes even disconnected. The Leiden algorithm not only solves this problem by producing better connected communities, but it is also faster. The code that we used was implemented in the igraph package (Csardi & Nepusz, 2006) for the statistical software R (R Development Core Team, 2013).

Clustering of prey-based and predator-based modules

As the fourth and fifth clustering methods, we clustered the nodes of every module that was found by maximising the prey modularity and the predator modularity of the food web, as in Guimerà et al. (2010). In this case, the modularity of the food web is expressed as to how much 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 (Roger Guimerà, Sales-Pardo, & Amaral, 2007) for prey modularity

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

or in the following one for predator modularity

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

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 by using the rnetcarto package (Doulcier & Stouffer, 2015) implemented in R. This finds the community structure of the network by using the optimisation method of simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983).

Clustering of groups

As a sixth clustering method, we clustered the nodes inside the modules found by the group model of Allesina & 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 retrieving the food web is:

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

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.

Because of the high number of possible module arrangements, it is not possible to explore them all. To find the best possible solution that our computation power allows us to find, we used the algorithm of Sander, Wootton, & Allesina, 2015. This 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 as a Markov chain Monte Carlo (MCMC) with multiple chains running all at once (Sander et al., 2015).

Connecting the clusters and assigning interaction strength

The connection of the clusters followed a similar approach to the one described in Martinez (1991). We used five methods to decide whether there was a link between two clusters. The

first one produces the maximum connectance and is known as maximum linkage method (NMAX). Here, a cluster has a connection to another cluster if it has at least one link going from one of its nodes to the nodes of the second cluster. The second one produces the minimum connectance and is known as minimum linkage method (NMIN). In this case, a cluster is connected to another only if all its nodes have a connection to all the nodes of the other cluster. The other three methods produce an intermediate connectance. They consider a link from a cluster to another only if at least 25%, 50%, or 75% of possible connections from the first cluster to the second are realised.

The weight of the link was then calculated in four different ways: as the minimum weight, the maximum weight, the mean weight, and the sum of the weights of the links going from the members of the first cluster to the ones of the second cluster.

Indices of importance

For each food web, we calculated the indices of importance before and after the aggregation. The indices of importance of a node after the aggregation was defined as the one of its cluster. Let's 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 food web and 8 in the aggregated food web. The importance indices we used belonged to the following families.

Degree centrality (DC)

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

$$DC_i = \sum_{j=1}^{n} A_{ij} \quad (5)$$

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 in 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, Zalesky, & Bullmore, 2016)

$$wDC_i = \sum_{j=1}^{n} A_{ij} \quad (6)$$

Closeness centrality (CC)

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

$$CC_i = \frac{1}{\sum_{j=1}^{n} d(i,j)} \quad (7)$$

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

Betweenness centrality (BC)

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

$$BC_i = \sum_{i \neq m \neq n} \frac{\sigma_{mn}(i)}{\sigma_{mn}} \quad (8)$$

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 (s)

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, Senánszky, Libralato, & Jordán, 2018)

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

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). By following the same method but in a top-down direction we obtain the contrastatus (s_i')

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

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

$$\Delta s_i = s_i - s_i'. \quad (11)$$

The computation of the status, contrastatus and net status needs to be performed on a network without cycles. See supporting information for the algorithm used to create a directed acyclic graph (DAG).

Keystone index (K)

The keystone index was firstly introduced by Jordán, Takacs-Santa, & Molnar (1999) and 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 (Ferenc Jordán, Liu, & Davis, 2006). Unlike the status index, which only considers the distance between a node and all the other nodes, the keystone index takes into consideration 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, as the status 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

$$K(i) = K_b(i) + K_t(i) \quad (12)$$

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 is calculated as

$$K_b(i) = \sum_{j=1}^{n} \frac{1}{m(i)(j)} + \frac{K_b(j)}{m(i)(j)} \quad (13)$$

where j is a predator of i and $m(i)(j)$ is the number of preys of j. $\frac{K_b(j)}{m(i)(j)}$ is the fraction of bottom-up effects of j that are caused by i. The $K_b(j)$ of top predators is set as 0. The top-down keystone index is calculated exactly as the bottom-up keystone index, but with the direction of the links inverted.

When calculating the keystone index, we might be also interested in how much of the effect of a species on the community 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

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

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

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

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

where $K_{b,dir}$ is the directed component of the bottom-up effect, $K_{t,dir}$ is the directed component of the top-down effect, $K_{b,indir}$ is the indirected component of the bottom-up

effect, and $K_{t,indir}$ is the indirected component of the top-down effect. The directed and indirected components of the bottom-up index are calculated as

$$K_{b,dir}(i) = \sum_{j=1}^{n} \frac{1}{m(i)(j)} + \frac{1}{m(i)(j)} \quad (17)$$

$$K_{b,indir}(i) = \sum_{j=1}^{n} \frac{1}{m(i)(j)} + \frac{K_b(j)}{m(i)(j)} \quad (18)$$

The direct and indirect components of the top-down effect are calculated in the same way, but with the direction of the links inverted.

