functionInk: An efficient method to detect functional groups in multidimensional networks reveals the hidden structure of ecological communities

May 31, 2019

Alberto Pascual-García[1, 2,*] and Thomas Bell[1]

(1) Department of Life Sciences. Silwood Park Campus. Imperial College London, Ascot, United Kingdom
(2) Current address: Institute of Integrative Biology. ETH-Zürich, Zürich, Switzerland
(*) Correspondence: alberto.pascual.garcia@gmail.com.

Abstract

Complex networks have been useful to link experimental data with mechanistic models, becoming widely used in modern science. More recently, the increasing amount and complexity of data, in particular in biology, prompted the development of multidimensional networks, where dimensions reflect the multiple qualitative properties of nodes, links, or both, classifying them into types. As a consequence, traditional quantities computed in single dimensional networks should be adapted to incorporate this new information. A particularly important problem is the detection of communities, namely sets of nodes sharing certain properties, which reduced the complexity of the networks, hence facilitating its interpretation. The two traditional approximations to this problem come either through the identification of *modules* (communities more densely connected within their members than with nodes belonging to other communities) or of structurally equivalent communities (sets of nodes connected with the same neighbours, even if they are not connected themselves), that we call *guilds*. The relevance of this distinction is notable in biology, where we aim to differentiate between trophic levels, guilds, or concepts such as functional groups or ecotypes. In this work, we argue that structural equivalence represents an appropriate definition of the function of a node in the network, and we exploit this fact to show that it is possible to detect modules and guilds that, in this way, can be understood as different kinds of functional groups. We call our method functionInk (functional linkage), a method capable of objectively finding simultaneously both modules and guilds, and to determine which is the most relevant kind of functional group for a given network. Notably, it is computationally efficient handling large multidimensional networks, since it does not require an optimization procedure nor tests for robustness. The method is freely available at: HTTPS://GITHUB.COM/APASCUALGARCIA/FUNCTIONINK.

1 Introduction

Networks have played an important role in the development of ideas in ecology, particularly in understanding food webs [1], and flows of energy and matter in ecosystems [2]. However, modern ecological datasets are becoming increasingly complex, notably within microbial ecology, where multiple types of information (taxonomy, behaviour, metabolic capacity, traits) on thousands of taxa can be gathered. A single network might therefore need to integrate different sources of information, leading to connections between nodes representing relationships of different types, and hence with different meanings. Advances in network theory have attempted to develop tools to analyses these more sophisticated networks, encompassing ideas such as multiplex, multilayer, multivariate networks, reviewed in [3]. There could therefore be much value in extending complex networks tools to ecology in order to embrace these new concepts.

Broadly speaking, a network represents how a large set of entities *share* or *transmit* information. This definition is intentionally empty-of-content to illustrate the challenges we face in network analysis. For instance, a network in which information is shared may be built relating genes connected if their sequence similarity is higher than certain threshold. In that case, we may capture how their similarity diverged after an evolutionary event such as a gene duplication. On the other hand, networks may describe how information is transmitted, as in an ecosystem in which we represent how biomass flows through the trophic levels or how behavioural signals are transmitted among individuals. We aim to illustrate with these examples that, when building networks that consider links of different nature (e.g. shared vs. transmitted information) or different physical units (e.g. biomass vs. bits) care should be taken in extrapolating methods from single-dimensional to multidimensional complex networks.

A particularly relevant problem in mutiplex networks is the detection of "communities", which are defined in network theory as being sets of nodes sharing similar topological properties. Perhaps the most widely adopted definition of community is the one considering sets of nodes more densely connected within the community than with respect to other communities, often called *modules* [4]. Strategies to detect modules explore trade-offs in quantities like the betweeness and the clustering coefficient [4], as in the celebrated Newman-Girvan algorithm [5]. Generalizing traditional quantities like the clustering coefficient to multidimensional networks is difficult. Consider, for instance, that a node A is linked with a node B and this is, in turn, linked with a node C, being both links of a certain type. If A is then linked with node C with a different type of link, should the triangle ABC considered in the computation of the clustering coefficient?

An approximation to such generalization was the derivation of the Newman-Girvan modularity definition, which considers the dynamics of the flux of information in the network (in a statistical sense) through a Laplacian dynamics [6]. This method was extended to consider multilayer networks [7] but it is unclear whether it can be extended to networks considering *any* type of edges. It would likely be possible if, for the flux of information of interest, the different types of edges have a clear meaning on how their presence affect informational fluxes. If the links types are, however, qualitatively very distinct and have no explicit relationship with the fluxes dynamics, the application of a dynamical model would likely be meaningless.

Another alternative is given by the notion of structural equivalence. Two nodes are said to be structurally equivalent if they share the same number (and type, if the network is multidimensional) of links. This framework was originally developed for social systems [8], where the role of the nodes (social agents) is important, and it is encoded in the nodes' interactions. This notion also goes hand in hand with the Eltonian classic definition of niche, which emphasizes species function rather than species habitat [9]. We note that, since two nodes can be structurally equivalent even if they are not connected themselves, communities determined under this definition may be quite different to modules, in which members of the same module are tightly connected by definition. This latter community structure is known as disassortative mixing [10], and has received comparatively less atention than the "assortative" situation, leading to modules, perhaps with the exception of bipartite networks [11, 12]. We call to this second class of communities guilds, inspired in the ecological meaning in which species may share similar ways of exploiting resources (i.e. similar links) without necessarily sharing the same niche (not being connected themselves), emphasizing the functional role of the species [13]. Nevertheless, we note that the nodes within modules are also structurally equivalent. Therefore, the notion of structural equivalence seems to open an avenue to identify and distinguish both guilds and modules. Both types of communities can then be understood as *functional groups* —in the eltonian sense– and this is the name we adopt here. We reserve in this way the term community for a more generic use, since other types of communities beyond functional groups may exist.

In this work, we show that a modification of the community detection method developed by Ahn et al. [14], leads to the identification of two quantities we call internal and external partition densities, which allow the identification of modules and guilds, respectively. For a set of nodes joined within a community by means of their structural equivalence similarity, the partition densities quantify whether their similarities come from connections linking them with nodes outside the community (external density) or within the community (internal density). One of the main challenges in this work is to determine the structural equivalence similarity threshold above which nodes are considered to belong to the same community, a problem that, in the literature, led to definitions such as regular equivalence [8], which proves to be problematic, as illustrated in Fig. 1. Notably, our method brings absolute maximum values for internal and external partition densities, allowing us to objectively determine optimal cut-offs for the structural equivalence similarities. In addition, the method has several advantages that make it particularly suitable for analysing large networks. First, we generalized the method to consider an arbitrary number of link types, which makes it suitable for the analysis of multidimensional networks. Second, the method is deterministic, and hence it does not require costly optimality or tests for robustness [10], whose improvement has attracted much attention in recent years [15].

We call our method functionInk (functional linkage), emphasizing the fact that the number and types of links of a node determine its functional role in the network. We illustrate its use by considering complex biological examples, for which we believe the notion of functional role is particularly relevant. For instance, modules may be of interest if we aim to detect sets of species within the same trophic level, with competitive interactions within the set and other types of interactions like prey-predator or mutualistic, between sets. In this case, the functionInk method is able to accomodate the different types of interactions (predator-prey, competitive, mutualisms) in these networks. In other circumstances, we may be interested in identifying biological entities having a similar function despite not interacting themselves. This kind of community would be closer to the notion of guild than to the concept of a trophic level [13]. We show in the examples that, combining the external and internal partition densities, we are able to identify the underlying type of dominant structures of the network (either towards modules or towards guilds). Moreover, selecting the most appropriate community definition in each situation provides results that are comparable to state-of-the-art methods. This versatility in a single algorithm, together with its low computational cost to handle large networks and its ability to work in multidimensional networks, makes our method suitable for any type of complex, multidimensional network.

Results

Structural equivalence similarity in multidimensional networks

We modified and extended the method presented in Ahn et al. [14] to consider i) different types of links, where types are classified according to their qualitative attributes and ii) the different expressions defined between nodes instead of between links. The latter modification has several technical advantages. Most notably it allows us to propose two quantities, the external and internal partition densities, which we use to identify guilds and modules, respectively.

The method starts by considering a similarity measure between all pairs of nodes, that quantifies the number and type of interactions they share, shown in Fig. 1. For simplicity, we present the derivation for two types of undirected interactions (for instance positive, +, and negative, $-$), and its extension for an arbitrary number of types is presented in Methods. We call $\{i\}$ the set of N nodes and $\{e_{ij}\}$ the set of M edges in a network. We call $n(i)$ the set of neighbours of i, that can be split into different subsets according to the types of links present in the network. We split the set of neighbours linked with the node i into those linked through positive relationships, $n_+(i)$, or through negative relationships, $n_-(i)$ (see Fig. 1, we follow here a notation similar to the one presented in [14], but note that $n(i)$ stands there for neighbours irrespective of the kind of edges).

This splitting of neighbours into different subsets according to their identity, is one of the modifications of the original method. Distinguishing link types induces a division in the set of neighbours of a given node into subsets sharing the same link type, shown in Fig. 2A. More specifically, in the absence of link types we define the Jaccard similarity between to nodes i and j as:

$$S^{(\mathrm{J})}(i,j) = \frac{|n(i) \cap n(j)|}{|n(i) \cup n(j)|} \tag{1}$$

where $|\cdot|$ is the cardinality of the set (the number of elements it contains). Generalizing this expression to two attributes (see Suppl. Methods for an arbitrary number of attributes) leads to

$$S^{(\mathrm{J})}(i,j) = \frac{|n_+(i) \cap n_+(j)| \cup |n_-(i) \cap n_-(j)|}{|n_+(i) \cup n_+(j) \cup n_-(i) \cup n_-(j)|}. \tag{2}$$

