UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

Combination of Topological Indices in Network Analysis: A Computational Approach

Catarina Gomes Lisboa Marcelo Gouveia

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Ferenc Jordán
Francisco Pinto

2019

Acknowledgments

I want to thank my external supervisor, Dr. Ferenc Jordán, that since the beginning was very responsive, helpful and truly trusted my (work) skills. It was a tough year, full of changes, but in the end, we managed to conclude this stage. I also want to thank him for all the extra-thesis support, because life is more than just work and we should make the most of it. I also want to thank my internal supervisor, Dr. Francisco Pinto, for all the help in all the critic moments and for all the support.

I'm also grateful for the possibility of doing the ERASMUS+ program, with some financial support, and for the help of the people that enabled to make the process true: my masters' coordinator – Dr. Octávio Paulo – and all the mobility department from FCUL.

To my co-workers, that helped me in critical moments with helpful inputs. A special thanks to: Ágnes Móréh and Anett Endrédi for the data provided and to Imre Sándor Piross for the technical help.

And last, but not least, to all my friends and family.

Abstract

Ecosystems, and particularly, food webs have been subject of many studies in the last years. This is a hot topic if we consider the anthropogenic pressures and modifications that are increasingly prominent and notorious nowadays. Theoretical biologists try to figure out, in predictive analyses, how ecosystems will react if a species suddenly disappears. These analyses are often supported by a variety of different mathematical tools. Different mathematical tools are intended to give different biological answers.

In fact, food webs are usually complex relationships of interactions and thus, it is very common, in any systems ecology work, to simplify these interactions. A common way to do it is to reduce the system under focus to a network form. In network representation, we usually have entities connected by links. These links can be weighted, directed, looped, and thus, allow us to add some detail to the analysis. In biology, a food web can also be depicted as a network: species or groups of species are depicted as entities and their trophic interactions are depicted as links.

In general network analysis, there are a myriad of different centrality indices. Centrality indices allow us to identify the critical nodes in a network. In ecology these indices are also used and have been suggested to identify key organisms and to quantify their importance in food webs. All of these indices provide some information related to the centrality of a node in a network, but each of them is different. They express different aspects of being in the centre. Their relationships and also their biological meanings are often unclear.

Another common way to evaluate the importance of a species to the food web is through dynamical simulations of system behaviour. We can test, based in mathematical expressions such as logistic models, how the whole system will behave, i.e., how all the species will react if we take one species almost to its extinction.

This dissertation will analyse the correlation between centrality indices and the outcome of dynamical simulations, namely, community responses. The assumption is that the disturbance of more central species will generate larger community responses in the system. The goal is to understand if centrality indices are good enough to make predictions (closer to the ones we can get performing community response simulations). It is a major challenge to use simpler structural indicators to predict the outcome of much more complicated simulations. We approach these goals by using machine learning techniques to combine k centrality indices out of n in such a way that the correlation between the structural node centrality rank and the simulated node importance rank is the strongest. We ask which centrality indices should be chosen and how exactly they should be combined to best predict simulated food web dynamics. We also evaluate to what extent these correlations can be improved.

Keywords: Systems Biology, Modelling, Network Analysis, Dynamical Analysis, Machine Learning

Resumo

Muitos sistemas biológicos têm visto um grande declínio desde que o ser humano entrou em cena há muitos milhares de anos. Estes sistemas têm sofrido particulares perturbações com o aumento da população mundial e o estilo de vida moderno. Vários ecossistemas têm sido atacados com a constante destruição de habitats, pesca excessiva ou aumento de poluição, que leva, inevitavelmente à decadência e extinção de muitas espécies e ecossistemas.

De modo a tentar compreender como os ecossistemas funcionam, quais as interações entre os organismos e os sistemas abióticos que os circundam, tem-se tentado reduzir tais sistemas, geralmente complexos e intrincados, a simplificações que podem ser avaliadas de modo mais objetivo.

Desta forma, a modelação de sistemas ecológicos é cada vez mais utilizada para perceber como as espécies se encontram interligadas nos seus habitats. Reduzir sistemas complexos a representações matemáticas como redes permite-nos quantificar as partes dentro do todo. Um exemplo desta modelação biológica consiste em reduzir cadeias alimentares complexas a redes simples, representadas por nós (ou vértices) ligados por arcos (ou arestas) – grafos. Neste tipo de representação é possível adicionar algum nível de detalhe como, por exemplo, pesos, direções ou ciclos. Esta simplificação constitui uma ferramenta importante para se estudar interações topológicas e dinâmicas em sistemas biológicos.

A modelação ecológica tem por base noções matemáticas. Neste caso, utilizaram-se conceitos derivados do estudo de grafos. Através destes conceitos é possível simular o comportamento de sistemas através de simulações e análises computacionais quantitativas.

As redes ou teias alimentares são simplificações usadas para representar interações tróficas entre organismos. As relações presa-predador permitem-nos entender a dinâmica e a resiliência das comunidades: interações alimentares dão-nos noções sobre taxas de vitalidade, de crescimento e de mortalidade. Por exemplo, se a população sob análise tem mais presas à sua disposição, terá tendência a crescer; no entanto, se a mesma população tiver de confrontar um grande número de predadores, provavelmente irá decair, visto que mais indivíduos serão ingeridos.

A aplicação de estatística de redes, importada de áreas como a matemática, física e informática, a estas representações biológicas, permite-nos avaliar os ecossistemas. Existe uma miríade de índices de centralidade (ou índices topológicos) utilizados. Em ecologia, estes índices têm sido usados e sugeridos como métodos de identificação de organismos chave de modo a quantificar a sua importância em cadeias alimentares ou ecossistemas. Estes índices dão alguma informação sobre a centralidade de um nó na rede. No entanto, cada um é diferente – expressam aspetos diferentes de centralidade – e, quando aplicados a sistemas biológicos, podem representar também diferentes significados.

Espécies classificadas como mais importantes terão, de um ponto de vista de uma ecologia mais funcional, uma maior importância de conservação. Isto acontece porque, se essa população sofrer níveis críticos de extinção, ou se for completamente extinta do(s) ecossistema(s), pode levar a perturbações em cascata, isto é, provocar a extinção de muitas outras espécies – direta ou indiretamente dependentes – ou até, em casos mais drásticos, o desaparecimento de toda a comunidade em que se inserem. No entanto, a aplicação destes índices para expressar relações biológicas e o seu significado ainda é pouco claro.

Para além disto, estes índices permitem avaliar redes biológicas, de um modo teórico, a diferentes escalas. Índices topológicos globais, usados também em outras áreas de análise de grafos, permitem avaliar as redes como um todo. No entanto, estes índices oferecem menos detalhe no que toca a como cada indivíduo se interliga com os restantes nessa rede. A perspetiva local, do outro lado da escala, permite-nos entender melhor como cada indivíduo está conectado na rede. O problema é que, de uma perspetiva local, não temos compreensão sobre interações mais afastadas na rede, ou seja, sobre efeitos diretos ou indiretos que um indivíduo pode estar a causar aos outros indivíduos. Devido a isso, uma variedade de diferentes índices de escala intermédia têm surgido e sido utilizados. Estes, procuram acrescentar algumas dessas informações previamente ausentes – permitem analisar a posição topológica de cada espécie, mas com alguma perspetiva sobre os impactos diretos e ou indiretos que essa espécie tem no resto da comunidade.

A representação simplificada destes sistemas pode ser vantajosa, em termos de análises teóricas mais objetivas. No entanto, utilizar índices estáticos – como é o caso dos índices topológicos – para avaliar sistemas que se encontram em constante mudança, pode ser falacioso: a vida não é estática e muito menos as complexas redes de interação entre espécies. Todas as espécies têm a sua taxa de crescimento, mortalidade e vitalidade, diferentes ciclos de reprodução, e, geralmente dependem de outras espécies para a sua sobrevivência. Para além disso, estas taxas variam ao longo do ciclo de vida de cada espécie. Deste modo, para entender como o tempo e as flutuações topológicas afetam as populações, outros tipos de análises matemáticas podem ser realizados. Estas análises, intituladas de análises dinâmicas (ou simulações), são frequentemente descritas por sistemas ordinários de equações diferenciais, que têm em conta diferentes parâmetros populacionais, adequados a cada espécie.

No entanto, bancos de dados com informações relativas ao histórico de vida, informações demográficas e de interação de espécies, necessárias para parametrizar modelos de redes ecológicas raramente estão disponíveis, devido à sua grande complexidade de recolha, quer em termos de esforço quer em termos de escalas temporais – é difícil recolher dados relativamente a cada uma das espécies dentro das comunidades, durante todo o seu ciclo de vida (visto que pode ser um período relativamente extenso). Por este motivo, análises dinâmicas são muitas vezes preteridas relativamente a análises estáticas.

Neste trabalho, pretendeu-se investigar, através da análise de clusters e correlações, quais as semelhanças entre alguns dos índices topológicos disponíveis e utilizados em modelação ecológica. Também se analisou como podemos utilizar estes índices, de maneira individual ou combinada, para prever quais são as espécies-chave (ou espécies críticas) numa rede alimentar. Utilizou-se, para isso, um algoritmo genético com regressão simbólica, e, como elemento alvo para esta comparação, uma simulação dinâmica. A simulação dinâmica usada tenta prever a resposta das espécies de uma comunidade quando uma delas é perturbada – simulação de resposta da comunidade. Para o uso desta simulação como "alvo", partiu-se do pressuposto que esta seria a maneira mais correta de avaliar a importância de cada espécie dentro da rede. Devido à escassez de informação relativa a teias alimentares reais, utilizaram-se 1000 cadeias alimentares hipotéticas, constituídas por 15 espécies, 3 espécies basais e 4 espécies predadoras de topo.

Os resultados obtidos fornecem novas maneiras de avaliar a ordem de importância de cada espécie, de acordo com simulações biológicas complexas. Podemos usar índices estruturais (topológicos) simples ou algumas combinações desses índices, variando dos mais simples aos mais complexos. Também foi percetível que o grau (D) e o índice de importância topológica ponderada de 5 etapas (WI^5) foram os que mais emergiram nas combinações de índices obtidas. Estes resultados são interessantes se considerarmos que o grau é um índice simples, baseado em interações diretas. O índice de importância topológica ponderada de 5 etapas, por sua vez, é um índice mais complexo, ponderado,

e que considera, também, interações indiretas. Além disso, descobrimos que este índice está simetricamente correlacionado com os resultados da simulação, contrariamente ao grau, que não se encontra correlacionado. São índices totalmente diferentes e, por isso, é interessante e conveniente combiná-los: permitem informações complementares e adequadas.

Acreditamos que maneiras mais concisas e eficientes de identificar espécies-chave em redes ecológicas serão essenciais para o futuro da ecologia de sistemas que visa alcançar prioridades ou regulamentos objetivos de conservação e gestão de ecossistemas. A nossa abordagem, baseada na maximização do poder preditivo de análises estruturais, pode ser um grande passo em direção a pesquisas e análises rápidas e simples, mas bastante realistas sobre redes alimentares e espécies-chave nos ecossistemas.

Palavras-chave: Biologia de Sistemas, Modelação, Análise de Redes, Análise Dinâmica, Aprendizagem Automática

Table of Contents

List of Tables

List of Figures

List of Abbreviations

D_i – Degree

wD_i – Weighted Degree

BC_i – Betweenness Centrality

CC_i – Closeness Centrality

TI_i^n – Topological importance index

WI_i^n – Weighted topological importance index

s_i – Status index

s'_i – Contra-status index

Δs_i – Net status index

K_i – Keystone index

$K_{bu,i}$ – Keystone index for bottom-up effects

$K_{td,i}$ – Keystone index for top-down effects

$K_{dir,i}$ – Keystone index for direct effects

$K_{indir,i}$ – Keystone index for indirect effects

3N – In a ranked network, it means the first three nodes (i.e. the first three most important species, according to the rank used) in a network

4N – In a ranked network, it means the first four nodes (i.e. the first four most important species, according to the rank used) in a network

6N – In a ranked network, it means the first six nodes (i.e. the first six most important species, according to the rank used) in a network

Chapter 1

Introduction

1.1 Ecology and Ecosystems

Ecology (from the Greek: οἶκος - "house" or "environment"; -λογία - "study of") is the biological science that studies biotic and abiotic interactions between living organisms and their surroundings in a specific time and space. It was first coined by Haeckel in 1866.

Ecologists study biodiversity, distribution, biomass, ecosystems among other topics. An ecosystem is a community of species – living organisms – interacting with their environment – non-living components. These interactions are always escorted by energy and nutrient fluxes. Ecologists can study ecosystems in different details' level: ranging from individual organisms or species (organisms with specific traits that can mate and produce fertile offspring), to populations (comprising organisms from the same species), to communities (composed by different populations)[1].

Ecosystems depend upon internal and external factors. Internal factors are usually associated to the type and quantity of species interacting in the ecosystem. External factors are frequently linked to climate, soil and topography. It is also important to mention that these systems are dynamic and, therefore, constantly changing, adapting and evolving[1,2].

From an anthropogenic perspective, these systems are important since they provide us natural resources, e.g., water and air, and natural services, e.g., the nutrient cycle or air purification – cycles that are dependent upon the interaction of species in these habitats[2].

Since we are crossing an era of unprecedent changes in ecosystems, is thus primordial understanding how they work and how species interact between themselves.

1.2 Networks

Networks (or graphs – used interchangeably in this work) can be generally used to represent overall real-life problems since they can reduce a system to a simple interaction of points and edges. In fact, they are used in the more diverse fields such as, social networks, telecommunication networks or biological networks[2-7].

A suitable way to study an ecosystem or its parts is using graph theory. Graphs allow us to represent the interactions between species and their abiotic surroundings. Furthermore, graph analysis is becoming more and more refined and thus, allow us to address different biological questions with more accurate and, simultaneously, meaningful answers[8].

We can simplify and reduce a system to a graphical network, generally, by considering the parts of the system we want to study as specific entities, connected to each other. Entities are represented by nodes and connections between them by arcs. This is a simplistic way to represent a system. As a result, some information about the system is always lost. However, it is still one of the best ways to understand and predict the behaviour of a complex system.

1.3 Mathematical Concept

Mathematically, a graph is defined by a set of points (nodes, vertices or junctions) connected by lines (edges, arcs, branches). Formally, a graph is defined as G = (N, A) where N is a finite set and $A \subseteq N \times N$. N elements are denoted by nodes and A elements by edges or arcs, whether the graph is directed or undirected, respectively. In case of an undirected graph, an edge between *i* and *j* is represented by {i, j} (in this case, edges {i, j} and {j, i} are the same). In case of a directed graph, an arc from i to j is represented by (i, j)[4].

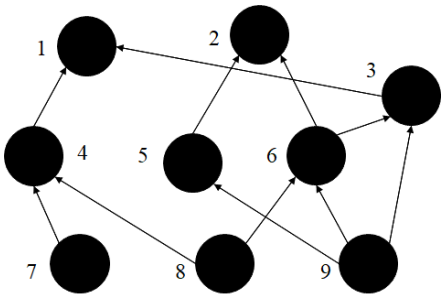


Figure 1.1. Small network representing a simple, directed graph.

Graphs can also be represented in a visual form, as in Figure 1.1. The directed graph of Figure 1.1 corresponds to the graph G = (N, A) where N = {1, …, 9} and A ={(1,3), (1,4), (2,5), (2,6), (3,6), (3,9), (4,7), (4,8), (5,9), (6,8), (6,9)}.

As noted before, graphs can be directed or undirected. An undirected graph "similar" to the previous directed graph can be defined as follows:

N = {1, …, 9} and E ={{1,3},{1,4},{2,5},{2,6},{3,6},{3,9},{4,7},{4,8},{5,9},{6,8},{6,9}}.

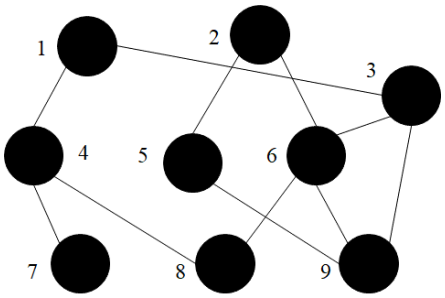


Figure 1.2. Small network representing a simple, undirected graph.

Another common way to represent graphs is through an adjacency matrix. The adjacency matrix A, for the undirected graph with nodes i to j, is constructed according the following rules:

$$A_{i,j} = \begin{cases} 0, & if\ nodes\ i\ and\ j\ are\ not\ connected, \\ n, & the\ number\ of\ n\ edges\ connecting\ i\ and\ j. \end{cases} \quad (1.1)$$

Fig. 1.2, as an undirected graph would be denoted by:

$$A_{1,9} = \begin{pmatrix} 001100000 \\ 000011000 \\ 100001001 \\ 100000110 \\ 010000001 \\ 011000011 \\ 000100000 \\ 000101000 \\ 001011000 \end{pmatrix} \quad (1.2)$$

A similar definition holds for directed graphs.

It is also important to notice that if we want to provide more information regarding the system or graph we study, we can consider other properties. They can be weighted and unweighted, whether we want to consider that some interactions are quantitatively more important than others. If not, they are binary, represented by zeros and ones (like in matrix $A_{1,9}$). Graphs can contain multi-edges, more than one connection between the same nodes for example; self-loops, when a node is, simultaneously, the provider and receiver of the information; and they can also include cycles, e.g., when $A \rightarrow B, B \rightarrow C, C \rightarrow A$.

