Food web aggregation: effects on key positions

Emanuele Giacomuzzo[1] and Ferenc Jordàn[1,2]

[1] *Centre for Ecological Research, Budapest, Karolina 26, 1113, Hungary*
[2] *Stazione Zoologica Anton Dohrn, Napoli, 80122, Italy*

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, even the most highly resoluted food webs contain big aggregates (e.g., "bacteria", see Martinez 1991). At the same time, even the least resoluted 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 and 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 ((Paine, 1988; Hall and Raffaelli, 1993).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 and 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).

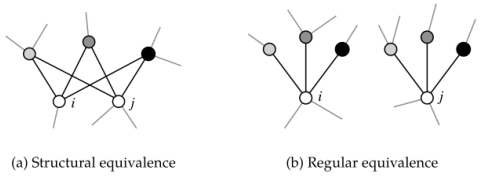
Similarity can be understood mathematically (equivalent network positions) and biologically (similar trophic habits). Yodzis and Winemiller (1999) and Luczkovich et al. (2003) tried to answer this question by borrowing two definitions from social networks. Yodzis and Winemiller (1999) borrowed the concept of structural equivalence – where two nodes are similar when sharing a high number of neighbours – and called the aggregation of structurally equivalent species" trophospecies". 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 share ecological roles.

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 multi-species processes (e.g. connecting benthic and pelagic organisms). Aggregating the modules of a food web has been suggested already by Allesina and 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 and Pascual (2009) and then extended by Sander et al. (2015). Modularity maximisation was firstly 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 by which nodes inside modules interact more among themselves than with nodes of other modules. The second one, which we will refer to as prey-based modularity, is the degree by which nodes inside modules tend to interact with the same predators. The third one, which we will refer to as predator-based modularity, is the degree by which nodes inside modules tend to interact with the same preys.

The positional importance of species differs in both highly-aggregated and highly-resoluted networks. 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.

In this paper, we investigate how these different aggregation methods maintain the relative importance of species, as a proxy of network structure. To compute the importance of species we used 15 of the most used centrality indices used in keystone species research. Our investigation was carried out on the data of the plankton food web of the Gulf of Naples (Figure 1), sampled at the Long-Term Ecological Research station MareChiara (LTER-MC) (Ribera d'Alcalà et al., 2004)). This is composed of 63 different nodes (see Table 1 of D'Alelio et al. (2016) for the species assemblage). The node number 59 had no connection to other nodes, so it had been deleted.

See Figure 1.



(a) Structural equivalence (b) Regular equivalence

Figure 1: Two types of similarity indices highly used in social networks, which have been applied also to food webs. As you can see, two nodes are regularly equivalent if they are connected to similar nodes (b) and structurally equivalent if they are connected to the same exact nodes (a). Two nodes that are structurally equivalent are also regularly equivalent, but not the other way around. For example, two nurses are regularly equivalent becuase they have the same connections to other personell in the hospital such as doctors, other nurses, receptionists, patients and so on. If the personell they are in contact not only is similar but it's the same exact persons, then they are also structurally equivalent.

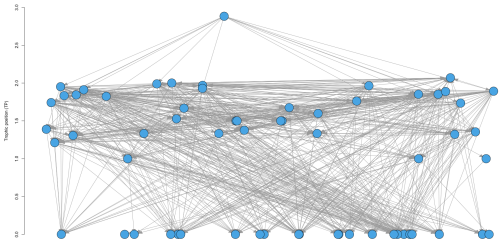


Figure 2: The plankton food web of the Gulf of Naples before the aggregation of its nodes (data from D'Alelio et al. (2016)). The self-loops have been omitted from the figure to make it clearer.

Methods: clustering techniques

Hierarchical clustering of nodes according to their Jaccard similarity index

As a first clustering method, we clustered structurally equivalent nodes as in Yodzis and Winemiller (1999), using the Jaccard similarity index as a measure of structural equivalence. The clustering algorithm can be found in the appendix.

Hierarchical clustering of nodes according to their REGE index

As a second clustering method, we clustered regularly equivalent nodes as in Luczkovich et al. (2003), using REGE index as a measure of regular equivalence. The clustering algorithm cna be found in the appendix.