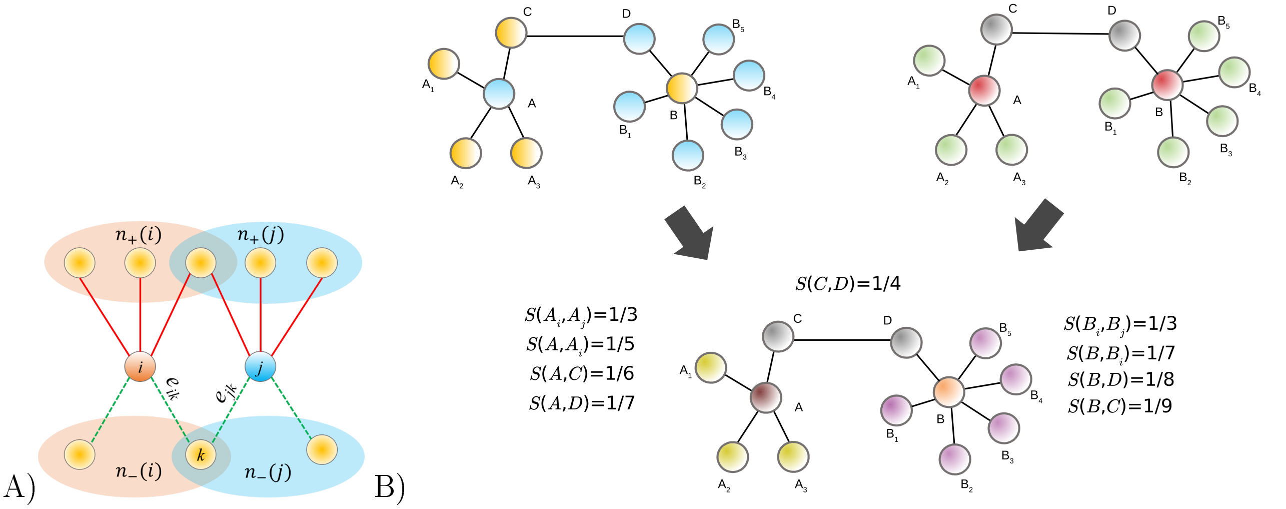

Figure 1: **Illustration of the method.** (A) The similarity between nodes i and j is computed considering the neighbours that each node has and the types of interactions linking them. In this example, continuous links stand for positive interactions, determining the set of neighbours $n_+(i)$ and $n_+(j)$, and negative ones are shown with dotted links connecting the sets $n_-(i)$ and $n_-(j)$. In Ref. [14] the similarity computed in this way would be assigned to the links e_{ik} and e_{jk}. (B) Under the blockmodelling approach, nodes that are regularly structurally equivalent are classified in two communities (blue and yellow, top-left network). The method of Guimerá and Amaral determines communities through their topological role (top-right network) having central networks (A and B nodes), peripheral (A1-A3 and B1-B5) and connectors (C and D). functionInk (bottom network) defines communities joining nodes sharing approximately the same number and type of links. All non-zero Jaccard similarities $S^{(\mathrm{J})}(n(i), n(j))$ of the example are shown. Clustering these similarities will lead to different partitions and, stopping at $S^{(\mathrm{J})} = 1/4$, would lead to communities being the intersection of those found in the above networks, highlighting the potential to identify communities whose nodes are joined through a local notion of function that encapsulates more global topological features. Figure adapted from [16].

Accounting for the weight of the edges can be made with the generalization of the Jaccard index provided by the Tanimoto coefficient [17], $S^{(\mathrm{T})}(i,j)$, presented in Methods.

We note the particular case in which i and j are only connected between them which, with the above definition, means that they do not share any neighbours. This is problematic, because we want to distinguish this situation from the one in which they do not share any nodes, for which we get $S(i,j)=0$. On the other hand, if they share a connection with respect to a third node, we want to distinguish the situation in which all three nodes are connected (a perfect transitive motif, the archetype of a module) from a situation in which they only share a neighbour but they are not themselves connected (the archetype of a guild). If we consider that a node is always a neighbour of itself, i.e. $\{i\} \in n(i)$, in both situations $S(i,j)=1$ and we cannot distinguish these cases. Therefore, we take the convention that, for a connection between two nodes $|n(i) \cap n(j)| = 1$ and $|n(i) \cup n(j)| = 2$. In Fig. 1 we illustrate the computation of this similarity with a simple example.

Identification of optimal similarity cut-offs

Once the similarity between nodes is computed, the next objective is to cluster the nodes using a similarity measure in order to identify communities (see Methods). Clustering is performed in a stepwise manner, where nodes that are increasingly dissimilar in their links are iteratively partitioned into communities. A critical question in clustering procedures is to identify the clustering step for which the optimal partition is achieved [18]. This question is often addressed by proposing a measure that monitors the clustering and that has a well defined maximum or minimum determining a threshold to stop the clustering. In [14], they proposed a quantity called the partition density, whose maxima determines the optimal clustering threshold (that we recover for completeness in Methods). We reconsider the definition of partition density because it was originally defined over edge partitions. We develop a similar measure that defines partition densities across nodes, and which adds a new dimension to the investigation of node partitioning. To develop this measure, we note that, when we join nodes into a partition, we are concluding that these nodes approximately share the same number and type of connections, but we actually do not know whether they are connected between themselves or not. We therefore redefine the partition density differentiating between the contribution of links density arising from the connections *within* the community from connections shared with respect to external nodes, i.e. *between* partitions.

Given a node i, we differentiate between those neighbours that are within the same community to which the node belongs, that we call $n^{\mathrm{int}}(i)$ (where int stands for “interior”), and those in the exterior of the community, $n^{\mathrm{ext}}(i)$, hence $n(i) = n^{\mathrm{int}}(i) \cup n^{\mathrm{ext}}(i)$ (See Fig. 2). For a singleton (a cluster of size one) $n^{\mathrm{int}}(i) = \{i\}$ and $n^{\mathrm{ext}}(i) = \emptyset$. Similarly, the set of edges $m(i)$ linking the node i with other nodes can be also split into two sets: the set linking the node with neighbours within its community $m^{\mathrm{int}}(i)$, and those linking it with external nodes $m^{\mathrm{ext}}(i)$.

Given a partition of nodes into T communities, our method identifies, for each community, the total number of nodes within it, n_c^{int}, and the total number of links connecting these nodes m_c^{int}. In addition, it computes the total number of nodes in other communities with connections to the nodes in the community C (n_c^{ext}) through a number of links m_c^{ext}. Clearly, to identify n_c^{ext} neighbours requires at least n_c^{ext} links, and the number of links in excess, $m_c^{\mathrm{ext}} - n_c^{\mathrm{ext}}$, contributes to the similarity of the nodes in the community through external links. Therefore, we quantified the fraction of links in excess out of the total possible number $(m_c^{\mathrm{ext}} - n_c^{\mathrm{ext}})/n_c^{\mathrm{ext}}\left(n_c^{\mathrm{int}} - 1\right)$. The weighted average of this quantity through all communities leads to the definition of external partition density:

$$D^{\mathrm{ext}} = \frac{1}{M}\sum_c \frac{m_c^{\mathrm{ext}}}{2}\frac{(m_c^{\mathrm{ext}} - n_c^{\mathrm{ext}})}{n_c^{\mathrm{ext}}\,(n_c^{\mathrm{int}} - 1)}, \tag{3}$$

where M is the total number of edges. We follow now a similar reasoning now considering the internal links, but we should acknowledge that in a community created linking nodes through the similarity measure we propose, it may happen that $n_c^{\mathrm{int}} > 0$ even if $m_c^{\mathrm{int}} = 0$. Therefore, any link is considered a link in excess, leading to the following expression for the internal partition density:

$$D^{\mathrm{int}} = \frac{1}{M}\sum_c m_c^{\mathrm{int}} \frac{2m_c^{\mathrm{int}}}{n_c^{\mathrm{int}}(n_c^{\mathrm{int}} - 1)}. \tag{4}$$

Finally, we define the total partition density as the sum of both internal an external partition densities:

$$D^{\mathrm{total}} = D^{\mathrm{int}} + D^{\mathrm{ext}},$$

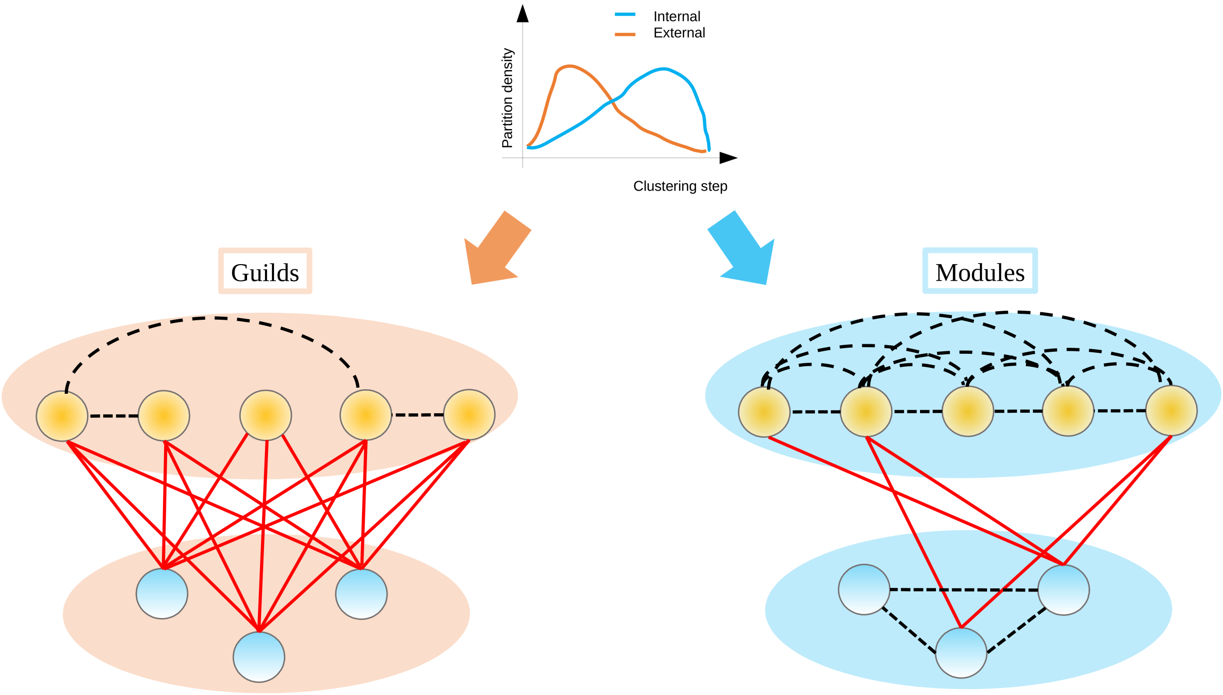

Figure 2: **Definition of guilds and modules.** For each set of nodes belonging to the same community c (n_c^{int}, nodes within the same shaded area) we consider the number of links within the community (black dashed links, called m_c^{int} in the main text) out of the total number of links of the community, to compute the internal partition density. We identify modules as those communities that maximize the internal partition density (right figure). We then compute the density of links connecting nodes external to the community (m_c^{ext}, solid red lines linking nodes in other partitions) out of the total to estimate the external partition density. We call guilds to those communities obtained maximizing this quantity.

and hence, if all the fractions in D^{int} and D^{ext} are equal to one, i.e. all possible links in excess are realized, D^{total} equals to one. Since at the beginning of the clustering the communities have a low number of members, most of the contribution towards D^{total} comes from D^{ext} while, in the last steps where the communities become large, D^{int} will dominate. All three quantities will reach an absolute value along the clustering and, if one of them clearly achieves a higher value, it will be indicative that one kind of functional group is dominant in the network. If that is the case, the maximum of D^{total} (which is always larger or equal to $\max(\max(D^{\text{int}}), \max(D^{\text{ext}}))$), will be close to one of these. If neither D^{ext} nor D^{int} clearly dominates, D^{total} will peak at an intermediate step between the two partial partition densities, suggesting that this intermediate step is the best candidate of the optimal partition for the network, and communities determined at this intermediate point will be called, generically, functional groups.

Plant-pollinator networks

To illustrate the use of the method we start analysing a synthetic example. In ecological systems, species are often classified into communities according to their ecological interactions, such as in mutualistic networks of flowering plants and their animal pollinators. These networks are characterized by intra and interspecific competition within the pool of plants and within the pool of animals, and by mutualistic relationships between plants and animals, leading to a bipartite network.

