Food web aggregation: effects on key positions

Emanuele Giacomuzzo[1] and Ferenc Jordàn[1]

[1] *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 25 of the most used centrality indices used in keystone species research. Our investigation was carried out on 86 Ecopath with Ecosim food web models. By having been constructed with the same methodology (see Okey (2004)), they were easy to

compare. These models were freely available on the EcoBase database (Colléter et al., 2013). The way we selected the food webs to be included in our analysis was the number of nodes: we selected only the food webs with at least 14 nodes. See a table of these food webs in Tables S1 from Heymans et al. (2014).

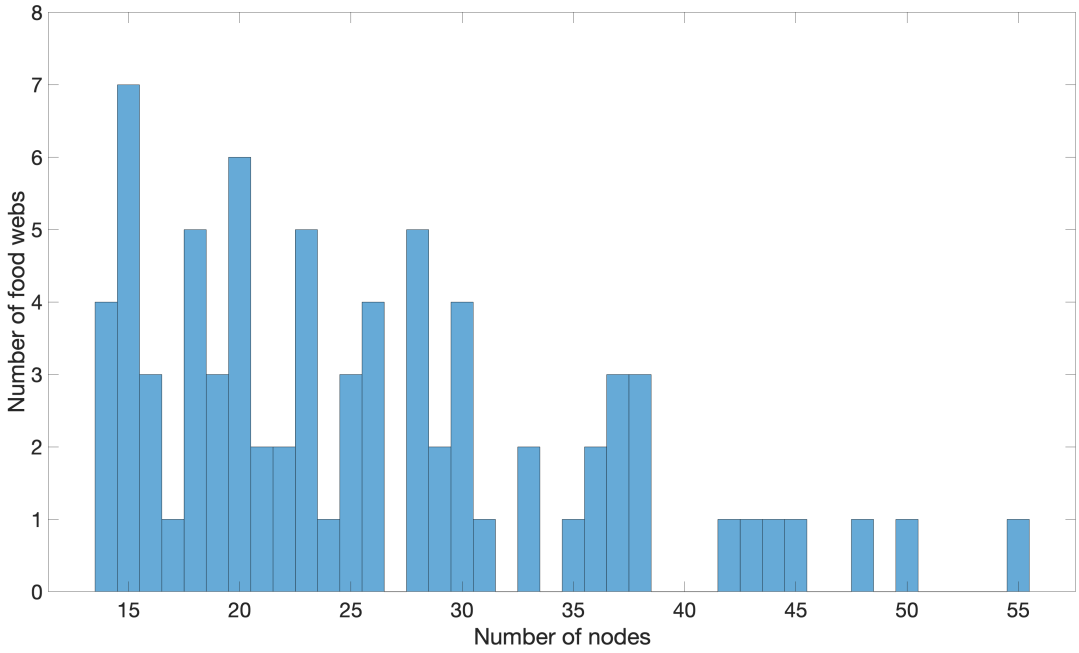

Figure 1: Size of the food webs we used in our study.

Methods: clustering techniques

To cluster similar nodes, we used the following clustering techinques.

Hierarchical clustering with Jaccard 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. See Appendix A for the clustering algorithm.

Hierarchical clustering with 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. See Appendix B 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 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 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 fastes for community detection (Traag et al., 2019). However, it tends to produce communities that are arbitrarly poorly connected from 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 and Nepusz, 2006) for R (R Development Core Team, 2013).

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

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

or in the following one for predator-modularity

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

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 this type of modules by using the rnetcarto package (Doulcier and Stouffer, 2015) for R. This finds the community structure of the network by using simulated annealing (Kirkpatrick et al., 1983).

Clustering of 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 retrieving the food web is:

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

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 arrangments, 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 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 the clusters and assigning interaction strength

The wiring of the food web followed a similar approach to the one describe in Martinez (1991). We used five methods to decide whether there was a link between two clusters. The first method produced the maximum connectance and is known as maximum linkage. Here, a cluster had a connection to another cluster if it had at least one link

going from one of its nodes to the nodes of the second cluster. The second one produced the minimum connectance and is known as minimum linkage. This time, a cluster was connected to another only if all its nodes had a connection to all the nodes of the other cluster. The other three methods produced an intermediate connectance. They considered a link from a cluster to the other only if at least 25%, 50% or 70% of possible connections from the first cluster to the second were 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.

