Structural and dynamical interdependencies in complex networks at meso- and macroscale: nestedness, modularity, and in-block nestedness.

A thesis submitted by

María José Palazzi Nieves

in fulfillment of the requirements for the degree of
Doctor of Philosophy in Networks and Information Technologies

Barcelona, 2020

Universitat Oberta de Catalunya

Structural and dynamical interdependencies in complex networks at meso- and macroscale: nestedness, modularity, and in-block nestedness.

A thesis submitted by

María José Palazzi Nieves

in fulfillment of the requirements for the degree of
Doctor of Philosophy in Networks and Information Technologies

Supervisors: Javier Borge-Holthoefer
Albert Solé-Ribalta

Barcelona, 2020

... "El viaje no acaba nunca. Sólo los viajeros acaban. E incluso éstos pueden prolongarse en memoria, en recuerdo, en relatos...
El objetivo de un viaje es sólo el inicio de otro."
José Saramago.

Abstract

A pervasive amount of real systems can, in a broad sense, be said to be *complex*, i.e., they are made of a large number of interacting components, and display collective behavior that can not be inferred from the behaviors of the individual parts. A particular approach to study them is to consider them as *networks*, with its components represented as nodes and the interactions between them as links. Systems like the financial market, the brain, and the Internet are frequently studied as complex networks. Under this view, we find a large body of research focused on putting to light the interplay between *structure* and *dynamics*: exploring how the dynamical behavior of a complex network is constrained by the nature of the interactions between its elements, as well as by the topology of such interactions. These analyses are usually performed at three different scales: the *microscale* based on single node properties, the *macroscale* that explores global properties of the whole network, and the *mesoscale* based on the properties of groups of nodes. Nonetheless, most studies so far have focused exclusively on either of them, despite the increasing evidence suggesting that networks often exhibit structures at several scales of organization. Thus, understanding the interrelations between the macro- and mesoscale structures represents a major challenge.

In this thesis, we apply structural network analysis to a variety of synthetic and empirical systems at meso- and macroscale. The thesis organizes in three main parts:

In the first part (Chapter 1), we provide a brief introduction to complex networks, its origins, and significant breakthroughs. We introduce the associated literature focused on the study of nested (at macroscale), and modular, and in-block nested structural arrangements (at mesoscale), its definitions, dynamical implications, metrics, etc.

In the second part, we focus on an examination of the structural properties of the in-block nestedness and its relationship with nestedness and modularity. We start with an empirical, analytical, and numerical exploration, proving that the in-block nestedness objective function lacks a resolution limit (Chapter 2). Then, in Chapter 3, we disentangle the effects that nestedness, modularity, and in- block nestedness, impose on each other, in uni- and bipartite settings. Through an analytical, numerical, and empirical study, we show that purely structural constraints forbid interactions to be completely modular and completely nested at the same time, with ample room, however, for a wide range of intermediate possibilities.

We devote part three of the thesis to perform a combination of empirical work and theoretical modeling that will help us to better understand some of the mechanisms that enable the emergence of nested, modular, or in-block nested patterns. Specifically, in Chapter 4, we analyze the patterns of interaction between the developers and files than compose a set of open-source projects (OSS) aiming to identify their relevance in the OSS communities. These analyses show that indeed the interaction patterns of the projects evolve into internally organized blocks. Moreover, the sizes of those blocks are bounded, and such size is compatible with the Dunbar number, regardless of the total size of the projects. Finally, in Chapter 5, we introduce an ecology-inspired modeling framework that explains the structural transitions between the observed nestedness and modularity in real information ecosystems. The model builds on the idea that the network structure is driven by an optimization process, aiming at maximizing the visibility of the involved actors. In addition, we present evidence that these systems exhibit a remarkable structural resilience or elasticity to environmental changes.

We complete the thesis in Chapter 6, where we offer the main conclusions derived from the realization of this work and a discussion about possible directions for future works.

Resumen

Una gran cantidad de sistemas tales como el mercado financiero, el cerebro y el Internet son, en un sentido amplio, *sistemas complejos*. Es decir, están formados por una gran cantidad de elementos que interactúan entre sí, y exhiben un comportamiento colectivo que no puede ser inferido partir de las propiedades de sus elementos aislados. Éstos sistemas suelen estudiarse representándolos como *redes*, donde los elementos que los componen constituyen sus *nodos*, y las interacciones entre ellos constituyen los *enlaces*. La investigación en redes, se enfoca principalmente en el estudio de sus propiedades *dinámicas* y *topológicas*. Especificamente, en explorar cómo el comportamiento dinámico de una red está influenciado por la naturaleza de las interacciones entre sus elementos, y por la topología de las mismas. Estos análisis son usualmente realizados a tres escalas: la *microescala*, basada en las propiedades de los nodos individuales, la *macroescala* que explora las propiedades globales de toda la red y la *mesoescala* basada en las propiedades de grupos de nodos. Sin embargo, la mayoría de los estudios se centran en una escala a la vez, a pesar de la creciente evidencia que sugiere que las redes a menudo exhiben estructura a múltiples escalas de organización.

En esta tesis, estudiamos las propiedades estructurales de redes, sintéticas y empíricas a escala múltiple. La tesis está organizada en tres partes:

En la primera parte (Capítulo 1), presentamos una breve introducción al campo de las redes complejas. En particular, repasaremos la literatura enfocada al estudio de patrones estructurales anidados (o nested), en la macroescala, y patrones modulares y anidados en bloque (in-block nested), en la mesoescala.

En la segunda parte, nos enfocaremos en estudiar a profundidad las propiedades estructurales de los patrones in-block nested y su relación con los patrones anidados y modulares. Empezamos el Capítulo2 con una demostración empiríca, numérica y analítica de que la función objetivo para la caracterización de patrones in-block nested carece de un límite de resolución similar al observado en funciones objetivo similares. Luego, en el Capítulo 3, mediante un estudio analítico, numérico y empírico, aplicado tanto a redes unipartitas como a bipartitas, demostramos que existen restricciones inherentes que prohíben que, a nivel general, los patrones de interacción en redes sean completamente modulares y completamente anidados al mismo tiempo, dejando espacio para un amplio rango de configuraciones intermedias.

Dedicamos la tercera parte de la tesis a realizar una combinación de trabajo empírico y de modelado teórico para explorar en detalle algunos de los mecanismos que permiten la emergencia de patrones estructurales. Específicamente, en el Capítulo 4, analizamos los patrones de interacción entre los desarrolladores y los archivos que componen diversos proyectos de software de código abierto. Mediante estos análisis demostramos que, los patrones de interacción de los proyectos evolucionan hacia una estructura de bloques, que están organizados internamente de forma jerárquica. A su vez, probamos que los tamaños de los subgrupos están delimitados, independientemente del tamaño total de los proyectos. Dicho tamaño es compatible con el número de Dunbar, reportado en diversos entornos sociales. Finalmente, en el Capítulo 5, introducimos un modelo dinámico que adapta una serie de conceptos ampliamente usados en ecología, para explicar como se llevan a cabo las diversas transiciones entre los patrones anidados y modulares, observadas en redes de información reales. Presentamos evidencia de que dichos patrones son estructuralmente resilientes y elásticos ante perturbaciones del entorno.