Topological importance (TI)

The topological importance (TI) of a node represents its potential to create bottom-up effects on other species, up to a certain number of steps that we can set. It was first introduced to host-parasitoid networks by Müller, Adriaanse, Belshaw, & Godfray (1999) and then to food webs by Jordán, Liu, & van Veen (2003). If topological importance takes interaction strength into consideration, we refer to it as weighted topological importance (WI). See supporting information for the algorithm for the computation of topological importance and weighted topological importance.

Species uniqueness (STO)

Species uniqueness (STO) represents how redundant the strong interactions of a node are. It was first introduced by Lai, Liu, & Jordán (2015) and can be considered as an extension of the trophic field overlap (TO) (see Jordán, Liu, & Mike (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 having to choose a single threshold, TO is calculated based on multiple thresholds. Each of these TO is then summed to give species uniqueness. If we calculate the topological overlap taking into consideration interaction strength, we can calculate weighted species uniqueness ($wSTO$). See supporting information for the algorithm for the computation of trophic field overlap.

Statistical analysis

The combination of the six clustering techniques, five linkage methods and four ways of determining interaction strength produced 120 ways of aggregating food webs. For each of these aggregation methods, we studied their effects on the indices of importance. In particular, we studied the correlation between the ranking of the nodes before and after the aggregation. This correlation was calculated by 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. This required us to convert τ_B using the Fisher z-transformation (Fisher, 1915). For each fisher's z mean, we found its 95% confidence interval by bootstrapping (DiCiccio & Efron, 1996). The fisher's z means, and 95% confidence intervals were then back transformed to τ_B. 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 (Figure 1). The median size of the aggregated food web compared to the original one was 74.5% (IQR=10.8%) for the Jaccard index, 73% (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 (Figure 2 and 6).

Correlation of indices of importance before and after the aggregation

The correlation between the ranking before and after the aggregation can be seen in Figure 3. By focusing only on the clustering method and ignoring the linkage method and the interaction strength method, we can select the best clustering for each combination of centrality indices and clustering methods. This provides us with a clearer heat map. See Figure 4. Ranking the clustering algorithms in Figure 4 produces Table 1. Density modularity always ranked as the worst clustering algorithm. Prey-based modules and group model were always 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 Jaccard index and REGE index were always 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. Status index and keystone index were maintained better either by Jaccard or by REGE according to which index of those two families we considered. 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

The aggregation of nodes in food webs is an old problem, dating back at least to the beginning of the 1980s (e.g., Pimm (1982)). Multiple studies investigated the effect of different types of aggregation on food web properties, such as connectance, food chain length and ratio between bottom, intermediate and top species (Hall & Raffaelli, 1991; Martinez, 1991, 1993; Sugihara, Bersier, & Schoenly, 1997; Sugihara, Schoenly, & Trombla, 1989). Not much attention, however, has been given to the effects that aggregation has on our ability to find keystone species inside the food web. Essington & Plagányi, (2014) and Plagányi & Essington, (2014) tried to investigate this topic. The former by studying changes in degree centrality and in the ratio between the consumer species biomass and the total consumer biomass. The latter by the SURF index (an index used to find important forage fish species). These, however, focused on indices that are rarely used in keystone species research and did not compare the effect of different aggregation methods.

In this paper, we study the effects of aggregation on several indices used in network ecology (Estrada, 2007; Olmo Gilabert et al., 2019). The aggregation methods that we compared were the widely used hierarchical clustering using Jaccard index and REGE index, as well as directed modularity maximisation, prey-based modularity maximisation, predator-based modularity maximisation, and clustering through the group model. The latter four are normally used for community detection, but we explored the possibility of using them in the future for data aggregation.

Our results show that different aggregation methods maintain the relative importance of species in different ways. Therefore, they have different potential of changing the keystone species of the food web (Figure 5). Except for the contrastatus index (s'), hierarchical clustering with the Jaccard index and hierarchical clustering with the REGE index outperformed the other methods. When choosing between these two methods, however, we need to consider that not all indices have the same power to predict keystone species. Gouveia, Móréh, & Jordán (2021) looked at topological indices and how their findings correlated with the findings of the dynamical index keystoneness (Libralato, Christensen, & Pauly, 2006). They found that the most reliable topological index was the weighted degree centrality (wDC). This 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, as it maintains wDC and WI^5 the best.

The choice of the aggregation algorithm, however, boils down to our research question. The particular index we are interested in might drive the choice between the Jaccard index and the REGE index. Another factor influencing our choice is the resulting network size. How much do we need to aggregate? In case we might want to have higher aggregation, we might consider using the group model, which produces high aggregation, but performs way better than density-based modularity (Figure 6). But also, we might be interest in a particular species role – in that case we should aggregate according to that specific role. Each method reveals some kind of biological similarities between nodes. The key point is to know the effects of these aggregation procedures and to keep them in mind when applying them and evaluating the properties of the aggregated network.

In this paper, we focus on a well-defined set of food webs that are methodologically strictly comparable (all created by the EwE methodology). Studying aggregation would be more exciting for larger food webs 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 major 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 major step towards predictive food web modelling.