Methods: centrality indices

For each food web, we calculated the centrality indices before and after the aggregation. The centrality index of a node after the aggregation process equaled the one of its cluster. For example, let's consider a hypothetical aggregation. Before the aggregation, the node "hake" has a degree centrality of 5. Through the aggregation process, this happens to be aggregated into a fish cluster. The degree centrality of this fish cluster is 8. The degree centrality of hake was 5 before the aggregation and 8 after the aggregation.

Degree centrality (DC)

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

$$DC_i = \sum_{j=1}^{n} A_{ij} \tag{5}$$

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 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)

$$nDC_i = \frac{DC_i}{n-1} \tag{6}$$

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).

$$WDC_i = \sum_{j=1}^{n} A_{ij} \tag{7}$$

Closeness centrality (CC)

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

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

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

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

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 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 status index family, the keystone index of a species i ($K(i)$) is calculated by considering the bottom-up and the top-down effects separately 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 indirect component. The indirect component takes into consideration the bottom-up effects of the predator and direct component does not

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

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

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

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

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 reported in Appendix C (Jordán, 2009).

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)

$$a_{1,ji}=\frac{A_{ij}}{weighted\,indegree_{j}} \tag{22}$$

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

Trophic field overlap (TO) species uniqueness (STO)

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 can be found in Appendix D (Jordán et al., 2018).

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))

$$a_{1,ij}=\frac{A_{ij}}{D_{j}} \tag{23}$$

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 Lai et al. (2015).

Methods: statistical analysis

The combination of clusterings (6 methods), linkages (5 methods) and interaction strength determinations (4 methods) produced 120 aggregation methods. For each of these aggregation methods, we studied their effects on centrality indices. More in particular, for each centrality index we studied how the nodes were ranked before and after the process. It was possible to study the difference in the these two rankings by using Kendall's tau b (τ_B) - a version of Kendall's rank correlation coefficient that makes adjustments for ties (**?**). For each aggregation method and for each centrality index, we found the mean τ_B across all food webs. This required us to convert τ_B using the Fisher z-transformation (Fisher, 1915), find the mean and back-transform it. For each mean τ_B we found its confidence interval by bootstrapping (DiCiccio and Efron, 1996). τ_B and bootstrapping were implemented in the Statistics and Machine Learning Toolbox for MATLAB (Mathworks Inc., 2019).

Results

Clustering of the food webs

The 86 food webs had a median of 25.5 nodes (IQR = 16.0), with a minimum of 14 nodes and a maximum of 55 nodes. See Figure **??**. They produced a median number of 0.76 (IQR = 0.11) for the Jaccard index, 0.73 (IQR = 0.07) for the REGE index, 0.16 (IQR = 0.08) for the density modularity, 0.35 (IQR = 0.02) for the prey modularity, 0.16 (IQR = 0.08) for the predator modularity and 0.16 (IQR = 0.07) for the group model. See Figure **??**.

Comparison between the original and the clustered rankings

The results of the comparison between the ranking of nodes in the original and the clustered food webs can be seen in Figure **??** - **??**.