To investigate the performance of our method and, in particular, the influence of the topological properties into the partition density measures, we generated a set of artificial mutualistic networks with diverse topological properties, following the method presented in [19]. For the mutualistic interactions, we focused on two properties: the connectance κ_{mut}, which is the fraction of observed interactions out of the total number of possible interactions, and the nestedness ν as defined in [20] (see Methods), which codifies the fraction of interactions that are shared between two species, averaged over all pairs of species. We selected these measures for their importance in the stability-complexity debate in mutualistic systems [19], and the similarity between thenestedness (which, in the definition we adopt here, represents the mean ecological overlap between species) and the notion of structural equivalence we considered. For the competition matrices, we considered random matrices with different connectances, κ_{comp}, since it is difficult to estimate direct pairwise competitive interactions experimentally, and it is frequently modelled with a mean field competition matrix.

In all networks, the set of plants and animals are joined in the very last step of the clustering irrespective of the clustering method used, indicating that our similarity measure is appropriate and that the method is robust with respect to the clustering method selected. As expected, the curves monitoring the external and internal partition densities depends on the properties of the networks. We illustrate this finding in Fig. 3, where we have selected two networks with contrasting topological properties. One of the networks has high connectance within the pools and low connectance and nestedness between the pools. The internal partition density peaks at the last step minus one (i.e. where the two pools are perfectly separated) consistent with the definition of modules, where the intra-modules link density is higher than the inter-modules link density. On the other hand, the second network has intra-pool connectance equals to zero, and very high connectance and nestedness between the pools (see Fig. 3). We selected a $\kappa_{\text{comp}} = 0$ for simplicity in the network representation, but similar results are obtained for low values of κ_{comp}, see for intance Suppl. Fig. 10. In this second network (see Fig. 3, right panel), only the external partition density peaks and, at the maximum, the communities that we identified clearly reflect the structural equivalence of the nodes members in terms of their connectance with nodes external to the group, as we expect for the definition of guilds. The ecological information retrieved for guilds is clearly distinct from the information retriedved for the modules, being the former related to the topology of the network connecting plants and animals. We observe that guilds identify specialist species clustered together, which are then linked to generalists species of the other pool: a structure typical of networks with high nestedness.

The method identified several interesting guilds and connections between them. For instance, generalists Plant 1, Animal 1 and Animal 2 (and to a lesser extent Plant 2) have a low connectivity between them but, being connected to many specialists, determine a region of high vulnerability, in the sense that a directed perturbation over these species would have consequences for many other species. This is confirmed by the high betweeness of these nodes (proportional to the size of the node in the network). In addition, the algorithm is able to identify more complex partitions of nodes into clusters. As an example of this, Animal 16 (torquoise) is split from Animals 10 and 11 (cyan), which form a second cluster, and from Animals 15, 18 and 19 (light pink) that are joined into a third cluster, despite of the subtle connectivity differences between these six nodes. Finally, it also detects clusters of three or more species that have complex connectivity patterns which, in this context, may be indicative of functionally redundant species (e.g. red and blue clusters).

Examples with other intermediate properties are analyzed in the Suppl. Figs. 8 and 9. Broadly speaking, either the internal or the total partition density maximum peaks at the last step minus one, allowing for detection of the two pools of species. Nevertheless, the method fails to find these pools if the within-pools connectance is very low, since the network becomes highly dissasortative (see Suppl. Fig. 10). The relative magnitude of the external vs. internal partition density depends on the connectance between the pools of plants and animals and on the connectance within the pools, respectively (see Suppl. Fig. 8). Interestingly, networks for which the nestedness is increased being the remaining properties the same, generated an increase in the external partition density (see Suppl. Fig. 9). These examples illustrate how the external partition density is sensitive to complex topological properties, in particular to an increase in the dissassortativity of the network.

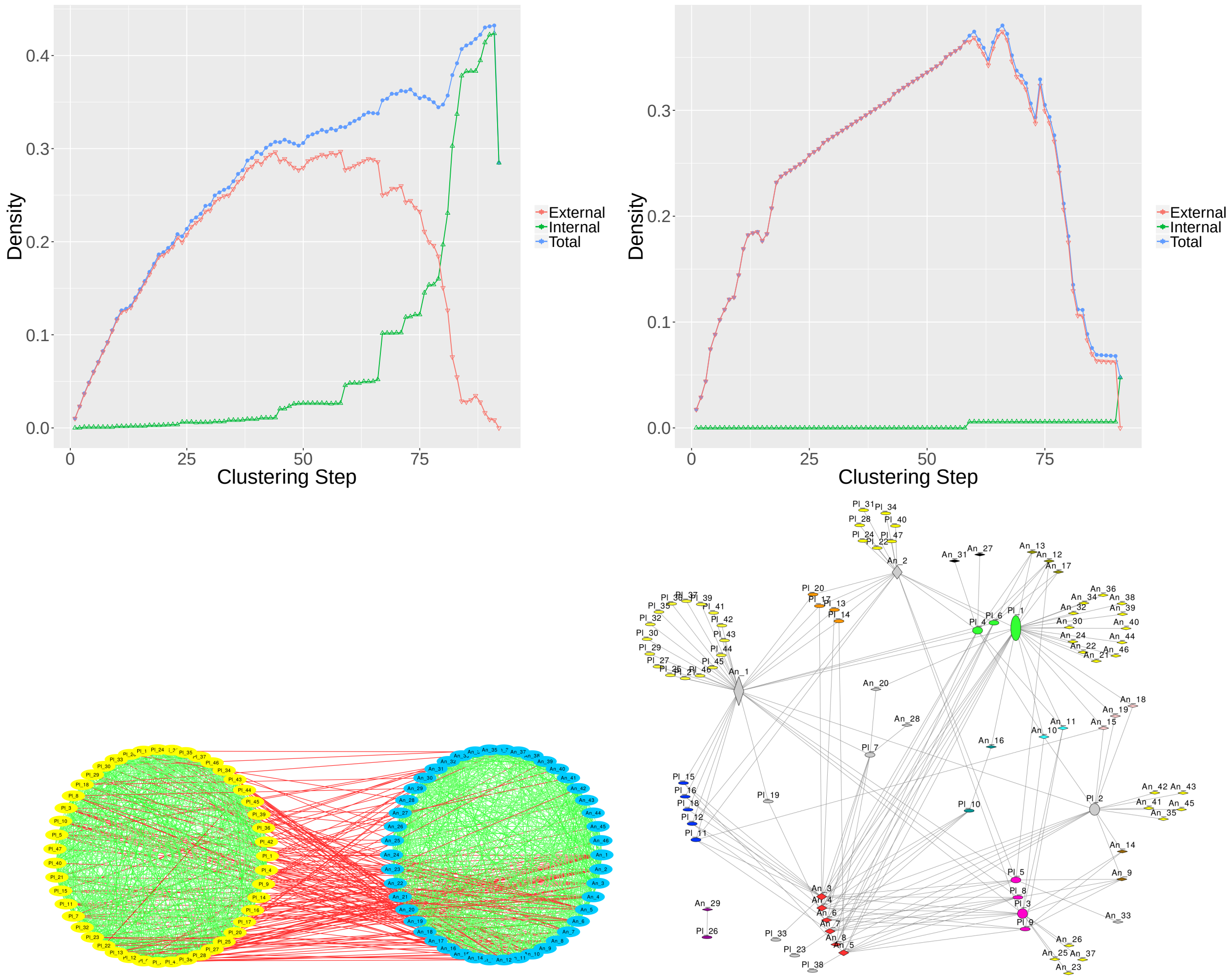

Figure 3: **Synthetic mutualistic networks**. (Top left) Partition densities for a network with $\kappa_{\text{comp}} = 0.5$, nestedness $\nu = 0.15$ and $\kappa_{\text{mut}} = 0.08$ and (top right) for a network with $\kappa_{\text{comp}} = 0$, nestedness $\nu = 0.6$ and $\kappa_{\text{mut}} = 0.08$. The high density of competitive links in the first network makes the internal partition density dominate, leading to two modules representing the plant-pollinator pools (bottom left network), while reducing the density of competitive links to zero in the second network makes the external partition density to dominate, finding guilds (bottom right). In the networks, plants are labelled “Pl” and animals are labelled “An”. Nodes are coloured according to their functional group. In the network finding guilds (bottom right), specialist species are yellow, single species clusters are grey, and the size of the nodes is proportional to their betweeness.

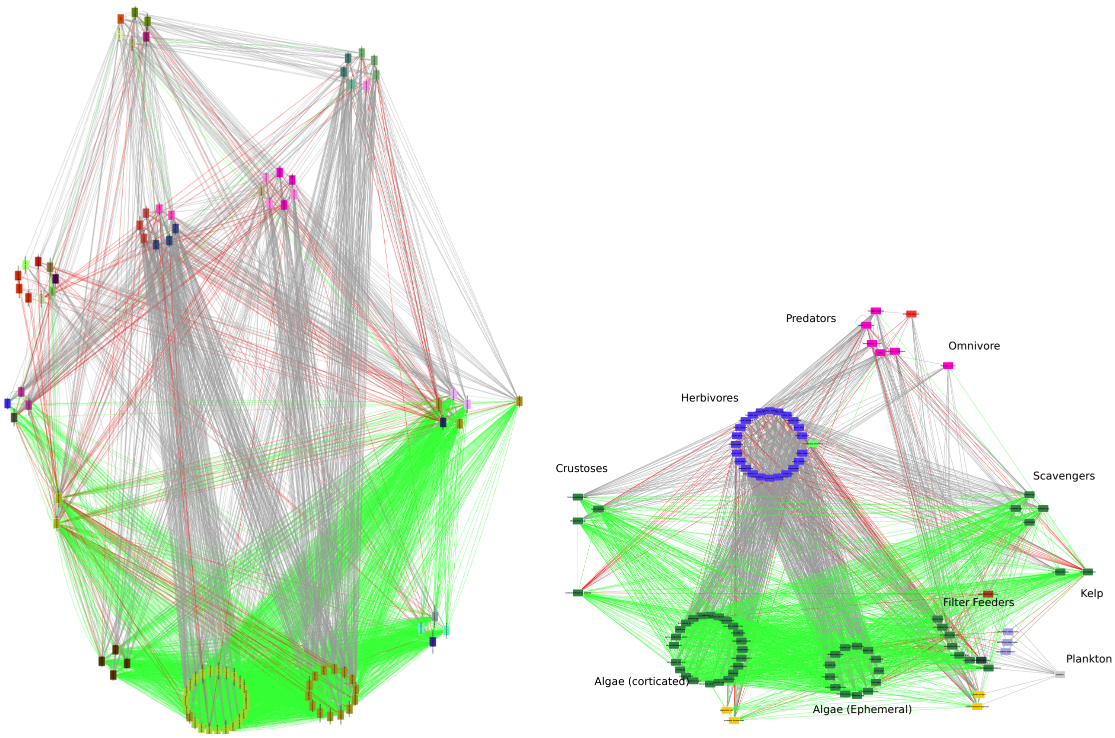

Figure 4: **Analysis of a trophic network.** Trophic networks with links representing trophic (grey), non-trophic positive (red), and negative (green) interactions. (Left) Nodes are grouped according to the classification found in [21] (reference classification), and coloured by the guilds found with functionInk at the maximum of the external partition density. (Right) Nodes are grouped according to the trophic levels and coloured by the modules found by functionInk (see Main Text for details). The modules separate the three main levels, predators, herbivores and basal species, although it separates these into finer subgroups (filter feeders) and plankton, which is an orphan module.