In conclusion, we have shown that different aggregation methods maintain differently the relative importance of species in a food web. Hierarchical clustering with Jaccard index and hierarchical clustering with REGE index have been shown to be the best at doing this. The choice between these two algorithms should depend upon the type of importance index we are interested in maintaining. Future research should be carried out on larger food webs, dynamical indices and determining what the best linkage method and new interaction strength method are.

Supplementary material

The code used for this analysis is available at https://github.com/Emanuele-Giacomuzzo/Data_aggregation. This work resulted also in the MATLAB toolbox "Food Web Analysis Toolbox" available at https://uk.mathworks.com/matlabcentral/fileexchange/89907-food-web-analysis-toolbox.

References

Agresti, A. (2012). Analysis of Ordinal Categorical Data: Second Edition. In *Analysis of*

Ordinal Categorical Data: Second Edition. https://doi.org/10.1002/9780470594001

Allesina, S., & Pascual, M. (2009). Food web models: A plea for groups. *Ecology Letters*, *12*(7), 652–662. https://doi.org/10.1111/j.1461-0248.2009.01321.x

Arenas, A., Duch, J., Fernández, A., & Gómez, S. (2007). Size reduction of complex networks preserving modularity. *New Journal of Physics*, *9*(6), 176. https://doi.org/10.1088/1367-2630/9/6/176

Bauer, B., Jordán, F., & Podani, J. (2010). Node centrality indices in food webs: Rank orders versus distributions. *Ecological Complexity*, *7*(4), 471–477. https://doi.org/10.1016/j.ecocom.2009.11.006

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borgatti, S. P., & Everett, M. G. (1993). Two algorithms for computing regular equivalence. *Social Networks*, *15*(4), 361–376. https://doi.org/10.1016/0378-8733(93)90012-A

Colléter, M., Valls, A., Guitton, J., Morissette, L., Arreguín-Sánchez, F., Christensen, V., … Pauly, D. (2013). EcoBase: A repository solution to gather and communicate information from EwE models. In *Fisheries Centre Research Reports*.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*.

D'Alelio, D., Libralato, S., Wyatt, T., & Ribera D'Alcalà, M. (2016). Ecological-network models link diversity, structure and function in the plankton food-web. *Scientific Reports*, *6*(1), 1–13. https://doi.org/10.1038/srep21806

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–228. https://doi.org/10.1214/ss/1032280214

Doulcier, G., & Stouffer, D. B. (2015). *Rnetcarto: Fast network modularity and roles*

computation by simulated annealing.

Endrédi, A., Senánszky, V., Libralato, S., & Jordán, F. (2018). Food web dynamics in trophic hierarchies. *Ecological Modelling*, (368), 94–103. https://doi.org/10.1016/j.ecolmodel.2017.11.015

Essington, T. E., & Plagányi, É. E. (2014). Pitfalls and guidelines for "recycling" models for ecosystem-based fisheries management: Evaluating model suitability for forage fish fisheries. *ICES Journal of Marine Science*, *71*(1). https://doi.org/10.1093/icesjms/fst047

Estrada, E. (2007). Characterization of topological keystone species. Local, global and "meso-scale" centralities in food webs. *Ecological Complexity*, *4*(1–2), 48–57. https://doi.org/10.1016/j.ecocom.2007.02.018

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, *10*(4), 507–521. https://doi.org/10.2307/2331838

Fornito, A., Zalesky, A., & Bullmore, E. T. (2016). Fundamentals of Brain Network Analysis. In *Fundamentals of Brain Network Analysis*. https://doi.org/10.1016/C2012-0-06036-X

Geyer, C. J. (1991). *Markov chain Monte Carlo maximum likelihood*.

Gouveia, C., Móréh, Á., & Jordán, F. (2021). Combining centrality indices: maximizing the predictability of keystone species in food webs. *Ecological Indicators*, *126*, 107617. https://doi.org/10.1016/j.ecolind.2021.107617

Guimerà, R., Stouffer, D. B., Sales-Pardo, M., Leicht, E. A., Newman, M. E. J., & Amaral, L. A. N. (2010). Origin of compartmentalization in food webs. *Ecology*, *91*(10), 2941–2951. https://doi.org/10.1890/09-1175.1

Guimerà, Roger, Sales-Pardo, M., & Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review E - Statistical, Nonlinear, and Soft*

Matter Physics, *76*(3), 1–8. https://doi.org/10.1103/PhysRevE.76.036102

Hall, S. J., & Raffaelli, D. (1991). Food-Web Patterns: Lessons from a Species-Rich Web. *The Journal of Animal Ecology*, *60*(3), 823–841. https://doi.org/10.2307/5416

Hall, S. J., & Raffaelli, D. G. (1993). Food Webs: Theory and Reality. In *Advances in Ecological Research* (Vol. 24). https://doi.org/10.1016/S0065-2504(08)60043-4

Harary, F. (1961). Who eats whom? *General Systems*, *6*, 41–44.

Harary, Frank. (1959). Status and Contrastatus. *Sociometry*, *22*(1), 23. https://doi.org/10.2307/2785610

Jaccard P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37–50.

