Format the main text as single column, double line spaced text with page- and line-numbers.

Table of Contents

[Abstract 1](#_Toc67600376)

[Keywords 1](#_Toc67600377)

[Introduction 1](#_Toc67600378)

[Material and methods 2](#_Toc67600379)

[Methods: clustering techniques 2](#_Toc67600380)

[Hierarchical clustering with Jaccard index 2](#_Toc67600381)

[Hierarchical clustering with REGE index 2](#_Toc67600382)

[Clustering of density-based modules 2](#_Toc67600383)

[Clustering of prey-based and predator-based modules 3](#_Toc67600384)

[Clustering of groups 3](#_Toc67600385)

[Methods: connecting the clusters and assigning interaction strength 4](#_Toc67600386)

[Methods: centrality indices 4](#_Toc67600387)

[Degree centrality (DC) 4](#_Toc67600388)

[Closeness centrality (CC) 5](#_Toc67600389)

[Betweenness centrality (BC) 5](#_Toc67600390)

[Status index (s) 5](#_Toc67600391)

[Keystone index (K) 6](#_Toc67600392)

[Topological importance (TI) 7](#_Toc67600393)

[Trophic field overlap (TO) and species uniqueness (STO) 7](#_Toc67600394)

[Trophic position (TP) 7](#_Toc67600395)

[Methods: statistical analysis 7](#_Toc67600396)

[Results 8](#_Toc67600397)

[Clustering of the food webs 8](#_Toc67600398)

[Comparison between the original and the clustered rankings 8](#_Toc67600399)

[Discussion 8](#_Toc67600400)

[Acknowledgements 10](#_Toc67600401)

[Supplementary material 10](#_Toc67600402)

[References 10](#_Toc67600403)

[Figures and tables with captions 14](#_Toc67600404)

# Abstract

# Keywords

Keystone species, centrality indices, species role, trophospecies, trophic role, ecological networks, structural equivalence, regular equivalence, group model, modularity, data aggregation, species importance.

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 & 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 answer this question 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” 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 & 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 extended by Sander, Wootton, & Allesina (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 called prey-based modularity, is the degree by 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 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, Jordán, & Podani, 2010) or hypothetical importance values (Mills, Doak, & Soulé, 1993) provide macroscopic descriptors of ecosystems.

In this paper, we investigate how these different aggregation methods maintain the relative importance of species according to 25 of the most used centrality indices. These centrality indices are the most widely used in keystone species research. Our investigation was carried out on 76 Ecopath with Ecosim food web models available on the EcoBase database (Colléter et al., 2013). By having been constructed with the same methodology (see Okey (2004)), they were easy to compare. 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 the supporting information.

# Material and methods

## Clustering techniques

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

### Hierarchical clustering with Jaccard index

As a first clustering method, we clustered structurally equivalent nodes as in Yodzis & Winemiller (1999). We used Jaccard similarity index (Jaccard P., 1912) as a measure of structural equivalence. See Appendix A 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 REGE index (Borgatti & Everett, 1993). 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, Duch, Fernández, & Gómez, (2007), which is a generalisation of the Newman-Girvan modularity (Newman, 2004)

where is the modularity of the network, L is the number of links in the network, is the element of the adjacency matrix of the directed binary network (links go from j to i), is the indegree of i, is the outdegree of j, is the module of i, is the module of j and 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 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 & Nepusz, 2006) for the statistical software R (R Development Core Team, 2011).

### 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 (Roger Guimerà, Sales-Pardo, & Amaral, 2007) for prey-modularity

or in the following one for predator-modularity

where is the number of outgoing links that i and j have in common and 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 & Stouffer, 2015) for R. This finds the community structure of the network by using 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:

where is the food web with number of nodes and number of links, is the vector containing the probabilities of a connection between and within clusters, is the probability that a node inside the group connects to another node inside the group , is the number of links connecting nodes belonging to the group to nodes belonging to the group , is the number of nodes in the cluster , and is the number of nodes in the cluster .

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 (), also known as parallel tempering (Geyer, 1991), with a Gibbs sampler (Yildirim, 2012). 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 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 (NMAX). 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 (NMIN). 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 75% 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.

## 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 equalled the one of its cluster. Let’s consider the following example. 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 () of a node is the number of links it has (Wasserman & Faust, 1994)

where is the number of nodes in the food web, and 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 (), 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, Zalesky, & Bullmore, 2016)

### Closeness centrality (CC)

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

where is the shortest path between node and .

### Betweenness centrality (BC)

The betweenness centrality () 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)

where is the total number of shortest paths going from to and is the total number of these paths passing through .

### 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)

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

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

### Keystone index (K)

The keystone index was firstly introduced by Jordán, Takacs-Santa, & Molnar (1999) and inspired by the status index. As 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)

where is its bottom-up keystone index of species and the top-down keystone index of species .