1.1 Trophic networks

We tested our method in a comprehensive multidimensional ecological network of 106 species distributed in trophic layers with approximately 4500 interactions, comprising trophic and non-trophic interactions (approximately 1/3 of the interactions are trophic) [21]. This network was analysed looking for communities extending a stochastic blockmodelling method [10] to deal with different types of interactions [21]. The estimation of the parameters of the model through an Expectation-Maximization algorithm requires a heuristic approximation, and hence it is needed to test the robustness of the results found. Here we show that, in this example, our method is comparable with this approximation, and it has the advantage of being deterministic. Moreover, the simplicity of the method allows us to handle large networks with arbitrary number of types of links and to evaluate and interpret the results, as we show in the following.

Our method finds a maximum for the internal density when there are only three clusters. Previous descriptions of the network identified three trophic levels in the network (Predators, Herbivores and Basal species). The latter are further subdivided into subroups like (e.g. Kelps, Filter feeders), and there are some isolated groups like one Omnivore and Plankton. To match these subgroups we observed that the total partition density reaches a maximum close to the maximum of the external partition density (step 69) and maintains this value along a plateau until step 95 (see Suppl. Fig. 11). We analysed results at both clustering thresholds finding that, at step 95, we obtain modules with a good agreement with the trophic levels, shown in Fig. 4. On the other hand, at step 69 we find a larger number of communities some of which fit the definition of modules and others the definition of guilds.

To shed some light on the information obtained from this second network, we compared the classification obtained by Kefi et al. [21] (in the following reference classification) and our method, shown in Fig. 4. We computed a number of similarity metrics between the classification we obtained at each step of the agglomerative clustering with functionInk and the reference classification (see Methods). In Fig. 5, we show that the similarity between both classifications is highly significant (Z-score > 2.5) and is maximized when the external partition density is also maximized, i.e. at step 69. This is particularly apparent for the Wallace 01, Wallace 10 and Rand indexes (see Fig. 7 and Suppl. Fig. 12). Clusters in the reference classification were also interpreted as functional groups in the same sense proposed here [21] , supporting our arguments to use the external partition density as a quantity to detect guilds.

Nevertheless, there are some discrepancies between both classifications. In particular, although there is a complete correspondence between the two largest clusters in both classifications, there are a number of intermediate clusters in the reference classification whose members are classified differently in our method. To illustrate these discrepancies, we plotted a heatmap of the Tanimoto coefficients of members of four clusters of intermediate size containing discrepancies, showing their membership in both the reference and the functionInk classification with different colours (see Fig. 4). The dendrograms cluster rows and columns computing the Euclidean distance between their values. Therefore, this illustration is very similar to functionInk, and the clusters must be consistent, representing a powerful way to visually inspect results. Indeed, the blocks found in the heatmap are in correspondence with functionInk clusters, as expected, but we observe some discrepancies with the reference classification. For instance, the cluster found by the reference classification containing several *Petrolishtes* species, joins species that have low similarity regarding the number and type of interactions as measured by the Tanimoto coefficients, while functionInk joins together the three species with high similarity, leaving aside the remainder species. Of course, we cannot discard that the method used in the reference classification captures other properties justifying the differences. But it is immediately apparent the advantages provided by the simplicity our method, which permits validation through visual inspection of the consistency of the classification.

1.2 Microbial networks

We discuss a last example of increasing importance in current ecological research, which is the inference of interactions among microbes sampled from natural environments. We considered a large matrix with more than 700 samples of 16S rRNA operative taxonomic units (OTUs) collected from rain pools (water-filled tree-holes) in the UK [23, 24] (see Suppl. Methods). We analysed $\beta-$diversity similarity of the communities contained in the matrix with the Jensen-Shannon divergence metric [25], further classifying the communities automatically, leading to 6 disjoint clusters (see Methods). Next, we inferred a network of significant positive (co-occurrences) or negative (segregations) correlations between OTUs using SparCC [26], represented in Fig. 6 (see Methods). Applying functionInk to the network of inferred correlations, we aimed to understand the consistency between the results of functionInk (modules and guilds) and the $\beta-$diversity-classes. The rationale is that, by symmetry, significant co-occurrences and segregations between OTUs should reflect the similarity and dissimilarity between the communities, hence validating the method.

Contrasting with the trophic network analysed in the previous example, the external partition density achieves a low relative value and brings a poor reduction of the complexity of the network, suggesting optimal clustering after just 22 clustering steps (see Suppl. Fig. 13). The internal partition density achieves a higher value, hence suggesting that modules are more relevant than functional groups in this network. Two large modules are apparent, see Fig. 6, with a large number of intracluster co-occurrences (continuous links) and interclusters segregations (dotted links). Note that this is quite different to what is found in macroscopic trophic networks, where pools of species (e.g. prey) have within module competitive (segregating) interactions, while between-modules interactions can be positive (for predators) or negative (for prey). In addition, the total partition density peaks at a much higher value and it seems to be able to split some of the large modules into smaller motifs, some of which were identified as guilds and have clearly distinctive connectivities, that we analyse in further detail. Since some of these motifs have characteristics closer to those of guilds while others are closer to modules, we refer to them generically as functional groups.

There is reasonable agreement between the functional groups found at the maximum of the total partition density and the $\beta-$diversity-classes shown in Fig. 6. Since it was shown in [23] that the $\beta-$diversity-classes might be related to a process of ecological succession driven by environmental variation, the functional groups are likely driven by environmental preferences rather than by ecological interactions, likely explaining the large number of positive co-occurrences. This speaks against a näive interpretation of correlation networks in microbial samples as ecological interactions, unless environmental preferences are controlled [27]. However, the detection of networks complements the information that $\beta-$diversity-classes provides, since it is possible to individuate the key players of these classes. For instance, only two OTUs from the green functional group have an important number of

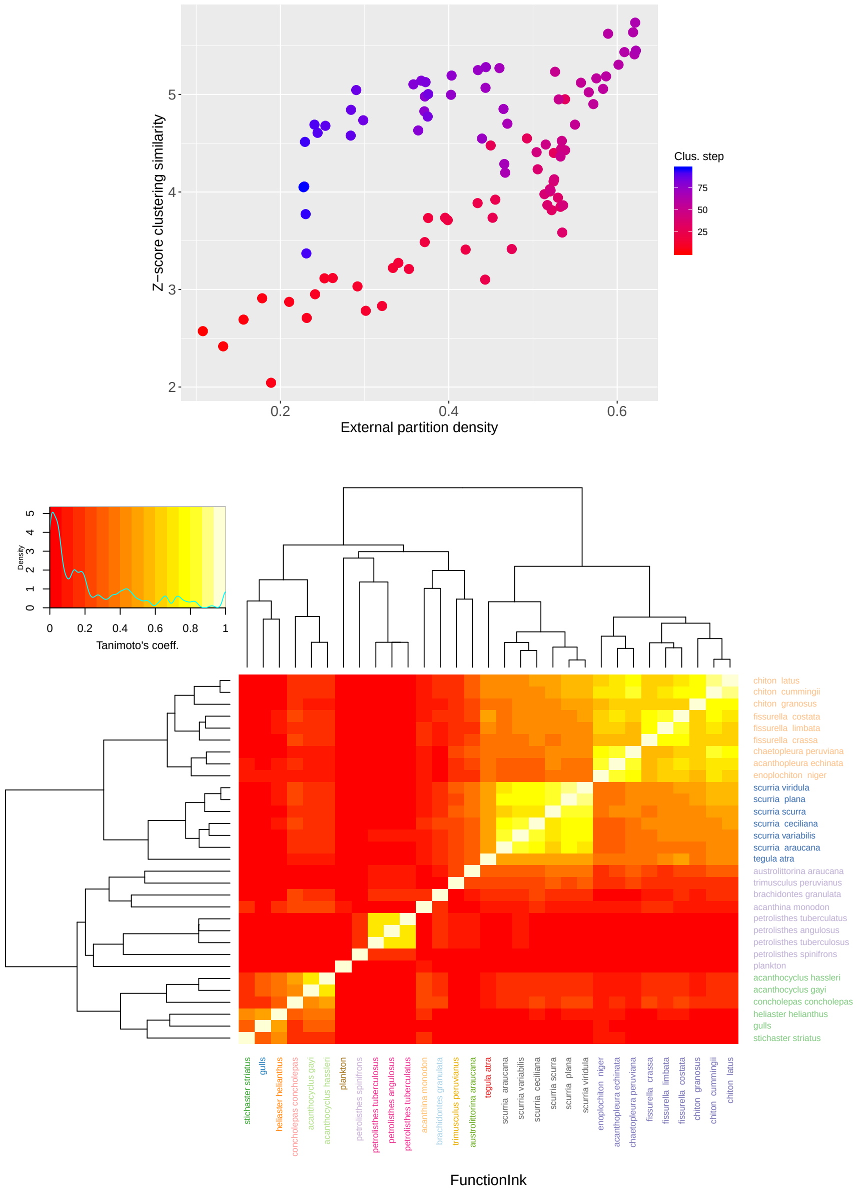

Figure 5: **Analysis of guilds.** (Top) Z-score of the Wallace 10 index [22], measuring the similarity between the reference classification and the functionInk method at each clustering step. The similarity with the reference classification is maximized around the maximum of the external partition density. (Bottom) Comparison of clusters 1, 4, 7 and 9 in the reference classification, whose members were classified differently by functionInk. Colours in the names of species in rows (columns) represent cluster membership in the functionInk (reference) classifications. The heatmap represents the values of the Tanimoto coefficients, and the dendrograms are computed using Euclidean distance and clustered with complete linkage. The heatmap blocks of high similarity are in some cases inconsistent with the reference classification. Given the simplicity of the interpretation of these coefficients, it is difficult to explain the clustering of distant members by the reference classification.

co-occurrences with members of the red functional group, and only one of them has a significant segregation with respect to a member of the blue functional groups. These two highly abundant segregating OTUs are *Pseudomonas putida* and *Serratia fonticola*, both of which were shown to dominate two of the $\beta-$diversity-classes [23]. The functional groups hence allow us to easily identify their most important partners and, more in general, to analyse in detail how clusters of communities are structured.

Discussion

We presented a novel method for the analysis of multidimensional networks, with nodes containing an arbitrary number of link types. The method extends and generalizes the method proposed by Ahn et al. [14] and presents a number of advantages with respect to other approximations. First, we developed the method to work with nodes instead of with links (which was the case for the original method [14]), which we find more intuititive, and allows the interpretation of the communities and its analysis with current visualization software. From an ecological perspective, we were also interested in the functional role of the nodes. In this sense, the definition of a species function in the network is straightforward for nodes, adopting the definition of structural equivalence used in social networks. This notion underlies both the similarity measure definition and the rationale behind both the clustering and our definition of external partition density. Working with this definition we defined two measures of nodes partitioning. While the internal partition density is very similar to the definition provided in [14] (see Suppl. Methods), the external partition density brings a new dimension, being similar in spirit to the search of structural equivalent clusters in social networks [8]. This allowed us to propose a clear differentiation between modules (determined by the maximum of the internal partition density) and guilds (determined by the maximum of the external partition density). Although the method might not be able to achieve the generality of approximations such as the mixture models proposed in [10], which aims to find any arbitrary structure in the network, such approximations are far from being generalized to an arbitrary number of link types, as we presented here. In addition, these are statistical approximations requiring either heuristics to find a solution for the parameters –and hence a unique optimal solution is not guaranteed–, or a computationaly costly sampling of the parameter space. Our method relies on a deterministic method whose results are easily inspected, given the simplicity of the similarity metric used and the partition density functions proposed to monitor the optimal clustering.