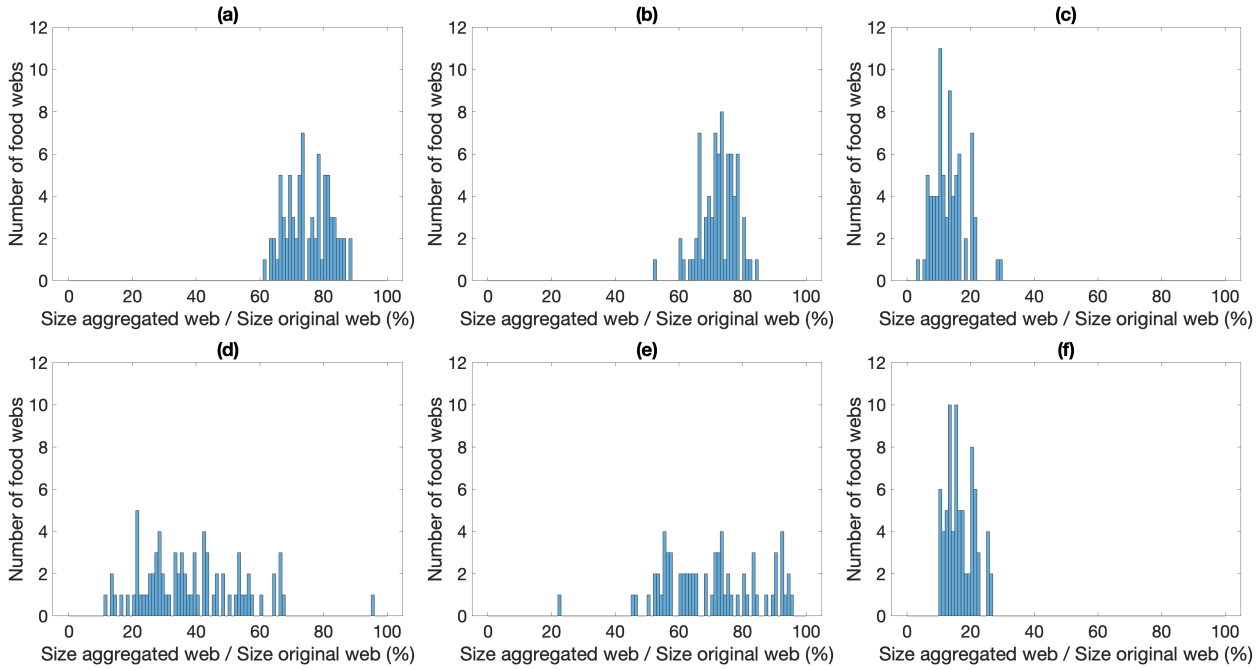

Figure 2: Size of the clusters produced by the different clustering methods. (a) = hierarchical clustering with Jaccard index, (b) = hierarchical clustering with REGE index, (c) = maximisation of density-based modularity, (d) = maximisation of prey-based modularity, (e) = maximisation of predator-based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

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. Ferenc Jordàn was supported 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`.