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 et al. (2007), which is a generalisation of the Newman-Girvan modularity (Newman, 2004):

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

where Q is the directed modularity of partition P, L is the number of links in the network, A_{ij} is the element of the adjacency matrix of a directed, binary network (links go from j to i), $k_i^{in}k_j^{out}/L$ is the probability of having an edge between i and j, k_i^{in} is the indegree of i and k_j^{out} is the out-degree of j), m_i is the module of i, and δ is the Kronecker delta (Kozen and Timme, 2007).

The number and composition of the modules were found by using the spectral optimisation algorithm of Leicht and Newman (2008). This algorithm is based on the principle that if we keep subdividing every module into the two modules that maximise modularity the most, we reach a point when a further subdivision would not increase the modularity anymore, giving us the most accurate modules.

Clustering of prey-based and predator-based modules

As 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 (Guimerà et al., 2007) for prey-modularity

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

or in the following one for predator-modularity

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

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. For simplicity, we used the same spectral optimisation algorithm to maximise also this type of modularity. However, simulated annealing might give faster results (Guimerà et al., 2007).

Group model

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 an Erdős-Rényi random graph. For an arbitrary number of groups k, the probability of retrieving the food web is:

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

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.

Becuase of the high number of possible module arrangments, it is not possible to explore them all. To find the best possible solution our computation power allows us to find, we used the algorithm of Sander et al. (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).

Methods: connecting modules

The wiring of the food web followed a similar approach to the one describe in (Martinez, 1991): two clusters were connected only if a certain number of links between their nodes was realised. The minimum number was one, where two clusters were connected if there was also just one connection between a member of the first cluster and a member of the second cluster. The maximum number was 100%, where two clusters were connected only if there was a connection between every member within a cluster and every member of the other cluster. The numbers in between went from 1% to 99% with a growth rate of 1%. 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 between the memebers of the first and the second cluster. This produced 404 food webs for each aggregation method (101 link percentages x 4 weight methods = 404 wiring options).

For each centrality index, we compared the ranking of the nodes between the original food web and the aggregated food webs. This was done by using Kendall's tau b (τ_B), a version of the Kendall rank correlation coefficient that makes adjustments for ties (Agresti, 2012). The τ_B was calculated through the package "Mann-Kendall Tau-b with Sen's Method (enhanced)" available for MATLAB (Burkey, 2021). This allowed us to select the food web with the best link ratio and the best weight method for every combination of centrality index and aggregation (e.g., the status index for group model).

Methods: centrality indices

Degree centrality (DC)

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

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

where DC_i is the degree centrality of the node i, 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. It can be normalised by dividing it by the total number of possible connections that a node could have (Wasserman and Faust, 1994)

$$nDC_i = \frac{DC_i}{n-1} \quad (6)$$

where n-1 is the maximum number of connections the node can have (the minus one shows that a node cannot have a connection to itself).

Another type of degree centrality that we considered was the weighted degree centrality (wDC), often referred to as node strength. Its formula, as well as the formula of its normalised version, are 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)

$$WDC_i = \sum_{j=1}^{n} A_{ij} \quad (7)$$

Closeness centrality (CC)

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

$$CC_i = \frac{1}{\sum\limits_{j=1}^{n} d(i,j)} \quad (8)$$

where d(i,j) is the distance between node i and j. It can be normalised as follows (Wasserman and Faust, 1994)

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

Betweenness centrality (BC)

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

$$BC_i = \sum_{i \neq m \neq n} \frac{\sigma_{mn}(i)}{\sigma_{mn}} \tag{10}$$

where σ_{mn} is the total number of shortest paths going from s to t and $\sigma_{mn}(i)$ is the total number of these paths passing through i. It can be normalised with the following equation (Wasserman and Faust, 1994)

$$nBC_i = \frac{BC_i}{(n-1)(n-2)/2} \tag{11}$$

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 et al., 2018)

$$s_i = \sum_{j=1}^{n} d(i,j) \tag{12}$$

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 controstatus (s'_i)

$$s'_i = \sum_{j=1}^{n} d(i,j) \tag{13}$$

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

$$\Delta s_i = s_i - s'_i \tag{14}$$

Keystone index (K)

The keystone index was firstly introduced by Jordán et al. (1999) and inspired by the status index. As the net status index, the keystone index of a species i (K(i)) is calculated by considering separately the bottom-up (like the status index), as well as the top-down (like the controstatus index) effects of a node Jordán et al. (2006)

$$K(i) = K_b(i) + K_t(i) \tag{15}$$

where $K_b(i)$ is its bottom-up keystone index of species i and $K_t(i)$ the top-down keystone index of species i. 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 bottom-up effect of a certain node $(K_b(i))i$ is then calculated in the following way

$$K_b(i) = \sum_{j=1}^{n} \frac{1}{m(i)(j)} + \frac{K_b(j)}{m(i)(j)} \tag{16}$$

where j is a predator of I,$m(i)(j)$ is the number of preys of j, and $\frac{K_b(j)}{m(i)(j)}$ 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 effect of a certain node $K_t(i)$ is calculated exactly as $K_b(i)$, but with the direction of the links inverted.