Besides these briefly mentioned properties, there are also a variety of other, for example, regarding graphs structure. Approaching them is out of scope of this work but they can be consulted in more detail in a variety of technical or introductory book texts related to graph theory[4].

1.4 Graph Theory: Brief history

Graph theory was discovered independently in different times and places. It started as a mathematical tool to try to solve a series of different physical and real problems. Its foundation is usually attributed to Euler (1707-1782) since he tried to solve a topology problem called the Königsberg Bridge Problem. In this problem, Euler proved that it is impossible to start and end in the same point, without repeating the same path, if we have two islands and two banks of a river connected by seven bridges. This was a real-life problem and, in order to solve it, Euler represented the four different pieces of land as points, and the bridges as lines, producing a graph[9].

With the course of the years, this area became more and more useful to manage problems in a diversity of other subjects and, therefore, different definitions and methods to classify topologies and patterns arose. Social sciences were of great contribution for the development of measures that could help to analyse networks. These measures are now widely used in all areas, including biology[4].

1.5 Evaluate Node Position: Centrality

One important and useful group of measures that arose within graph theory was the one that allow us to evaluate the importance of a node (or edge) in the considered network. This group of measures is known as centrality measures.

There is a myriad of different mathematical measures in this group, each of which, based in different assumptions and concepts. In spite of that, they all stand for the same principle: identify which is the most important node for the network and helping us to define what it means to be central in a network[4]. Usually this measures account for the topology of the network and, therefore, they are also considered as topological, structural or positional indices (these names will be used interchangeably).

Mathematically, if we consider $G_1 = (N_1, E_1)$ and $G_2 = (N_2, E_2)$ as two isomorphic graphs (directed or undirected, weighted or unweighted) where N_1 and N_2 are node sets and E_1 and E_2 edge sets respectively, a real-valued function F will be considered a structural index if and only if:

For each $n \epsilon N_1 \implies F_{G_1}(n) = F_{G_2}(M(n))$, where $F_{G_1}(n)$ denotes the value of $F(n)$ in $G_1$and M is a mapping function from N_1 to N_2.

A centrality index, by definition, is a function C that is a structural index and allows us to derive an order for the set of nodes or edges. By this rank or order we can say which are the most important vertices for the network[10].

1.6 Global vs local

Centrality indices can be considered as ranging from a global to a local spectrum, crossing the meso-scale level, in between both[11].

By using a global index, like, connectivity or link density, for example, we are able to characterize the topology of the whole network. However, we lose information about the specific position – and importance – of each node in the network. A local index, on the opposite side of the spectrum, such as the degree of a node (number of links that are directed assigned to it) provides information only about that specific node[12] without considering the rest of the network or secondary connections.

Due to this, some meso-scale metrics emerged – so one could analyse the importance of a node, regarding its direct or indirect interactions in the network. The meso-scale perspective considers that the strength of indirect effects decreases with the length of the pathway[11]. One example of these metrics is the positional keystone index (explained further in more detail) which comprises the neighbours of the neighbours of a node and, thus, provides more information about the network structure in which the node is embedded[13].

1.7 Ecological Networks

As aforementioned, ecological systems provide important sources for life and human economies. Therefore, knowing how species interact between themselves in their context can be very useful.

Indeed, nowadays, we are witnessing a never-seen rate of human direct or indirect modifications in the environment[14] and thus, it is important to try to predict the deeds of an ecosystem and understand its intricate connections.

Ecologists can rely on countless methodologies and different ways to subset communities in ecosystems in an effort to explain its interactions[2,8].

One of these methodologies consists of zoom in food webs in the ecosystems and transform their biological information in graph representation: food webs can be represented as diagrams of trophic interactions between species in an ecosystem, depicting which species eat which others[15]. They can include patterns of material and energy flow in communities[2].

This informational reduction allows us to study, and quantitatively evaluate, the interactions and properties of each food web based on mathematical, computational and statistical methods. Plus, this simplicity allow us to overcome problems concerning data collection for food webs, which happen quite often[16].

1.8 Food Web History

Food webs, also nominated as "food cycles" was a concept widely spread by Charles Elton[17]. Elton emphasized that "Every animal is closely linked with a number of other animals living round it, and these relations in an animal community are largely food relations". Elton goes further in his idea about food cycles, describing patterns in how organisms are related, a concept coined as "Pyramid of Numbers". Elton stood that most food webs had many organisms on their bottom trophic levels and subsequently fewer on the upper ones. This concept is now known as Eltonian Pyramid.

These concepts were later developed by scientists like Raymond Lindeman[18], Robert May[19], John Lawton[20] and Stuart Pimm[21].

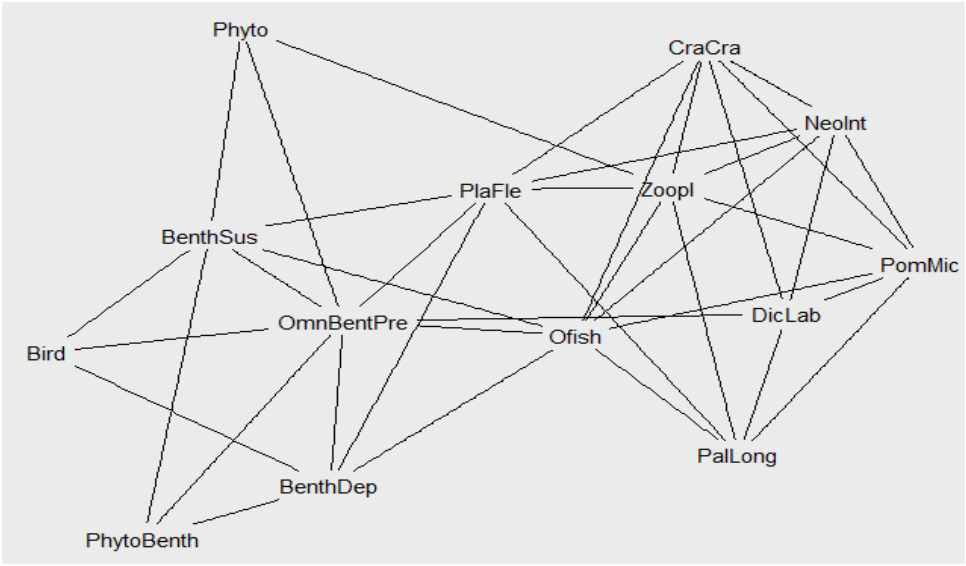
Lindeman started to consider the trophic dynamics and the energy transfers in ecosystems due to these dynamics. He realized that energy flows in food webs starting in a light form – assimilated by producers – and then passes through animal consumers and bacterial decomposers with some losses. He recognized that decomposers transform organic substances back to inorganic matter, ready to use by autotrophic organisms again.

In 1972, May started to use a theoretical approach applied to ecological systems to understand whether populations with n animals and l interactions would be stable or not.

Pimm and Lawton developed these ideas even further, using food chains (food webs in which is predator as only one prey) and food webs as a structure to understand and study population dynamics through predictions[2].

1.9 Food Webs

Usually, food webs are depicted as diagrams represented with species (e.g. orca), or functional groups of species (e.g. benthic invertebrates), linked by arrows or lines that represent the trophic interactions between species: when arrows are used, energy flow is represented from the resource to the consumer, while when lines are used, the population dynamical effects are represented between the prey and the predator (the predator does have an effect on the prey, even if this does not follow the direction of energy). Examples of food webs can be observed in Figures 1.2 – 1.4.



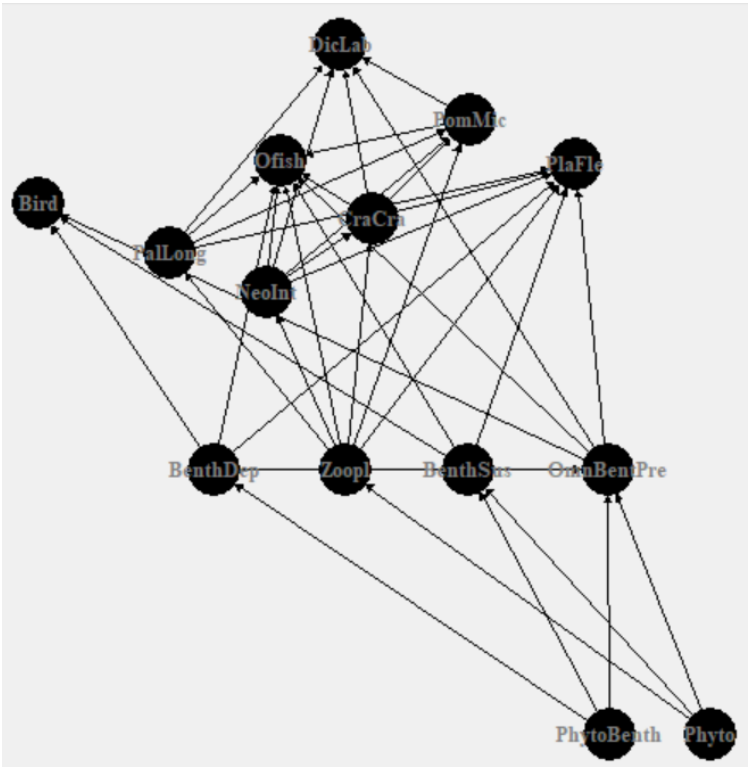


Figure 1.3. and **Figure 1.4.** Examples of representations of a simplified real food web: Seine Estuary food web. Each node represents a species (or group of species) and each link represents the trophic interaction between species. In Fig. 1.2. links are undirected, unweighted and nodes are randomly distributed. In Fig. 1.3 links are directed, unweighted and nodes are distributed

according to the respective trophic level. DicLab – *Dicentrarchus labrax* (fish); PomMic – *Pomatoschistus microps* (suprabenthos); Ofish – Other fishes; PlaFle – *Platichthys flesus* (fish); Bird – Birds; CraCra – *Crangon crangon* (suprabenthos); PalLong – *Palaemon longirostris* (suprabenthos); NeoInt – *Neomysis integer* (suprabenthos); OmnBentPre – Omnivorous & benthic predators; BenthDep – Benthic deposit feeders; BenthSus – Benthic suspension feeders; Zoopl – Zooplankton; Phyto – Phytoplankton; PhytoBenth – Phytobenthos[22].

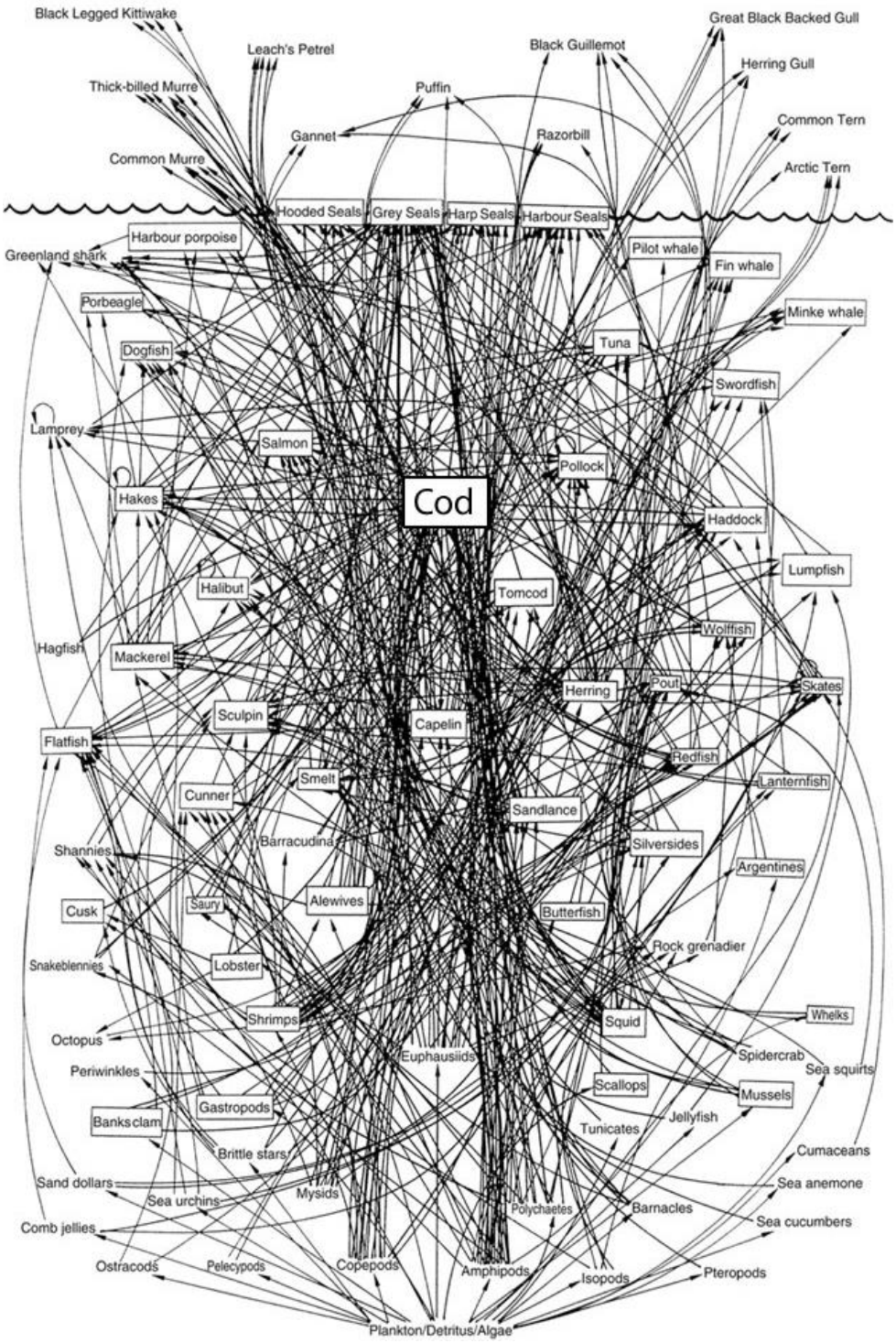


Figure 1.5. Simplified food web for the Northwest Atlantic. Each node represents a species (or group of species) and each arrow points to the predator species. Image from David Lavigne, National Science and Engineering Research Council.

1.9.1 Basic Concepts

Food webs have a typical organization: they usually start with *primary producers* or *basal species*, at the bottom of the food web. They are followed by *herbivores* or *omnivores* and then by animals that eat herbivores or omnivores: carnivores, or predators. Animals not consumed by any other are defined as *top predators*.

Basal species grow and develop using inorganic nutrients, water and energy from sunlight – photosynthesis – or from chemicals – chemosynthesis. Due to this, they are considered as *autotrophic* or *chemotrophic*, respectively.

Herbivores, omnivores, predators, decomposers and detritivores are *heterotrophic* organisms: they feed on organic substrates to obtain nutrients and energy. Herbivores feed on plants. Omnivores can feed either on plants or animals. Predators are usually associated to carnivores: animals that eat other animals[23]. The difference between decomposers and detritivores is that decomposers can break down matter without ingesting it. Detritivores must ingest and digest the organic dead matter using internal processes.

Food webs can be grouped in different ways: we can either consider them regarding the ecosystem they are depicting (e.g., detrital food webs, fresh water food webs) or we can gather them in different categories, such as:

- Source webs – all relationships in these food webs rise from only one food source, i.e., they only contain one element in their basis (basal species or basal trophic group).
- Sink webs – all the trophic interactions depicted descend from only one "sink", i.e., the top predator or top trophic group.
- Community webs – all the feeding interactions in the community. This concept is hard to materialize since the limits of a community are often difficult to establish or it can generate dauntingly complex webs.

1.9.2 Structure

1.9.2.1 Trophic Positions

To better understand some of the concepts above-mentioned, one should define the trophic organization of a food web.

The trophic level of a species or group of species is an abstract definition that helps us distinguishing subgroups of species within the community that acquire energy similarly. Thus, we have *basal species* or *primary producers* in the first position, conventionally attributed as trophic level one (but sometimes can also be considered as zero[8]). They are usually followed by *herbivores* at level 2, *predators* or *omnivores* at level 3 or higher, and *top-predators* (that can also be omnivores), "finishing" the chain, usually at level 4 or 5. There are also *decomposers* and *detritivores* that obtain their energy from all dead species of all trophic levels, and usually, due to this, they are not assigned to any trophic level.

Besides basal species – that produce directly their energy – the rest of the species can feed in more than one trophic level, making it hard to objectively attribute one species to one trophic level[2].

1.9.2.2 Keystone Species: Positional Importance

In a trophic network, there are species, or groups of species that are critical for that network. These species are considered as keystone species. Their removal or perturbation can imply severe destabilization for all the community, with loss of other species[24]. A formal definition of keystone species is "one whose impact on its community or ecosystem is large, and disproportionately large relative to its abundance"[25].

However, quantifying the importance of a species experimentally is a delicate process due to the spatiotemporal ranges intrinsically associated. These cause rising difficulties regarding the execution of objective methodologies related to field manipulation or laboratory experiments[8]. To counter this, a variety of theoretical methods arose. Nowadays, we can predict which are the most important species for the network, based on diverse topological indices – broadly used in graph theory –, or dynamical analysis – usually more specific to the field of study.