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 is then calculated in the following way

where is a predator of , is the number of preys of , and is the fraction of bottom-up effects of that are caused by . The of top predators is set as 0. The top-down effect of a certain node is calculated exactly as , 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

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

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

### 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, Adriaanse, Belshaw, & Godfray (1999) and then to food webs by Jordán, Liu, & van Veen (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 () – if instead of using the degree () we use the weighted degree () (Scotti, Podani, & Jordán, 2007)

where Aij is the element of the adjacency matrix of the weighted directed network.

### Trophic field overlap (TO) and species uniqueness (STO)

The trophic field overlap (TO) represents how redundant the strong interactions of a node are. It was first introduced by Jordán, Liu, & Mike (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, Endrédi, Liu, & D’Alelio, 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)).

To avoid having to choose a 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, Liu, & Jordán (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 fractional positions. It can be calculated through the following formula

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

## 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 25 centrality indices. 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 () - a version of Kendall's rank correlation coefficient that makes adjustments for ties (Agresti, 2012). For each combination of aggregation method and centrality index, we found the mean across all food webs. This required us to convert 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 . 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. See 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. See Figure 2.

## Correlation of centrality indices before and after the aggregation

The correlation between the ranking before and after the aggregation can be seen in Figure 3. If we focus only on the clustering method and we ignore the linkage method and the interaction strength method, we can select the best clustering for each of combination of centrality indices and clustering methods. This would give us a clear explanation of what the best aggregation for each of the indices are. See Figure 4. If we now rank the clustering algorithms for each of the centrality indices, we get Table 1.

Density-modularity always ranked as the worst clustering algorithms. Prey-based modules and group model also always ranked as either fourth or fifth. The predator modules ranked as best algorithm to maintain trophic position and was ranked as first algorithm sharing a tie with the REGE index for the controstatus. Excluding the controstatus and trophic position, the Jaccard index and the REGE index always ranked as 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 took into consideration.

# Discussion

An example food web is the one in Figure 5 and 6, taken from T. A. Okey (2004). This is the food web of the West Florida Shelf and the largest network we used (55 nodes).

When we have large food webs, we need to aggregate them to better understand them, especially if we want to simulate them dynamically. The best way of comparing different aggregation methods would be to check which one is the one that maintains the dynamics the best. However, we did not have enough time to see what is the best one in that sense. This is why we decided to use structure as a proxy for dynamics. We used centrality indices as a proxy for network structure. Centrality indices tell us about many things of nodes. This is why we used them, as a proxy for structure. So, we can actually say that this is a study where we check whether large food webs can be aggregated into smaller ones by maintaining the properties of the nodes.

Different centrality indices have different predictive power to find keystone species. The most reliable indices (e.g. keystoneness index, see Libralato, Christensen, & Pauly, (2006)) are the ones based on a dynamical model of the network. However, building a good dynamical model is really time consuming and might not be feasible for multiple networks. This is why often researchers use topological indices, such as the ones studied in this work. To check the reliability of topological indices, Gouveia, Móréh, & Jordán (2020) tested how much they agreed with the dynamical index keystoneness (KS). They found that the most reliable topological index was the weighted degree (wDC). It could predict the most important species 70.06% of the times. It was followed by the 5-step weighted topological importance (WI5). A combination of wDC and WI5 increased this percentage to 78.42\%. It seems like a good aggregation should then maintain these two indices. Another index that a good aggregation method should maintain is the trophic level. This can be calculated, for example, as the trophic position in this study. Trophic level is associated with many biological properties of species. For example, the direction by which a perturbation propagates to (Curtsdotter et al., 2011). Or the risk of extinction of a species (Binzer et al., 2011). Or the success of a species to colonise a new environment (Holt, 2010). (Add: (Gouveia et al., 2020) didn't take STO and wSTO into consideration, as well as the trophic level.)

By looking at the results in Figure 5, we see clear differences between algorithms. Density modularity is the worst aggregation algorithm. This is probably because it produces communities based on other factors (such as phylogeny, body mass, and habitat structure, see Rezende, Albert, Fortuna, & Bascompte (2009)) rather than trophic links between modules. Predator modularity, prey modularity and the group model also performed poorly, considering they performed worse than REGE and Jaccard across all centrality indices (except for predator modularity that performed the best with regards to trophic position). Jaccard and REGE seem to be the best aggregation methods. They both maintained well the rankings of all centrality indices. Jaccard maintained the greatest number of centrality indices slightly better than REGE. However, REGE maintained better the pattern of the two most important indices according to Gouveia et al. (2020), wDC and WI5. In light of these findings, we suggest using either the Jaccard index or the REGE index for aggregation in keystone species research. The clusters produced by these two similarity indices can be useful to answer different questions. The Jaccard index maintains all the information about a certain network, but it doesn't allow you to compare different food webs (Luczkovich et al., 2003). The REGE index produces a controlled loss of information but it allows to compare between different food webs (Luczkovich et al., 2003).