The bottom-up and the top-down effects can also be split into their direct and an indirect component. The indirect component takes into consideration the bottom-up effects of the predator and direct component does not

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

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

The direct and indirect components of the top-down effect are calculated in the same way, but with the direction of the links inverted. The direct and indirect keystone indices of a node are the sum of its direct/indirect bottom-up effects and its direct/indirect top-down effects

$$K_{direct}(i) = K_{b,direct} + K_{t,direct} \tag{19}$$

$$K_{indirect}(i) = K_{b,indirect} + K_{t,indirect} \tag{20}$$

The keystone index not only is the sum of its top-down and bottom-up effects, but also the sum of its direct and indirect effects

$$K(i) = K_{dir}(i) + K_{indir}(i) \tag{21}$$

Topological importance (TI)

The topological importance 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 et al. (1999) and then to food webs by Jordán et al. (2003).

The algorithm of its computation is as follows (Jordán, 2009):

1. *Compute the one-step matrix.*

   In the one-step matrix, if the energy flows from a prey to the predator, then the effect of the prey on the predator is the reciprocal of the indegree of the predator

   $$a_{1,ij} = \frac{A_{ij}}{D_j} \tag{22}$$

2. *Compute the n-step matrices.*

   In the higher steps matrices, a node influences another node at a higher trophic level by summing the effects of every path that connects the two nodes. The effect of every path is the multiplication of the inverse of the outdegree of every node along the path. For a visual explanation see Figure 3. It can be calculated as follows

   $$A\,(n) = A^n_{(1)} \tag{23}$$

3. *Calculate topological importance*

   The topological importance of a node i (TI_i) can be calculated through the following formula

   $$TI_i = \frac{\sum_{m=1}^{N} \sum_{j=1}^{n} a_{m,ji}}{N} \tag{24}$$

   where N is the total number of steps considered, m is the step number, n is the total number of nodes, and $a_{m,ji}$ is the effect of species i on species j at m number of steps.

Topological importance can be also used for weighted networks - giving us weighted topological importance (WI) – if instead of using the degree (D) we use the weighted degree (WD) (Scotti et al., 2007)

$$a_{1,ji} = \frac{A_{ij}}{weighted\,indegree_j} \tag{25}$$

where A_{ij} is the element of the adjacency matrix of the weighted directed network.

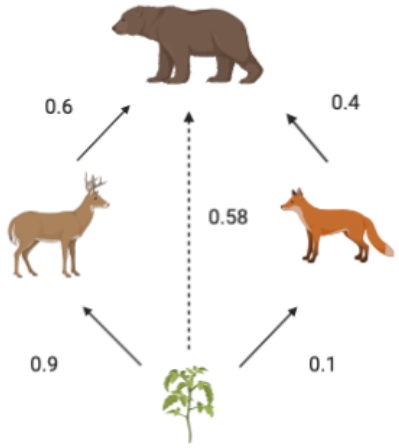


Figure 3: Topological importance (TI) of a species on another. The plant and the bear are not connected, so there is no direct effect from the plant to the bear. However, indirect effects reach the bear from the plant through two paths and the final effect is the sum of these effects. The first one is through the deer and the second is through the fox. The strength of these paths is the product of the direct effects composing the path. The first path has an effect on the bear that is 0.9*0.6=0.54, the second one has an effect on the bear that is 0.1*0.4=0.04. Summing the effects through these two 2-step paths connecting the plant with the bear, we get the 2-step effect of the plant on the bear: 0.54+0.04=0.58. Figure created with BioRender.com.