Currently, one challenge for theoretical ecologists is to choose which network centrality indices or dynamical simulations perform better or are more adequate to solve specific problems or questions[8]. If simple and fast structural analysis can reach the predictability of complicated and data-intensive dynamical simulations, conservation management can be more efficient.

1.10 Problem

1.10.1 Similarity Between Centrality Indices

As previously mentioned, there are a variety of different positional indices, each of which based in different definitions and with its own mathematical construction. But, if we have so many options to analyse our data, how can we choose or, at least, be sure that we are using indices that will bring new information to our study?

One of the current problems in graph theory – and graph theory applied to biology – is weather to use each index. There's no specific formula to answer this problem yet. Some indices, very closely derived, such as status and contrastatus (explained in more detail later), despite of their similar mathematical constructions, can provide very different information about the same node or network in general[26].

One way to avoid redundant information (or to reinforce the importance of a node in a network) is to study the pairwise correlation between different indices: the most correlated indices will characterize a network likewise and, therefore, won't bring much new information for the analysis[27].

Some studies based in small sets of networks were made in order to better understand this similarity[12,28]. These studies give us some insight and guidelines about the most related indices, but they can be specific to the networks used rather than general.

Here, we attempt to widen these studies, using more networks and comparing the rank order provided by 18 centrality indices, through spearman correlation and clustering: the more redundant indices will cluster together.

1.10.2 Structure to Dynamics

These topological indices are useful to evaluate static food webs. However, real-world networks are not static: they change through time and space. For instance, the topology of a food web can change if there are two differential preys for a predator: at one moment, one of them can be more available in nature, and thus, will be the preferential one. But this can easily change if this species starts eventually to be scarcer[8]. Thus, dynamical approaches, that study, across a chronological time line, changes in size or proportion of entities (species or groups of species) and also their connections (interactions), are usually useful to understand the behaviour of the whole system, or merely to evaluate the effects in the network if one entity is deleted, for example.

Unfortunately, dynamical studies have also their flaws: they can be either too simplistic to model the real-world dynamics[3] or, when more complex and accurate, they are usually time and effort demanding, since they require much more specific data about the network[29].

Due to this, it would be very convenient if we could predict the outcome of dynamical models based solely in topological indices: 1) topological indices that characterize the importance of a node in a network (even though statically) are easier to obtain, 2) we would get more accurate results, not only based in the connections or the hierarchy of the network (as structural properties directly enable us to infer) but also based in how entities, in a specific context, interact between themselves and how they are connected in the same system, 3) we could infer the behaviour of a wider diversity of networks, even in fields where there's still a lack of available experimental data as in the case of food webs.

With this work, we aim to tie the knot between both topological methods and dynamical simulations in a mathematical expression. This expression intends to avoid performing dynamical simulations but reach similar answers using just a few of the available topological indices.

Chapter 2

Data and Research Methodology

2.1 Data

We used 1000 randomly generated networks composed of 15 species: 12 consumers and 3 basal species. The maximum number of top predators was set to 4 and 36 links were randomly generated between species. The basal species #1, #2 and #3 were not perturbed for the community response simulation. Therefore, these species were not considered for this analysis. Each network was considered undirected[8]. In the case of weighted topological indices, weights for the arcs AB were generated as $\frac{1}{nr\ of\ A's\ preys}$, if A eats B.

With these data, a matrix $M_{12n\times k}$ was constructed: being n the number of networks (each composed of 12 nodes) and k the respective 18 topological indices and the community response simulation's results. This matrix was then processed in two different ones: $M_{r_{12n\times k}}$ and $M_{ro_{12n\times k}}$. $M_{r_{12n\times k}}$ containing the real (original) values derived by each index and the community response simulation in each column, compressed in the range [0,1]. We used the function "normalize" from the "BBmisc" package[30] with the default method parameter "range", using R[31]. $M_{ro_{12n\times k}}$, containing the rank order values, i.e., in this matrix the real values were replaced by the respective node rank order (from 1 to 12). To nodes in the same network, with the same index value, a random order was assigned.

Figure 1.3 and Figure 1.4 were generated using the "igraph" package[32], using R.

2.2 Research Methodology

2.2.1 Topological indices

To access the positional importance of nodes in a network one can use a range of different network indices (some of these indices are more local or global, some are for weighted or directed networks and some are used to evaluate hierarchies, as previously mentioned).

2.2.1.1 Degree and weighted degree (D_i, wD_i)

The most local network centrality index is the degree of a node (D). It represents the number of other nodes directly connected to it. In a food web, the degree of a node i (D_i) is the sum of its preys and

predators. In weighted networks, the weighted degree of node i (wD_i) is the sum of weights on links adjacent to node i[33].

2.2.1.2 Betweenness centrality (BC_i)

Betweenness centrality is considered a global index since it quantifies the portion of shortest paths crossing a given node i.

Considering $i \neq j \neq k$ three different nodes, d_{jk} as the total number of shortest paths between nodes j and k, and $d_{jk}(i)$ as the number of these shortest paths that cross node i, we can represent betweenness centrality as[11,33]:

$$BC_i = \sum_j \sum_k \frac{d_{jk}(i)}{d_{jk}}, i \neq j \neq k. \quad (2.1)$$

Since this index scales with the number of nodes and edges, if we want to have $BC_i \in [0, 1]$ we can divide this measure by the number of pairs of nodes not including node i: $(N-1)(N-2)$, if we are considering directed graphs. If we are considering undirected graphs, we can divide by $(N-1)(N-2)/2$ (only one direction is considered)[34].

$$BC_i = \frac{\sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}}{(1/2)(N-1)(N-2)} = \frac{2\sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}}{(N-1)(N-2)}, \; i \neq j \neq k. \quad (2.2)$$

Where, $j < k$ reinforces that unidirectional state. N is the number of nodes in the network.

Biologically speaking, if BC_i is large for a species or trophic group i, deleting this species will affect quickly and directly the whole web. This happens because the node is incident to many shortest paths in the network[35].

2.2.1.3 Closeness centrality (CC_i)

The closeness centrality index is also considered a global index since it evaluates the network as a whole. It is the inverse of the normalized average length of the shortest path between node i and all other nodes[33,36]. It quantifies how close a node i is to others[11,35].

$$CC_i = \frac{N-1}{\sum_{j=1}^{j=N} d_{ij}}, i \neq j \quad (2.3)$$

d_{ij} is the length of the shortest path between nodes i and j in the network[33].

The biological meaning is close to BC_i: the larger the value for CC_i, the more the deletion of this group will affect the other groups directly.

2.2.1.4 Positional importance based on indirect chain effects (TI_i^n, WI_i^n)

TI_i^n and WI_i^n measure the topological importance considering the n-step-long indirect effects, in unweighted and weighted networks, respectively[36]. These indices can be considered as local, meso or global scale. If $n = 1$ it is a local index. If $n \leq biggest\ path$ it is considered a meso-scale index. If the maximum number of steps is considered, the whole network is accounted: it's considered a global index. We used $n = 1$, $n = 3$ and $n = 5$.

To derive these indices, let $a_{n,ij}$ be the effect of j on i, if j is connected to i in n steps. One step is each connection between direct neighbours. We consider first the case with n = 1 as follows:

$$a_{1,ij} = \frac{1}{D_j} \tag{2.4}$$

To define $a_{n,ij}$ let P_k be a path from i to j with n arcs, where P_k is the path $i = i_1, i_2, \dots, i_{n-1}, i_n = j$.

The contribution of this path P_k to $a_{n,ij}$ is equal to $a_{1,i1i2} \times a_{1,i2i3} \times \dots \times a_{1,in-1in} = C(P_k)$.

Finally, $a_{n,ij} = \sum_{k \in \{\text{all paths from } j \text{ to } i \text{ with } n \text{ arcs}\}} C(P_k)$.

The total n-step effects of node (species) i is the sum of its effects on every other species j[37].

$$\sigma_{n,i} = \sum_{j=1}^{j=N} a_{n,ij} \tag{2.5}$$

For the calculation of the topological importance of node i when effects are considered up to n steps we normalize i's n-step effects with the total number of steps we want to consider (n)[38]:

$$TI_i^n = \frac{\sum_{m=1}^{m=n} \sum_{j=1}^{j=N} a_{m,ij}}{n} = \frac{\sum_{m=1}^{m=n} \sigma_{m,i}}{n}. \tag{2.6}$$

We can only consider steps up to the biggest path in the network, e.g., if the biggest path is 3, $n \leq 3$.

We made the simplifying assumption that community-level effects spread both bottom-up and top-down, with equal strength in both directions, so we used only undirected links (i.e., undirected graphs).

Biologically, these indices preview that the specie(s) with higher attributed value will be the ones with more direct and indirect effects in all the network. Thus, if they are deleted from the network it will, more likely, generate cascade events. This index considers how many interactions that particular species has in the network. Direct and indirect effects have the same importance in unweighted graphs, whether in weighted ones it will depend in the strength of the connection.

2.2.1.5 Status index and its components (s_i, s'_i, Δs_i)

Considering the food web as a directed acyclic graph (DAG), the status is the sum of distances from node i to each other nodes[37]. Reverting the direction of the links, the same calculation will give the contrastatus of each node (s'_i). These indices were primarily used in sociology[26] but rapidly transposed to biological use, applied first, to food webs[39].

Δs_i is called the net status of node i and it's the difference between status and contrastatus:

$$\Delta s_i = s_i - s'_i \tag{2.7}$$

These indices rank the "power" of a species in a network. Status and net status usually point out the top-predators as the "most powerful" ones and basal species as the "most powerless", while contrastatus is the opposite: it considers the basal species as the powerful ones[39]. It was observed in some cases, that the net status can define more accurately who is the most important node for the network than the status or contrastatus[26,39].

2.2.1.6 Keystone index and its components (K_i, $K_{bu,i}$, $K_{td,i}$, $K_{dir,i}$, $K_{indir,i}$)

The keystone index and its components were derived from the status indices, previously discussed[40]. Hence, these indices were primarily developed to find the keystone species in a web, based solely on their position according to trophic interactions, i.e., disregarding their secondary interactions (competition, mutualism, etc.). Keystone index considers secondary interactions between species since the calculation of the value for each node accounts the sequent chain of nodes and their connections. In fact, although these indices are based in global indices, they are of meso-scale type (keystone index and its components, contrary to status and its derivations, account for the neighbours of the neighbours and consider a decreasing importance of the effects of the neighbouring nodes with the increase of the path length).

The keystone index of a species i is defined as[35]:

$$K_i = K_{bu,i} + K_{td,i} = \sum_{c=1}^{n} \frac{1}{d_c}\left(1 + K_{bu,c}\right) + \sum_{e=1}^{m} \frac{1}{f_e}\left(1 + K_{td,e}\right) \tag{2.8}$$

$$K_i = K_{indir,i} + K_{dir,i} = \underbrace{\left(\sum_{c=1}^{n} \frac{K_{bu,c}}{d_c} + \sum_{e=1}^{m} \frac{K_{td,e}}{f_e}\right)}_{K_{indir,i}} + \underbrace{\left(\sum_{c=1}^{n} \frac{1}{d_c} + \sum_{e=1}^{m} \frac{1}{f_e}\right)}_{K_{dir,i}} \tag{2.9}$$

Where $K_i, K_{bu,i}, K_{td,i}, K_{dir,i}, K_{indir,i}$ stand for five indices that can be considered individually: the keystone index (bidirectional) of species i, the bottom-up and top-down keystone indices, and the keystone indices accounting for direct and indirect-effects, respectively. n represents the number of direct predators of species i, d_c is the number of prey species of its c^{th} predator and $K_{bu,c}$ is the bottom-up keystone index of the c^{th} predator. Similarly, m is the number of direct preys of species i, f_e is the

number of predators of this e^{th} prey being considered and $K_{td,e}$ the top-down keystone index of e^{th} prey. Equations (2.8) and (2.9) were rearranged in order to show their meaning[35].

Biologically, as the bottom-up and top-down indices consider the interactions of a species in the food web, we know that the removal of an important species will lead to more disconnections in both directions of the food web. Therefore, these indices count the number of species that will be disconnected after the removal of species i.

The keystone index, K, loses that specific information, since it's the sum up of both effects (either indirect and direct or top-down and bottom-up), and thus, only refers to the importance of a species in maintaining the trophic flow, in a more broad and general sense[13].

2.2.2 Networks dynamics

Studying the dynamical behaviour of a food web and particularly, simulating how the whole system will behave when a particular species is perturbed can retrieve relevant information about the importance of a species or group of species since some of the species can cause large community responses[24]. However, approaching food web dynamics – dynamic sensitivity analyses – is not easy since many communities are known to be complex tangles of interactions. To model such dynamics, we usually need some measurements and labour effort, not always easy to obtain.

To model the dynamics of the hypothetical food webs used for this work, the following system of differential equations was applied[35,41]:

$$\frac{dN_i}{dt} = r_i N_i \left(1 - \frac{N_i}{K_i}\right) + \sum_{\rho=\text{resources}} N_i \varepsilon_{i\rho} \frac{N_\rho^h \omega_{i\rho}}{N_0^h + \omega_{i\rho} Q_{i\rho}} - \\ - \sum_{c=\text{resources}} N_c \varepsilon_{ci} \frac{N_i^h \omega_{ci}}{N_0^h + \omega_{ci} Q_{ci}} - d_i N_i \tag{2.10}$$

Here, N_i means the abundance of species i, r_i is the rate of increase, K_i the carrying capacity of the logistic model, d_i the mortality rate of consumer species i, $\omega_{i\rho}$ is species i's relative consumption rate when consuming species ρ, N_0 is the half-saturation density, $Q_{i\rho}$ is the sum of the abundances of the resources i can consume. For more information about how this dynamics was performed, one can consult the original works[35,41]. The difference here was that we were only interested in single-species perturbations and how their community reacted to those perturbations.

2.2.2.1 Community response

As in the already mentioned work[8], the community response (CR_j) of species j was measured as the average of the population variations, of all species, after the perturbation of j (without considering self-effects):

$$CR_j = \sum_{i=1}^{n} \left| \frac{N_{i(j)}^t}{N_i^t} - 1 \right| / 14, \quad (i \neq j) \tag{2.11}$$

$N_{i(j)}^t$ is the population size of species i, at time t in a simulation where species j was perturbed and N_i^t is the population size of species i at time t without perturbations, i.e., in a reference simulation.

2.2.3 Single Index Correlation

Since we still don't know how the topological indices described are correlated with this dynamical simulation[35], two simple analyses were made to compare each structural index with the simulation: a metric one, using the normalized outcome values of the topological indices and the simulation ($M_{r_{12n\times k}}$), and an ordinal one, considering the rank order of each node for each structural index and the community response simulation ($M_{ro_{12n\times k}}$). To better understand the similarities between indices, Spearman Correlation was applied to both matrices, using "spearmanr" function from "scipy.stats"[42] in Python[43].

2.2.4 *N*–Index Correlation

To compare the *N*-indices, we also performed an UPGMA classification, using all the topological indices' results, and excluding the community response values, in order to understand which are the most related, i.e., which indices are more redundant and thus don't bring new information to the analysis.

We performed this analysis again in both matrices ($M_{r_{12n\times k}}$ and $M_{ro_{12n\times k}}$). The distances between indices i and j were calculated using $d_{ij} = 1 - |\rho_{ij}|$, where ρ_{ij} is the Spearman Correlation between indices i and j for each network. The next step was to perform a consensus dendrogram using the majority rule – only the groups with a 50% appearance in the original set of 1000 networks will appear in this final dendrogram.

This statistical analysis was performed using R[31], package "ape"[44].

2.2.5 Combination of indices

In theory, all topological indices used in this work have their biological importance, ranging either from local to global characterizations of species in the network, based either in neighbourhood and or distance. On the other hand, a lot of different dynamical approaches have been applied to describe the role of species in the ecosystems, although these approaches are usually more time and effort consuming[8].

Using some of the topological indices to try to predict the same results obtained throughout dynamical approaches would be of added value inasmuch as static food webs and the embedded species importance are easier to assess.

Here, we try to address a call made by other studies[27] in the attempt of replacing complicated, time and effort consuming, sometimes, even hard to compute formulae – as are dynamical simulations – with simpler and easier ones, without much information loss. We also attempt to better explain the similarity between centrality indices and how they are related to dynamical analyses, since there's still a lack of bridging information.

2.2.6 Program used

In order to find a mathematical expression that could predict the community response values, based in the used topological indices, we used "gplearn" version 0.4.0 – "Genetic Programming (GP) with symbolic regression" – a free open source code[45], adapted. This code extends the "scikit-learn"[46] machine learning library available for Python[43].

Roughly, this algorithm works by 1) applying one of the basic mathematical operations (namely addition, subtraction, multiplication, division) to two of the independent variables (in our case, structural indices), all chosen randomly. Then, 2) a prediction for the community response values is calculated, based in this randomly generated mathematical formula. This prediction is then 3) compared to the real values of the community response simulation ("Fitness"). The program works by generating an initial number of mathematical random formulae (chosen by the user) and, after, from this population a part is chosen ("Selection"). 4) The best individual – mathematical formula that performed better in predicting the community responses' values – will act like a "parent" for the next generation.

This individual can undergo through mutations ("Evolution") in order to generate the next generation (with the same initial population size).

This process is then repeated for g generations or until the parents reach perfection (which in programming means accuracy = 100%).