Finalmente el Capítulo 6, contiene las Conclusiones en las cuales se discuten los resultados obtenidos en esta Tesis y se plantean diversos problemas que quedan abiertos para estudio futuro.

Resum

Una gran quantitat de sistemes com el mercat financer, el cervell, i nternet són, en un sentit ampli, *sistemes complexos*. És a dir, estan format per una gran quantitat d'elements que interactuen entre sí, i exhibeixen un comportament col·lectiu que no es pot inferir des de les propietats del seus elements aillats. Sovint, aquests sistemes s'estudien mitjançant *xarxes*, on els elements constituents són els *nodes*, i les interaccions entre ells es representen amb *enllaços*. La recerca en xarxes, s'enfoca principalment en l'estudi de les seves propietats dinàmiques i topològiques. Especialment, en explorar com el comportament dinàmic d'una xarxa està definit per la naturalesa de les interaccions entre el seus elements, i per la topologia de les mateixes. Aquest anàlisis sovint es fan en tres escales: la microescala, que es basa en les propietats de els nodes individuals, la macroescala que explora les propietats globals de tota la xarxa i la mesoescala que es basa en les propietats de grups de nodes. No obstant això, la majoria dels estudis es centren només en una escala, tot i la creixen evidencia que suggereix que les xarxes sovint exhibeixen estructura a múltiples escales d'organització.

En aquesta tesis, estudiarem les propietats estructurals de les xarxes, sintètiques i empíriques, a escala múltiple. La tesi s'organitza en tres parts: En la primera part (Capítol 1), es presenta una breu introducció al camp de les xarxes complexes. En particular, repassarem la literatura enfocada a l'estudi dels patrons estructurals anidats (o nested en anglès), a la macroescala, modulars i anidats a blocs (in-block nested), a la mesoescala.

La segona part, es centra en estudiar en profunditat les propietats estructurals dels patrons in-block nested i la seva relació amb els patrons anidats i modulars. Començaren el Capítol 2 amb una demostració empírica, numèrica i analítica de com la funció objectiu, emprada per a la caracterització de patrons in-block nested manca d'un límit de resolució similar a l'observat en altres funcions objectiu similars. Després, en el Capítol 3, mitjançant un estudi analític, numèric i empíric, aplicat tant a xarxes unipartitas com a bipartitas, demostrarem que existeixen restriccions inherents que prohibeixen que, a nivell general, els patrons d'interacció en xarxes siguin completament modulars i completament nested a el mateix temps, deixant espai per a un ample rang de configuracions intermèdies.

Dediquem la tercera part de la tesi a realitzar una combinació de treball empíric i de modelatge teòric per explorar en detall alguns dels mecanismes que permeten

l'emergència de aquests patrons estructurals. Específicament, en el Capítol 4, analitzarem els patrons d'interacció entre desenvolupadors de software i els fitxers que componen diversos projectes de software de codi lliure. Mitjançant aquests anàlisis demostrarem que, els patrons d'interacció dels projectes evolucionen cap a una estructura de blocs, que estan organitzats internament de forma jeràrquica. Al mateix temps, trobem que les mides dels subgrups estan delimitades, independentment de la mida total dels projectes. Aquesta mida és compatible amb el nombre de Dunbar, reportat en diversos entorns socials. Finalment, en el Capítol 5, introduïm un model dinàmic que adapta una sèrie de conceptes àmpliament usats en ecologia, per explicar com es duen a terme les diverses transicions entre els patrons anidats i modulars, observats en xarxes de informació reals. Presentem evidències que aquests patrons son estructuralment robustos i elàstics davant pertorbacions de l'entorn.

Finalment, el Capítol 6 conté les conclusions en les quals es discuteixen els resultats obtinguts en aquesta Tesis y es plantegen divers problemes que queden oberts per a estudis futurs.

Agradecimientos

He d'agrair en primer lloc, als meus directors de tesi, el Dr. Javier Borge i el Dr. Albert Solé. Moltes gràcies per donar-me l'oportunitat d'emprendre aquest projecte junt a vosaltres. Gràcies pel vostre suport, paciència i tots els consells que m'hen donat durant aquest tres anys.

A mi esposo, por su apoyo y comprensión durante estos años de carrera; gracias por su amor incondicional y por su ayuda en esta tesis.

A mi madre y a mi abuela, quiénes me ha dado todo; gracias por hacer de mí la mujer que hoy soy. A mis hermanas, quienes han estado conmigo en las buenas y malas, siempre apoyándome y dándome ánimos para seguir adelante. A mis sobrinos, porque sin importar las circunstancias o la distancia, siempre son mi motivo para sonreír. A mi familia extendida, mis suegros y cuñados, por recibirme como uno más en su hogar.

A mis primos, Amanda, Amara, Carlos, Daniela, Flavia, Juan Carlos, y Manuelito. Por los años de risas, anécdotas y apoyo incondicional, sin importar la distancia.

A mis compañeros del grupo de Sistemas Complejos: Cristina, Daniel y Nello. A mis compañeros del Internet Interdisciplinary Institute (IN3): Rafael, Leandro y Santiago. A los salineros anónimos: Óscar y David. Gracias por las risas, las cervezas, buenas comidas y momentos compartidos. Sin duda alguna, su compañía ha hecho este trayecto mucho más llevadero.

A mi familia por elección: Ana Julia, Carlos E., Carolina, Ricardo, Sara y Viviana. Por su compañía, apoyo y cariño durante todos estos años. No me alcanzan las palabras para agradecerles por estar siempre ahí para mi, en las buenas y las no tan buenas.

Por último, pero no menos importante, a mi amado Neko, aunque no sea capaz de leer estas líneas.

Son muchas más las personas a las que me encantaría agradecer su amistad, consejos y compañía a lo largo de este camino, pero no me alcanzaría el espacio para nombrarlos a todos.

Gracias.