Beyond these technical advantages, we illustrated the versatility of functionInk using several ecological examples. The relative value between the internal and external partition density, immediately yields information on whether the network is dominated by modules, guilds, or intermediate structures. This allows for increasing flexibility in the analysis of the networks, and for a more nuanced interpretation of network structure and species' roles in the ecosystem. For both mutualistic and trophic networks, the internal partition density correctly finds the trophic layers, justifying the success of the original method [14]. Our extension recovered the functional groups as determined by Kefi et al. [21] through the external partition density, and the visual inspection reflects a good consistency with the definition we proposed for functional groups in terms of structural equivalence. Moreover, in the mutualistic networks, we showed that the functional groups discovered in this way was sensitive to changes to high-order topological properties such as the nestedness.

The analysis of the microbial network was dominated by modules rather than guilds. Interestingly, these modules had intra-cluster positive correlations, contrary to what would be expected in a macroscopic trophic network, where competitive interactions would be dominant between members of the same trophic layer. We selected in this example for further exploration the functional communities found at the maximum of the total partition density, with some groups having properties closer to those of guilds and others closer to modules. The communities that we identified were in good agreement with the functional communities found using $\beta-$diversity similarity [23], supporting the consistency of the method. Interestingly, it was found in [23] that similar $\beta-$diversity-classes were driven by environmental conditions. Although co-occuring more often in the same environment may be indicative of a higher probability of interaction [28], the most economical hypothesis is that they co-occur because they share similar environmental preferences, and hence it cannot be disentangled the type of interaction (if any) unless the environmental variables are under control.

functionInk requires the computation of Jaccard or Tanimoto coefficients, whose computational cost scales as N^2, being N the number of nodes. However, the similarity coefficients only need to be computed once, and then the clustering with different methods and posterior analysis are at most order N, making the method suitable for large networks. The method is available in the address (HTTPS://GITHUB.COM/APASCUALGARCIA/FUNCTIONINK) and, importantly, although we developed it with ecological networks in mind, it can be applied to any kind of

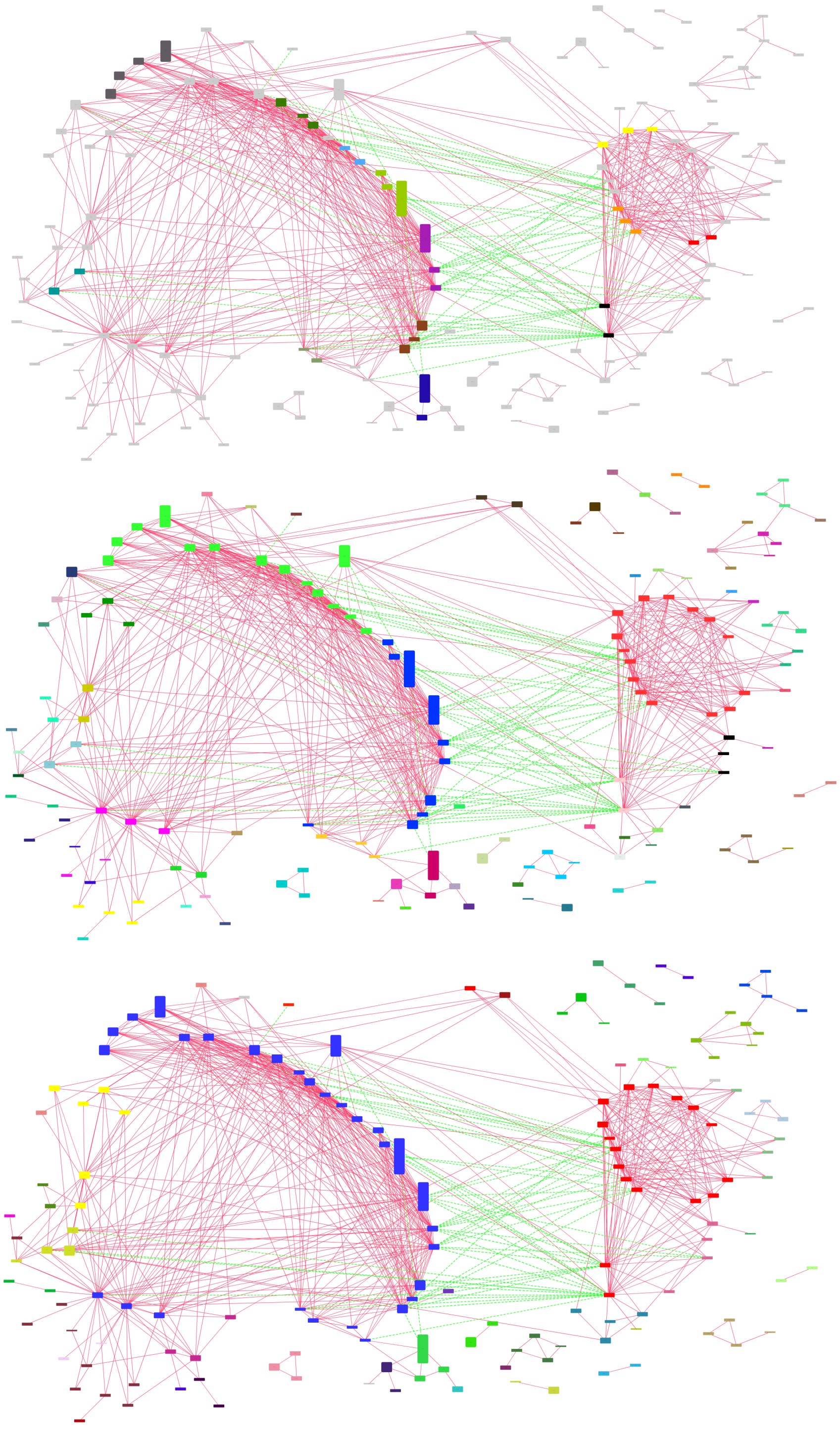

Figure 6: **Comparison of functional groups in the microbial network.** Network of significant co-occurrences (continuous links) and segregations (dotted links) at the species level (nodes). Colours indicate functional group membership, which was determined by the maximum of the external (top), total (middle) and internal partition densities (bottom). Orphan nodes are coloured grey in the top figure for clarity. The higher value of the internal partition density (see Suppl. Fig. 13) suggests that a modular structure is the more appropriate to describes the functional groups. This is confirmed by the low number of guilds (top figure) and the good agreement between the global topological structure and the modules (bottom figure).

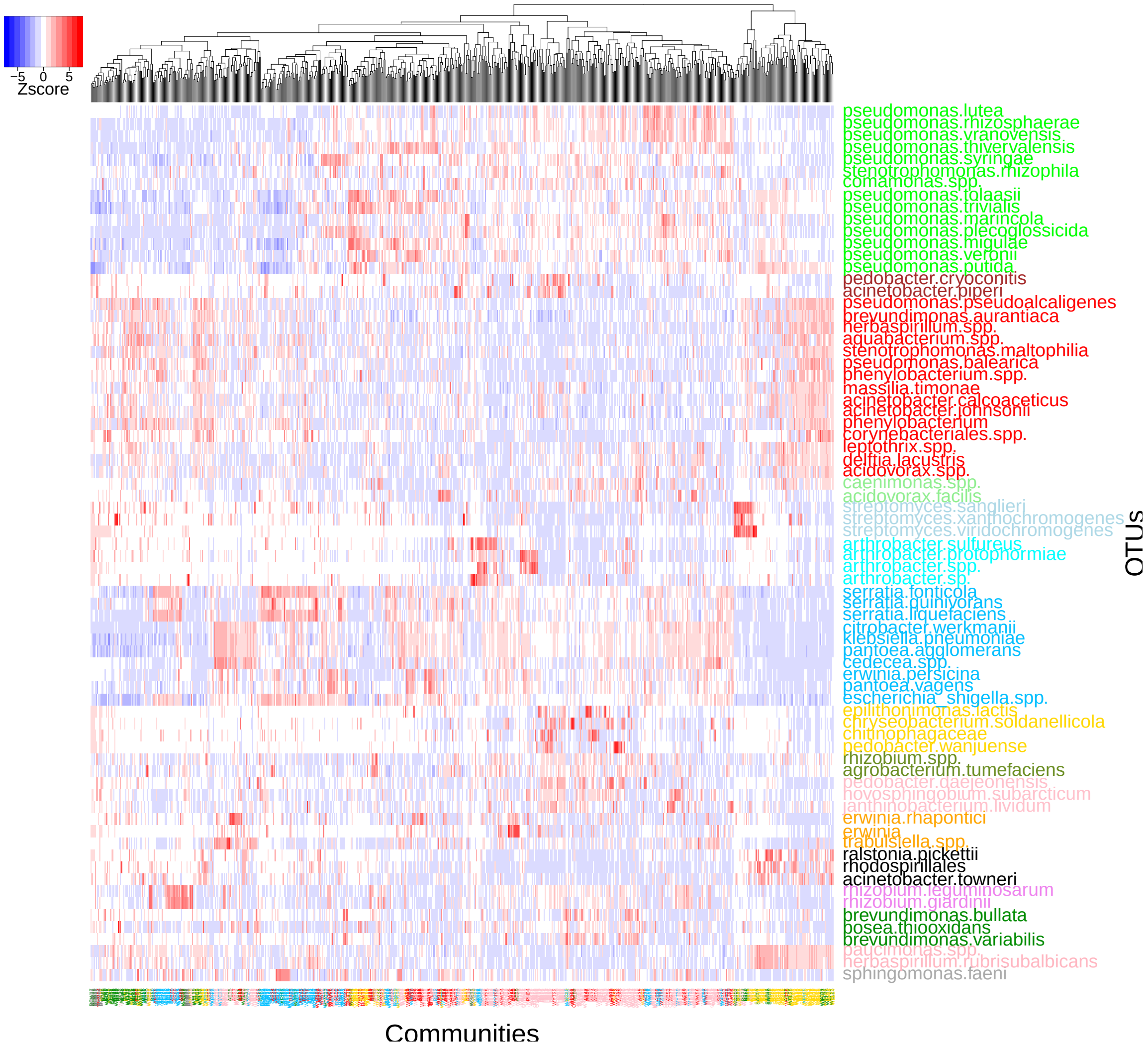

Figure 7: **Analysis of a microbial network.** Heatmap representing the z-score of the log-transformed abundances of the OTUs (see Methods). Species are coloured according to their functional group membership obtained at the maximum of the total partition density. Samples are coloured according to one of the six community classes found in [23]. Heatmap blocks show segregation and co-occurrence between modules, further mapping the $\beta-$diversity classes.

network.