2.2.6.1 Parameters used

The Symbolic Regressor algorithm was trained using the values from both matrixes ($M_{r_{12n\times k}}$ and $M_{ro_{12n\times k}}$). An attempt of performing a grid search was made but due to computational and time constraints different parameters were applied instead, to induce more variability in the generated results.

We performed 10000 iterations in each of which we assigned random values based in a uniform distribution to the "mutation" parameters (see Appendix A). Due to program constraints, the sum of these variables couldn't surpass one. Each final mathematical formula obtained was the best result of an evolutionary process of 100 generations, chosen from half of a population that started with an initial population size of 1000, 5000 and 10000. Spearman correlation, using a cross validation test, was used to access the quality of the results. The data was partitioned into ten subsets of equal size. One of these ten subsets was used as the test data set, while the other nine were used as the training data set. The

cross-validation process was then repeated ten times, with each of the ten subsets used once as the test set. The results of all iterations were then averaged to produce a single estimation for the accuracy.

The parameters used can be accessed in more detail in Appendix A.

Chapter 3

Results and Discussion

In this section we present the results of the single correlation and cluster analysis. We also present the best combinations of *k*-indices obtained, i.e., the best mathematical formulae obtained to predict the community response values (dependent variable). The most common mathematical formulae obtained in the overall results are also showed. Results for both metric and ordinal data are showed and commented.

3.1 Cluster analysis

The consensus dendrogram (Figure 3.1) was the same, for both matrixes' datasets, so only one figure is showed. This dendrogram confirms the two big groups previously observed: correlated {wD, K_{bu}, WI^5, WI^3, K_{td}, s, Δs, WI^1, s'} and uncorrelated indices {K_{indir}, K_{dir}, K, BC, TI^1, TI^3, TI^5, D, CC} – see the two first branches formed (at 75%, for example). In addition, if we consider a distance of 40% (see Appendix B) we can form four groups. Distances between 30% and 35% allow us to form five groups: {WI^1, WI^3}, {Δs, s', K_{td}}, {K_{bu}, wD, s, WI^5}, {K, K_{dir}, K_{indir}}, {TI^3, TI^5, TI^1, BC, CC, D} – we will focus our analysis in these five groups.

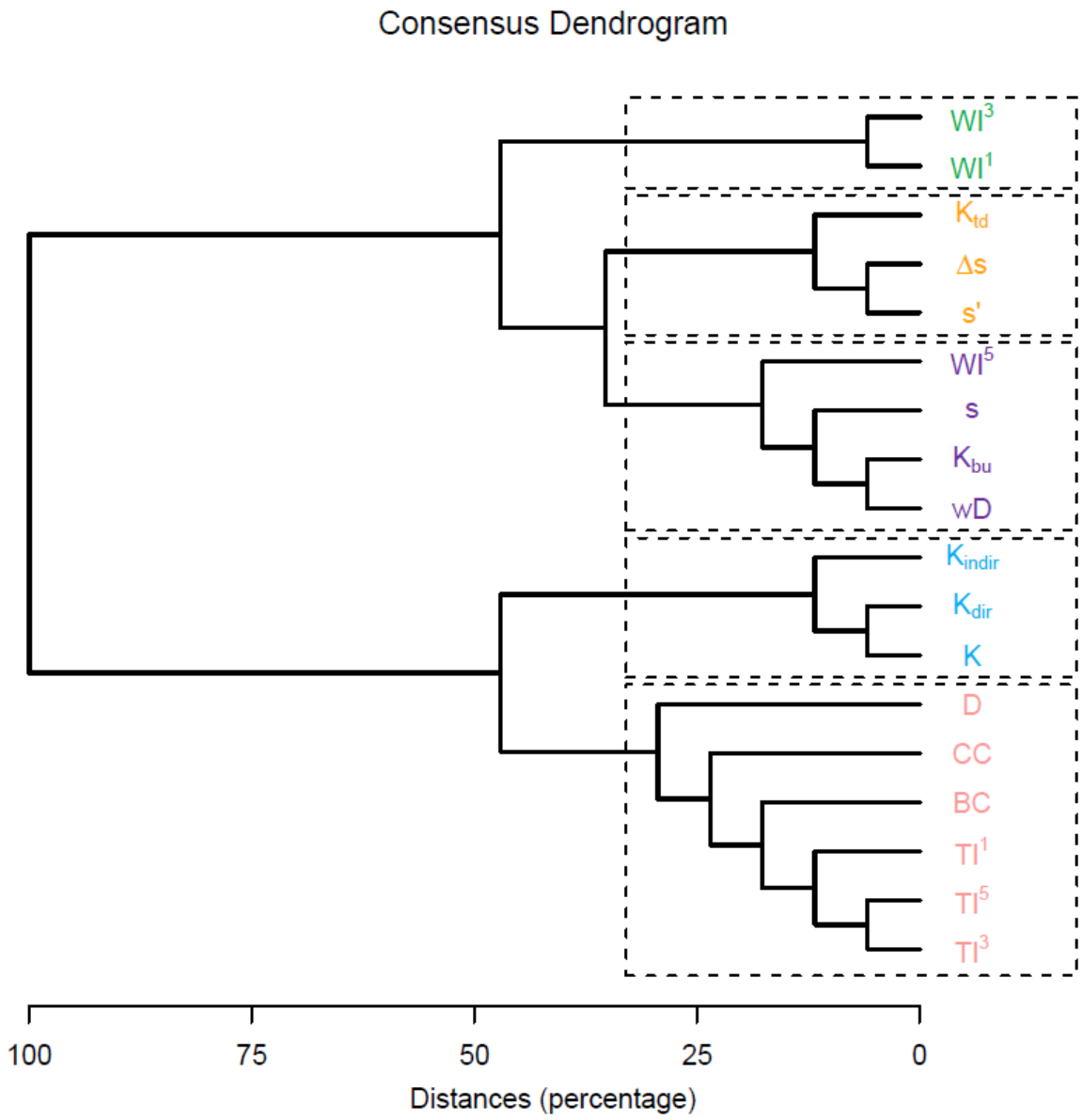


Figure 3.1. Consensus dendrogram between topological indices. Between distance 35% and 30% (a distance of 32,5% was used to select groups), five groups can be formed: {WI^1, WI^3}, {Δs, s', K_{td}}, {K_{bu}, wD, s, WI^5}, {K, K_{dir}, K_{indir}}, {TI^3, TI^5, TI^1, BC, CC, D}.

These groups show redundant information: indices within same groups are highly correlated to each other, due to their mathematical construction or to the final attributed results regarding nodes' importance in networks. Consequently, if we want to simply evaluate a food web in a static way, we can use one index from each group and, therefore, we can reduce analysis' complexity, without losing much information.

3.2 Single correlation

Table 3.1 and Table 3.2 show the Spearman correlation for each of the 18 topological indices related to the community response. Table 3.1 was calculated using the metric data ($M_{r_{12n\times k}}$). Table 3.2 was calculated using ordinal data ($M_{ro_{12n\times k}}$). For better results' visualization, indices were coloured as in the five groups presented in Figure 3.1.

Table 3.1. Spearman correlation and respective *p*-values related to all indices used and the simulation of the "community response" – metric values.

Indices	Spearman correlation (ρ) [c]	*p*-value [e]	Weighted	Directed
wD	**-0.7006**	**0.00**	**yes**	**no**
WI^5	**-0.6773**	**0.00**	**yes**	**no**
K_{bu}	**-0.6755**	**0.00**	**no**	**yes**
WI^3	**-0.6645**	**0.00**	**yes**	**no**
WI^1	**-0.6002**	**0.00**	**yes**	**no**
K_{td}	**0.5986**	**0.00**	**no**	**yes**
s	**-0.5866**	**0.00**	**no**	**yes**
Δs	**-0.5806**	**0.00**	**no**	**yes**
s'	**0.5438**	**0.00**	**no**	**yes**
K_{indir}	0.3226	0.00	no	yes
K	0.2665	0.00	no	yes
K_{dir}	0.1524	0.00	no	yes
BC	-0.1401	0.00	no	no
TI^1	-0.1381	0.00	no	no
TI^3	-0.1059	0.00	no	no
TI^5	-0.1009	0.00	no	no
D	-0.0548	0.00	no	no
CC	-0.0330	0.00	no	no

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [c] Spearman correlation (ρ) was calculated for each of the 18 topological indices versus the community response results; [e] p-values were derived for each Spearman correlation.

Table 3.2. Spearman correlation and respective *p*-values related to all indices used and the simulation of the "community response" – ordinal values.

Indices	Spearman correlation (ρ) [c]	*p*-value [e]	Weighted	Directed
wD	**-0.6884**	**0.00**	**yes**	**no**
K_{bu}	**-0.6701**	**0.00**	**no**	**yes**
WI^5	**-0.6690**	**0.00**	**yes**	**no**
WI^3	**-0.6565**	**0.00**	**yes**	**no**
K_{td}	**0.6169**	**0.00**	**no**	**yes**
s	**-0.6014**	**0.00**	**no**	**yes**
Δs	**-0.5931**	**0.00**	**no**	**yes**
WI^1	**-0.5916**	**0.00**	**yes**	**no**
s'	**0.5650**	**0.00**	**no**	**yes**
K_{indir}	0.3433	0.00	no	yes
K	0.2804	0.00	no	yes
K_{dir}	0.1569	0.00	no	yes
TI^1	-0.1396	0.00	no	no
BC	-0.1352	0.00	no	no
TI^3	-0.1123	0.00	no	no
TI^5	-0.1071	0.00	no	no
D	-0.0689	0.00	no	no
CC	-0.0459	0.00	no	no

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [c] Spearman correlation (ρ) was calculated for each of the 18 topological indices versus the community response results; [e] *p*-values were derived for each Spearman correlation.

Indices {wD, K_{bu}, WI^5, WI^3, K_{td}, s, Δs, WI^1, s'} correlated, in both cases, with community response values ($\rho \geq 0.5$, *p*-value ≤ 0.05) while {K_{indir}, K_{dir}, K, BC, TI^1, TI^3, TI^5, D, CC} were not correlated. Only two of the correlated indices were positively correlated: the keystone index based on top-down approach and the contrastatus {K_{td}, s'}. All indices correlated better with the rank order data than with the metric data. However, {wD, WI^5, K_{bu}, WI^3, WI^1} performed better when applied to metric data – increasing the correlation in about 0.85%.

It is also noticeable that the order of the indices (from the most correlated one to the least correlated one) was slightly different whether we used ordinal and metric values: {wD, K_{bu}, WI^5, WI^3, K_{td}, s, Δs, WI^1, s'} versus {wD, WI^5, K_{bu}, WI^3, WI^1, K_{td}, s, Δs, s'}. Indices {TI^1, BC} were also switched in both tables. In addition, we can observe that the group of indices {K_{td}, s, Δs, s'} performed better when applied to ordinal data. These differences are probably due to the random order assigned to nodes

when draws between values were found – since the Spearman correlation is based in the correlation between data ranks.

3.3 Combination of k – Indices

The mathematical expressions obtained can be consulted in Tables 3.3 – 3.5. Only the top 20 results are shown. Table 3.3 shows the mathematical expressions with the best Spearman correlation combinations' values, Table 3.4 displays the most frequent results (out of the 30 000 generated in total: 10 000 for each initial population size) and Table 3.5 shows the indices' groups (or "families", within the top 20 most frequent). Results for ordinal data can be consulted in Appendix B.

Table 3.3. Best mathematical expressions, derived from the algorithm used, according to absolute Spearman correlation results.

Results		Spearman correlation (ρ) [c]	Relative frequency (percentage) [r]
$K_{indir}{}^2 \times WI^5 \times s \times (CC - K_{dir})(K_{dir} - 0.028) + + \left(\frac{WI^5}{CC} - \frac{D}{WI^5}\right)\frac{K_{bu} \times WI^3}{wD^2} - (D - \Delta s)(\Delta s - K) \times \times (wD \times TI^1)\left(\frac{BC \times wD^2}{D} - WI^5 \times \Delta s + TI^{3\,2}\right)$	(3.1)	0.7842	0.003
$\left(\frac{TI^5}{WI^5} - WI^3 - BC\right)\left(\frac{s}{wD} + K_{td} - BC\right) \times \times \left(TI^5 \times WI^3 + K_{td} - TI^1 + \frac{WI^3 \times D}{WI^{5\,2}}\right)$	(3.2)	0.7818	0.003
$wD + \frac{WI^5}{D + CC \times s}$	(3.3)	0.7801	0.003
$\frac{\Delta s}{s} \times wD \times TI^5 + \frac{WI^5}{TI^5}$	(3.4)	0.7801	0.003
$(CC + 1)WI^5 + wD - \frac{TI^5}{WI^5} - - \left(\frac{K_{bu}}{wD} + \Delta s - BC\right)\left(WI^3 \times WI^5 + \frac{D}{WI^3}\right)$	(3.5)	0.7799	0.003
$\frac{WI^5}{TI^3} + TI^1 + BC - \frac{D}{WI^5}(\Delta s + K_{td})$	(3.6)	0.7791	0.003
$TI^5 + wD - 1 - (K - BC)(TI^3 + K_{dir}) - - \left(\frac{wD}{K_{bu}} - \frac{D}{WI^3} + \frac{WI^5 \times CC}{TI^3 \times TI^5}\right)$	(3.7)	0.7789	0.003

$\frac{(s + TI^5)(wD - D)}{BC \times wD + TI^1 \times WI^5} + (WI^5 + \Delta s) \times$ $\times (0.833TI^1)(CC - WI^3 + K_{td} \times D)$	(3.8)	0.7788	0.003
$WI^5 \times 0.547 - CC \times K_{td} - \left(\frac{D}{0.804}\right) \times$ $\times (\Delta s - D) - 1 - wD \times K - 0.099TI^5 - \frac{WI^3}{D}$	(3.9)	0.7788	0.003
$\frac{D}{WI^5}(\Delta s + K_{indir}) - (TI^3 + CC)\frac{wD}{WI^5}$	(3.10)	0.7786	0.003
$\frac{WI^5}{D} - (K_{bu} \times TI^1) + \frac{BC}{\Delta s} \times \frac{wD}{TI^5}$	(3.11)	0.7786	0.003
$\frac{(TI^3 + CC)wD}{K_{bu}} - D\left(\frac{1}{WI^5} + CC\right)$	(3.12)	0.7784	0.003
$\left(s' + wD - \frac{D}{WI^5}\right)(TI^1 \times CC - K_{bu} - 0.916)$	(3.13)	0.7782	0.003
$\left(\frac{WI^5}{D} + K_{bu}\right)\frac{wD}{K_{bu}}$	(3.14)	0.7781	0.003
$\left(\frac{WI^5}{D \times K_{bu}} + 1\right)wD$	(3.15)	0.7781	0.003
$(TI^1 + WI^1)(s - 0.314) + wD\left(1 + \frac{1}{D}\right) - s$	(3.16)	0.7781	0.003
$\frac{0.990D}{0.593WI^5}(\Delta s - wD + s' \times K)$	(3.17)	0.7770	0.003
$(1 + 0.661K_{td} - TI^3 \times D)(WI^5 \times s \times K \times$ $\times TI^1 - \frac{D}{WI^5} + s' + K_{td})$	(3.18)	0.7770	0.003
$(BC - s + WI^5 \times wD) - (TI^3 + WI^3) \times$ $\times (WI^1 - CC) - \left(\frac{D}{WI^1} + \Delta s + K_{td}\right)\left(\frac{WI^1}{WI^5} - s \times K\right)$	(3.19)	0.7770	0.003
$(\Delta s - K_{dir})(\Delta s - TI^5) - (CC \times wD) - \frac{WI^5}{D}$	(3.20)	0.7770	0.007
Average *		0.7789 ± 0.0667	0.003

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [c] Spearman correlation (ρ) is the Spearman correlation when testing the performance of the obtained formula in 10% of the data (test data); [r] Relative frequency in percentage is the number of times each result appeared among all results (i.e., among the 30 000 iterations executed), in percentage. * Average is the average of the values showed. Standard deviation for the Spearman correlation of these mathematical expressions was the average of the standard deviations when applied to 90% of data versus all data. Metric data was used in order to obtain these results.

Table 3.4. Most frequent mathematical expressions derived by the algorithm used.