Trophic field overlap (TO)

The trophic field overlap (TO) represents how redundant the strong interactions of a node are. It was first introduced by Jordán et al. (2009). It is the number of times that it and another node interact strongly with the same predator. The algorithm for its computation is the following one (Jordán et al., 2018):

1. Compute the one-step matrix as in toplogical importance

2. Compute the n-step matrix as in topological importance

3. Compute the average effect matrix. The average effect matrix ($E(n)$) represents the effect of each node on the other nodes average by the number of steps

$$E_n = \frac{1}{n}\sum_{i=1}^{n} A_{(i)} \tag{26}$$

4. Compute the interactor matrix. Compute the so-called interactor matrix (M_T), whose values tell us whether the interaction between two nodes is weak (W) or strong (S). To do this, we need to define a threshold over which a certain interaction is strong.

5. Compute the topological overlap matrix. Compute a matrix with how many times two species interact strongly with the same predator, called the topological overlap matrix.

6. Compute the trophic field overlap (TO) The trophic field overlap (TO) of a node is calculated by summing the elements of the rows of the topological overlap matrix.

Trophic field overlap can be also used for weighted networks – giving us weighted trophic field overlap (WO) – if instead of using the degree (D) we use the weighted degree, (e.g., Xiao et al. (2019))

$$a_{1,ij} = \frac{A_{ij}}{D_j} \tag{27}$$

Finally, to avoid the bias of choosing a wrong threshold, we chose multiple thresholds and summed the TO of a species i for each of these thresholds. This gave us the species uniqueness (STO), an index that was firstly introduced by mei Lai et al. (2015).

Trophic position (TP)

The trophic position of a node is the mean length connecting it to the producers of the ecological community (its energy source). It was firstly introduced by Levine (1980), as a generalization of the earlier use of integer trophic levels to include cycles and fractional positions. It can be calculated through the following formula

$$TP_i = \sum_{k=0}^{\infty} k \cdot p_i(k). \tag{28}$$

where k is a certain path length and $p_i(k)$ is the probability that species i will reach the energy produced by the autotrophs via a path of length k. TP equals 0 for producers, it equals 1 for herbivores and larger values for omnivores and carnivores.

Results

Clustering

The result of clustering gave 39 clusters for the Jaccard index hierarchical clustering (see Figure 4), 14 clusters for the REGE index hierarchical clustering (see Figure 5), 13 clusters for the prey modularity, ... clusters for the predator modularity, 11 clusters for the density modularity and 8 clusters for the group model. When running the algorithm for the group model, we set as a random seed 587184, 100 Markov chains running at once with 100.000 Monte Carlo steps. This gave us the best partition of the food web at the step number 352.

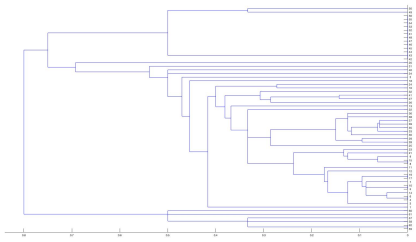


Figure 4: Dendrogram of Jaccard.

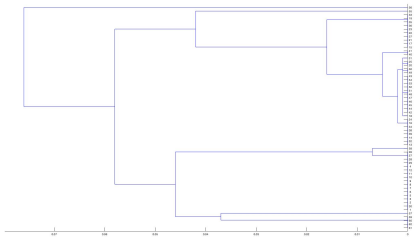


Figure 5: Dendrogram of Rege.

Food web wiring

The best link percentgae can be seen in Figure 6. The best weight was always the minimum weight, except for nwDC in the REGE hierarchical clustering (which was the mean weight) and nwDC in the group model(which was the sum of the weights). See the result of the Kendall tau b in Figure 7.

Discussion

A possible future direction is using new algorithms for clustering. These could be either hierarchical clustering with different types of similarity indices - such as automorphic equivalence (Wasserman and Faust, 1994) or Katz similarity (Newman, 2018) - or methods of directed

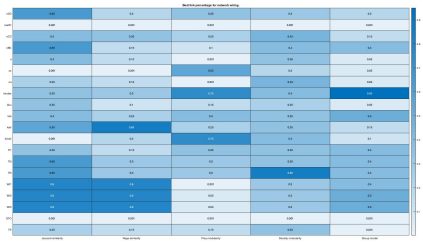


Figure 6: Best link percentage for the wiring of food webs.

network clusterings. For the latter type of clustering, check out the excellent review of Malliaros and Vazirgiannis (2013). Some examples of these algorithms are the semi-supervised learning (Zhou et al., 2005), the two-step random walk (Huang et al., 2006), the mixture models (Newman and Leicht, 2007; Ramasco and Mungan, 2008; Wang and Lai, 2008), the infomap (Rosvall and Bergstrom, 2008), the Link Rank algorithm (Kim et al., 2010), and the maga method (Zhan et al., 2011). Zhou et al. (2005); Huang et al. (2006); Wang and Lai (2008); Kim et al. (2010) and Zhan et al. (2011) have never been cited by the ecological literature and their application to food webs might reveal to be useful.