Publications

Main thesis publications:

1. **M.J. Palazzi**, J. Cabot, J.L. Cánovas Izquierdo, A. Solé-Ribalta and J. Borge-Holthoefer. "Online division of labour: emergent structures in Open Source Software". **Scientific Reports 9**, 13890 (2019). DOI: 10.1038/s41598-019-50463-y

2. **M.J. Palazzi**, J. Borge-Holthoefer, C.J. Tessone and A. Solé-Ribalta. "Macro- and mesoscale pattern interdependencies in complex networks". **Journal of the Royal Society Interface 16**, 20190553 (2019). DOI: 10.1098/rsif.2019.0553

3. M. S. Mariani, **M.J. Palazzi**, A. Solé-Ribalta, J. Borge-Holthoefer and C.J. Tessone. "Absence of a resolution limit in in-block nestedness". In press in **Communications in nonlinear science and numerical simulation**. Preprint: arXiv:2002.08265

4. **M.J. Palazzi**, A. Solé-Ribalta, S. Meloni., V. Calleja-Solanas, C.A. Plata, S. Suweis and J. Borge-Holthoefer. "Resilience and elasticity of co-evolving information ecosystems". Under Review. Preprint: arXiv:2005.07005

Other publications related to the PhD thesis:

1. C. A. Plata, E. Pigani, S. Azaele, V. Calleja-Solanas, **M.J. Palazzi**, A. Solé-Ribalta, S. Meloni, J. Borge-Holthoefer, S. Suweis. "Neutral Theory for competing attention in social networks". Under Review. Preprint: arXiv:2006.075865

Additional publications not related to the PhD thesis:

1. P. Garcia-Canadilla, J.F. Rodriguez, **M.J. Palazzi**, A. Gonzalez-Tendero, P. Schönleitner, V. Balicevic, et al. "A two dimensional electromechanical model of a cardiomyocyte to assess intra-cellular regional mechanical heterogeneities". PLoS ONE 12(8): e0182915 (2017). DOI: 10.1371/journal.pone.0182915.

2. **M.J. Palazzi**, M. G. Cosenza. "Amplitude death in coupled robust-chaos oscillators". The European Physical Journal Special Topics, **223**, 2831–2836 (2014) (2014). DOI: 10.1140/epjst/e2014-02296-5.

Contents

Part I

Introduction

Chapter 1

Complex networks: micro-, meso- and macroscale structure.

1.1 Introduction to complex networks

What do neurons interacting inside our brain, people moving around a city, birds flocking during migration, and proteins interacting inside cells have in common? They all can be seen as systems composed of several elements, which behavior cannot be fully understood without considering the interactions of its constituent parts, and whose patterns of connection are neither regular nor random. Complexity science is an interdisciplinary field that aims to understand the structural and dynamical properties of such systems: how they arise, how they evolve over time, and how the dynamics that take place on these systems operate. For example, by studying how their connectivity patterns influences the emergence of collective behaviors, such as synchronization [1], or its effect in other processes like epidemic spreading [2], to name a few. In the study of complex systems, *networks* play a central role: many complex systems can be represented as networks in which a node represents an entity (e.g., a person, a protein), and a link represents an interaction between these two entities (e.g., a friendship or a physiochemical interaction). At the same time, the many types of interactions that take place in a complex system (e.g. time-changing, stochastic, or nonlinear), can all be described within a network formalism [3–6].

The study of networks dates back to the 18th century with the development of graph theory, a branch of discrete mathematics initiated by the work of Leonhard Euler on the solution to the Königsberg bridge problem. Euler proved that the problem had no solution and laid the foundations of graph theory. Regular graphs were deeply used since then, and it was not until the late 1950's, when Paul Erdös and Alfréd Rényi, combined concepts of graph theory and probability theory to describe networks with complex topology as random graphs [7–9], triggering an intensive amount of studies

focusing on them [10–12].

Later on, in 1998 Watts and Strogatz, inspired by the work of the psychologist Stanley Milgram [13, 14], who introduced the concept of *small-world*, found that many real networks–social, biological or technological– were characterized by a low characteristic path length, but also by a high clustering coefficient that could not be captured by traditional approximations based on regular or random graphs. They defined the networks with this mixed properties as *small-world* networks, and introduced a model that recovers such properties [15]. Shortly after, in 1999, Barabási-Albert discovered that the distribution on the number of links of a network's nodes is highly heterogeneous or *scale-free*, following a power law [16]. They proposed a model to generate networks with such type of degree distribution by recurring to *preferential attachment* processes, where nodes are sequentially added to the network, attaching them to the existing nodes with a probability proportional to their number of connections. The development of the models introduced in these two papers, along with the increased power of modern computers, the availability of large datasets, and the development of powerful data analysis tools, have marked the beginning of modern network theory. From that moment, we have witnessed the evolution of the field, and ever since, many other contributions have increased our understanding of networked systems.

Roughly speaking, a network can be characterized at three different scales: micro- (at the level of single nodes or links), meso- (groups of few nodes), and macro-scale (the network as a whole). Many different metrics have been proposed for each one of these scales, and all of them have proven to be essential to determine many network properties such as robustness [17] and resilience to attacks [18], and dynamical processes like information spreading [19, 20]. In the ecological literature for example, macro-scale features such as gradient [21], spatial turnover [22], checkerboard [23, 24], and segregation [25] patterns have been deeply explored. At the meso-scale, core-periphery [26, 27] or combined [21] structures have also attracted the focus of researchers.

But undoubtedly, two particular patterns have concentrated special attention of the researchers, especially within the ecological context: nestedness and modularity. Modularity [28–32], a meso-scale pattern [33, 34], that considers the organization of species as a set of cohesive subgroups. It assumes that the species within the group interact among them with larger frequency than with species belonging to other groups [35], and is a widespread organizational structure [36–40]. Nestedness [41, 42], a prominent macro-scale pattern, that was first described by ecologists, and that quantifies to what extent specialists nodes (low connectivity nodes) interact with proper subset of those nodes interacting with generalists nodes (high connectivity nodes) [43, 44]. It stands as a frequent emergent structural arrangement, that has also been observed in Ecology [41, 45], in Economy [46–48] and information systems [49]. For this reason, these two patterns– nestedness and modularity– will be explored in depth in Sections 1.4.3 and 1.5.1, respectively.

Not less important, the presence of compound modular-nested –or in-block nested– structures on empirical networks has also been assessed [21, 50–53]. The evidence from these studies suggests that, since a system can integrate properties from distinct organizations at different scales, in-block nested patterns may play a prominent role in the dynamical processes of many systems. However, a full understanding of the conditions for the emergence of in-block nested patterns, and its relationship with nestedness and modularity remains unstudied, and is one of the problems that will be addressed in this thesis. Unraveling the relationships between these patterns may shed light on the dynamical trade-offs that either arrangement can facilitate.

This thesis aims to explore the relationship between nested, modular and in-block nested organizational patterns, i.e, how they affect each other. We will investigate how the transitions between nested and modular structural configurations occur, and to what extent this hybrid in-block nested pattern is indeed a transitional organization between them. Last but not least, we will perform extensive structural analysis at multiple scales, to a variety of empirical networks coming from different domains, with a particular focus on co-evolutive networks.

The rest of the Chapter will be devoted to present a primer of the minimal aspects that will be used throughout the thesis. Thus an extensive summary of the advances of the field during the last two decades could be useful for a full understanding of the remaining chapters, that goes beyond the scope of this thesis. For extensive reviews and books focusing on the structure and dynamics of complex networks, we refer the reader to [4, 54–56].