Results		Spearman correlation (ρ) [c]	Relative frequency (percentage) [r]
$\frac{WI^5}{D}$	(3.21)	0.7616	15.027
$\frac{D}{WI^5}$	(3.22)	0.7616	6.733
$\frac{WI^5}{TI^5}$	(3.23)	0.7586	2.897
$\frac{TI^5}{WI^5}$	(3.24)	0.7586	2.157
$\frac{WI^5}{TI^3}$	(3.25)	0.7546	1.257
$\frac{TI^3}{WI^5}$	(3.26)	0.7546	0.983
$\frac{WI^3}{D}$	(3.27)	0.7540	0.733
$\frac{D}{WI^3}$	(3.28)	0.7702	0.593
$wD + \frac{WI^5}{D}$	(3.29)	0.7702	0.573
$\frac{WI^3}{D} + wD$	(3.30)	0.7694	0.557
$\frac{TI^5}{WI^3}$	(3.31)	0.7539	0.445
$\frac{WI^3}{TI^5}$	(3.32)	0.7538	0.363
$\frac{wD}{D}$	(3.33)	0.7535	0.330
$BC - \frac{D}{WI^5}$	(3.34)	0.7529	0.290
$\frac{D}{WI^5} - TI^3$	(3.35)	0.7686	0.276
$\frac{D}{WI^5} - BC$	(3.36)	0.7686	0.270
$TI^3 - \frac{D}{WI^5}$	(3.37)	0.7677	0.267

$\frac{D}{WI^5} - TI^1$	(3.38)	0.7676	0.260
$\frac{D}{WI^5} - K_{dir}$	(3.39)	0.7660	0.237
$TI^5 - \frac{D}{WI^5}$	(3.40)	0.7677	0.233
Average *		0.7617 ± 0.0016	1.724

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [c] Spearman correlation (ρ) is the Spearman correlation when testing the performance of the obtained formula in 10% of the data (test data); [r] Relative frequency in percentage is the number of times each result appeared among all results (i.e., among the 30 000 iterations executed), in percentage. * Average is the average of the values showed. Standard deviation for the Spearman correlation of these mathematical expressions was the average of the standard deviations when applied to 90% of data versus all data. Metric data was used in order to obtain these results.

Table 3.5. Most frequent mathematical "families" of indices derived from Table 3.4.

Results	Relative frequency (percentage) [r]
WI^5, D	21.760
WI^5, TI^5	5.054
WI^5, TI^3	2.240
WI^3, D	1.326
WI^3, TI^5	0.808
wD, WI^5, D	0.573
WI^5, BC, D	0.560
WI^3, wD, D	0.557
WI^5, D, TI^3	0.543
wD, D	0.330
WI^5, D, TI^1	0.260
WI^5, K_{dir}, D	0.237
WI^5, D, TI^5	0.233

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [r] Relative frequency in percentage is the number of times each result appeared among all results (i.e., among the 30 000 iterations executed), in percentage.

The top 20 mathematical combinations used a range of 4 to 14 different indices in the same expression (Table 3.3). These results showed significant increases in correlation values when compared to single-correlations: an average of 77.89% versus 62.53%.

The mathematical expression with the best Spearman correlation showed an increase of about 8.36% when compared to the best correlation using a single index (78.42% versus 70.06%). Furthermore, an average, of 3.5 of the groups formed in the consensus dendrogram were used (and the median of groups used was of 4).

We also analysed the most frequent mathematical expressions derived from the algorithm used. In this scenario, only two or three indices were combined and Spearman correlation was significantly better than single-index correlations. For instance, Spearman correlation obtained was 77.02% (for two different mathematical expressions – one relying in two indices and the other one in three). The loss of information here is not significant since the previously best correlation obtained was of 78.42% (and 70.06% for the single correlation), which implies that this approximation is valid to predict our "target values" more than using only one topological index. The average correlation was also slightly worse than the previous one (76.17% vs 77.89%) but this difference is also not significant and allow us to conclude that, in general, we can use these simpler mathematical expressions (Table 3.4) to get good predictions without much information loss and with a decrease of effort and computational time.

Table 3.5 shows the most common groups of indices. We can see that the "purple family" was present in almost every expression (in Table 3.4). This family was more frequently married with the pink, green and blue. Note that green, yellow and purple families represent community response correlated indices while blue and pink represent not correlated indices. Consequently, when using metric data one can use correlated and uncorrelated indices together to have strong and fast predictions (regarding community response values).

For these predictions, we can rely on 13 "families" that mix 9 indices (see Table 3.5): $\{WI^5, D\}$, $\{WI^5, TI^5\}$, $\{WI^5, TI^3\}$, $\{WI^3, D\}$, $\{WI^3, TI^5\}$, $\{wD, WI^5, D\}$, $\{WI^5, BC, D\}$, $\{WI^3, wD, D\}$, $\{WI^5, D, TI^3\}$, $\{wD, D\}$, $\{WI^5, D, TI^1\}$, $\{WI^5, K_{dir}, D\}$, $\{WI^5, D, TI^5\}$. This means that in order to simplify our approach and analysis we can rely in some combination of these nine indices. The most frequent index was the degree (D), it occurred in 10 out of 13 families, followed by weighted topological importance 5-step-long (WI^5), present in 9. Topological indices TI^5, WI^3, wD, TI^3, TI^1, BC and K_{dir} were also important. It is also interesting to highlight that the combination of indices $\{WI^n, TI^n\}$ was present in half of these mathematical expressions and the combination of $\{WI^n, D\}$ in 13 out of the most frequent top-20 expressions showed.

Chapter 4

Conclusions

The use of ecological models is being increasingly used to better understand interactions between organisms in ecosystems. Reducing complex systems to mathematical representations such as networks allow us to quantify the parts within the whole and thus, food web representations fill an important place to study topological and dynamical interactions. Since ecological modelling is a mathematical concept we can, through simulations and quantitative computational analysis, try to understand how complex biological systems are connected and how their species interact with each other[12,47].

Food webs are representations used to depict in a simple way, trophic interactions between organisms, from a micro to a macro level[48]. Thus, prey-predator relationships allow us to understand the dynamics and resilience of communities: feeding interactions give us insights related to vital rates, rates of growth and rates of mortality. These rates depend on the animals that are eaten or being eaten; if the population under analysis is eating more preys, the population grows; if the population is being eaten by predators, it is decaying[49].

Applying network statistics to food webs allow us to have some perception about the global scale – global topological indices provide us information to understand the network as a whole but give less insights related to each individual in the network. The local perspective, in the other side of the scale, allow us to better understand how each individual is connected in the network. The problem is that, from a local perspective, we don't have any grasps about further interactions, i.e., about direct or indirect effects that that individual may be causing to the rest of the network. Due to this, a variety of mesoscale indices emerged. These indices add some of these missing information – they allow us to look to each species' topological position with some insights about its interactions within the community[12].

However, these representations are of continuously changing systems: life is not static and much less the intricate nets where species interact. Every species has their rate of growth, mortality and vitality and these are usually dependent on other species. As a result, considering only a static perspective may not be very useful for the sake of the knowledge inferred about these complex systems. In order to understand how time and topological fluctuations affect populations, a number of different dynamic analysis raised[8]. These dynamics are often described by ordinary differential or difference equation systems, that take different population parameters into account.

Nonetheless, databases holding life history, demographic and species interaction information necessary to parameterize ecological network models are rarely available[50]. Due to these, hypothetical medium sized food webs were generated and used as real ones. Our work focused in trying to understand how well some topological indices available and already used in ecological modelling can predict which are the keystone species (critic species in a food web) when compared to each other and when compared to a dynamical simulation.

Understanding the dynamics of food webs, particularly, which species are keystones (or of major importance) is also a greater concern for conservation biology that aims to be more functional[28]. To understand which are the species that, in case of extinction, will originate a collapse in the food web is primordial nowadays, since, if we can't save all species from habitat loss we can, at least, prevent the

extinction of the most (or group of) important ones, in order to avoid a cascade loss[51]. Thus, understanding the structure and dynamics of a network is primordial to infer knowledge.

In the literature there are studies suggesting the use of different topological network indices to characterize the importance of a species in a community and there are some also comparing their performance[11,12,27-29]. There are also some dynamical analyses that have been made on food webs in order to predict which are these key species in a community and how the others react to their perturbations[29,52].

In addition, there are studies on the relationship between structural centrality and simulated importance[24]. However, to our knowledge, this is the first attempt to combine different centrality indices and to test the correlation between these combined indices and simulated importance.

Our results showed that weighted topological indices 1 and 3-step long are in the same branch of the dendrogram obtained which may indicate that taking small paths from the focusing species into account (1 to 3 in this case) is not very different. The Δs, s' and the K_{td} might be related because mathematically, they all derive from the status index.

We found that wD, WI^5 and K_{bu} are the most reliable topological indices, when we want to find the critic species in a food web (when we are comparing to the dynamical analysis used), with an accuracy rate of almost 70%. It is important to notice that these dynamical analyses tell us how the population of an organism responds when each of the others in the food web are disturbed, i.e., when they almost reach extinction. It is also interesting to note that the weighted degree is a local index – comprising only the species directly attached to the species being considered – but it considers different weights for these species. The weighted topological importance 5-step-long and the bottom-up keystone indices are mesoscale. These results are not surprising since it makes sense, that when analysing the importance of a species, we should consider the importance (weights) of the species directly attached to it. On the other hand, it also makes sense to look to closer species' interactions, in a shorter to a medium level – since the species to which it is connected, will suffer more if they are directly dependent on it.

It is also noticeable that the degree (D) and the 5-step-long weighted topological importance index (WI^5) were the ones that appeared the most in all results. This is interesting if we consider that the degree is a simple, direct index that provides the direct interactions and the 5-step-long weighted topological importance index is a complex, weighted index that considers also indirect interactions. Furthermore, we found that this index is symmetrically correlated with the simulation results, unlike the degree, that is not correlated at all. These are totally different indices, and this is why it is interesting and convenient to combine them: they provide complementary and adequate information.

The results obtained provide new ways to achieve the order of importance of each species obtained through complex biological simulations. We can either use simple structural indices, or some combinations of these indices, ranging from simpler ones to more complex. More complex combinations allow to increase even more the accuracy of the results.

Future studies could focus on 1) use more food webs in order to get more accurate combinations 2) apply the simple indices or combinations obtained to real food webs to identify which are the keystone species and check their biological role in those networks 3) apply the same approach to other network types (e.g. metabolic networks). We believe that more concise and efficient ways to identify keystone species in ecological networks will be essential for the future of systems-based ecology that aims to

achieve objective conservation priorities or regulations to manage ecosystems. We suggest that our machine learning-based approach to maximize the predictive power of structural analysis can be a major step towards simple and fast, yet quite realistic research on food webs.

References

1 Begon, M., Townsend, C. R. & Harper, J. L. *Ecology: From Individuals to Ecosystems*. (Blackwell Publishing, 2006).

2 Morin, P. J. *Community Ecology.* (Wiley-Blackwell, 2011).

3 Ingalls, B. *Mathematical Modelling in Systems Biology: An Introduction*. (MA, USA:MIT Press, 2012).

4 Newman, M. E. J. *Networks - An Introduction*. (Oxford University Press Inc., 2010).

5 Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).

6 Warren, P. H. Spatial and Temporal Variation in the Structure of a Freshwater Food Web. *Oikos* **55**, 299-311 (1989).

7 Zhang, D., Yin, J., Zhu, X. & Zhang, C. Network Representation Learning: A Survey. *IEEE Transactions on Big Data*, 1-25 (2018).

8 Jordán, F. & Scheuring, I. Network ecology: Topological constraints on ecosystem dynamics. *Physics of Life Reviews*, 139-172 (2004).

9 Harary, F. *Graph Theory*. (Addison-Wesley Publishing Company, Inc., 1969).

10 Ghasemi, M., Seidkhani, H., Tamimi, F., Rahgozar, M. & Masoudi-nejad, A. Centrality Measures in Biological Networks. *Curr. Bioinform.* **9**, 1-17 (2014).

11 Estrada, E. Characterization of topological keystone species. Local, global and 'meso-scale' centralities in food webs. *Ecol. Complex.* **4**, 48-57 (2007).

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

13 Jordán, F., Scheuring, I. & Vida, G. Species positions and extinction dynamics in simple food webs. *Journal of Theoretical Biology* **215**, 441–448 (2002).

14 Ceballos, G., Ehrlich, P. R. & Dirzo, R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E6089–E6096 (2017).

15 Pimm, S. L., Lawton, J. H. & Cohen, J. E. Food web patterns and their consequences. *Nature* **350**, 669–674 (1991).

16 Morin, P. J. & Lawler, S. P. Food Web Architecture and Population Dynamics: Theory and Empirical Evidence. *Annual Review of Ecology and Systematics* **26**, 505–529 (1995).

17 Elton, C. S. *Animal ecology*. (Macmillan Co., 1927).

18 Lindeman, R. L. The Trophic-Dynamic Aspect of Ecology. *Ecol. Ecol. Soc. Am.* **23**, 399–417 (1942).

19 May, R. M. Will a Large Complex System be Stable? *Nature* **238**, 413–414 (1972).

20 Pimm, S. L. & Lawton, J. H. Number of trophic levels in ecological communities. *Nature* **268**, 329-331 (1977).

21 Pimm, S. L. Food Web Design and the Effect of Species Deletion. *Oikos* **35**, 139 (1980).

22 Rybarczyk, H. & Elkaïm, B. An analysis of the trophic network of a macrotidal estuary: The Seine Estuary (Eastern Channel, Normandy, France). *Estuar. Coast. Shelf Sci.* **58**, 775–791 (2003).

23 Cozzens, M. M. in *Algebraic and Discrete Mathematical Methods for Modern Biology* 29–50 (Rutgers University, 2015).

24 Livi, C. M., Jordán, F., Lecca, P. & Okey, T. A. Identifying key species in ecosystems with stochastic sensitivity analysis. *Ecological Modelling* **222**, 2542-2551 (2011).

25 Paine, R. T. *et al.* Challenges in the Quest for Keystones. *Bioscience* **46**, 609–620 (1996).

26 Harary, F. Status and Contrastatus. *American Sociological Association* **22**, 23–43 (1959).

27 Jordán, F., Okey, T. A., Bauer, B. & Libralato, S. Identifying important species: Linking structure and function in ecological networks. *Ecological Modelling* **216**, 75-80 (2008).

28 Jordán, F., Benedek, Z. & Podani, J. Quantifying positional importance in food webs: A comparison of centrality indices. *Ecological Modelling* **205**, 270–275 (2007).

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

30 Lang, M., Bossek, J., Horn, D., Richter, J. & Surmann, D. Package ‘ BBmisc ’. (2017).

31 Team, R. C. R: A Language and Environment for Statistical Computing. (2018).

32 Csardi, G. & Nepusz, T. The igraph software package for complex network research. (2006).

33 Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications*. (Cambridge University Press, 1994).

34 Freeman, L. C. A Set of Measures of Centrality Based on Betweenness. *American Sociological Association* **40**, 35-41 (1977).

35 Móréh, Á., Endrédi, A. & Jordán, F. Additivity of pairwise perturbations in food webs: Topological effects. *Journal of Theoretical Biology* **448**, 112–121 (2018).

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

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

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

39 Harary, F. Who Eats Whom? *General Systems* **6**, 41-44 (1961).

40 Jordán, F., Takacs-Sánta, A. & Molnár, I. A reliability theoretical quest for keystones. *Oikos* **86**, 453–462 (1999).

41 Móréh, Á., Jordán, F., Szilágyi, A. & Scheuring, I. Overfishing and regime shifts in minimal food web models. *Community Ecology* **10**, 236–243 (2009).

42 Jones E, Oliphant E, P. P. and others. SciPy: Open Source Scientific Tools for Python. (2001).

43 Rossum, G. van. Python 3. (2017).

44 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in {R}. *Bioinformatics* **35**, 526-528 (2018).

45 Stephens, T. gplearn Documentation Release 0.3.0., 45 (2016).

46 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

47 Serra, H. Using Ecological Modeling to Enhance Instruction in Population Dynamics and to Stimulate Scientific Thinking. *Creative Education* **2**, 83-90 (2011).

48 Ulanowicz, R. E. *Growth and Development Ecosystems Phenomenology*. (Springer-Verlag 1986).

49 Fussmann, G. F. in *Complex Population Dynamics: Nonlinear Modeling in Ecoloty, Epidemiology and Genetics* 1-20 (World Scientific Publishing Co. Pte. Ltd., 2007).

50 Beas-Luna, R. *et al.* An online database for informing ecological network models. *PLoS One* **9**, 1-9 (2014).

51 Allesina, S. & Bodini, A. Who dominates whom in the ecosystem? Energy flow bottlenecks and cascading extinctions. *J. Theor. Biol.* **230**, 351–358 (2004).

52 McDonald-Madden, E. *et al*. Using food-web theory to conserve ecosystems. *Nat. Commun.* **7**, 1-8 (2016).

Appendix A Python Script

Python script used to adapt “gplearn” algorithm to the specific studied problem.