Methods

Generalization of the Jaccard and Tanimoto coefficients to an arbitrary number of link types

Consider a network with a set $\{i\}$ of N nodes and a set $\{e_{ij}\}$ of M links. These links are classified into Ω types labelled with the index $\alpha = (1, ..., \Omega)$. These types would typically account for differential qualitative responses of the nodes properties due to the interactions. For example, if we consider that the nodes are species and the property of interest is the species abundances, the effect of cooperative or competitive interactions on the abundances can be codified using two different types of links: positive and negative. If these relations are inferred through correlations between abundances, we could use a quantitative threshold (for instance a correlation equal to zero) to split the links into positive and negative correlations. In general, we may use a number of qualitative attributes or quantitative thresholds in the weights of the links to determine different types of links.

We call $n(i)$ the set of neighbours of i, and we split these neighbours into (at most) Ω different subsets according to the types of links present in the network. The Jaccard coefficient defined in 2 can be extended (already considering similarities between nodes) as:

$$S^{(\mathrm{J})}(n_i, n_j) = \frac{\bigcup_{\alpha=1}^{\Omega} |n_\alpha(i) \cap n_\alpha(j)|}{\bigcup_{\alpha=1}^{\Omega} |n_\alpha(i) \cup n_\alpha(j)|}. \tag{5}$$

Accounting for the weight of the edges can be made with the generalization of the Jaccard index provided by the Tanimoto coefficient [17]. We first introduce the method without differentiating between different types of neighbours. Consider the vector $\boldsymbol{a}_i = \left(\tilde{A}_{i1}, \ldots, \tilde{A}_{iN}\right)$ with

$$\tilde{A}_{ij} = \frac{1}{k_i} \sum_{i' \in n(i)} w_{ii'} \delta_{ij} + w_{ij} \tag{6}$$

where w_{ij} is the weight of the edge linking the nodes i and j, $k_i = |n(i)|$ and δ_{ij} is the Kronecker's delta ($\delta_{ij} = 1$ if $i \neq j$ and zero otherwise). Determining the quantity $W_{ij} = \boldsymbol{a}_i \boldsymbol{a}_j = \sum_k \tilde{A}_{ik} \tilde{A}_{kj}$, the Tanimoto similarity is defined as

$$S^{(\mathrm{T})}(e_{ik}, e_{jk}) = \frac{W_{ij}}{W_{ii} + W_{jj} - W_{ij}}. \tag{7}$$

Working with link types requires a generalization of the above expression. Consider for the moment two types related with a positive $w_{ij} > 0$ or a negative $w_{ij} < 0$ weight of the links. The term $\tilde{A}_{ii} = 1/k_i \sum_{i'} w_{ii'}$ is the average of the strengths of the edges connected with node i, and it is desiderable to keep this meaning when considering two types to properly normalize the Tanimoto similarity. This is simply achieved redefining $\tilde{A}_{ij}$ as

$$\tilde{A}_{ij} = \frac{1}{k_i} \sum_{i' \in n(i)} \mathrm{abs}(w_{ii'}) \delta_{ij} + w_{ij}. \tag{8}$$

On the other hand, the similarity is essentially codified in the term W_{ij} that we now want to redefine to account for two types of interactions in such a way that only products $\tilde{A}_{ik} \tilde{A}_{kj}$ between terms with the same sign contribute to the similarity. This is achieved with the following definition, which generalizes the Tanimoto coefficient

$$W_{ij} = \sum_k \tilde{A}_{ik} \tilde{A}_{kj} \delta(\mathrm{sgn}(\tilde{A}_{ik}) - \mathrm{sgn}(\tilde{A}_{kj})) \tag{9}$$

where $\mathrm{sgn}(\cdot)$ is the sign function and $\delta(a - b)$ is the Dirac delta function ($\delta(a - b) = 0$ if $a \neq b$). Generalizing to an arbitrary number of types can be achieved by defining a variable μ_{ij} that returns the type of the link, i.e. $\mu_{ij} = \alpha$ with α being a factor variable which, for the example of positive and negative links, is codified by the sign of the links' weight. We finally generalize the expression 9 as follows

$$W_{ij} = \sum_k \tilde{A}_{ik}\tilde{A}_{kj}\delta(\mu_{ik} - \mu_{kj}). \tag{10}$$

Finally, the generalization of the external and internal partition densities to consider multiple types of links simply requires us to correctly classify the neighbours of each node accounting for the different types $n(i) = \bigcup_{\alpha=1}^{\Omega} n_\alpha^{int}(i) \cup \bigcup_{\alpha=1}^{\Omega} n_\alpha^{ext}(i)$. Similarly, the set of edges $m(i)$ linking the node i with other nodes must be also split into sets according to the different types $m(i) = \bigcup_{\alpha=1}^{\Omega} m_\alpha^{int}(i) \cup \bigcup_{\alpha=1}^{\Omega} m_\alpha^{ext}(i)$. The expressions for the internal and external partition densities remain otherwise the same.

Original definition of partition density

For completenes, we present the definition of partition density presented in [14]. In short, the method starts building a similarity measure between any pair of links sharing one node in common. Two links will be similar if the nodes that these two links do not share have, in turn, similar relationships with any other node, shown in Fig. 1. From this similarity measure, edges are clustered and an optimal cut-off for the clustering is found monitoring a measure called *partition density* (which in this paper we call *internal partition density*). The optimal classification found at the cut-off, determines groups of links that are similar because they connect nodes that are themselves similar in terms of their connectivity. Therefore, the nodes are classified indirectly, according to the groups that their respective links belong, and a node may not belong to a single community but to several communities if its links belong to different clusters. This is claimed to be an advantage with respect to other methods (in particular for high density networks) as membership to a single cluster is not enforced. At every step of the clustering it is obtained a partition $P = P_1, ..., P_C$ of the links into C subsets. For every subset, the number of links is $m_c = |P_c|$ and the number of nodes that these edges are linking is $n_c = | \cup_{e_{ij} \in P_c} \{i, j\}|$. The density of links for the cluster C is then

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)} \tag{11}$$

where the normalization considers the minimum $(n_c - 1)$ and maximum $(n_c(n_c - 1)/2)$ number of links that can be found in the partition. The diference with respect to Eq. 4, is that a term $(n_c^{int} - 1)$ is now subtracted. The reason is that, in Ahn *et a*l. method, clustering with links implies that two nodes in the same cluster must share links. But, according to our definition of function, two nodes may be structurally equivalent even if there is no interaction between them.

The partition density D is then given by the average of the density of links for all the partitions, weighted by the number of links

$$D = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \tag{12}$$

where M is the total number of links. It was shown that when using agglomerative clustering, this function achieves a maximum which determines the optimal partition [14].

Clustering algorithm

After computing the similarity between nodes with the method presented in the Results, the algorithm clusters nodes using one of three hierarchical clustering algorithms: average linkage [29], single linkage and complete linkage. Starting from each node being a separate cluster, at each step t all algorithms join the two most similar clusters A and B, and compute the similarity between the new combined cluster and all other clusters C in a way that depends on the clustering algorithm.

Single linkage is the most permisive algorithm, because the similarity it assigns to the new cluster is the maximum similarity between the two clusters joined and clusters C:

$$S^{t+1}(AB, C) = \max\left(S(A, C), S(B, C)\right).$$

where t labels the step of the algorithm, A and B are the clusters that are joined, AB denotes the new composite cluster, and C is any other cluster. On the other hand, complete linkage is the most restrictive, assigning the minimum similarity

ν	κ_{mut}	κ_{comp}
0.15	0.08	0
0.35	0.16	0.15
0.6	0.28	0.5

Table 1: **Topological properties of the bipartite networks analysed.** Different combinations of nestedness (ν), intra-pools connectance κ_{comp} and inter-pools connectance κ_{mut} were analysed.

$$S^{t+1}(AB, C) = \min\left(S(A, C), S(B, C)\right).$$

Finally, average linkage assigns an intermediate value computed as the weighted average similarity with the two joined clusters

$$S^{t+1}(AB, C) = \frac{n_A S^t(A, C) + n_B S^t(B, C)}{n_A + n_B}$$

being n_A and n_B the number of elements that A and B contain, respectively. Identification of the two pools of plants and animals is indendent of the clustering method used, but the maximum of the external partition density is achieved earlier for single linkage and later for complete linkage; we found a good compromise between the number and the size of the clusters working with average linkage, but the clustering method could be selected according to information known from the links. In our experience, single linkage is easily dominated by the giant cluster in high density networks in which modules are prevalent (rather than for guilds). The appropriate clustering method should be guided by the research question. For instance, if gene homology is explored, it is probably more appopriate to use single linkage (as a relative of one gene's relative is also its relative, i.e. transitivity is automatically fulfilled [30]). On the other hand, if we analyse well-differentiated functional similarity, it might be more appropriate to be conservative and use complete linkage.

Plant-pollinator networks and topological properties

We selected six plant-pollinator networks artificially generated in [19] with known topological properties, summarized in Table. We consider as topological properties the connectance (fraction of links) of the mutualistic matrix, the connectance of the competition matrices, and the definition of nestedness provided in [20]. Given a mutualistic matrix $A^{(\text{P})}_{ik}$ representing presence-absence of interaction between the set of plants, indexed by i, and the set of animal species, indexed by k, we compute the degree of a species as $n^{(\text{P})}_i = \sum_k A^{(\text{P})}_{ik}$ (see Ref. [20] in Supplementary Material). A similar definition would apply for animals $n^{(\text{A})}_k = \sum_i A^{(\text{P})}_{ik}$. Next we define the ecological overlap between two species of plants i and j as the number of insects that pollinate both plants:

$$n^{(\text{P})}_{ij} = \sum_k A^{(\text{P})}_{ik} A^{(\text{P})}_{jk},$$

a definition that is equivalent to the Jaccard similarity used in this work. Summing over every pair of plants and normalizing leads to the definition of nestedness:

$$\nu^{(\text{P})} = \frac{\sum_{i<j} n^{(\text{P})}_{ij}}{\sum_{i<j} \min(n^{(\text{P})}_i, n^{(\text{P})}_j)}.$$

A symmetric definition applies for animals, so we take as final definition of nestedness $\nu = \max(\nu^{(\text{P})}, \nu^{(\text{A})})$.

Trophic networks

We downloaded the network and metadata provided in [21] and compared the clusters found with those obtained by functionInk. After computing the Tanimoto coefficients as explained above, we cluster the nodes and retrieve the classification found at each step. We then computed five indexes (Rand, Fowlkes and Mallows, Wallace 10, Wallace 01 and Jaccard), implemented in the R PCI function of the PROFDPM package [22]. In order to assign a significance value for the different indexes we obtained, for each index x, a bootstrapped distribution with mean $\bar{x}_{(\mathrm{B})}$ and standard deviation $\sigma_{(\mathrm{B})}$, resampling with replacement the samples and recomputing the indexes 10^3 times. Next we computed 10^3completly random classifications, obtained by shuffling the identifiers relating each sample with one of the classifications, and retrieving the maximum $x_{(\mathrm{R})}$. We finally verified that the random value was significantly different from the bootstrapped distribution by computing the z-scores:

$$z = \frac{\mathrm{abs}(x_{(\mathrm{R})} - \bar{x}_{(\mathrm{B})})}{\sigma_{(\mathrm{B})}},$$

which we considered significant if it was higher than 2.5. Heatmaps were generated with the HEATMAP.2 function in R package GPLOTS.