1.2 Graphs

Formally, networks are usually represented as graphs Fig. 1.1, composed of nodes (vertices) connected by links (edges) that represent an interaction between the nodes. A graph $G(\mathcal{M}, \mathcal{L})$ consists of a pair of sets $\mathcal{M}$ and $\mathcal{L}$, where $\mathcal{M} = \{m_1, m_2, m_3 \dots m_N\}$ is the set of nodes and $\mathcal{L} = \{l_1, l_2, l_3 \dots l_K\}$ is the set of links that represent the relations of a particular type between pair of nodes. If the links have a direction, pointing from one vertex to another, we say that the graph is *directed*. Additionally, in some situations, the edges among the nodes can have different strengths, in such case, we say that the graph is *weighted*. Moreover, if the graph contains links connecting the same pair of nodes, we have a *multigraph* (this type of graph will not be considered in this thesis). The two nodes that identify a link are the *endpoints* of the link, and these two nodes are said to be *adjacent* or *neighbors*. For undirected graphs, the number of nodes N is the order of the graph, and the number of links K ranges from zero to $N(N-1)/2$. The graph size if given by the total number of edges, and can be considered *sparse* if $K \lll N^2$, or considered *dense* if $K = \mathcal{O}(N^2)$. If $K = N(N-1)/2$, i.e., all vertices are connected to one another by one link, the graph is *complete*, denoted by K_N.

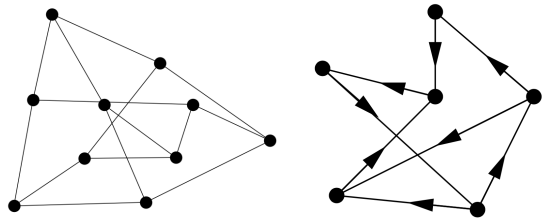
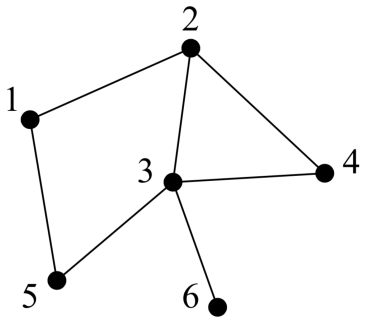


Figure 1.1: Graphical representation of a undirected (left) and a directed (right) graph [56].

1.2.1 Adjacency Matrix of a Graph

All the information of a graph $G = (\mathcal{M}, \mathcal{L})$ can be represented by the adjacency matrix $\mathbf{A}$, which is an $N \times N$ square matrix with elements $A_{ij} = 1$ when there is a link between nodes i and j, and zero otherwise. The matrix *fill* is equal to the total sum of the links in the graph, and its connectance is the number of actual edges expressed as a proportion of the total number of possible edges. For the case of weighted graphs, the adjacency matrix $\mathbf{W}$, is an $N \times N$ matrix whose entry ω_{ij} indicate the weight of the link connecting nodes i and j. As an example, the adjacency matrix of the network in the figure below is given by Eq. 1.1



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \tag{1.1}$$

1.2.2 Bipartite Graphs

Another important type of graphs (or networks) are bipartite graphs. A graph G is *bipartite* if the node set $\mathcal{M}$ is composed of two disjoint subsets $\mathcal{M}_1$ and $\mathcal{M}_2$, each link represents the interaction of a node of $\mathcal{M}_1$ with a node of $\mathcal{M}_2$ and interactions between nodes of the same subset are not allowed see Fig. 1.2. The graphs composed of only one set of nodes are called *unipartite*.

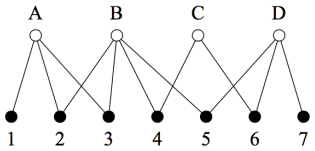


Figure 1.2: Graph representation of a bipartite network [56].

The adjacency matrix of a *bipartite graphs* with nodes of types r and b, has a block off-diagonal form of

$$\mathbf{A} = \begin{bmatrix} \mathbf{O}_{r\times r} & \tilde{\mathbf{A}}_{r\times b} \\ (\tilde{\mathbf{A}}^T)_{b\times r} & \mathbf{O}_{b\times b}, \end{bmatrix} \tag{1.2}$$

where $\mathbf{O}$ is the all-zero matrix, and $\tilde{\mathbf{A}}$ is the incidence matrix. During this thesis, part of our attention will be devoted to the characterization of the structural pattern of these type of networks.

1.3 Microscale characterization of complex networks

The description and characterization of complex networks can be performed at different scales. The lower level, the microscale, regards study of the role and properties of the network's nodes. For this, different measures can be considered: the node degree and the betweenness, to name a few [56]. In this section, we will review some of these micro and macroscale measures. The higher and middle levels of description, respectively, the macro- and mesoscales, will be covered in the following sections.

1.3.1 Node degree

One of the simplest centrality measures is the node degree. The degree k_i of a node i is defined as the number of links that the node is connected to:

$$k_i = \sum_{j=1}^{n} A_{ij}. \tag{1.3}$$

In directed graphs, the node has two types of degree: the number of links that start at node i or out-degree k_i^{in} and the number of links that end at i or in-degree of the node k_i^{out}; then, the total degree is defined as $k_i = k_i^{out} + k_i^{in}$. The degree sequence is a list of the node degrees of the graph.

In weighted networks, the degree k_i of a node i is generalized to the notion of strength. The strength of the node combines the information regarding the number and weights of links incident in such node and is defined as

$$s_i = \sum_{j\in\mathcal{M}} \omega_{ij}. \tag{1.4}$$

Although simple, this centrality measure, can be useful to identify the presence of the influential nodes–or hubs- in a netwok, e.g., the nodes with high connectivity.

1.3.2 Betweenness

One of the most significant node centrality measures is the betweenness centrality. Proposed by Freeman [57], betweenness centrality was first introduced in the context of social networks and measures the relevance of a given node by counting the number of shortest paths between other pairs of nodes that go through it. The shortest path

between pair of nodes is simply the shortest route that connects them when moving along the links.

The betweenness of a node i is measured as the fraction of shortest paths passing through the node, and is defined according to the equation:

$$B(i) = \sum_{j,k \in \mathcal{M}, j \neq k} \frac{n_{jk}(i)}{n_{jk}}, \tag{1.5}$$

where n_{jk} is the total number of shortest paths connecting nodes j and k, and $n_{jk}(i)$ is the number of shortest paths connecting nodes j and k that pass through node i. Commonly, betweenness is normalized by dividing it through some factor, that depends on the total number of nodes, usually $(N-1)(N-2)/2$, N^2 or simply N.

1.3.3 Assortativity

Another interesting property that can be useful to characterize networks is the assortativity. This property is assessed by considering the correlations between two nodes connected by a link. A common way to determine such degree correlation is by computing the Pearson coefficient r of the degrees at both ends of the links [58]. Then, a network is said to be assortative if $r > 0$ and disassortative if $r < 0$; when there is no correlation between node links we should expect a coefficient $r = 0$. The assortativity of networks depends on its type, for example, social networks are usually assortative, while biological networks tend to be disassortative.