Future work should be focused on developing new algorithms for the aggregation of species in food webs. It should follow five steps:

1. Developing a singular index that tells us the goodness of the aggregation.
2. Deciding what is the best method to assign the interaction strength of the link between the clusters.
3. Deciding what is the best percentage of realised links between clusters to decide that there is a link between them.
4. Deciding what are the indices we should use. Are more indices better or worse? Should we combine all indices? Or should we develop different aggregation methods according to the type of pattern we want to preserve (e.g., redundancy, capacity of spreading an effect)?

This research, however, is complicated by our lack of understanding of keystone species. It is not clear what the best centrality indices are. Also, it seems like trophic links might not be the only interactions important in determining keystone species. Donohue et al. (2017) found that secondary extinctions were better modelled when they took into consideration competition. This might mean that a multilayer approach (see Pilosof, Porter, Pascual, & Kéfi (2017)) could be a better way of finding keystone species. So, we would argue that developing algorithms of aggregation should go hand in hand with this type of research. Another possible direction is checking how different data aggregations influence dynamical indices instead of topological indices. For example, the keystoneness index (KS) (Libralato et al., 2006) or eigenvector centrality (Allesina & Pascual, 2009b).

An interesting way of using aggregation, however, could be to find keystone species. It has been suggested by Bond (1994) that keystone species might be the species that cannot be aggregated. This makes us wonder whether it would be possible to use aggregation to find keystone species in topological networks. And whether they would work better than centrality indices. We always need to remind ourselves that aggregated food webs bias our analyses. Ultimately, the question should be also whether we should aggregate our data. Aggregating entities in ecology has been found to always bias models and produce worse results. For example, Woodward et al., (2010) showed how using the Allometric Diet Breadth Model (ADBM) (Owen L. Petchey, Beckerman, Riede, & Warren, 2008) to predict the links between nodes predicted 52\% of the links when building a species-based food web. This number increased to 83\% when considering size-based classes. Another example is from Petchey & Gaston (2002). When species were put into functional groups, the community showed high functional redundancy. When single species were considered singularly, the community showed low functional redundancy. We might argue that the best way of working with food webs is not aggregating. However, the problem is that it is never possible to do this. This is because we have a limited resolution power when we sample and the data that have been collected in the past and are available for our analyses have low resolution, at least in some part of the network. This could be solved by the way we sample. Ideally, we should develop technologies and methods that allow us to sample a food web in its highest resolution possible. However, this might be really expensive, and ecology might not have enough money to finance this type of research. This should be especially directed toward the lowest trophic levels of the food web, for example plankton in aquatic food webs. Aggregation research should also inform sampling methods and how to deal with missing data ( see Patonai & Jordán (2017).

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, Eklöf, Borrvall, & Ebenman (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.

The problem with aggregation algorithms is that they might give you a solution where certain clusters are not connected to each other.

Keystone index and status index can’t be calculated if there are cycles in the graph. The way that we deleted the cycles was the following one: find cycles. Start from the smallest cycles, a.k.a. the ones with the smallest number of nodes in the cycle. Now check their trophic level. If there are species in the cycle that have a trophic level difference larger than 0.2 eliminate the connection going between the species that had the highest difference in trophic position. If they have a difference in trophic level that is less than 0.2 eliminate the connection with the lowest interaction strength. Once you have eliminated all the cycles with this size, repeat the process, until no simple cycles are left, and the graph has become a directed acyclic graph (DAG). Sometimes, it is not possible, however, to calculate trophic level. In this case, we eliminated the connections with the lowest interaction strength.

When trophic level cannot be calculated for a certain network (why?) we just transformed the graph into a DAG.

How should we deal with non-connected clusters?

# 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 code used for this analysis is available at <https://github.com/Emanuele-Giacomuzzo/Data_aggregation>.

# References

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

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

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.

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

Endredi, A., Senanszky, V., Libralato, S., and Jordàn, 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.

Geyer, C. J. (1991). Markov Chain Monte Carlo Markov Chain Monte Carlo. (5):1–6.

Guimera, 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.

Guimera, 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 Controstatus. Sociometry, 22(1):23.

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

Hays, W. L. (1981). Statistics for Psychologists.

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).

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

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

Jordàn, F., Endredi, A., Liu, W. C., and D’Alelio, D. (2018). Aggregating a plankton food web: Mathematical versus biological approaches. Mathematics, 6(12).

Jordàn, 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.

Jordàn, 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.

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

Kendall, M. G. and Smith, B. B. (1939). The Problem of $m$ Rankings. The Annals of Mathematical Statistics.

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

Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E., and Taylor, W. W. (2003). Compartments revealed in food-web structure. Nature, 426(6964):282.

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 Jordàn, F. (2015). A trophic overlap-based measure for species uniqueness in ecological networks. Ecological Modelling, 299:95–101.

Muller, 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. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America.

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

Petchey, O. L., Eklof, 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 Jordàn, F. (2007). Weighting, scale dependence and indirect effects in ecological networks: A comparative study. Ecological Complexity, 4(3):148–159.

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.

Yildirim, I. (2012). Bayesian Inference: Gibbs Sampling. Department of Brain and Cognitive Sciences, 14627:1–6

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