Another possible direction is checking how different data aggregations influence the dynamics and not only the structure of the network.

The repercussion of food web aggregation on sampling have been also investigated by Patonai and Jordán (2017). Considering that it would be difficult to know in advance what would be the connections between different species in the food web (I mean, we can kind of know, but not exactly I guess. Actually in some cases we can, so it is no problem. But in other cases it might be more difficult). A good future direction would be trying to understand when do species have the same trophic role. What is the biology behind this? Can we predict by using functional traits what are the species that have the same trophic role?

Also it needs to test to what extent are centrality indices a good proxy for food web structure.

I think that it is really important for a certain aggregation method to maintain the trophic position of the different species. This is because the trophic level of a species is associated with many information about its role inside a food web. Trophic level is a proxy for many things. I think that if a method doesn't maintain the trophic level of a species, it can't be a reliable one for maintaining the structure of the network.

It is also important to remember that it seems like species who are unique trophospecies seem to be really important for secondary extinction. In particular, (Petchey et al., 2008) found that they are particularly vulnerable to secondary extinctions when trying to model their dynamic food web. The fact that the concept of trophospecies not only is important to understand which ones are the most vulnerable species in the system, but it seems also important in something related to how different species are related to each other. If, as someone said but I don't remember who, keystone species are the ones that are unique in their trophospecies, it means that secondary extinctions happen only if you hit the network close or on keystone species. This is because keystone species, if my interpretation is correct, are not only the most important, but also the most vulnerable. This might be due to the fact that they have such low abundance as well. Keystone species might be also the ones that we cannot aggregate in a food web (Bond, 1994). So, at this point, maybe data aggregation would reveal keystone species. The fact that the aggregation of according to Jaccard not only can reveal the keystone species, but also would maintain the relative importance of the nodes, seems like a great way of aggregating data. Wait: I need to check whether the species that are unique in their cluster also have high centrality indices.

A problem is always the one from the fact that food webs are more resolved at higher than at lower trophic levels. For example, a node at the highest trophic level might represent a single species of shark, meanwhile a node at the bottom trophic level might represent hundreds of phytoplankton species ().

Understanding how to aggregate data might also help us with dealing with missing data.

We would not suggest to use modularity maximisation and the group model for data aggregation.

One of the things that needs to be defined as well is: what is the relevance of the clusterings at this point?

For my networks maybe I should have also networks that are really different between each other in terms of size and in terms of habitat.

Acknowledgements

We would like to thank Wei-Chung Liu for providing the code for computing some centrality indices and Stefano Allesina & Elizabeth Sander for providing the code for the computation of the group model. Support by H2020 AtlantECO.

Supplementary material

The adjacency matrix of the food web of the Gulf of Naples, as well as the code used to analyse it is available at `https://github.com/Emanuele-Giacomuzzo/Data_aggregation`.