1.4 Macroscale characterization of complex networks

The higher level of description, the macroscale, can be represented by the statistical properties of the networks. For example, by studying average quantities like the mean degree, the degree distribution, or by the identification of macroscale connectivity patterns, like nested and core-periphery patterns. Some of these macroscale pattern are described in more detail below:

1.4.1 Degree distribution.

The degree distribution of a graph $P(k)$ is defined as the fraction of nodes in the graph that have a degree k. In directed graphs, we have two distributions, $P(k^{in})$ and $P(k^{out})$. Another way of obtaining information on how the degree is distributed among the nodes of a network is by the calculation of the n-moments of the distribution:

$$\langle k^n \rangle = \sum_k k^n P(k), \tag{1.6}$$

where the first moment $\langle k \rangle$ is the average degree of the graph G. Similarly, for weighted graphs we can obtain its strength distribution $P(s)$ and the average strength $\langle s \rangle$.

The degree distribution is one of the most fundamental network properties since it provides us with some insights about the network structure. For example, in a random

graph with edge probability p the degree k_i of a node i follows a binomial distribtution (Poisson, for large N), which is very different from the degree distributions often observed in real networks.

1.4.2 Core-Periphery Structure

A relveant macroscale pattern found in real networks [59–61] is the Core-Periphery structure (CP). The idea that networks can present a core and a periphery was first discussed by the end of the 70's. Nevertheless, it was not until 1999, when Borgatti and Everett [26], formally introduced the concept of CP structure in networks. In a CP structure, some nodes are part of a highly connected core and the others are part of a sparsely connected periphery. The nodes from the core are well-connected to nodes from the periphery, but the peripheral nodes are not connected to each other see Fig. 1.3.

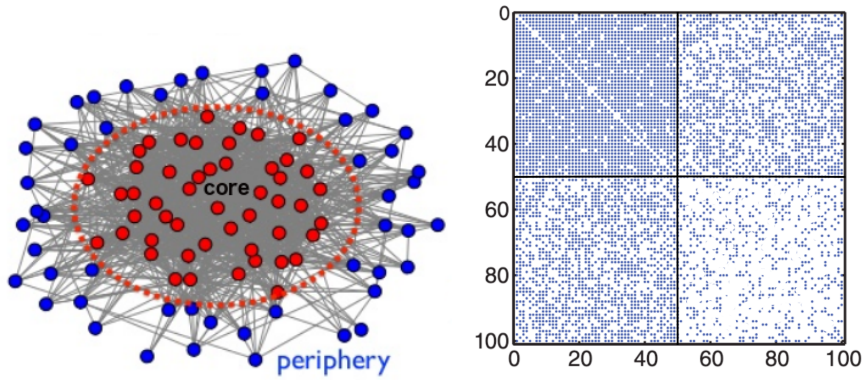


Figure 1.3: Graph representation (left) and adjacency matrix (right) of networks with core-periphery structure [62].

In [26], the authors quantify the level of CP structure of a network by comparing it to a model that consists of a perfect CP matrix. Such as:

$$\begin{aligned} \rho &= \sum_{ij} A_{ij}\delta_{ij} \\ \delta_{ij} &= \begin{cases} 1, & \text{if } c_i = \text{core or } c_j = \text{core}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{1.7}$$

After the work of Borgatti and Everett, different methods for detecting CP structures have been developed [27, 62, 63]. For example, Zhang et al. [63], introduced a method for detecting core-periphery structures based on methods of statistical inference. The method aimed to find the parameters of a stochastic block model that produce the best fit with respect to the real data using a combination of expectation-maximization and belief propagation algorithms. Additionally, Rombach et al. [27], based on the formulation of [26], developed a method for studying core-periphery structure in weighted networks.

1.4.3 Nested Structure

An important structural pattern that has been found in biological [41, 45], economical [46, 47] and social [49] systems is a nested pattern. The concept of nestedness was first introduced by ecologists, as a way to describe the structured patterns of distribution of species in different types of landscapes [43] and was brought in to the context of complex networks over a decade ago [45]. In structural terms, a perfect nested pattern is observed when the set of neighbours nodes with low number of interactions (specialists nodes) are a subset of those with larger degree (generalists nodes), see Fig. 1.4.

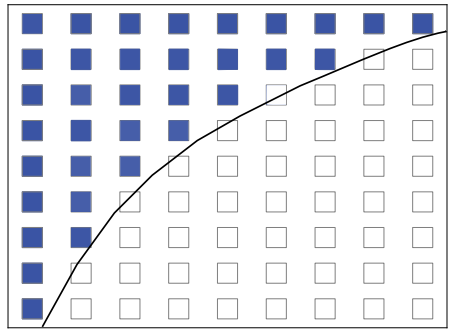


Figure 1.4: Representation of the adjacency matrix of a network with perfect nested structure. Rows and columns have been rearranged to highlight the nested property with all the interactions lying above the curve of perfect nestedness (black solid line).

The widespread observation of nestedness in a variety of systems, has triggered a large amount of research trying to unveil the mechanisms behind its emergence, its relationship with other network properties, its implications in the system's dynamics and its relationship to other commonly observed network properties. On one side, it has been suggested that nested arrangements promote the persistence of mutualistic ecological systems, i.e. increase in abundances [64–66]; but, at the same time, it minimises the system's local asymptotic stability [66–68]. Moreover, it has been shown that nestedness emerges as a result of an individual's fitness optimization process [66], while other works suggest that nestedness is a consequence of the assemblage rules of the systems [69, 70]. On another note, scholars have explored the relation between nestedness and other network properties. For example, nestedness and disassortativity have been found to be correlated [71], and this degree heterogeneity have been identifed as a determinant factor of nestedness [71, 72]. Exploring some of the mechanisms for the emergence of nested patterns constitutes part of the focus of this thesis, see Chapter 5.

1.4.4 Metrics to quantify nestedness

Among the all the research studies devoted to nested patterns, many efforts have also been devoted to the technical aspect, i.e, developing appropriate methods to measure the level of nestedness of a given system [41, 53, 68, 73], and how to properly

assess its statistical significance [23, 72, 74]. From an algebraic perspective, the spectral properties of perfect nested graphs have been studied by mathematicians [75–77], which later facilitated the proposal of a robust detection method [68], in which Staniczenko *et* al. quantified nestedness with respect to the maximum eigenvalue of binary and weighted graphs' adjacency matrices. In a different tradition, ecologists have also dedicated many efforts to quantify nested structures in real systems. In first place, there are measures based on counting misplaced relations to complete a perfect upper triangular nested structure in the adjacency matrix, such as the Nested Temperature (NT) measure, introduced by Atmar and Patterson [41]. To overcome some pitfalls around placement-based measures, Almeida-Neto *et* al. [73] developed overlap metrics, like the Node Overlap and Decreasing Fill (NODF), which considers the amount of common neighbors between every two pair of nodes in matrix A, alongside with its weighted version [78–80], or different extentions of NODF overlap metric, like the global nestedness fitness $\mathcal{N}$, introduced by Solé-Ribalta *et* al.[53]. In the following, we provide a brief description of some of the measures developed for the characterization of nested patterns. For an extensive review on the literature covering nestedness, see [42].