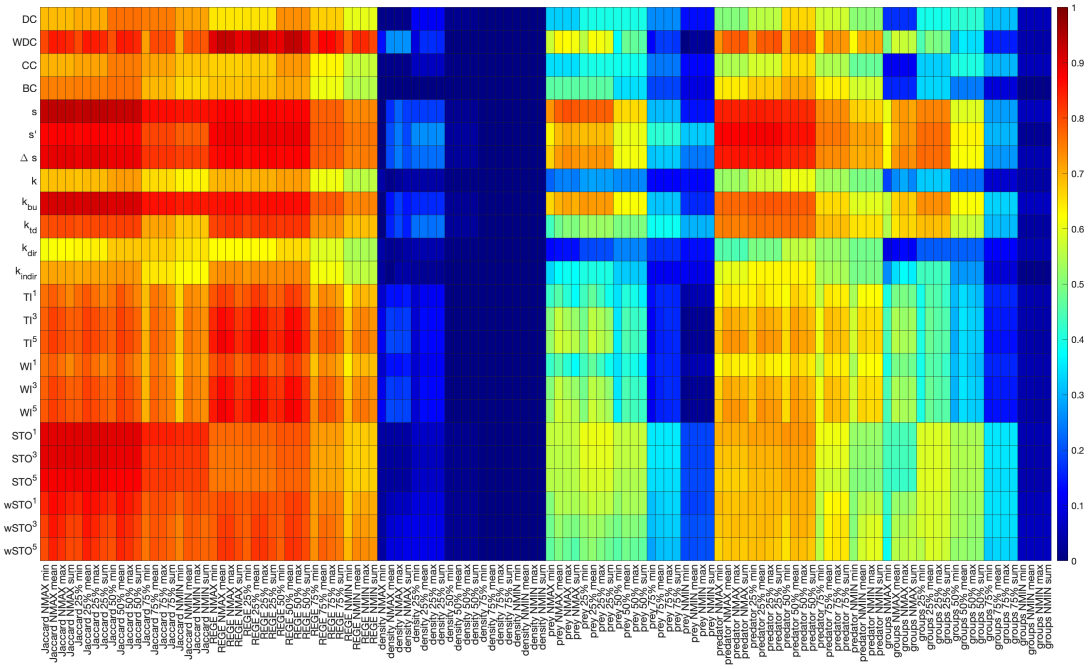

Figure 3: Heat map describing the kendall's rank correlation () between the ranking of the nodes in the original food web and in the aggregated food web. On the x axis, there is the aggregation method, which is described by three components. The first component is the clustering algorithm. The second one is the linkage method. The third one is the method by which we determined the interaction strength. On the y axis, there are the centrality indices.Jaccard = Hierachical 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, 25

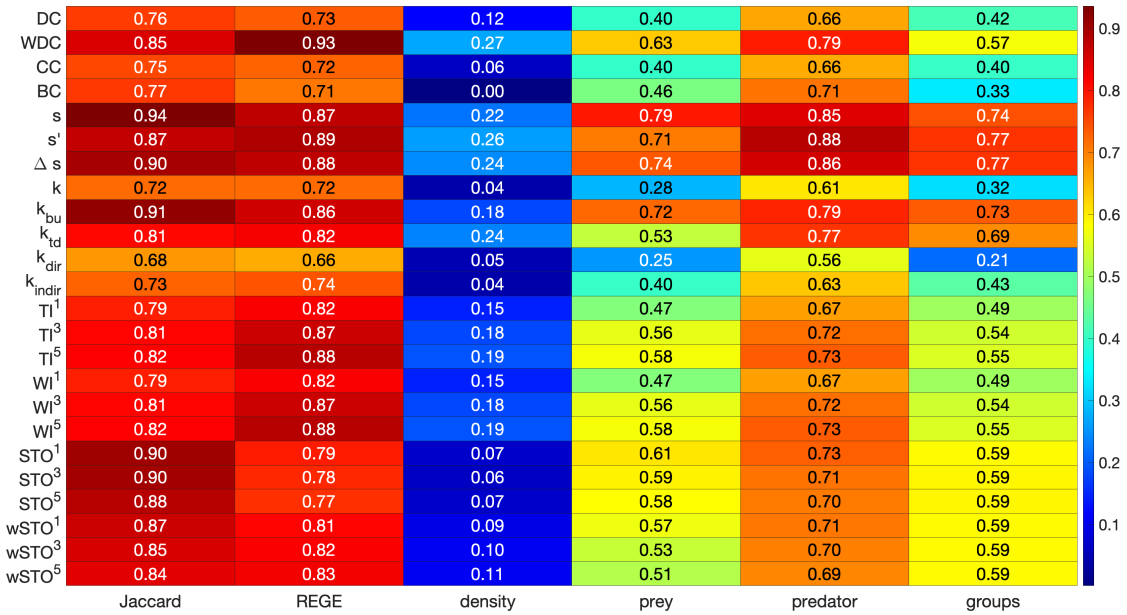

Figure 4: Heat map of the best Kendall's rank correlation coefficient for each combination of clustering methods and centrality indices. The best correlation is selected across linkage methods and methods of determining interaction strength.

References

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*, 9(6):176.

Bauer, B., Jordán, F., and Podani, J. (2010). Node centrality indices in food webs: Rank orders versus distributions. *Ecological Complexity*, 7(4):471–477.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10.

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.

Colléter, M., Valls, A., Guitton, J., Morissette, L., Arreguín-Sánchez, F., Christensen, V., Gascuel, D., and Pauly, D. (2013). EcoBase: A repository solution to gather and communicate information from EwE models. Technical report.

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

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228.

Doulcier, G. and Stouffer, D. B. (2015). Rnetcarto: Fast network modularity and roles computation by simulated annealing.