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
@author: Catarina Gouveia
"""
# Importing the libraries
import pandas as pd
from gplearn.genetic import SymbolicRegressor
from sklearn.model_selection import KFold
import random
from scipy.stats import spearmanr
from numpy.random import RandomState

# Allows the results to be reproducible
random_state = RandomState(seed = 201819)

# Importing the dataset:
path = r"C:\ " # Define the path

# CHANGE THIS - realvalues / rankorder
data = pd.read_csv(path+r"totaldata_realvalues.csv")

# Choosing our data features and matrix of target variable
X = data[data.columns[:-1]] #1st to 18th are our features
y = data[data.columns[-1:]] #last column is our target variable
y_vector = data[data.columns[-1:]].values.ravel()
```

```
var = {}
results = {}
i = 0
for j in range(10000):

    population_size = 1000

    # Atributing random values to the next 4 different variables:
    p_crossover = round(random.uniform(0.5, 0.9), 1)
    p_hoist_mutation = round(random.uniform(0.01, 0.1), 2)
    p_point_mutation = round(random.uniform(0.01, 0.1 - p_hoist_mutation), 2)

    #p_hoist_mutation + p_point_mutation can't exceed 0.1 - otherwise can cause trouble
    p_subtree_mutation = abs(round(random.uniform(0.01, 1 - p_crossover - p_hoist_mutation - p_point_mutation), 2))

    if p_crossover + p_hoist_mutation + p_point_mutation + p_subtree_mutation > 1:
        p_hoist_mutation = 0

    results[j] = {}
    print(j)
    # Splitting the data in 10 folds
    kf = KFold(n_splits = 10, random_state = random_state, shuffle = False)

    while population_size <= 10000:
        all_est_gp = []
        for train_index, test_index in kf.split(data):
            X_train, X_test = X.iloc[train_index], X.iloc[test_index]
            y_train, y_test = y.iloc[train_index].values.ravel(), y.iloc[test_index].values.ravel()
            # This X_train and y_train will be used to fit and generate the model
```

```
 # gplearn model
est_gp = SymbolicRegressor(generations = 100,
          p_crossover = p_crossover,
            p_hoist_mutation = p_hoist_mutation,
          p_point_mutation = p_point_mutation,
            p_subtree_mutation = p_subtree_mutation,
          population_size = population_size,
          function_set = ('add', 'sub', 'mul', 'div'),
          tournament_size = 0.5 * population_size,
          n_jobs = -1,
          low_memory = True,
          metric = 'spearman',
          parsimony_coefficient = 'auto',
            random_state = random_state
            )

est_gp.fit(X_train, y_train)
# New line to append the 10 different solutions to each subset
all_est_gp.append((est_gp.__str__(), str(est_gp._program.raw_fitness_)))
to_calc_test_averageTest = []
to_calc_test_averageTrain = []
to_calc_test_averagepTest = []
to_calc_test_averagepTrain = []

pred3 = est_gp.predict(X_train)

##################################################
 #Correlation: train data
 # (Must be the same of the one that comes out from the algorithm)
scoreTrain_Algorithm, pval_algorithm = spearmanr(y_train, pred3)
```

```
##################################################

# Calculating predictions using X_test, X_train (should be exactly the same
# than the real ones) and with all the data, X (the predictions coming out from the
# ones that are not used to train should be different)
pred = est_gp.predict(X_test) # est_gp is your trained gp
pred2 = est_gp.predict(X_train) # Used as control, to verify
pred_alldata = est_gp.predict(X) # Used as control, to verify

##################################################
scoreTrain, pTrain = spearmanr(y_train, pred2) # Used as control, to verify
 #pTrain contains p-values
scoreTest, pTest = spearmanr(y_test, pred)
score_alldata, p_alldata = spearmanr(y_vector, pred_alldata) # Used as control, to verify
 ##################################################

all_est_gp.append((scoreTrain, pTrain))
all_est_gp.append((scoreTest, pTest))
all_est_gp.append((score_alldata, p_alldata))

 # Save results in dictionary format
results[j]['p_crossover'] = est_gp.p_crossover
results[j]['p_hoist_mutation'] = est_gp.p_hoist_mutation
results[j]['p_point_mutation'] = est_gp.p_point_mutation
results[j]['p_subtree_mutation'] = est_gp.p_subtree_mutation
#results[j]['random_state'] = est_gp.random_state.get_state()
results[j]['population size: '+str(est_gp.population_size)] = {}
results[j]['population size: '+str(est_gp.population_size)] = [all_est_gp]
for train_index, test_index in kf.split(data):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index].values.ravel(), y.iloc[test_index].values.ravel()
```

```
            to_calc_test_averageTest.append(scoreTest)
            to_calc_test_averageTrain.append(scoreTrain)

            to_calc_test_averagepTest.append(pTest)
            to_calc_test_averagepTrain.append(pTrain)

        averageTest = sum(to_calc_test_averageTest)/10
        averagepTest = sum(to_calc_test_averagepTest)/10
        averageTrain = sum(to_calc_test_averageTrain)/10
        averagepTrain = sum(to_calc_test_averagepTrain)/10

        # This X_train and y_train will be used to fit and generate the model
        results[j]['population size: '+str(est_gp.population_size)].append(
        [ #Algorithm fitness [est_gp.__str__(), str(est_gp._program.raw_fitness_), #Algorithm fitness
        (scoreTrain_Algorithm, pval_algorithm), # It has to be the same of the one from the algorithm
        (to_calc_test_averageTest[0], to_calc_test_averagepTest[0]),
        # The Spearman that comes out directly from the data test
        #  (the only data not used to produce the model)
        (averageTest, averagepTest), # Average Spearman derived from the 10 != test sets derived from the Kfold
        (averageTrain, averagepTrain), # Average Spearman derived from the 10 != train sets derived from the Kfold
        (score_alldata, p_alldata)] # It will be similar to the average since all data is also being used)

    if population_size == 1000:
        population_size = 5000
    elif population_size == 5000:
        population_size = 10000
    else:
        population_size += 10000
```