1. *The nestedness temperature T:* the first, and one of the most popular metrics for quantifying nestedness, introduced by Atmar and Patterson [41]. To compute this measure, the calculator performs a three steps process: in first place, an isocline of perfect nestedness is calculated for the matrix, a curve drawn from the lower-left corner of the matrix to the upper-right, with a curvature defined by matrix fill, see Fig 1.4. Then, the rows and columns of the matrix are reordered by their marginal totals in a way that minimizes its temperature. Finally, For all the missing interactions above the isocline or all the observed interactions below the isocline, a global distance to the isocline is computed, and all the values are averaged. The final temperature of the matrix will be the sum of these distances. The score should be zero for a highly nested matrix and 100 for a non-nested one. It should be stress here, that this type of placement-based measures, are sensitive to the algorithms employed to ordering the nodes of the matrices and to determine the line of perfect nestedness.

2. *Node overlap and decreasing fill (NODF):* introduced by Almeida-Neto et al. [73], this measure is independent of row and column order. The NODF is calculated in the following way: for a matrix with N rows and M columns, for any pair of row (i, j) or column nodes (l, m), the number of common interactions (overlap) between them in the adjacency matrix is computed as $O_{ij} = \sum_k A_{ij}A_{ik}$. Finally, the NODF of the whole matrix is:

$$NODF = \frac{1}{\mathbf{K}}\left\{\sum_{ij}^{N}\left[\frac{O_{ij}}{k_j}\Theta(k_i - k_j)\right] + \sum_{lm}^{M}\left[\frac{O_{lm}}{k_m}\Theta(k_l - k_m)\right]\right\}, \tag{1.8}$$

where $\mathbf{K} = [N(N-1)+M(M-1)]/2$ is a normalization over the number of all possible pairs, $\Theta(\cdot)$ stands for the Heaviside step function[1]; and is used to encapsulate the decreasing fill condition.

3. *Spectral radius:* recently, Staniczenko et al. [68] stated that the level of nestedness of a network could be given by the spectral radius $\rho(A)$ of the adjacency matrix of the network, i.e., the largest eigenvalue of the matrix. To demonstrate such claim, the authors built a set of bipartite binary matrices with fixed size and number of interactions, and considered all the possible permutations of such interactions along each matrix. Then, they computed the spectral radius for the set and reported that those matrices with perfectly nested distribution of the interactions had a higher spectral radius than most other matrices, see Fig 1.5.

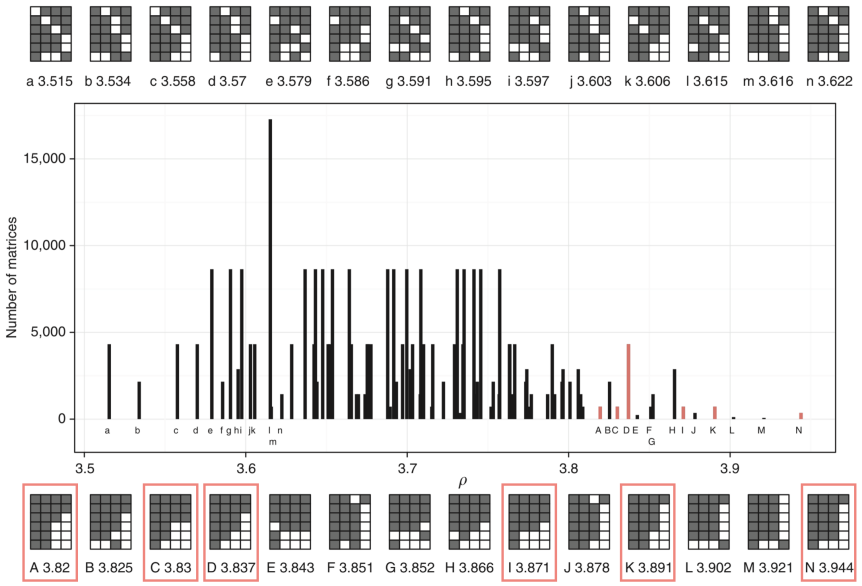


Figure 1.5: Spectral radius ($\rho(A)$, largest eigenvalue) distribution for several bipartite matrices with different internal organization. There are six perfectly nested matrices (orange solid squares) that have higher spectral radius than most other matrices. The maximum spectral radius is found for matrix N, and all matrices with spectral radius greater than that of matrix A are either perfectly nested or very close to being perfectly nested (bottom series). Matrices with the lowest spectral radius depart most severely from perfect nestedness (top series) [68].

4. *Global nestedness fitness $\mathcal{N}$:* is a NODF-like descriptor introduced by Solé-Ribalta

[1] which is zero when its argument is negative, and 1 otherwise

et al [53] that takes into account a null model correction, and is defined as:

$$\mathcal{N} = \frac{2}{N+M}\left\{\sum_{ij}^{N}\left[\frac{O_{ij}-\langle O_{ij}\rangle}{k_j(N-1)}\Theta(k_i-k_j)\right]+\sum_{lm}^{M}\left[\frac{O_{lm}-\langle O_{lm}\rangle}{k_m(M-1)}\Theta(k_l-k_m)\right]\right\}, \quad (1.9)$$

in a similar way as in the NODF metric, here, $O_{ij} = \sum_k A_{ik}A_{jk}$ (or O_{lm}) accounts for the amount of commonly shared neighbours between row or column node pairs (a.k.a overlap); $k_i = \sum_k A_{ik}$ corresponds to the degree of node i (and simililarly for node j); and $\Theta(\cdot)$ is the Heaviside step function, that ensures that O_{ij} is only considered when $k_i \geq k_j$. Additionally, O_{ij} is conveniently corrected by a null model that discounts the expected change of each node have to share a neighbour [53], namely, the expected overlap $\langle O_{ij}\rangle$. Assuming no correlation between neighbouring nodes of i and j the probability of sharing a particular neighbour only depends on the degree of i and j and on size of the network, $(k_i k_j)/N^2$. Hence, the average overlap is $\langle O_{ij}\rangle = \sum_{k=1}^{N}(k_i k_j)/N^2 = (k_i k_j)/N$. The presence of a null model term enforces $\mathcal{N} \in [0,1)$. Note that Eq. 1.9 follows closely the NODF metric with the exception that this metric includes a null model term, $\langle O_{ij}\rangle$, and the normalization term weights the contribution of rows and columns nodes in a linear way (instead of a quadratic form as in NODF). This equation is equally valid for unipartite networks, simply imposing that the sets of rows and columns nodes are equal. This nestedness metric will be employed when needed throughout this thesis.