Jordán, F., Liu, W. C., & van Veen, F. J. F. (2003). Quantifying the importance of species and their interactions in a host-parasitoid community. *Community Ecology*, *4*(1), 79–88. https://doi.org/10.1556/ComEc.4.2003.1.12

Jordán, Ferenc, Liu, W.-C., & Davis, A. J. (2006). Topological keystone species: measures of positional importance in food webs. *Oikos*, *112*(3), 535–546.

Jordán, Ferenc, Liu, W. C., & Mike, Á. (2009). Trophic field overlap: A new approach to quantify keystone species. *Ecological Modelling*, *220*(21), 2899–2907. https://doi.org/10.1016/j.ecolmodel.2008.12.003

Jordán, Ferenc, Takács-Sánta, A., & Molnár, I. (1999). A Reliability Theoretical Quest for Keystones. *Oikos*, *86*(3), 453. https://doi.org/10.2307/3546650

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Kozen, D., & Timme, M. (2007). Indefinite summation and the Kronecker delta. *Ecommons.Cornell.Edu*.

Lai, S. M., Liu, W. C., & Jordán, F. (2015). A trophic overlap-based measure for species

uniqueness in ecological networks. *Ecological Modelling*, *299*, 95–101. https://doi.org/10.1016/j.ecolmodel.2014.12.014

Libralato, S., Christensen, V., & Pauly, D. (2006). A method for identifying keystone species in food web models. *Ecological Modelling*, *195*(3–4), 153–171. https://doi.org/10.1016/j.ecolmodel.2005.11.029

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. *Journal of Theoretical Biology*, *220*(3), 303–321. https://doi.org/10.1006/jtbi.2003.3147

Martinez, N. D. (1991). Artifacts or Attributes ? Effects of Resolution on the Little Rock Lake Food Web. *Ecological Monographs*, *61*(4), 367–392.

Martinez, N. D. (1993). Effects of Resolution on Food Web Structure. *Oikos*, *66*(3), 403–412.

Mathworks Inc. (2019). Matlab Statistics and Machine Learning Toolbox. Retrieved from 2015 website: https://www.mathworks.com/products/statistics.html

Mills, L. S., Doak, M. E., & Soulé, D. F. (1993). The keystone-species concept in ecology and conservation. *BioScience*, *43*(4), 219–224.

Müller, C. B., Adriaanse, I. C. T., Belshaw, R., & Godfray, H. C. J. (1999). The structure of an aphid-parasitoid community. *Journal of Animal Ecology*, *68*, 346–370. https://doi.org/10.1046/j.1365-2656.1999.00288.x

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *69*(6), 5. https://doi.org/10.1103/PhysRevE.69.066133

Okey, Thomas A. (2004). *Simulating community effects of sea floor shading by plankton blooms over the West Florida Shelf*.

Okey, Thomas Anthony. (2004). *Shifted community states in four marine ecosystems: some potential mechanisms*. University of British Columbia, Canada.

Olmo Gilabert, R., Navia, A. F., De La Cruz-Agüero, G., Molinero, J. C., Sommer, U., & Scotti, M. (2019). Body size and mobility explain species centralities in the Gulf of California food web. *Community Ecology*, *20*(2), 149–160. https://doi.org/10.1556/168.2019.20.2.5

Paine, R. T. (1988). Road Maps of Interactions or Grist for Theoretical Development? *Ecology*, *69*(6), 1648–1654. https://doi.org/10.2307/1941141

Pimm, S. L. (1982). Food webs. In *Food webs*. London: Springer.

Plagányi, É. E., & Essington, T. E. (2014). When the SURFs up, forage fish are key. *Fisheries Research*, *159*, 68–74. https://doi.org/10.1016/j.fishres.2014.05.011

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7

Sander, E. L., Wootton, J. T., & Allesina, S. (2015). What Can Interaction Webs Tell Us About Species Roles? *PLoS Computational Biology*, *11*(7), 1–22. https://doi.org/10.1371/journal.pcbi.1004330

Sugihara, G., Bersier, L. F., & Schoenly, K. (1997). Effects of taxonomic and trophic aggregation on food web properties. *Oecologia*, *112*(2), 272–284. https://doi.org/10.1007/s004420050310

Sugihara, G., Schoenly, K., & Trombla, A. (1989). Scale invariance in food web properties. *Science*, *245*(4913), 48–52. https://doi.org/10.1126/science.2740915

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-41695-z

Wasserman, S., & Faust, K. (1994). *Social network analysis : methods and applications*. Cambridge University Press.

Yildirim, I. (2012). Bayesian Inference: Gibbs Sampling. *Technical Note, University of*

Rochester. Retrieved from http://www.mit.edu/~ilkery/papers/GibbsSampling.pdf

Yodzis, P. (1998). Local trophodynamics and the interaction of marine mammals and fisheries in the Benguela ecosystem. *Journal of Animal Ecology*, *67*(4), 635–658. https://doi.org/10.1046/j.1365-2656.1998.00224.x

Yodzis, P., & Winemiller, K. O. (1999). In Search of Operational Trophospecies in a Tropical Aquatic Food Web. *Oikos*, *87*(2), 327–340.