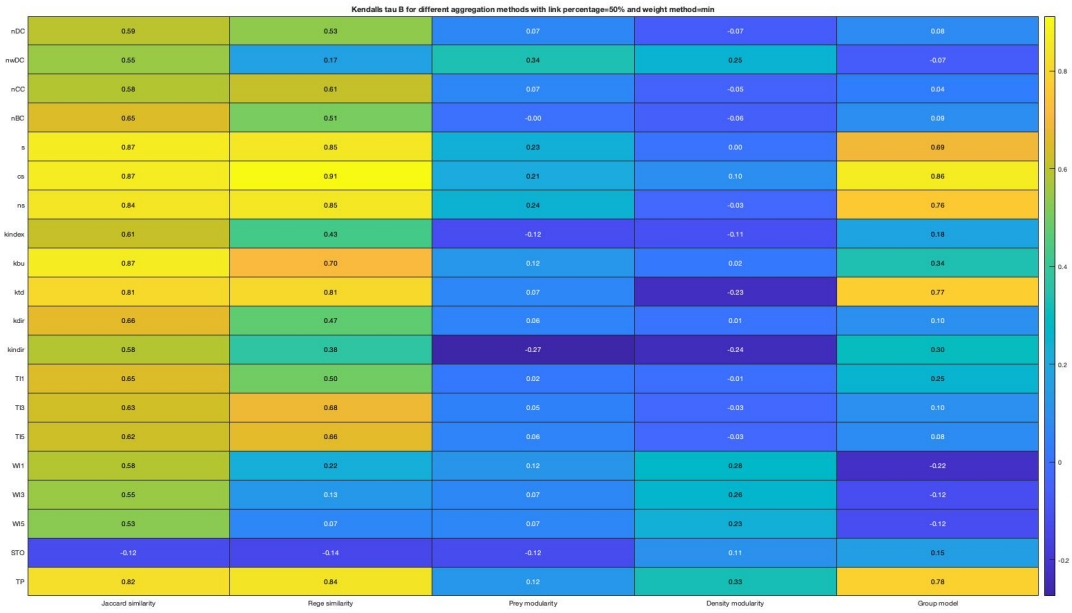


Figure 7: Heat map of how different methods of clustering preserve the relative importance of species inside the food web of the Gulf of Naples. The values represent the Kendall tau b.

References

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

Allesina, S. and Pascual, M. (2009). Food web models: A plea for groups. *Ecology Letters*, 12(7):652–662.

Arenas, A., Duch, J., Fernández, A., and Gómez, S. (2007). Size reduction of complex networks preserving modularity. *New Journal of Physics*.

Bauer, B., Jordán, F., and Podani, J. (2010). Node centrality indices in food webs: Rank orders versus distributions. *Ecological Complexity*.

Bond, W. J. (1994). Keystone Species. In *Biodiversity and Ecosystem Function*.

Borgatti, S. P. (2002). A Statistical Method for Comparing Aggregate Data Across A Priori Groups. *Field Methods*, 14(1):88–107.

Borgatti, S. P. and Everett, M. G. (1993). Two algorithms for computing regular equivalence. *Social Networks*, 15(4):361–376.

Burkey, J. (2021). Mann-Kendall Tau-b with Sen's Method (enhanced). Retrieved February 15, 2021.

D'Alelio, D., Libralato, S., Wyatt, T., and Ribera D'Alcalà, M. (2016). Ecological-network models link diversity, structure and function in the plankton food-web. *Scientific Reports*, 6(November 2015):1–13.

Endrédi, A., Senánszky, V., Libralato, S., and Jordán, F. (2018). Food web dynamics in trophic hierarchies. *Ecological Modelling*, 368:94–103.

Fornito, A., Zalesky, A., and Bullmore, E. T. (2016). *Fundamentals of Brain Network Analysis*.

Frigui, H. (2008). Clustering: Algorithms and applications. *2008 1st International Workshops on Image Processing Theory, Tools and Applications, IPTA 2008*.

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. (2007). Module identification in bipartite and directed networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(3):1–8.

Guimerà, R., Stouffer, D. B., Sales-Pardo, M., Leicht, E. A., Newman, M. E. J., and Amaral, L. A. N. (2010). Origin of compartmentalization in food webs. Appendix B: Confidence intervals for compartment properties. *Ecology*, 91(10):2941–2951.

Hall, S. J. and Raffaelli, D. G. (1993). *Food Webs: Theory and Reality*, volume 24.

Harary, F. (1959). Status and Contrastatus. *Sociometry*, 22(1):23.

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

Huang, J., Tingshao, Z., and Schuurmans, D. (2006). Web communities identification from random walks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Jordán, F. (2009). Keystone species and food webs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1524):1733–1741.

Jordán, F., Endrédi, A., Liu, W. C., and D'Alelio, D. (2018). Aggregating a plankton food web: Mathematical versus biological approaches. *Mathematics*, 6(12).

Jordán, F., Liu, W.-C., and Davis, A. J. (2006). Topological keystone species: measures of positional importance in food webs. *Oikos*, 112(July 2005):535–546.

Jordán, F., Liu, W. C., and Mike, Á. (2009). Trophic field overlap: A new approach to quantify keystone species. *Ecological Modelling*, 220(21):2899–2907.