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

Index	C1	C2	C3	C4	C5	C6
\\displaystyle DC \\\	0.76	0.73	0.66	0.42	0.4	0.12
\\displaystyle WDC \	0.93	0.85	0.79	0.63	0.57	0.27
\\displaystyle CC \	0.75	0.72	0.66	0.4	0.4	0.06
\\displaystyle BC \	0.77	0.71	0.71	0.46	0.33	0
\\displaystyle s \	0.94	0.87	0.85	0.79	0.74	0.22
\\displaystyle s' \	0.89	0.88	0.87	0.77	0.71	0.26
\\displaystyle \Delta s \	0.9	0.88	0.86	0.77	0.74	0.24
\\displaystyle k \	0.72	0.72	0.61	0.32	0.28	0.04
\\displaystyle k_{bu} \	0.91	0.86	0.79	0.73	0.72	0.18
\\displaystyle k_{td} \	0.82	0.81	0.77	0.69	0.53	0.24
\\displaystyle k_{dir} \	0.68	0.66	0.56	0.25	0.21	0.05
\\displaystyle k_{indir} \	0.74	0.73	0.63	0.43	0.4	0.04
\\displaystyle TI^1 \	0.82	0.79	0.67	0.49	0.47	0.15
\\displaystyle TI^3 \	0.87	0.81	0.72	0.56	0.54	0.18
\\displaystyle TI^5 \	0.88	0.82	0.73	0.58	0.55	0.19
\\displaystyle WI^1 \	0.82	0.79	0.67	0.49	0.47	0.15
\\displaystyle WI^3 \	0.87	0.81	0.72	0.56	0.54	0.18
\\displaystyle WI^5 \	0.88	0.82	0.73	0.58	0.55	0.19
\\displaystyle STO^1 \	0.9	0.79	0.73	0.61	0.59	0.07
\\displaystyle STO^3 \	0.9	0.78	0.71	0.59	0.59	0.06
\\displaystyle STO^5 \	0.88	0.77	0.7	0.59	0.58	0.07
\\displaystyle wSTO^1 \	0.87	0.81	0.71	0.59	0.57	0.09
\\displaystyle wSTO^3 \	0.85	0.82	0.7	0.59	0.53	0.1
\\displaystyle wSTO^5 \	0.84	0.83	0.69	0.59	0.51	0.11

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.

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

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

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

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. *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.

Heymans, J. J., Coll, M., Libralato, S., Morissette, L., and Christensen, V. (2014). Global patterns in ecological indicators of marine food webs: A modelling approach. *PLoS ONE*, 9(4).

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):336.

Jordán, F., Liu, W.-C., and Davis, A. J. (2006). Topological keystone species: measures of positional importance in food webs. *Oikos*, 112(3):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., Takács-Sánta, A., and Molnár, I. (1999). A Reliability Theoretical Quest for Keystones. *Oikos*, 86(3):453.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

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

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

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.

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

Mathworks Inc. (2019). Matlab Statistics and Machine Learning Toolbox.

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

Müller, C. B., Adriaanse, I. C., Belshaw, R., and Godfray, H. C. (1999). The structure of an aphid-parasitoid community. *Journal of Animal Ecology*, 68:346–370.

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

Okey, T. A. (2004). *Shifted community states in four marine ecosystems: some potential mechanisms*. PhD thesis, University of British Columbia, Canada.

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

R Development Core Team (2013). R: A Language and Environment for Statistical Computing.

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*, 11(2):33–40.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):1–12.

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

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*, 101:640–646.

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

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.

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

Appendices

A Hierarchical clustering with Jaccard index

1. *Compute similarity.*

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

   $$J_{ij} = \frac{a}{a+b+c} \tag{24}$$

   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):

     $$d_{(i \bigcup j),k} = \frac{d_{i,k} + d_{j,k}}{2} \tag{25}$$

     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):

     $$d_{(i \bigcup j),k} = \frac{|i| d_{i,k} + |j| d_{j,k}}{|i| + |j|} \tag{26}$$

     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.

C Topological importance computation

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

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

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 5. It can be calculated as follows

$$A\,(n) = A_{(1)}^{n} \tag{28}$$

3. *Calculate topological importance*

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

$$TI_i = \frac{\sum_{m=1}^{N} \sum_{j=1}^{n} a_{m,ji}}{N} \tag{29}$$

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.