1.4.5 Statistical significance of nestednes: null models

Last but not least, the debate regarding which one is the appropriate null model to assess nestedness statistical significance has not been less intense, mostly in the ecological community, where the study of nested was mostly performed over bipartite networks.

When randomizing a bipartite network, one has not only to be careful on the level of restriction of the constraints one wants to preserve, but also that such choice can be made independently for rows and columns nodes. Among the different possible options, one can choose between: exactly preserving the selected constraint, e.g., the nodes' degrees, or preserving it on average. The selection of which is the appropriate constraint, is context depending, and not a trivial task. Particularly concerning the risk of having false positives, or false negatives, if the employed null models are too loose or too restrictive [81–83]. For example, if the randomization is performed considering that all the interactions are equiprobable without considering the nodes' degrees [23, 74], or if they exactly preserved the rows and columns nodes degrees in the network [81].

Another disadvantage results in the dependence on the number of randomizations performed. Many randomizations might make the results more robust, but will increase the computational cost. To overcome some of these limitations, in recent years, a statistical physics framework that allows us to analytically compute the expected properties of networks that have been introduced [72, 84]. The expected values and the standard

deviation of the network properties are calculated by finding a probability distribution over an ensemble of networks, that maximizes the entropy of the network. Below, we review a couple of null models that have been introduced/employed to assess the statistical significance of nested patterns. We focus only on those that preserve, on average, the nodes' degree.

1. *Type II probabilistic model:* in this model, the interactions between a pair of nodes are set with a probability proportional to both nodes' degrees as follows [45],

$$p_{ij} = \frac{1}{2}\left(\frac{k_i}{N} + \frac{k_j}{M}\right), \tag{1.10}$$

where N stands for the total number of rows, and M for the number of columns, k_i and k_i correspond to the degree of nodes i and j, respectively. This model has been a popular choice for the generation of the random ensemble of networks, which allows assessing the significance of nested patterns by the computation of z-scores.

2. *Maximum-entropy models:* Under this framework [72, 84], one can calculate the expected values of the network properties, e.g., the nestedness [72] over a maximun entropy ensemble of uni- and bipartite graphs. Particularly, with this approach, one aims to find a probability distribution $P(G)$ over the ensemble of graphs G, with G^* the graph corresponding to the real network, that maximizes the Shannon-Gibbs entropy $S = -\sum_G P(G)\ln P(G)$, and that keeps the average nodes' degree fixed. This maximization has a solution given by the canonical distribution,

$$P(G) = \frac{e^{-H(G,\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})}, \tag{1.11}$$

where Z is the partition function, $H(G,\boldsymbol{\theta}) = \boldsymbol{\theta}\cdot\boldsymbol{C}(G)$ is the graph Hamiltonian, and $\boldsymbol{\theta}$ is a vector of Lagrange multipliers resulting from the maximization of the Shannon-Gibbs entropy, under the chosen constraints $\boldsymbol{C}$, e.g., the graph average degrees. The next step is to calculate the exact values of the Lagrange multipliers. Following the approach presented in [84], these multipliers are determined by imposing that the chosen constraints of the network are found in the ensemble with maximum probability. This is achieved by rewriting the log-likelihood of observing the real network as $L(\boldsymbol{\theta}) = -H(G^*,\boldsymbol{\theta}) - \ln Z(\boldsymbol{\theta})$ and maximizing this quantity in order to find the optimal variables $\boldsymbol{\theta}^*$ that define the ensemble. Once the parameters $\boldsymbol{\theta}^*$ are found, one can build the matrix containing the average probability of interaction corresponding to our empirical network $\langle A^*\rangle$. Finally, it is possible to derive an analytical expression of the first and second moments of the desired network property.

1.5 Mesoscale characterization of complex networks

The mesoscale is understood as substructures (or subgraphs) with distinctive interaction patterns that involve particular subsets of nodes. This section offers a brief description of some structural patterns identified in complex networks at the mesoscale level.

A rather ubiquitous type of mesoscale structure that has received most of the attention is community structure [36, 38–40, 85–87], where nodes organize forming groups, having many links among nodes of the same group and fewer links between nodes of different groups [35]. A typical example are social networks, where some individuals can be part of tightly connected groups, and some others can act as bridges between these groups. This simple intuition hides behind an NP-problem, provided that the number of possible ways to partition a graph scales faster than polynomial with respect to the network size: even for really small graphs, an exhaustive assessment of every partition's fitness becomes unfeasible.

In spite of its technical and inherent limitations, detecting community structure plays a fundamental role in deciphering the dynamical behaviour of many empirical systems. Among others, the role of community structure in the stability and biodiversity of mutualistic and competitive ecological systems has been a subject of study during the last decade [65, 86, 88, 89]. Outside ecology, scholars have investigated the role of communities in information diffusion and epidemic spreading [19, 20, 90, 91]. In parallel, some efforts have been devoted to exploring the possible co-existence between community structure and nestedness [51–53]. Although, the co-existence of community structure with other arrangements, like core-periphery patterns, has also been considered [92]. The latest aspect concerning the possible relation and/or combination of community structure with other architectural patterns will be explored in-depth in Chapter 3.

Thereby, the network science community has developed a rich collection of algorithms and methodologies to infer these communities from relational data. Hereafter, we present a brief review of some methods for community detection. For extensive reviews see [40, 93].

1. *Hierarchical clustering:* In some cases, the graphs may display groups of nodes at different levels, with small groups included in larger groups, i.e., a hierarchical structure. Therefore, in order to identify the multilevel community structure of the graph, the use of hierarchical clustering algorithms is useful [94]. Hierarchical clustering starts by defining and computing a similarity measure between pair of nodes and aims to identify groups of nodes with similarity according to two categories: *agglomerative algorithms*, in which the groups of nodes are iteratively merged if their similarity is high; and *divisive algorithms*, in which the groups are iteratively split by eliminating the links that connect nodes with low similarity.

One main advantage of the hierarchical clustering algorithms is that they do not require preliminary knowledge of the number and size of the clusters. Nonetheless, it lacks a way to discriminate between the different partitions obtained by the procedure, and the results are dependant on the specific similarity measure adopted.

2. *Graph spectral clustering:* generally speaking, spectral clustering includes all methods and techniques that partition the graph nodes by using the spectral properties of the graph [95]. Specifically, the main idea is to perform the clustering based on the eigenvalue spectrum of a Laplacian matrix[2]. The first step consists of computing the eigenvectors corresponding to the lowest eigenvalues of the Laplacian matrix. Then, all data points are projected to the eigen-space, and can be grouped in clusters by using standard partitional clustering techniques like k-means clustering [96]. This change of representation induced by the eigenvectors can make the cluster properties of the initial graph more distint in this eigenvector space, so that clusters can be more easily detected in this new representation.