```
i += 1
if i == 100: # Save results from 100 to 100 iterations
    i = 0
    var = results
    results_train = pd.DataFrame.from_dict(var, orient='index')
    results_train.to_csv(path+r'\filename.csv', index=False) # Save file
```

Appendix B Ordinal data matrix results

B1 *k* – Indices combination

B1.1 Top–20 Spearman correlation

Table B1 displays the results. These showed significant increases in correlation values when compared to single-correlation results: an average of 72.15% was obtained with the *k*-index combinations versus 62.80% when using one of the correlated indices (ordinal data).

The mathematical expression with the best Spearman correlation showed an increase of about 3.63% when compared to the best Spearman correlation using a single index (72.47% versus 68.84%).

What's more, the mathematical expressions obtained for the ordinal data considered in average, 3.95 groups (out of the 5 present in the clustering analysis) and most of the expressions included indices from 4 of these groups.

We also analysed the most frequent mathematical expressions derived from the algorithm used. In this scenario, only two or three indices were combined and Spearman correlations were significantly better than for single-index correlations. For instance, Spearman correlation for ordinal data increased to 71.85%, using three indices. In addition, the least correlated mathematical expression in this top-20 represented still an improvement in correlation when compared to the best obtained for single-correlation: 69.96% against 68.84% (see Table 3.2, in main text, and Table B1). However, the average correlation of this top-20 was slightly worse than the previous one (70.61% against 72.15%).

Table B1. Best mathematical expressions, derived from the algorithm used, according to absolute Spearman correlation results – ordinal data.

Results		Spearman correlation (ρ) [c]	Relative Frequency (percentage) [r]
$D - TI^5(1 + CC) - \frac{K_{indir} \times (TI^5 + s')}{K_{td}} + + \frac{(D - K)(s' \times K)}{K_{indir} \times WI^3} + + \left(\frac{wD}{0.015} - TI^5 \times WI^5\right) - \left(K \times wD - \frac{\Delta s}{TI^5}\right) - -(K_{td} - WI^1 - WI^5 - s') \times (2WI^3 - 0.051K_{bu})$	(B. 1)	0.7247	0.003

$TI^3 + WI^5 - \left(K_{td} + \frac{TI^3 \times CC}{s' + wD}\right)$	(B. 2)	0.7237	0.003
$\frac{s \times K_{indir}}{wD \times D} - 2WI^3 + K_{td} + TI^3 + + \left(\frac{s'}{WI^3} + \frac{WI^1}{0.999}\right) \times \frac{\Delta s - K_{bu}}{TI^3 + WI^5} + + \frac{0.268D(TI^3 - WI^5)}{0.598K_{bu}(s' + \Delta s)} - \frac{WI^1 - 2s' - TI^3}{CC \times K(K_{td} + \Delta s)}$	(B. 3)	0.7230	0.003
$D - wD + \Delta s - WI^1 - WI^3 + K_{td} + \frac{K_{bu} + s'}{WI^1 + CC}$	(B. 4)	0.7222	0.003
$\frac{wD \times WI^3 - \frac{WI^5}{BC}}{-0.223K_{td}(\Delta s + D)}$	(B. 5)	0.7218	0.003
$\Delta s - 2wD + K_{td} - \frac{K_{indir} + \Delta s}{K_{indir} + K_{dir}}$	(B. 6)	0.7216	0.003
$WI^1 + WI^3 + wD - \Delta s - K_{td} - 0.774 + + K_{bu} - CC$	(B. 7)	0.7215	0.003
$\frac{TI^1 \times TI^3}{WI^3 + s'} - WI^5 - wD + K_{td} - WI^3$	(B. 8)	0.7213	0.003
$\frac{TI^3}{WI^5} + \frac{K_{td}}{WI^3} + \frac{\Delta s + WI^3}{WI^5 + TI^5}$	(B. 9)	0.7212	0.007
$K_{td} + \Delta s + CC - WI^5 - 2wD + + \frac{s}{0.830(0.299 + WI^5)\left(\frac{K_{dir}}{WI^3} + TI^3 \times \Delta s\right)}$	(B. 10)	0.7211	0.003
$\frac{wD}{0.022} - \Delta s \times D - {K_{td}}^2 + BC + TI^3 - - \frac{TI^1 \times TI^3}{CC(CC - K_{bu})(WI^1 + K_{indir})K}$	(B. 11)	0.7211	0.003

$K_{td} - TI^5 - WI^1 + s' - \frac{wD}{0.220} + \frac{s'}{TI^3} - \frac{K_{dir} + TI^5}{\Delta s \times wD} + \frac{K_{bu}}{K_{td}}$	(B. 12)	0.7210	0.010
$K_{td} \times wD - K_{dir} + D - (WI^5 + s')(wD - 0.672)$	(B. 13)	0.7210	0.003
$\frac{wD}{K_{td} + 0.786\Delta s}$	(B. 14)	0.7207	0.003
$\frac{\Delta s + \frac{K_{td}}{0.755}}{wD}$	(B. 15)	0.7207	0.003
$\frac{-K_{td}(0.607 + WI^3)(K_{bu} + K_{dir})}{wD(CC + s)}$	(B. 16)	0.7207	0.003
$(WI^5 + s') \times \frac{wD}{K_{td}} - \frac{TI^1 \times CC}{WI^1 \times K_{td}}$	(B. 17)	0.7207	0.003
$2WI^1 - 2K_{td} - D + WI^3 + wD + s'$	(B. 18)	0.7206	0.003
$\frac{wD + CC}{wD + K_{bu}} + \frac{wD}{-0.013(\Delta s + K_{td})}$	(B. 19)	0.7206	0.003
$0.578 + wD + WI^3 - CC - \frac{K_{td}}{CC}$	(B. 20)	0.7203	0.003
Average *		0.7215 ± 0.0037	0.004

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [c] Spearman correlation (ρ) is the Spearman correlation when testing the performance of the obtained formula in 10% of the data (test data); [r] Relative frequency in percentage is the number of times each result appeared among all results (i.e., among the 30 000 iterations executed), in percentage. * Average is the average of the values showed. Standard deviation for the Spearman correlation of these mathematical expressions was the average of the standard deviations when applied to 90% of data versus all data. Ordinal data was used in order to obtain these results.

B1.2 Top–20 relative frequency

For ordinal data, we can observe (Table B2 and Table B3) that we just need to calculate the rank order for the nodes, based in five different indices in order to have good Spearman correlations. These indices were distributed in six families: {K_{td}, WI^5}, {WI^3, K_{td}}, {K_{td}, wD}, {K_{td}, wD, WI^5}, {K_{td}, Δs, wD}, {WI^3, K_{td}, wD} – K_{td} is always present, followed by wD (present in four out of the six families), but also WI^5, WI^3 and Δs are valuable indices.

Table B2. Most frequent mathematical expressions derived by the algorithm used – ordinal data.

Results		Spearman correlation (ρ) [c]	Relative frequency (percentage) [r]
$WI^5 - K_{td}$	(B. 21)	0.7034	5.697
$-(WI^5 - K_{td})$	(B. 22)	0.7034	5.630
$-(WI^3 - K_{td})$	(B. 23)	0.7027	2.380
$WI^3 - K_{td}$	(B. 24)	0.7027	2.337
$-(wD - K_{td})$	(B. 25)	0.7016	1.250
$wD - K_{td}$	(B. 26)	0.7016	1.230
$\frac{WI^5}{K_{td}}$	(B. 27)	0.7016	0.817
$\frac{K_{td}}{WI^5}$	(B. 28)	0.7016	0.770
$\frac{K_{td}}{WI^3}$	(B. 29)	0.7012	0.523
$\frac{WI^3}{K_{td}}$	(B. 30)	0.7012	0.493
$\frac{K_{td} + \Delta s}{wD}$	(B. 31)	0.7185	0.320
$wD + WI^5 - K_{td}$	(B. 32)	0.7113	0.310
$\frac{wD}{\Delta s + K_{td}}$	(B. 33)	0.7185	0.283
$wD + WI^3 - K_{td}$	(B. 34)	0.7109	0.267
$K_{td} - WI^5 - wD$	(B. 35)	0.7113	0.263
$\frac{K_{td}}{wD}$	(B. 36)	0.6996	0.257
$\frac{wD}{K_{td}}$	(B. 37)	0.6996	0.253
$\frac{wD \times WI^5}{K_{td}}$	(B. 38)	0.7104	0.247
$K_{td} - WI^3 - wD$	(B. 39)	0.7109	0.240
$\frac{K_{td}}{wD \times WI^5}$	(B. 40)	0.7104	0.237
Average *		0.7061 ± 0.0021	1.190

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [c] Spearman correlation (ρ) is the Spearman correlation when testing the performance of the obtained formula in 10% of the data (test data); [r] Relative frequency in percentage is the number of times each result appeared among all results (i.e., among the 30 000

iterations executed), in percentage. * Average is the average of the values showed. Standard deviation for the Spearman correlation of these mathematical expressions was the average of the standard deviations when applied to 90% of data versus all data. Ordinal data was used in order to obtain these results.

Table B3. Most frequent mathematical "families" of indices derived from Table B2.

Results	Relative frequency (percentage) [r]
K_{td}, WI^5	12.914
WI^3, K_{td}	5.733
K_{td}, wD	2.990
K_{td}, wD, WI^5	1.057
$K_{td}, \Delta s, wD$	0.603
WI^3, K_{td}, wD	0.507

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure 3.1); [r] Relative frequency in percentage is the number of times each result appeared among all the results (i.e., the 30 000 iterations executed), in percentage.

Table B3 shows the most common groups of indices. The "purple family" is, again, present in almost every expression (see also the analysis for metric data in the main text). In this case, this family was associated with yellow and green families. This means that, when using ordinal data, one should rely only on correlated indices.

The combination of indices {WI^n, K_{td}} occurred in 14 out of the top-20 mathematical formulae, and {wD, K_{td}} in 12 (Table B3).

Appendix C Consensus dendrogram analysis – Alternative Approach

Considering a distance of 40% for the clustering of groups, we can observe that four groups were formed (Figure C1): {WI^3, WI^1}, {Δs, s', K_{td}, K_{bu}, wD, WI^5}, {K, K_{dir}, K_{indir}}, {TI^3, TI^5, TI^1, BC, CC, D}. The first two depict the indices correlated to community response values and the other two clusters uncorrelated ones.

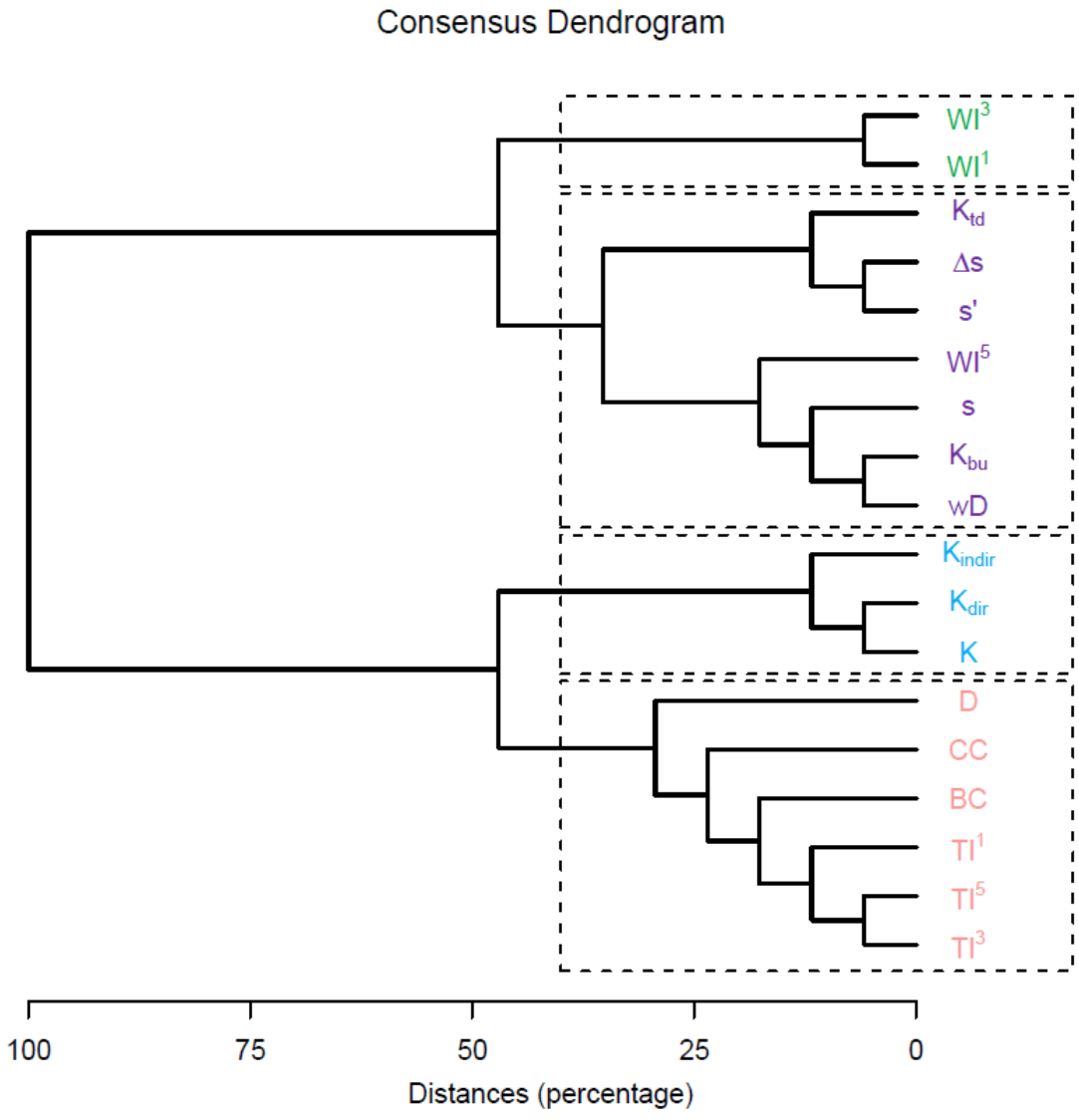


Figure C1. Consensus dendrogram between topological indices – four clusters. Using a distance above 35% (40% in this case), four groups can be formed: {WI^3, WI^1}, {Δs, s', K_{td}, K_{bu}, wD, WI^5}, {K, K_{dir}, K_{indir}}, {TI^3, TI^5, TI^1, BC, CC, D}.

C1 Families of indices in the top-20 more frequent

Relying our analysis in these four consensus dendrogram groups formed, the results for the most frequent mathematical families of indices are showed in Table C1 and Table C2, respective to metric and ordinal data.

Table C1. Most frequent mathematical "families" of indices derived from Table 3.4 – metric data.

Results	Relative frequency (percentage) [r]
WI^5, D	21.760
WI^5, TI^5	5.054
WI^5, TI^3	2.240
WI^3, D	1.326
WI^3, TI^5	0.808
wD, WI^5, D	0.573
WI^5, BC, D	0.560
WI^3, wD, D	0.557
WI^5, D, TI^3	0.543
wD, D	0.330
WI^5, D, TI^1	0.260
WI^5, K_{dir}, D	0.237
WI^5, D, TI^5	0.233

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure C1); [r] Relative frequency in percentage is the number of times each result appeared among all the results (i.e., the 30 000 iterations executed), in percentage.

Table C2. Most frequent mathematical "families" of indices derived from Table B2 – ordinal data.

Results	Relative frequency (percentage) [r]
K_{td}, WI^5	12.914
WI^3, K_{td}	5.733
K_{td}, wD	2.990
K_{td}, wD, WI^5	1.057
$K_{td}, \Delta s, wD$	0.603
WI^3, K_{td}, wD	0.507

Each index is represented with the respective family colour found in the dendrogram's clusters (Figure C1); [r] Relative frequency in percentage is the number of times each result appeared among all the results (i.e., the 30 000 iterations executed), in percentage.

We can conclude that related to the metric data, results don't change with this different clustering approach – see Table 3.5 and Table C1 (the yellow group – the one that was assimilated into the purple group – is not included in the most frequent families). Results obtained for ordinal data allow us to conclude that we can use, essentially, the "purple group" and, occasionally the weighted topological importance 3-step-long index (WI^3), belonging to the green group – see Table C2.

Appendix D Relative frequency of each index in the total unique mathematical expressions obtained

We also checked the relative frequency of each index in the total unique mathematical expressions obtained.

Comments on the results for the metric data are present in the main text (but see Table D1). Related to ordinal data, we found that the weighted degree (wD) and the keystone top-down indices (K_{td}) were the most frequent among all the results. They were followed by the 5 and 3-step-long weighted topological importance indices (WI^5 and WI^3). These results are also interesting if we note that wD and K_{td} are completely different: wD is a local, weighted, undirected index, while K_{td} is a meso, unweighted and directed index.

Table D1. Relative frequency, in percentage, of each index appearance in the total of different unique results obtained – metric data.

Indices	Relative frequency (percentage) [r]
WI^5	**80.39**
D	**75.00**
TI^5	**45.53**
WI^3	**40.53**
TI^3	**32.37**
wD	**30.73**
BC	**24.38**
TI^1	**22.19**
WI^1	**21.95**
CC	18.83
K_{bu}	18.83
Δs	18.00
s	17.64
K	17.47
K_{dir}	16.98
K_{indir}	14.97
s’	14.48
K_{td}	14.26

Each index is represented with the respective family colour found in the dendrogram’s clusters (Figure 3.1). We found 13082 different mathematical expressions for the metric data.

Table D2. Relative frequency, in percentage, of each index appearance in the total of different unique results obtained – ordinal data.

Indices	Relative frequency (percentage) [r]
wD	**74.75**
K_{td}	**73.21**
WI^{5}	**46.88**
WI^{3}	**38.73**
WI^{1}	**28.45**
K_{bu}	**24.00**
D	18.27
Δs	17.09
CC	14.82
TI^{5}	14.50
TI^{3}	13.48
s'	13.31
K	12.95
s	12.15
K_{indir}	11.53
K_{dir}	10.04
TI^{1}	9.83
BC	8.06

Each index is represented with the colour of each of the families found in the dendrogram (Figure 3.1). We found 13981 different mathematical expressions for the ordinal data.

Appendix E Single and combined indices performance when applied to three, four or six nodes

We decided to apply each individual index to check its performance when classifying the first three nodes (i.e., nodes #1, #2 and #3), the first four nodes (#1, #2, #3 and #4), the four middle nodes (#5, #6, #7, #8), the four last nodes (#9, #10, #11, #12). We did the same for the first six nodes versus the last six nodes. These results are showed in Table E1. The same was done for all the results (mathematical formulae) obtained either for metric and ordinal data. The average of these results is presented in Table E2.

Table E1. Spearman correlations derived from the results when applied to different groups of nodes in the networks.

Indices	Analysis	Three most important nodes	Four most important nodes	Four middle nodes	Four worse nodes	First six nodes	Last six nodes
WI^5	Rank order	-18.79	-26.77	-24.80	-48.49	-34.65	**-55.66**
	Real values	-43.24	-48.89	-29.00	-28.05	**-57.04**	-36.33
D	Rank order	-12.45	-14.45	-1.64	12.53	-13.04	7.68
	Rank order	15.90	14.28	1.71	-13.01	10.12	-12.09
TI^5	Rank order	-14.58	-16.81	-3.68	12.61	-15.38	7.91
	Real values	16.05	14.32	-4.29	-16.19	8.86	-15.58
WI^3	Rank order	-18.29	-26.27	-24.16	-47.20	-33.75	**-54.22**
	Real values	-42.38	-47.90	-28.47	-27.71	**-55.91**	-35.50
TI^3	Rank order	-14.92	-17.08	-3.87	12.58	-15.71	7.66
	Real values	15.92	14.19	-4.57	-16.69	8.61	-16.03
wD	Rank order	-17.78	-27.18	-25.96	**-50.98**	-35.71	**-57.50**
	Real values	**-51.71**	**-54.70**	-29.58	-28.55	**-60.28**	-37.52

BC	Rank order	-16.41	-18.06	-4.39	11.54	-16.75	6.21
	Real values	13.23	11.56	-5.47	-19.12	6.08	-18.22
TI^1	Rank order	-16.18	-18.10	-4.72	11.98	-17.34	6.63
	Real values	15.06	13.31	-6.39	-18.60	7.17	-18.25
WI^1	Rank order	-18.00	-24.48	-20.43	-36.79	-31.00	-45.46
	Real values	-34.18	-39.54	-26.38	-26.30	-48.19	-32.43
CC	Rank order	-12.49	-13.41	-1.04	13.80	-12.02	9.06
	Real values	15.19	14.08	0.56	-11.19	10.34	-10.13
K_{bu}	Rank order	-14.34	-24.30	-24.89	-48.94	-32.99	**-55.37**
	Real values	-49.77	**-52.44**	-27.54	-25.35	**-57.58**	-34.56
Δs	Rank order	-7.77	-17.75	-20.24	-39.42	-26.08	-46.19
	Real values	-32.65	-37.98	-18.96	-15.93	-44.87	-24.83
s	Rank order	-1.02	-19.14	-20.23	-45.94	-26.29	**-51.40**
	Real values	48.12	**-50.23**	20.02	-17.50	**-52.59**	-25.22
K	Rank order	-13.63	-15.24	-5.18	40.93	-11.96	42.37
	Real values	33.40	36.37	1.62	-14.19	38.57	-12.56
K_{dir}	Rank order	-16.80	-18.99	-1.90	34.69	-16.30	34.83
	Real values	30.56	32.99	0.09	-19.52	33.43	-17.31
K_{indir}	Rank order	-7.82	-9.83	-7.78	42.14	-7.25	44.47
	Real values	33.40	36.66	3.16	-7.11	39.60	-6.67
s'	Rank order	3.79	13.90	18.96	36.44	23.90	43.01

	Real values	27.83	33.48	15.49	11.11	40.35	20.73
K_{td}	Rank order	3.00	13.34	21.73	44.63	24.98	**50.13**
	Real values	36.58	40.88	19.00	12.34	46.82	23.90

Each index is represented with the colour of each of the families found in the dendrogram (Figure 3.1); Correlations were performed for each index values versus the community response values. Results are showed for: the first three most important nodes, the four most important nodes, the four middle nodes and the four worse nodes, and when applied to the first six and last six nodes. Networks were split according to the rank community response importance of their nodes. Bold values represent Spearman correlations higher than 50%.

Table E2. Average and standard deviation (percentage) of all results obtained related to their Spearman correlations.

	Three most important nodes	**Four most important nodes**	**Four middle nodes**	**Four worse nodes**	**First six nodes**	**Last six nodes**	**All**
Metric results	56.51 ± 1.44	60.19 ±1.32	36.24 ± 0.77	31.61 ± 1.81	66.12 ±1.39	43.39 ±1.28	76.03 ±0.50
Ordinal results	14.94 ± 2.34	24.95 ± 1.97	26.97 ± 0.72	52.59 ± 1.91	35.06 ± 1.36	58.88 ± 1.72	70.33 ± 1.60

Networks were split according to the rank community response importance of their nodes. Shaded cells represent Spearman correlations higher than 50%.

From all the singular indices (Table E1), only wD, K_{bu} and s are good enough to predict, by themselves, the most important nodes in a network: with predictions for three nodes being worse than for four and for six, respectively (3N < 4N < 6N), for wD and K_{bu}. s shows good results for 4N < 6N (its predictions for 3N are still below 50% for the Spearman correlation).