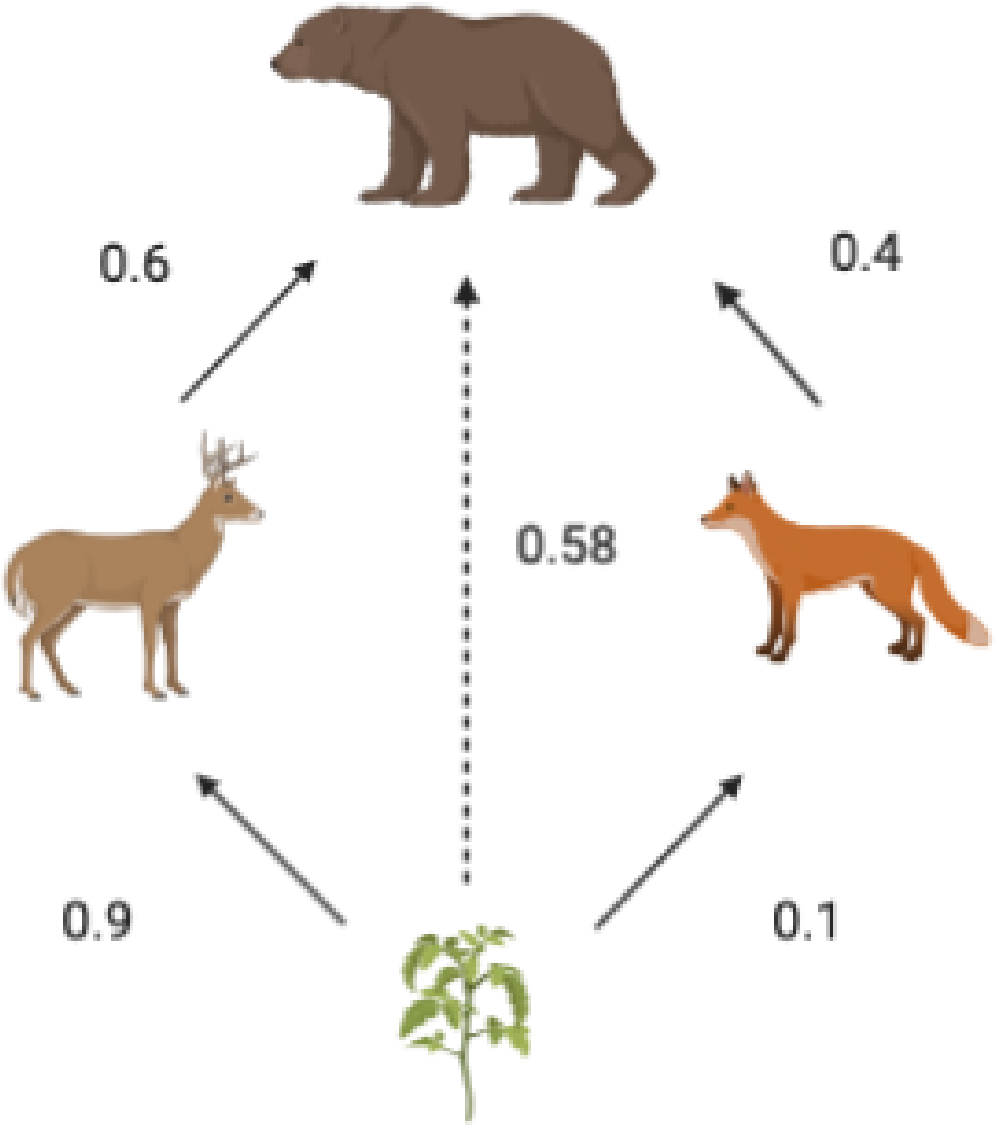

Figure 5: Topological importance (TI) of a species on another. The plant and the bear are not connected, so there is no direct effect from the plant to the bear. However, indirect effects reach the bear from the plant through two paths and the final effect is the sum of these effects. The first one is 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.

D Trophic field overlap computation

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

$$E_n = \frac{1}{n} \sum_{i=1}^{n} A_{(i)} \tag{30}$$

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.