Figures & table

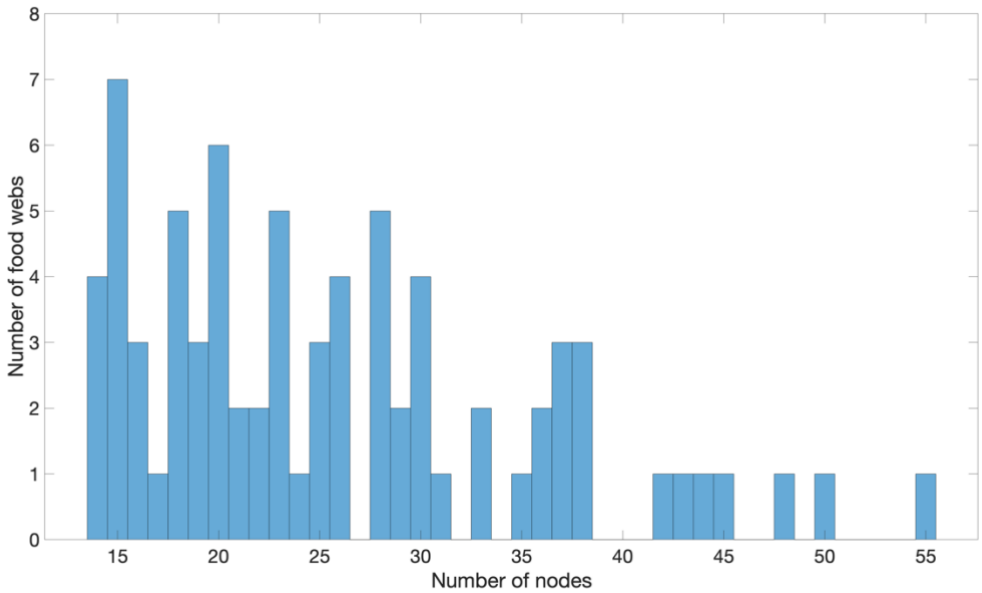

Figure 1. Size of the food webs we used in our study.

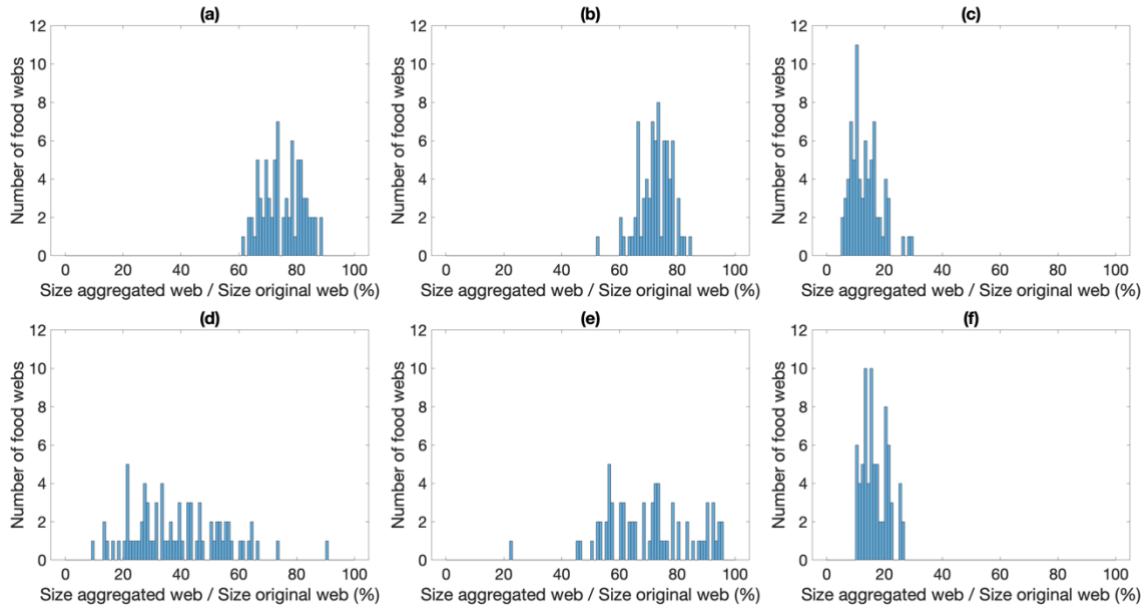

Figure 2. Size of the clusters produced by the different clustering methods. (a) = hierarchical clustering with Jaccard index, (b) = hierarchical clustering with REGE index, (c) = maximisation of density-based modularity, (d) = maximisation of prey-based modularity, (e) = maximisation of predator-based modularity, (f) = group model.