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

Jordán, F., Takacs-Santa, A., and Molnar, I. (1999). A Reliability Theoretical Quest for Keystones. *Oikos*, 86(3):453.

Kim, Y., Son, S. W., and Jeong, H. (2010). Finding communities in directed networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*.

Kozen, D. and Timme, M. (2007). Indefinite summation and the Kronecker delta. *ecommons.cornell.edu*.

Leicht, E. A. and Newman, M. E. (2008). Community structure in directed networks. *Physical Review Letters*, 100(11):1–4.

Levine, S. (1980). Several measures of trophic structure applicable to complex food webs. *Journal of Theoretical Biology*, 83(2):195–207.

Luczkovich, J. J., Borgatti, S. P., Johnson, J. C., and Everett, M. G. (2003). Defining and measuring trophic role similarity in food webs using regular equivalence. *Journal of Theoretical Biology*, 220(3):303–321.

Malliaros, F. D. and Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95–142.

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

mei Lai, S., chung Liu, W., and Jordán, F. (2015). A trophic overlap-based measure for species uniqueness in ecological networks. *Ecological Modelling*, 299:95–101.

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

Müller, C. B., Adriaanse, I. C., Belshaw, R., and Godfray, H. C. (1999). The structure of an aphid-parasitoid community. *Journal of Animal Ecology*.

Newman, M. (2018). Measures and metrics. In *Networks*, pages 304–339. Oxford University Press.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*.

Newman, M. E. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America*.

Paine, R. T. (1988). Road Maps of Interactions or Grist for Theoretical Development? *Ecology*.

Patonai, K. and Jordán, F. (2017). Aggregation of incomplete food web data may help to suggest sampling strategies. *Ecological Modelling*, 352:77–89.

Petchey, O. L., Eklöf, A., Borrvall, C., and Ebenman, B. (2008). Trophically unique species are vulnerable to cascading extinction. *American Naturalist*, 171(5):568–579.

Ramasco, J. J. and Mungan, M. (2008). Inversion method for content-based networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*.

Ribera d'Alcalà, M., Conversano, F., Corato, F., Licandro, P., Mangoni, O., Marino, D., Mazzocchi, M. G., Modigh, M., Montresor, M., Nardella, M., Saggiomo, V., Sarno, D., and Zingone, A. (2004). Seasonal patterns in plankton communities in pluriannual time series at a coastal Mediterranean site (Gulf of Naples): An attempt to discern recurrences and trends. *Scientia Marina*.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*.

Sander, E. L., Wootton, J. T., and Allesina, S. (2015). What Can Interaction Webs Tell Us About Species Roles? *PLoS Computational Biology*, 11(7):1–22.

Scotti, M., Podani, J., and Jordán, F. (2007). Weighting, scale dependence and indirect effects in ecological networks: A comparative study. *Ecological Complexity*, 4(3):148–159.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *TAXON*.

Von Luxburg, U. (2004). *Statistical Learning with Similarity and Dissimilarity Functions*. PhD thesis.

Wang, J. and Lai, C. H. (2008). Detecting groups of similar components in complex networks. *New Journal of Physics*.

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

White, D. (1980). Structural equivalences concepts and measurement tures. *Unpublished manuscript*.

White, D. (1982). Measures of global role equivalence. *Unpublished manuscript*.

White, D. (1984). REGGE: a REGular Graph Equivalence algorithm for computing prior to blockmodeling. *Unpublished manuscript*.

Xiao, Z., Wu, J., Xu, B., Zhang, C., Ren, Y., and Xue, Y. (2019). Uniqueness measure based on the weighted trophic field overlap of species in the food web. *Ecological Indicators*, 101(May 2018):640–646.

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

Zhan, W., Zhang, Z., Guan, J., and Zhou, S. (2011). Evolutionary method for finding communities in bipartite networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*.

Zhou, D., Schölkopf, B., and Hofmann, T. (2005). Semi-supervised learning on directed graphs. In *Advances in Neural Information Processing Systems*.

Appendices

A Hierarchical clustering with Jaccard similarity index

1. *Compute similarity.*

   Compute the Jaccard similarity between the nodes by using the following equation (Yodzis and Winemiller, 1999):

   $$J_{ij} = \frac{a}{a+b+c} \tag{29}$$

   where J_{ij} is the Jaccard similarity between node i and j, a is the number of preys and predators that i and j have in common, b is the number of preys and predators exclusively of i, and c is the number of preys and predators exclusively of j.