From the analysis of the results in Table E2, we can conclude that the mathematical expressions obtained using the real values of the indices, allow us to predict quite well the first more important nodes in a network. We can see that 3N < 4N < 6N – the more nodes used, the better the prediction. However, if we just want to evaluate the prediction of the first three nodes, it's still a good approach.

Related to the cocktails obtained from the nodes ordered according to their rank ("ordinal approach") we can conclude that they are better to evaluate the least important nodes in the network, with the last 4N < 6N. Again, the more nodes used, the better the prediction.

Moreover, we calculated the standard deviation for all Spearman correlations obtained when applied to the training set and to the testing set to check if they were biased due to the relatively small dataset used. The result was 0.46% ± 0.25% and 0.56% ± 0.25% respectively. These results show that our results were not under or overfitted since, the variance between the training and testing set was, in average, 0.1%.

We also checked the performance of each mathematical expressions obtained and previously presented when expressing the target values. We wanted to check whether they were good to predict which are the first three nodes, the first four, the middle four or last four, or the first six or last six nodes in the networks. With these – more detailed – information we confirmed the results previously seen in

Table E2. We found very different Spearman correlations – with big standard deviations in between (Tables E3 – E6). In Tables E3 – E6 we present the mathematical expressions that were accurate in more than 50% of the cases. We can infer from these that the first three nodes were more contributory to the final predictions, followed by the four, and the six first nodes, respectively. When more nodes were used, the accuracy of the predictions improved, in general, when the metric data was taken into account. It's also noticeable that not always the best cocktail – the one with higher accuracy – is the one that allows to get better accuracy for the prediction of the first few nodes.

When considering the ordinal data, we observed that the least important nodes in a network (according to the community response), were the more influential for the predictions: using the six last ranked nodes was more significant than using the first four. Indeed, if we aim to predict the first most important nodes, it's better to use metric data results. On the other hand, if we aim to predict the last ranked nodes we should use the ordinal data results. However, one should note that these cocktails performed better when evaluating the network as whole, than when evaluating only 3 to 6 nodes.

Table E3. Performance of the "best" mathematical expressions obtained using different groups of nodes from the networks – different partial datasets – metric data.

Results – Equation reference	Spearman correlation (ρ) [c]	Three most important nodes	Four best nodes	Four middle nodes	Four worse nodes	First six nodes	Last six nodes	Average Spearman correlation *	Average Standard deviation *
(3.1)	78.42	-16.97 ± 43.45	-21.45 ± 40.28	-36.17 ±29.87	-30.20 ±34.09	-37.55 ±28.89	-43.04 ±25.02	-30.90 ±9.16	33.60 ±6.48
(3.2)	78.18	30.67 ±33.59	34.76 ±30.70	35.74 ±30.01	32.73 ±32.14	46.94 ±22.09	43.34 ±24.64	37.36 ±5.82	28.86 ±6.37
(3.3)	78.01	**-58.16** ±14.04	**-61.78** ±11.48	-37.46 ±28.68	-33.65 ±31.37	**-67.82** ±7.21	-45.02 ±23.33	**-50.65** ±4.51	19.35 ±13.93
(3.4)	78.01	-3.33 ±52.81	-7.90 ±49.57	-34.27 ±30.93	-32.56 ±32.14	-20.82 ±40.44	-43.96 ±24.08	-23.81 ±9.85	35.43 ±15.95
(3.5)	77.99	-16.78 ±43.28	-22.07 ±39.54	-36.38 ±29.43	-33.19 ±31.68	-35.23 ±30.24	-44.10 ±23.96	-31.29 ±9.81	30.97 ±10.05
(3.6)	77.91	**-56.64** ±15.04	**-60.16** ±12.55	-37.83 ±28.34	-36.54 ±29.25	**-66.33** ±8.19	-47.61 ±21.43	**-50.85** ±5.61	19.95 ±12.20
(3.7)	77.89	-14.68 ±44.70	-20.30 ±40.72	-35.00 ±30.33	-32.39 ±32.17	-32.84 ±31.86	-43.73 ±24.15	-29.82 ±9.38	31.85 ±10.54
(3.8)	77.88	**-56.48** ±15.14	**-60.26** ±12.46	-37.29 ±28.71	-35.07 ±30.28	**-66.60** ±7.98	-46.64 ±22.09	**-50.39** ±5.92	20.30 ±12.79
(3.9)	77.88	**-52.26** ±18.11	**-56.81** ±14.89	-37.25 ±28.73	-35.99 ±29.62	**-64.70** ±9.31	-47.28 ±21.64	-49.05 ±9.83	20.84 ±11.21
(3.10)	77.86	**57.18** ±14.63	**60.60** ±12.20	36.96 ±28.92	36.47 ±29.27	**66.78** ±7.84	46.81 ±21.96	**50.8** ±8.78	20.04 ±12.69
(3.11)	77.86	**-57.79** ±14.19	**-61.00** ±11.92	-36.54 ±29.22	-34.53 ±30.64	**-66.70** ±7.89	-46.24 ±22.36	**-50.47** ±9.72	20.41 ±13.37

(3.12)	77.84	-14.62 ±44.70	-21.55 ±39.80	-35.79 ±29.73	-31.14 ±33.02	-35.10 ±30.22	-42.54 ±24.96	-30.12 ±10.18	31.55 ±10.26
(3.13)	77.82	**54.52** ±16.47	**58.70** ±13.52	37.82 ±28.28	37.38 ±28.59	**65.80** ±8.50	48.13 ±20.99	**50.39** ±5.45	19.98 ±11.45
(3.14)	77.81	-8.03 ±49.34	-12.88 ±45.91	-35.65 ±29.81	-31.26 ±32.92	-25.39 ±37.07	-42.41 ±25.04	-25.94 ±8.91	34.15 ±13.31
(3.15)	77.81	-8.03 ±49.34	-12.88 ±45.91	-35.65 ±29.81	-31.26 ±32.92	-25.39 ±37.07	-42.41 ±25.04	-25.94 ±7.91	34.15 ±13.31
(3.16)	77.81	**-60.68** ±12.11	**-63.80** ±9.91	-35.60 ±29.85	-33.51 ±31.32	**-68.31** ±6.72	-45.16 ±23.09	**-51.18** ±7.91	20.18 ±15.06
(3.17)	77.70	**57.61** ±14.21	**61.10** ±11.74	38.57 ±27.67	34.78 ±30.35	**67.36** ±7.31	46.26 ±22.23	**50.95** ±11.32	19.86 ±13.06
(3.18)	77.70	-12.16 ±46.35	19.16 ±0.41	-35.56 ±29.80	-34.47 ±30.57	-33.56 ±31.21	-44.98 ±23.14	-23.60 ±10.00	23.03 ±23.57
(3.19)	77.70	**-57.54** ±14.25	**-60.93** ±11.86	-36.60 ±29.34	-35.99 ±29.49	**-66.74** ±7.74	-46.91 ±21.77	**-50.79** ±13.05	20.04 ±12.95
(3.20)	77.70	**57.12** ±14.52	**60.75** ±11.95	37.74 ±28.22	36.48±29.11	**66.77** ±7.69	46.90±21.75	**50.96**±9.96	19.74±9.71

Networks were split according to the rank community response importance of their nodes. Spearman correlations were obtained using these partial networks. Standard deviations were obtained through the comparison of the correlation obtained for the whole network and the partial correlation, using metric values. [c] Spearman correlation when using all the metric data available. * Averages and respective standard deviations were calculated without considering the correlation result for the whole network (i.e., excluding [c]). Bold values show Spearman correlations higher than 50%.

Table E4. Performance of the most frequent mathematical expressions obtained using different groups of nodes from the networks – different partial datasets – metric data.

Results – Equation reference	Spearman correlation (ρ) [c]	Three most important nodes	Four best nodes	Four middle nodes	Four worse nodes	First six nodes	Last six nodes	Average Spearman correlation *	Average Standard deviation *
(3.21)	76.16	**-56.98** ±13.56	**-60.59** ±11.01	-36.08 ±28.34	-29.16 ±33.23	**-66.43** ±6.88	-41.85 ±24.26	-48.52 ±14.91	19.55 ±10.54
(3.22)	76.16	**56.98** ±13.56	**60.59** ±11.01	36.50 ±28.05	30.24 ±32.47	**66.43** ±6.88	42.80 ±23.59	48.92 ±14.48	19.26 ±10.24
(3.23)	75.86	**-57.23** ±13.17	**-60.79** ±10.65	-35.51 ±28.53	-29.88 ±32.51	**-66.40** ±6.69	-42.35 ±23.70	-48.70 ±14.84	19.21 ±10.49
(3.24)	75.86	**57.23** ±13.17	**60.79** ±10.65	35.51 ±28.53	29.93 ±32.48	**66.40** ±6.69	42.40 ±23.66	48.71 ±14.82	19.20 ±10.48
(3.25)	75.46	**-57.29** ±12.84	**-60.83** ±10.34	-34.64 ±28.86	-29.08 ±32.79	**-66.28** ±6.49	-41.59 ±23.95	-48.29 ±15.25	19.21 ±10.78
(3.26)	75.46	**57.29** ±12.84	**60.83** ±10.34	34.69 ±28.83	29.09 ±32.79	**66.28** ±6.49	41.60 ±23.94	48.30 ±15.23	19.21 ±10.77
(3.27)	75.40	**-55.47** ±14.09	**-59.29** ±11.38	-36.68 ±27.37	-28.14 ±33.41	**-65.55** ±6.96	-40.76 ±24.49	-47.65 ±14.60	19.62 ±10.32
(3.28)	77.02	**55.47** ±14.09	**-59.29** ±11.38	37.10 ±27.08	29.23 ±32.64	**65.55** ±6.95	41.71 ±23.82	28.30 ±44.85	19.33 ±10.01
(3.29)	77.02	**56.00** ±13.70	**59.78** ±11.03	36.46 ±27.52	29.33 ±32.56	**65.80** ±6.77	41.78 ±23.76	48.19 ±14.42	19.22 ±10.20
(3.30)	76.94	**-56.00** ±13.68	**-59.78** ±11.01	-36.46 ±27.50	-29.29 ±32.57	**-65.80** ±6.75	-41.72 ±23.78	-48.18 ±14.44	19.22 ±10.21
(3.31)	75.39	**-52.79** ±15.91	**-56.42** ±13.34	-33.41 ±29.61	-30.50 ±31.66	**-62.89** ±8.76	-42.78 ±22.98	-46.47 ±13.02	20.38 ±9.21

(3.32)	75.38	**-56.82** ±14.22	**-60.47** ±11.64	-36.59 ±28.53	-32.20 ±31.63	**-66.43** ±7.43	-44.04 ±23.26	-49.43 ±13.83	19.45 ±9.78
(3.33)	75.35	**-57.41** ±13.87	**-60.85** ±11.44	-35.68 ±29.23	-32.42 ±31.54	**-66.47** ±7.46	-44.31 ±23.13	-49.52 ±14.06	19.45 ±9.94
(3.34)	75.29	**-56.87** ±14.13	**-60.50** ±11.56	-36.59 ±28.48	-33.55 ±30.62	**-66.37** ±7.42	-45.01 ±22.52	-49.82 ±13.42	19.12 ±9.49
(3.35)	76.86	**56.89** ±14.06	**60.53** ±11.49	36.57 ±28.43	32.90 ±31.03	**66.40** ±7.33	44.56 ±22.78	49.64 ±13.63	19.19 ±9.64
(3.36)	76.86	**56.87** ±14.13	**60.50** ±11.56	36.59 ±28.48	33.55 ±30.62	**66.37** ±7.42	45.01 ±22.52	49.82 ±13.42	19.12 ±9.49
(3.37)	76.77	**-56.89** ±14.06	**-60.53** ±11.49	-36.57 ±28.48	-32.90 ±31.03	**-66.40** ±7.33	-44.56 ±22.78	-49.64 ±13.63	19.20 ±9.65
(3.38)	76.76	**56.85** ±14.08	**60.50** ±11.50	36.67 ±28.35	32.83 ±31.06	**66.38** ±7.34	44.48 ±22.83	49.62 ±13.62	19.19 ±9.63
(3.39)	76.60	**56.90** ±13.93	**60.59** ±11.32	36.93 ±28.05	32.19 ±31.40	**66.55** ±7.11	44.00 ±23.05	49.53 ±13.83	19.14 ±9.78
(3.40)	76.77	**-56.90** ±14.05	**-60.53** ±11.48	-36.54 ±28.45	-32.87 ±31.04	**-66.41** ±7.33	-44.56 ±22.78	-49.64 ±13.65	19.18 ±9.65

Networks were split according to the rank community response importance of their nodes. Spearman correlations were obtained using these partial networks. Standard deviations were obtained through the comparison of the correlation obtained for the whole network and the partial correlation, using metric values. [c] Spearman correlation when using all the metric data available. * Averages and respective standard deviations were calculated without considering the correlation result for the whole network (i.e., excluding [c]). Bold values show Spearman correlations higher than 50%.

Table E5. Performance of the "best" mathematical expressions obtained using different groups of nodes from the networks – different partial datasets – ordinal data.

Results – Equation reference	Spearman correlation (ρ) [c]	Three most important nodes	Four best nodes	Four middle nodes	Four worse nodes	First six nodes	Last six nodes	Average Spearman correlation *	Average Standard deviation *
(B. 1)	72.47	-15.85 ±40.04	-26.20 ±32.72	-28.36 ±31.19	**-57.47** ±10.61	-36.50 ±25.43	**-62.85** ±6.80	-37.87 ±18.56	24.47 ±13.12
(B. 2)	72.37	-17.11 ±39.07	-27.44 ±31.77	-28.05 ±31.34	**-52.54** ±14.02	-37.32 ±24.78	**-59.25** ±9.28	-36.95 ±16.15	25.04 ±11.42
(B. 3)	72.30	15.90 ±39.88	25.30 ±33.23	28.06 ±31.29	**56.09** ±11.46	36.17 ±25.55	**61.69** ±7.50	37.20 ±18.09	24.82 ±12.79
(B. 4)	72.22	14.93 ±40.51	24.11 ±34.02	28.67 ±30.79	**55.73** ±11.66	35.53 ±25.94	**61.96** ±7.44	36.82 ±18.43	25.06 ±12.98
(B. 5)	72.18	17.90 ±38.38	27.24 ±31.77	28.39 ±30.96	**53.36** ±13.31	36.87 ±24.96	**59.54** ±8.94	37.22 ±16.18	24.72 ±11.44
(B. 6)	72.16	14.45 ±40.81	24.96 ±33.37	29.02 ±31.21	**55.12** ±12.05	36.11 ±25.49	**61.00** ±7.90	36.78 ±18.01	25.14 ±12.80
(B. 7)	72.15	-12.24 ±42.36	-22.59 ±35.04	-28.71 ±30.72	**-56.96** ±10.74	-34.20 ±26.83	**-62.38** ±6.91	-36.18 ±19.68	25.43 ±13.91
(B. 8)	72.13	12.82 ±41.94	23.55 ±34.35	28.34 ±30.96	**57.33** ±10.47	35.03 ±26.23	**62.57** ±6.76	36.61 ±19.55	25.12 ±13.82
(B. 9)	72.12	16.23 ±39.53	25.87 ±32.70	28.19 ±31.07	**54.90** ±12.18	36.15 ±25.43	**60.90** ±7.93	37.04 ±17.47	24.81 ±12.35
(B. 10)	72.11	10.32 ±43.69	20.90 ±36.21	29.05 ±30.45	**56.81** ±10.82	33.39 ±27.38	**62.16** ±7.04	35.44 ±20.29	25.93 ±14.34
(B. 11)	72.11	-11.77 ±42.66	-20.80 ±36.28	-27.01 ±31.89	**-54.68** ±12.32	-31.62 ±28.63	**-59.71** ±8.77	-34.27 ±19.03	26.76 ±13.46

(B. 12)	72.10	19.09 ±37.49	27.86 ±31.28	27.71 ±31.39	**56.57** ±10.98	36.65 ±25.07	**62.06** ±7.10	38.32 ±17.27	23.89 ±12.21
(B. 13)	72.10	17.71 ±38.46	27.17 ±31.77	28.04 ±31.15	**54.61** ±12.36	37.29 ±24.61	**60.67** ±8.08	37.58 ±16.84	24.41 ±11.91
(B. 14)	72.07	-16.84 ±39.05	-26.97 ±31.89	-27.99 ±31.17	**-53.53** ±13.11	-36.94 ±24.84	**-59.86** ±8.63	-37.02 ±16.64	24.78 ±11.76
(B. 15)	72.07	16.82 ±39.07	26.97 ±31.89	27.97 ±31.18	**53.57** ±13.08	36.93 ±24.85	**59.87** ±8.62	37.02 ±16.66	24.78 ±11.78
(B. 16)	72.07	16.90 ±39.01	27.00 ±31.86	27.97 ±31.18	**53.69** ±12.99	36.55 ±25.11	**59.90** ±8.60	37.00 ±16.67	24.79 ±11.79
(B. 17)	72.07	-15.42 ±40.06	-25.75 ±32.75	-28.08 ±31.10	**-53.02** ±13.47	-36.52 ±25.13	**-59.55** ±8.85	-36.39 ±16.94	25.23 ±11.98
(B. 18)	72.06	-13.49 ±41.42	-23.07 ±34.64	-28.49 ±30.81	**-55.83** ±11.48	-34.65 ±26.45	**-61.75** ±7.30	-36.21 ±18.91	25.35 ±13.37
(B. 19)	72.06	16.67 ±39.16	26.89 ±31.93	28.06 ±31.11	**53.91** ±12.83	36.90 ±24.86	**60.07** ±8.48	37.08 ±16.81	24.73 ±11.89
(B. 20)	72.03	-16.21 ±39.47	-26.57 ±32.14	-28.11 ±31.06	**-55.47** ±11.71	-36.59 ±25.06	**-60.93** ±7.85	-37.31 ±17.51	24.55 ±12.38

Networks were split according to the rank community response importance of their nodes. Spearman correlations were obtained using these partial networks. Standard deviations were obtained through the comparison of the correlation obtained for the whole network and the partial correlation, using ordinal ranked networks. [c] Spearman correlation when using all the ordinal data available. * Averages and respective standard deviations were calculated without considering the correlation result for the whole network (i.e., excluding [c]). Bold values show Spearman correlations higher than 50%.

Table E6. Performance of the most frequent mathematical expressions obtained using different groups of nodes from the networks – different partial datasets – ordinal data.

Results – Equation reference	Spearman correlation (ρ) [c]	Three most important nodes	Four best nodes	Four middle nodes	Four worse nodes	First six nodes	Last six nodes	Average Spearman correlation *	Average Standard deviation *
(B. 21)	70.34	-12.29 ±41.05	-22.84 ±33.59	-26.66 ±30.89	**-52.40** ±12.69	-33.85 ±25.81	**-58.66** ±8.26	-38.88 ±15.86	22.25 ±11.21
(B. 22)	70.34	12.29 ±41.05	22.84 ±33.59	26.66 ±30.89	**52.40** ±12.69	33.85 ±25.81	**58.66** ±8.26	38.88 ±15.86	22.25 ±11.21
(B. 23)	70.27	12.27 ±41.01	22.96 ±33.45	26.75 ±30.78	**52.11** ±12.84	33.78 ±25.81	**58.43** ±8.37	38.81 ±15.68	22.25 ±11.09
(B. 24)	70.27	-12.27 ±41.01	-22.96 ±33.45	-26.75 ±30.78	**-52.11** ±12.84	-33.78 ±25.81	**-58.43** ±8.37	-38.81 ±15.68	22.25 ±11.09
(B. 25)	70.16	11.31 ±41.61	22.59 ±33.64	26.64 ±30.77	**51.80** ±12.99	33.84 ±25.69	**58.03** ±8.58	38.58 ±15.60	22.33 ±11.03
(B. 26)	70.16	-11.31 ±41.61	-22.59 ±33.64	-26.64 ±30.77	**-51.80** ±12.99	-33.84 ±25.69	**-58.03** ±8.58	-38.58 ±15.60	22.33 ±11.03
(B. 27)	70.16	-16.40 ±38.01	-26.08 ±31.17	-26.38 ±30.96	-49.70 ±14.47	-35.74 ±24.34	**-56.33** ±9.78	-38.85 ±13.71	22.14 ±9.69
(B. 28)	70.16	16.40 ±38.01	26.08 ±31.17	26.38 ±30.96	49.70 ±14.47	35.74 ±24.34	**56.33** ±9.78	38.85 ±13.71	22.14 ±9.69
(B. 29)	70.12	16.19 ±38.14	25.96 ±31.23	26.44 ±30.89	49.65 ±14.48	35.43 ±24.53	**56.29** ±9.78	38.75 ±13.72	22.18 ±9.70
(B. 30)	70.12	-16.19 ±38.14	-25.96 ±31.23	-26.44 ±30.89	-49.65 ±14.48	-35.43 ±24.53	**-56.29** ±9.78	-38.75 ±13.72	22.18 ±9.70
(B. 31)	71.85	17.09 ±38.72	27.03 ±31.69	27.84 ±31.12	**52.99** ±13.34	36.90 ±24.72	**59.52** ±8.72	40.86 ±14.76	21.92 ±10.44

(B. 32)	71.13	-14.59 ±39.99	-25.20 ±32.45	-27.19 ±31.07	**-53.61** ±12.39	-35.52 ±25.18	**-59.74** ±8.06	-40.25 ±15.63	21.83 ±11.05
(B. 33)	71.85	-17.09 ±38.72	25.12 ±32.54	-27.84 ±31.12	**-52.99** ±13.34	-36.90 ±24.72	**-59.52** ±8.72	-30.43 ±33.50	22.09 ±10.64
(B. 34)	71.09	-14.58 ±39.96	25.85 ±31.19	-27.26 ±30.99	**-53.41** ±12.50	-35.48 ±25.18	**-59.59** ±8.13	-29.98 ±33.84	21.60 ±10.69
(B. 35)	71.13	14.59 ±39.99	-25.85 ±31.19	27.19 ±31.07	**53.61** ±12.39	35.52 ±25.18	**59.74** ±8.06	30.04 ±33.91	21.58 ±10.75
(B. 36)	69.96	15.22 ±38.70	25.85 ±31.19	26.39 ±30.81	49.23 ±14.66	35.78 ±24.17	**55.74** ±10.06	38.60 ±13.48	22.18 ±9.53
(B. 37)	69.96	-15.22 ±38.70	-27.14 ±31.04	-26.39 ±30.81	-49.23 ±14.66	-35.78 ±24.17	**-55.74** ±10.06	-38.86 ±13.18	22.15 ±9.49
(B. 38)	71.04	-17.37 ±37.95	25.20 ±32.45	-26.97 ±31.16	**-51.57** ±13.76	-36.46 ±24.45	**-58.15** ±9.11	-29.59 ±33.00	22.19 ±10.40
(B. 39)	71.09	14.58 ±39.96	25.20 ±32.45	27.26 ±30.99	**53.41** ±12.50	35.48 ±25.18	**59.59** ±8.13	40.19 ±15.53	21.85 ±10.98
(B. 40)	71.04	-17.37 ±37.95	-27.13 ±31.04	-26.97 ±31.16	**-51.57** ±13.76	-36.46 ±24.45	**-58.15** ±9.11	-40.06 ±14.24	21.90 ±10.07

Networks were split according to the rank community response importance of their nodes. Spearman correlations were obtained using these partial networks. Standard deviations were obtained through the comparison of the correlation obtained for the whole network and the partial correlation, using ordinal ranked networks. [c] Spearman correlation when using all the ordinal data available. * Averages and respective standard deviations were calculated without considering the correlation result for the whole network (i.e., excluding [c]). Bold values show Spearman correlations higher than 50%.