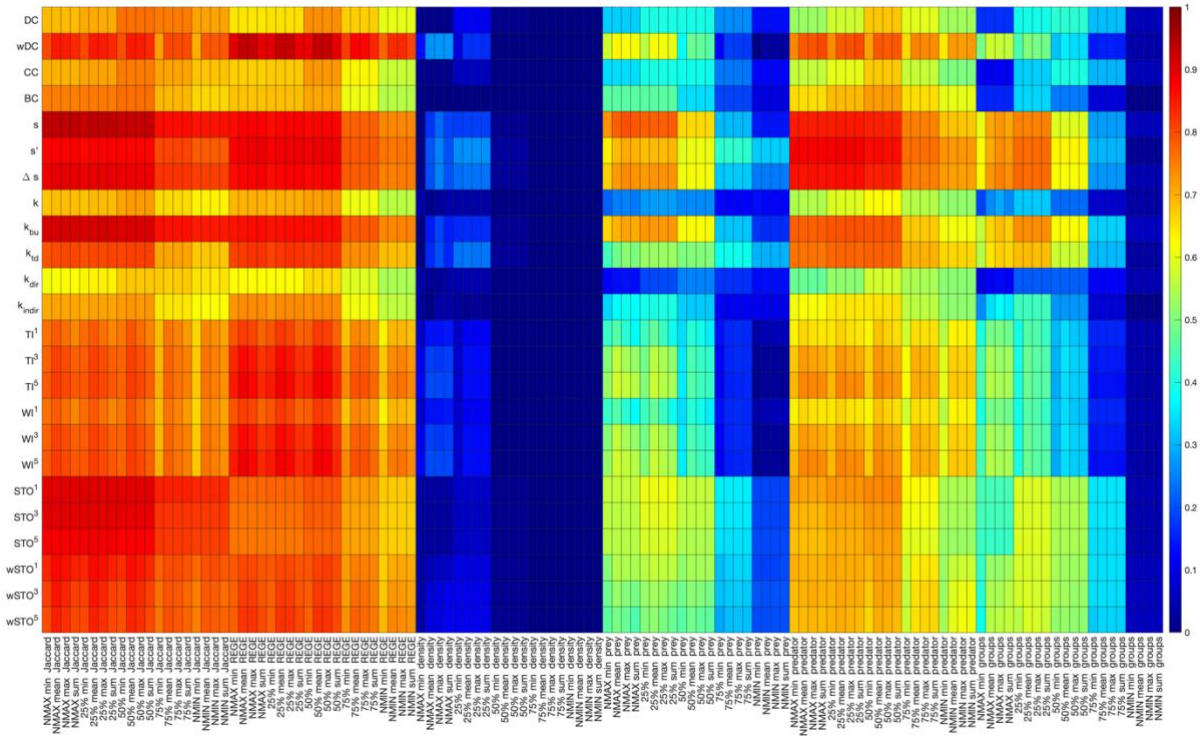

Figure 3. Kendall's rank correlation (τ) between the ranking of the nodes in the original food web and in the aggregated food web. The x-axis describes the aggregation method, which is composed of three components: (i) clustering algorithm; (ii) linkage method; and (iii) and interaction strength method. Jaccard = hierarchical clustering using Jaccard index, REGE = Hierarchical clustering using REGE index, density = clustering of density-based modules, prey = clustering of prey-based modules, predator = clustering of predator-based modules, groups = clustering of groups. NMAX = maximum linkage method, 25% = at least 25% of

links realised to consider a connection, 50% = at least 25% of links realised to consider a connection, 75% = at least 75% of links realised to consider a connection, NMIN = minimum linkage method. min = minimum interaction strength, max = maximum interaction strength, sum = sum of interaction strengths, mean = mean interaction strength. The y-axis describes the importance indices. DC = degree centrality, wDC = weighted degree centrality, CC = closeness centrality, BC = betweenness centrality, s = status index, s' = contrastatus index, Δs = net status index, k = keystone index, k_{bu} = bottom-up keystone index, k_{td} = top-down keystone index, k_{dir} = directed keystone index, k_{indir} = indirected keystone index, TI^1 = 1-step topological importance, TI^3 = 3-step topological importance, TI^5 = 5-step topological importance, WI^1 = 1-step weighted topological importance, WI^3 = 3-step weighted topological importance, WI^5 = 5-step weighted topological importance, STO^1 = 1-step species uniqueness, STO^3 = 3-step species uniqueness, STO^5 = 5-step species uniqueness, $wSTO^1$ = 1-step weighted species uniqueness, $wSTO^3$ = 3-step weighted species uniqueness, $wSTO^5$ = 5-step weighted species uniqueness.