Bacterial networks

We considered a public dataset of 753 bacterial communities sampled from rainwater-filled beech tree-holes (Fagus spp.) [24], leading to 2874 Operative Taxonomic Units (OTUs) at the 97% of 16 rRNA sequence similarity. These communities were compared with Jensen-Shannon divergence [25], and automatically clustered following the method proposed in Ref. [31] to identify enterotypes. The clusters found with this method in [23] were used to colour the community labels in Fig. 6.

The inference of the OTU network started quantifying correlations between OTUs abundances with SparCC [26]. To perform this computation, from the original OTUs we reduced the data set removing rare taxa with less than 100 reads or occuring in less than 10 samples, leading to 619 OTUs. Then, the significance of the correlations was evaluated bootstrapping the samples 100 times the data and estimating pseudo p-values for each of the $N(N-1)/2$ pairs. A relationship between two OTUs was considered significant and represented as a link in the network if the correlation was larger than 0.2 in absolute value and the pseudo p-value lower than 0.01. The network obtained in this way was analysed using Cytoscape [32].

Acknowledgements

We thank Michael Schaub for critical comments on an earlier version of the manuscript. The research was funded by a European Research Council starting grant (311399-Redundancy) awarded to T.B. T.B. was also funded by a Royal Society University Research Fellowship. APG was also funded by the Simons Collaboration: Principles of Microbial Ecosystems (PriME), award number 542381.

Supplementary Figures

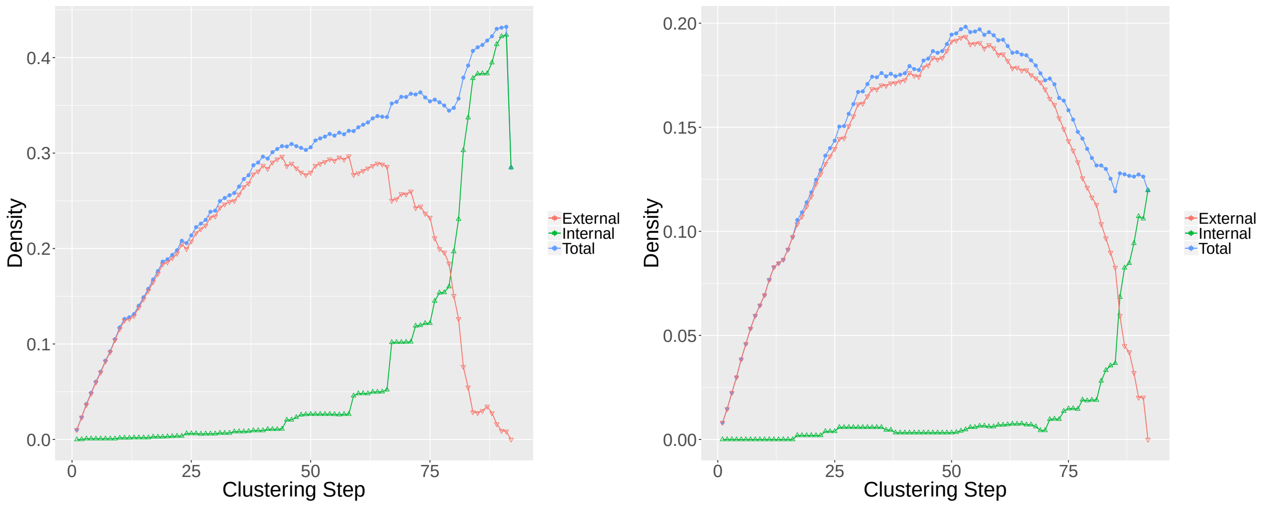

Figure 8: **Partition densities of synthetic mutualistic networks.** Networks with nestedness $\nu = 0.15$, $\kappa_{\text{mut}} = 0.08$, and $\kappa_{\text{comp}} = 0.5$ (left) or $\kappa_{\text{comp}} = 0.15$ (right). Changing the connectance change the relative value between the external and internal partition densities.

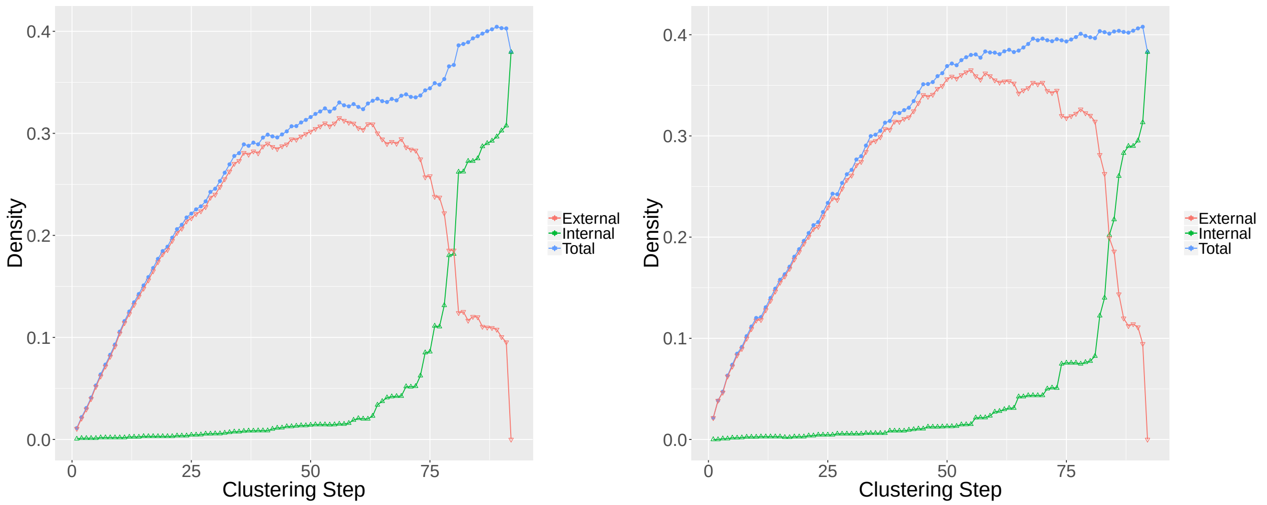

Figure 9: **Partition densities for synthetic mutualistic networks.** Networks with $\kappa_{\text{comp}} = 0.5$, $\kappa_{\text{mut}} = 0.28$ and $\nu = 0.35$ (left) or $\nu = 0.6$ (right). The high connectance of both networks make the internal partition density dominant, and two pools are detected through the total partition density. Nevertheless, the increase of the nestedness is detected through an increase in the internal partition density, which makes the second network more disassortative.

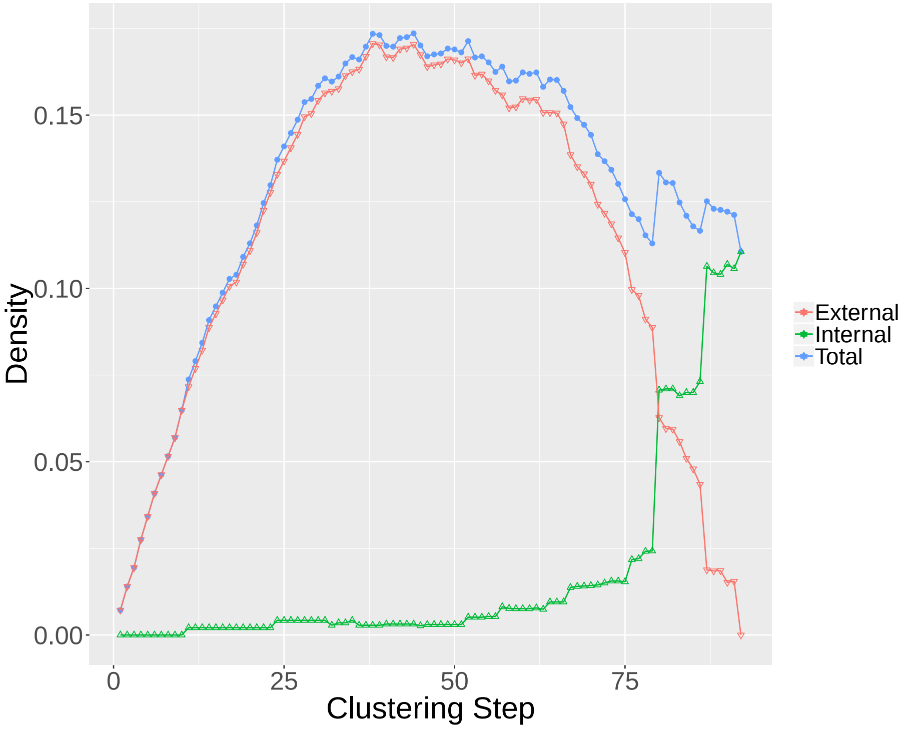

Figure 10: **Partition densities of synthetic mutualistic networks.** Network with nestedness $\nu = 0.05$, $\kappa_{\text{mut}} = 0.065$ and $\kappa_{\text{comp}} = 0.15$. The low connectance hinders the detection of the two pools of plants and pollinators.

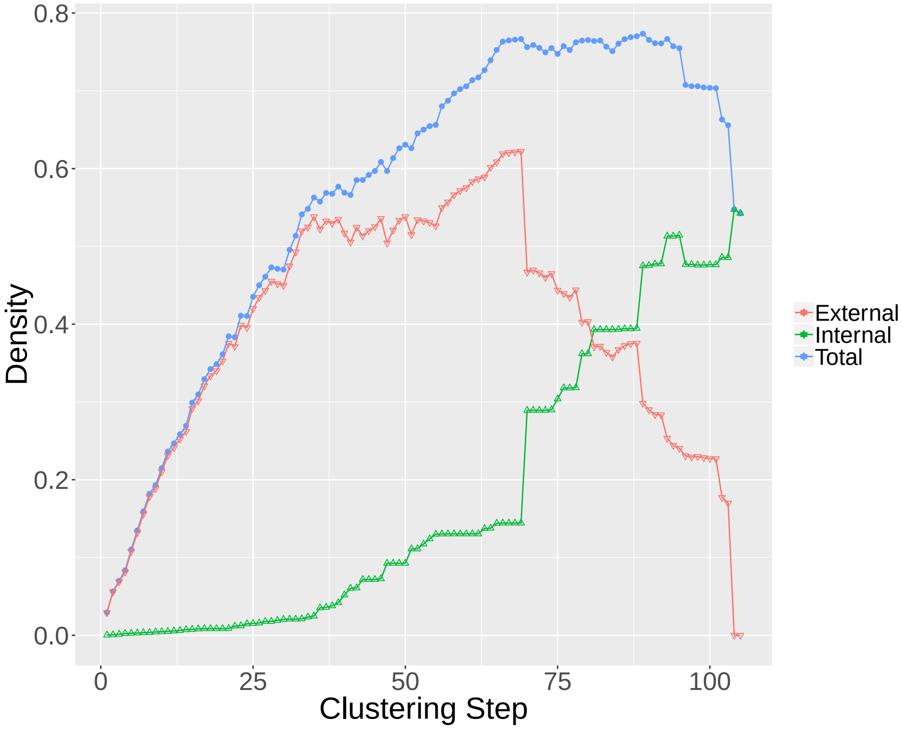

Figure 11: **Partition density of the trophic network.** The internal partition density peaks when there are three clusters, consistent with the existence of three trophic layers. The external partition density has a maximum at step 69, which is analysed in detail with respect to the reference classification found in [21].