2. *Build the dendrogram.*

   Find the two most similar elements and cluster them together (elements are intended as nodes or clusters. Of course, the first time we run this step all the elements are nodes). Repeat until you are left with only one item, which is the final dendrogram. During this process, the similarity between two clusters can be calculated in different ways, called linkage criteria. The ones that we used were

   - The similarity between the least similar nodes, one in each cluster, known as **single-linkage** (Frigui, 2008).
   - The similarity between the most similar nodes, one in each cluster, known as **complete linkage** (Frigui, 2008).
   - The mean similarity between the nodes inside the first item and the second item, known as the **weighted average distance(WPGMA)** (Sokal, 1958):

     $$d_{(i \bigcup j),k} = \frac{d_{i,k} + d_{j,k}}{2} \tag{30}$$

     where $d_{(i \cup j),k}$ is the distance between the cluster $i \bigcup j$ (cluster including i and j) and k, $d_{i,k}$ is the distance between i and k, and $d_{j,k}$ is the distance between j and k.
   - The mean similarity between the nodes inside the first item and the second item, but taking into consideration the average distance between the items inside the fist cluster; this is known as the**unweighted average distance (UPGMA)** (Sokal, 1958):

     $$d_{(i \bigcup j),k} = \frac{|i| d_{i,k} + |j| d_{j,k}}{|i| + |j|} \tag{31}$$

     where $|i|$ and $|j|$ are the mean distances between the elements inside i and j, respectively.

3. *Select the dendrogram.*

   After having produced a dendrogram for every linkage criteria, select the dendrogram with the highest cophenetic correlation (Sokal and Rohlf, 1962). This allows selecting the linkage criterion that produces the dendrogram that preserves the most faithfully the pairwise similarity between different elements.

4. *Cut the dendrogram.*

   Cut the dendrogram according to the maximum inconsistency of the branches, set at 0.01.

B Hierarchical clustering with REGE index

1. *Compute similarity.*

   Compute the similarity between nodes by using REGE, calculated by the homonym algorithm. This was originally developed in the unpublished work by White (1980, 1982, 1984) and firstly described in the literature by Borgatti and Everett (1993). It is available to be used in the software UCINET VI Borgatti (2002). The REGE algorithm is as follows (Jordán et al., 2018):

   (a) Set the maximum number of iterations. We set 3 iterations. Each iteration produces a matrix $R_{(t)}$ where t is the number of the iteration and every element $r_{(t)ij}$ is the regular equivalence between i and j at iteration t. The regular equivalence between nodes at iteration t=0 is always 1.

   (b) Starting from t=1, update the elements of the matrix following these sub-steps:

      i. For every predator k of species i, find the most similar predator m of species j according to $R_{(t)}$. Now, set $X_{i,k,j} = R_{(t)km}$.

      ii. For every predator m of species j, find the most similar predator k of species o according to $R_{(t)}$. Now, set $X_{j,m,i} = R_{(t)mk}$.

      iii. For every prey h of species i, find the most similar prey n of species j according to $R_{(t)}$. Now, set $Y_{i,h,j} = R_{(t)hn}$.

      iv. For every prey n of species j, find the most similar prey h of species i according to $R_{(t)}$. Now, set $Y_{j,n,i} = R_{(t)nh}$.

      v. Update the matrix R through the following equation

      vi. Increase t=t+1 and repeat step b until you reach the maximum number of iterations. The matrix of the maximum number of iterations contains the regular equivalence between nodes.

   (c) Increase t=t+1 and repeat step b until you reach the maximum number of iterations. The matrix of the maximum number of iterations contains the regular equivalence between nodes.

2. *Build the dendrogram.*

   The same as in the hierarchical clustering of nodes according to their Jaccard similarity index. During our analysis, we used the function linkage of MATLAB, which does not include the possibility of using a similarity matrix, so we converted the similarity matrices into dissimilarity ones. This was done by following what was written in Von Luxburg (2004). Namely, if the similarity function is normalised - takes values between 0 and 1 - and always positive, then $d = 1 - s$ where d is the dissimilarity measure and s is the similarity measure).

3. *Select the dendrogram.*

   The same as in the hierarchical clustering of nodes according to their Jaccard similarity index.

4. *Cut the dendrogram.*

   The same as in the hierarchical clustering of nodes according to their Jaccard similarity index.