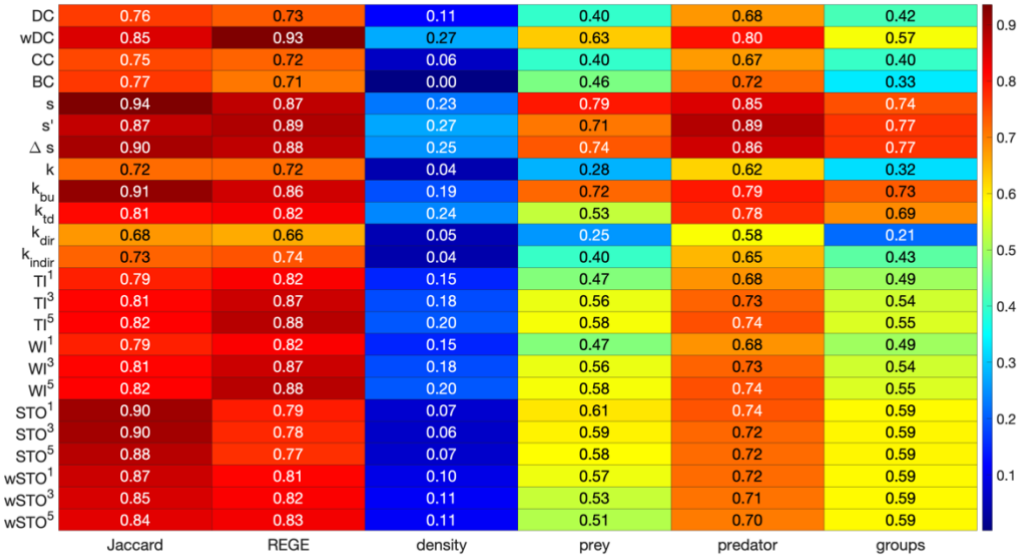

Figure 4. Heat map of the best Kendall's rank correlation coefficient for each combination of clustering methods and centrality indices. The best correlation is selected across linkage methods and methods of determining interaction strength. Jaccard = hierarchical clustering using Jaccard index, REGE = Hierarchical clustering using REGE index, density = clustering of density-based modules, prey = clustering of prey-based modules, predator = clustering of predator-based modules, groups = clustering of groups. For the confidence intervals associated with these values and for the linkage method & interaction strength method used, see supporting information.

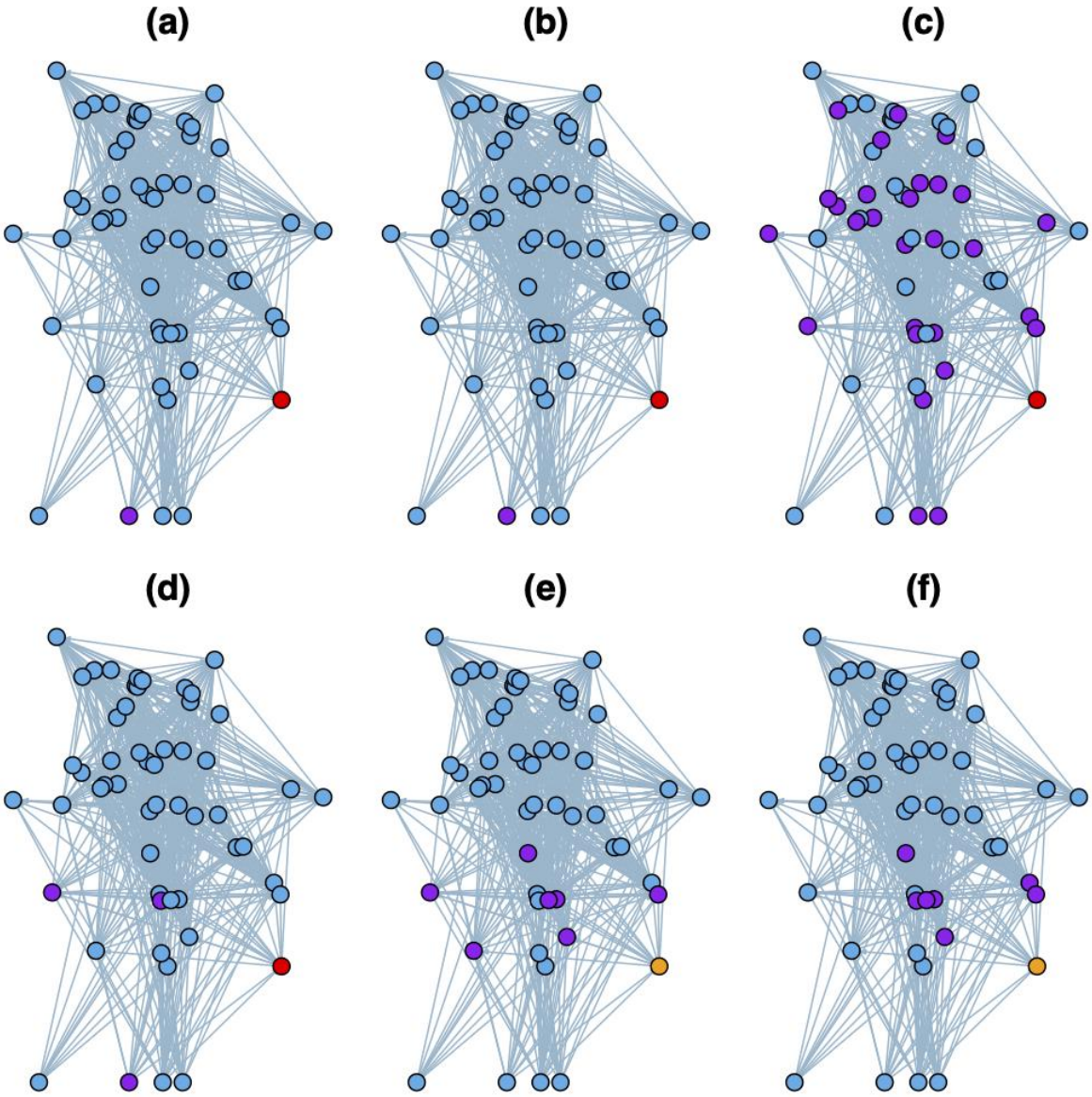

Figure 5. Effect of aggregation on most the most important nodes in a food web. Red: central before and central after. Orange: central before and not central after. Purple: not central before and central after. Blue: not central before and not central after. The self-loops are not included in the figure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