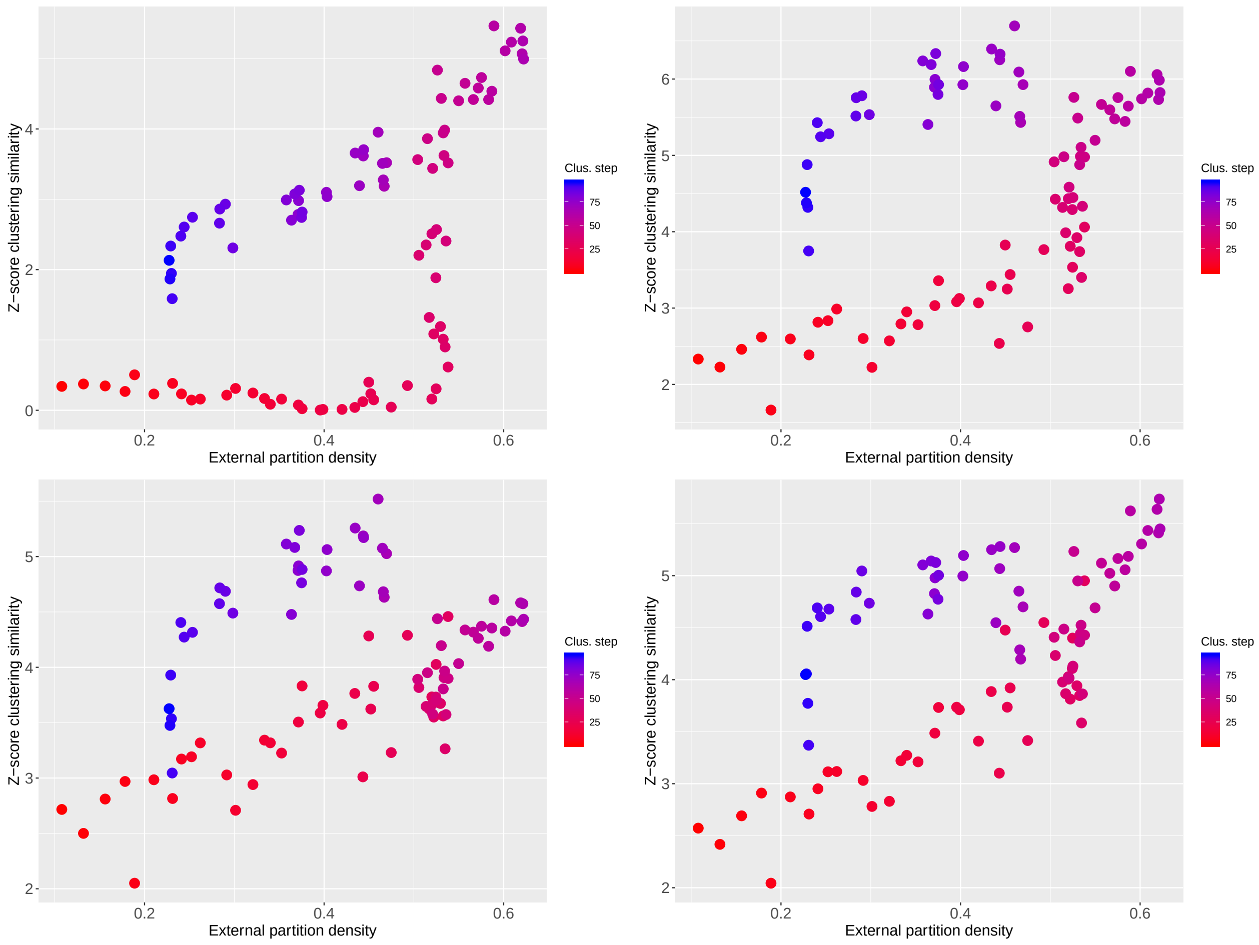

Figure 12: **Comparison between classifications of the trophic network.** Similarity between the reference classification found in [21] and the one found with functionInk is performed with the Z-score of a different indexes: Wallace 01 (Top left), Fowlkes and Mallows (Top right), Jaccard (Bottom left) and Rand (Bottom right). All indexes bring significant values and the maximum similarity is close to the maximum of the external partition density.

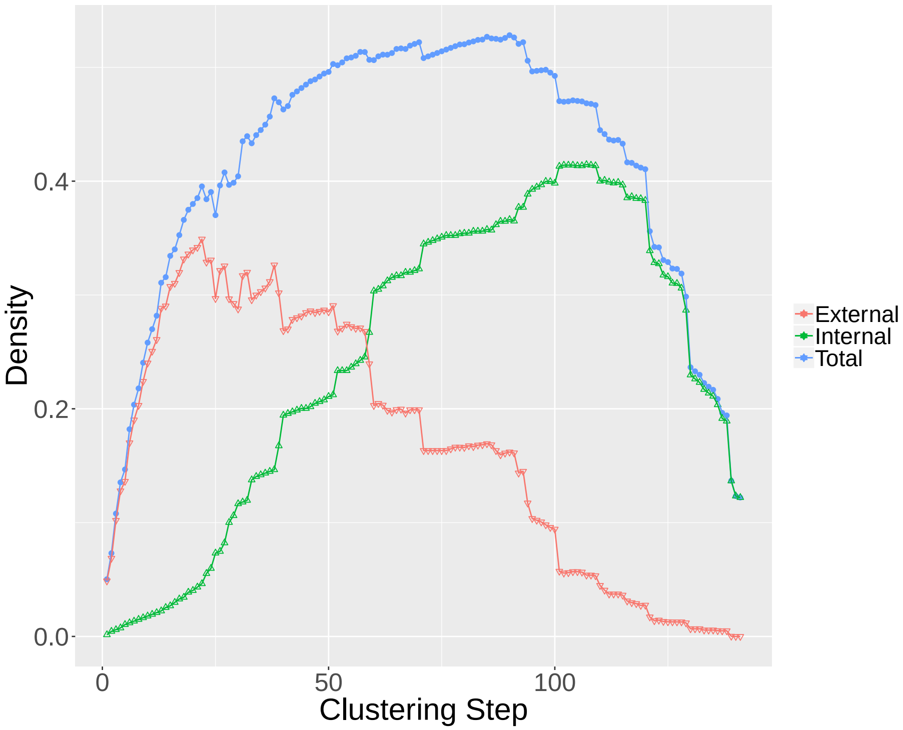

Figure 13: **Partition density of the microbial network.** The external partition density brings a poor reduction in the complexity of the network, with only 22 elements joined, while the internal partition density achieves a higher value and still a good number of clusters. Results suggest that modules are more relevant in this network given the high number of intra-cluster co-occurrences, later confirmed by visual inspection in the Main Text.

References

[1] J. E. Cohen and D. W. Stephens, *Food webs and niche space.* No. 11, Princeton University Press, 1978.

[2] R. MacArthur, "Fluctuations of animal populations and a measure of community stability," *Ecology*, vol. 36, p. 533, July 1955.

[3] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4, pp. 175–308, 2006.

[5] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[6] R. Lambiotte, J.-C. Delvenne, and M. Barahona, "Laplacian dynamics and multiscale modular structure in networks," *arXiv preprint arXiv:0812.1770*, 2008.

[7] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science*, vol. 328, no. 5980, pp. 876–878, 2010.

[8] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[9] C. S. Elton, *Animal ecology.* New York: The Macmillan Company, 1927.

[10] M. E. Newman and E. A. Leicht, "Mixture models and exploratory analysis in networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9564–9569, 2007.

[11] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[12] E. Estrada and J. A. Rodríguez-Velázquez, "Spectral measures of bipartivity in complex networks," *Physical Review E*, vol. 72, no. 4, p. 046105, 2005.

[13] D. Simberloff and T. Dayan, "The guild concept and the structure of ecological communities," *Annual review of ecology and systematics*, vol. 22, no. 1, pp. 115–143, 1991.

[14] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[15] M. Ganji, J. Chan, P. J. Stuckey, J. Bailey, C. Leckie, K. Ramamohanarao, and I. Davidson, "Image constrained blockmodelling: a constraint programming approach," in *Proceedings of the 2018 SIAM International Conference on Data Mining*, pp. 19–27, SIAM, 2018.

[16] R. Guimera and L. A. N. Amaral, "Cartography of complex networks: modules and universal roles," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 02, p. P02001, 2005.

[17] T. T. Tanimoto, "elementary mathematical theory of classification and prediction," 1958.

[18] A. Harrer and A. Schmidt, "Blockmodelling and role analysis in multi-relational networks," *Social Network Analysis and Mining*, vol. 3, no. 3, pp. 701–719, 2013.

[19] A. Pascual-García and U. Bastolla, "Mutualism supports biodiversity when the direct competition is weak," *Nature Communications*, vol. 8, p. 14326, Feb. 2017.

[20] U. Bastolla, M. A. Fortuna, A. Pascual-García, A. Ferrera, B. Luque, and J. Bascompte, "The architecture of mutualistic networks minimizes competition and increases biodiversity," *Nature*, vol. 458, pp. 1018–1020, Apr. 2009.

[21] S. Kéfi, V. Miele, E. A. Wieters, S. A. Navarrete, and E. L. Berlow, "How structured is the entangled bank? the surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience," *PLoS biology*, vol. 14, no. 8, p. e1002527, 2016.

[22] M. S. Shotwell *et al.*, "profdpm: An r package for map estimation in a class of conjugate product partition models," *J Stat Softw*, vol. 53, no. 8, pp. 1–18, 2013.

[23] A. Pascual-García and T. Bell, "Community-level signatures of ecological succession in natural bacterial communities," *bioRxiv*, p. 636233, 2019.

[24] D. W. Rivett and T. Bell, "Abundance determines the functional role of bacterial phylotypes in complex communities," *Nature microbiology*, p. 1, 2018.

[25] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, pp. 1858–1860, July 2003.

[26] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS computational biology*, vol. 8, no. 9, p. e1002687, 2012.

[27] A. Pascual-García, J. Tamames, and U. Bastolla, "Bacteria dialog with santa rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions?," *BMC microbiology*, vol. 14, no. 1, p. 284, 2014.

[28] M. T. Agler, J. Ruhe, S. Kroll, C. Morhenn, S.-T. Kim, D. Weigel, and E. M. Kemen, "Microbial hub taxa link host and abiotic factors to plant microbiome variation," *PLoS Biology*, vol. 14, no. 1, p. e1002352, 2016.

[29] R. R. Sokal, "A statistical method for evaluating systematic relationships," *Univ Kans Sci Bull*, vol. 38, pp. 1409–1438, 1958.

[30] A. Pascual-García, D. Abia, Á. R. Ortiz, and U. Bastolla, "Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures," *PLoS computational biology*, vol. 5, no. 3, p. e1000331, 2009.

[31] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, *et al.*, "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, 2011.

[32] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.