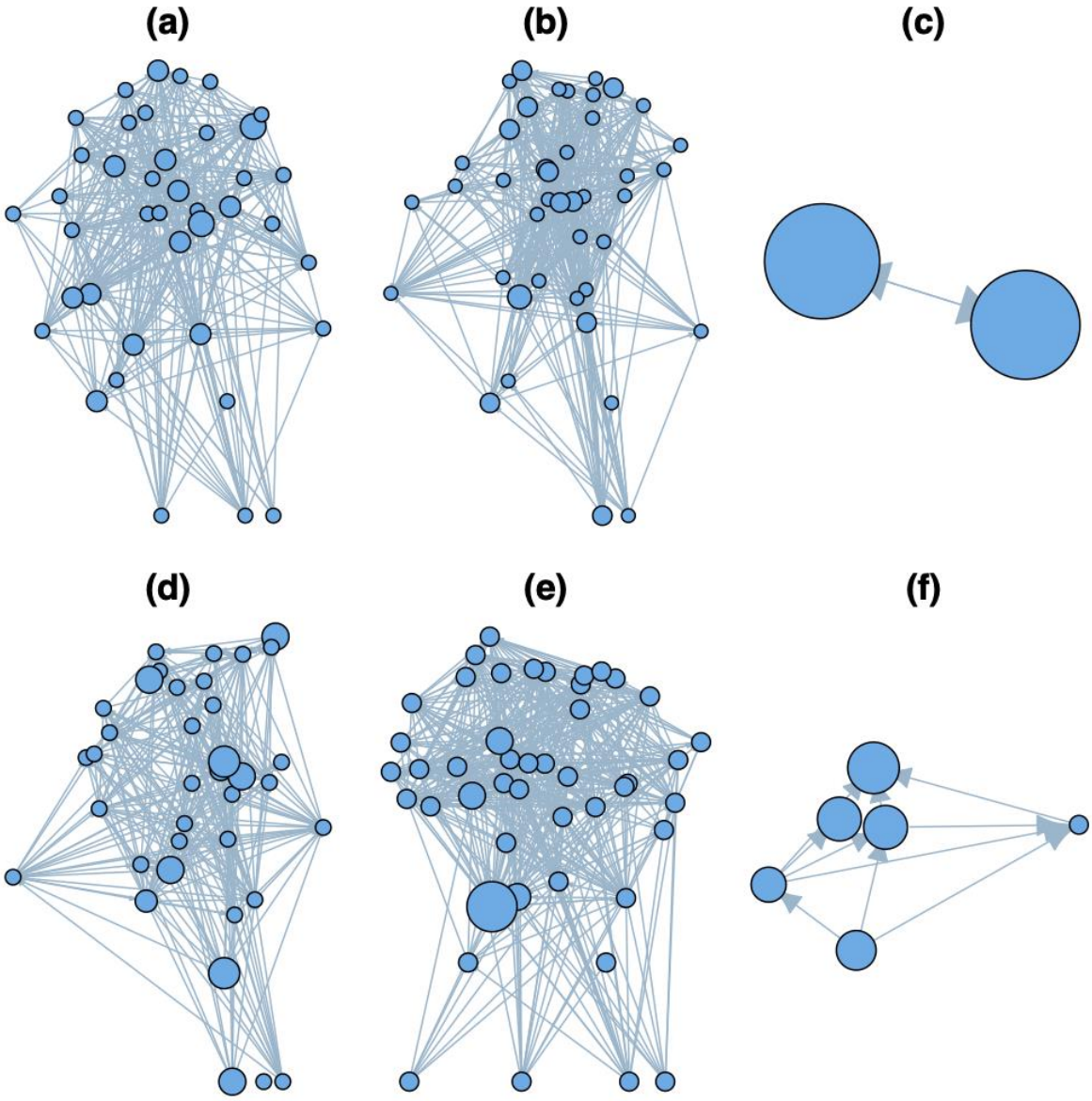

Figure 6. Aggregation of the food web of Figure 5 according to different clustering algorithms. (a) = hierarchical clustering using Jaccard index, (b) = hierarchical clustering using REGE index, (c) = clustering of density-based modules, (d) = clustering of prey-based modules, (e) = clustering of predator-based modules, (f) = clustering of 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 did not include self-loops.

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

Table 1. Best Kendall's correlation coefficients, as in Figure 4. This time, 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 = density-based modules, blue = predator-based modules, purple = groups produced by the group model. C1 = Best clustering, C2 = second best clustering, C3 = third best clustering, C4 = fourth best clustering, C5 = fifth best clustering, C6 = sixth best clustering.