3. *Methods based on statistical inference:* aimed at fitting a generative model to the network data. The advantage of these type of generative models is that can also be used to generalize the data and predict the occurrence of missing or spurious links in the network [97], and the capacity to inherently address issues of statistical significance. The stochastic block model (SBM) is by far the most used generative model of graphs with communities [97–100]. Starting from a set of N disconnected nodes, divided into B blocks, where g_i is the group to which node i belongs, the links between pairs of nodes are randomly placed, with a probability ω_{rs} that depends only on the groups r and s to which they belong. Thus, one can define $B \times B$ matrix Φ of parameters ω_{rs} that determine the probabilities of having links within and between every pair of groups. Then, the hypothesis is that the observed network was generated from the SBM, and the idea is to find the values of the model parameters that have been used in the generation. Thefore, in order to find the optimal division into communities, the aim is to discover a matrix Φ and the set of group memberships ($\mathbf{g}$), which maximize the likelihood of generating the observed network.

4. *Methods based on using a quality function:* in order to distinguish between “good" and “bad" clusterings, it is useful to require that the partitions satisfy some basic properties. Therefore, is covenient to have a quantitative criterion to assess the goodness of a graph partition. A quality function is a function that assigns a number to each partition of a graph. In this way, one can rank partitions based on the

[2]For undirected unweighter graph is defined as $L = D - A$, where A is the adjacency matrix and D is a diagonal matrix whose diagonal elements are equal to the degrees of each node.

score given by the quality function, after applying an heuristic method for its maximization. Examples of such functions are the map equation, introduced in [101] and the Girvan and Newman modularity [33].

Modularity is by far the most popular and studied quality function. Additionally, modularity optimization itself is nowadays one of the most popular methods for community detection, with plenty of literature available focusing on upgrading the computational time required for its optimization [29, 31], discussing its properties and limitations [102], exploring its relationship with other community detection methods, like SBM [100] and introducing different extensions of it [28, 30, 103, 104].

1.5.1 Modularity

Newman and Girvan modularity [33] was initially described for unipartite networks; and implies that nodes within a module (or block), are more likely to interact among themselves than across the rest of the network; Fig. 1.6. It is based on the idea that a randomly connected network is not expected to have a modular structure, so we can say that a group of nodes form a community if the number of links between them is higher than the expected number of links that the same group of nodes would have if the nodes in the network were at random (null model). Notwithstanding, it is important to point out that random networks can exhibit relatively high values of modularity, due to fluctuations in the establishment of links [105], and the statistical significance of modularity in a real network needs to be assessed.

The modularity measure can be written as

$$Q = \frac{1}{2L} \sum_{ij} \Big(A_{ij} - P_{ij} \Big) \delta(h_i, h_j), \tag{1.12}$$

where A_{ij} is the adjacency matrix of the network, L is the matrix fill, $\delta(h_i, h_j)$ is the Kronecker delta function that will be equal to one when nodes i and j belong to the same module (i.e. they have the same label) and zero otherwise. One of the strong points of Modularity is that by definition, it relies on the concept of null model, the term P_{ij} represents the expected number of links between nodes i and j in such null model. The null model proposed by Newman and Girvan consists of a randomized version of the original graph, where the links are rewired at random, but keeping the expected degree of each node as in the original graph as $P_{ij} = \frac{k_i k_j}{2L}$.

The fitness function in Eq. 1.12 can be rewritten in terms of the total contribution per community as

$$Q = \sum_{c=1}^{B} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right] \tag{1.13}$$

where B is the number of communities, l_c is the total number of links in community c, and d_c is the sum of the degrees of all nodes in such community.

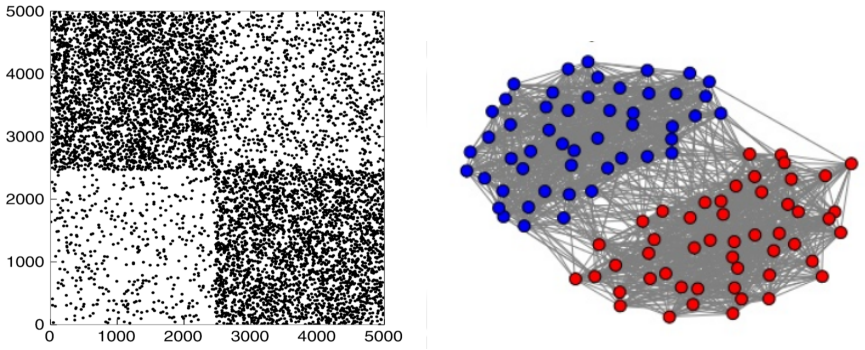


Figure 1.6: Graph (left) and adjacency matrix (right) representation of networks with community structure [93].

Furthermore, Eq. 1.12 and 1.13 can be adapted to account for the underlying nature of the network (weighted [28], signed [103], multilayer [104], etc.). Another interesting alternative, that will be employed along this thesis is the expression of modularity for bipartite graphs suggested by Barber [30]. This expression, addresses the specificities of bipartite networks, in which relations between pairs of nodes in the same set or guild are forbidden. For this, Barber took the adjacency matrix $\mathbf{A}$ of the bipartite graph (Eq. 1.2) and considered a null model matrix $\mathbf{P}$, which has a block off-diagonal form

$$\mathbf{P} = \begin{bmatrix} \mathbf{O}_{r\times r} & \tilde{\mathbf{P}}_{r\times b} \\ (\tilde{\mathbf{P}}^T)_{b\times r} & \mathbf{O}_{b\times b}, \end{bmatrix} \tag{1.14}$$

where $\mathbf{O}$ is the all-zero matrix and $\tilde{\mathbf{P}} = k_r k_b / L$. Then, the bipartite modularity matrix can be computed by $\tilde{\mathbf{B}} = \tilde{\mathbf{A}} - \tilde{\mathbf{P}}$, and Eq. 1.12 can be rewritten in its bipartite form as

$$Q = \frac{1}{L} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(\tilde{A}_{ij} - \tilde{P}_{ij} \right) \delta(\alpha_i^N, \alpha_j^M), \tag{1.15}$$

Finally, when expressed in terms of the sum over the number of modules, the bipartite modularity takes the form

$$Q = \sum_{d=1}^{B} \left[\frac{l_d}{L} - \left(\frac{k_d^r k_d^c}{L^2} \right) \right] \tag{1.16}$$

1.5.2 Modularity Optimization

As mentioned above, the modularity measure is one of the most popular methods to evaluate a network partition. The main problem is that a complete search over all the possible configurations having particularly high modularity values is usually unfeasible. As a result, several algorithms based on heuristic optimization methods have been introduced, such as greedy algorithms [31], simulated annealing methods [105], or extremal

optimization techniques [29]. Here, we briefly outline these approaches for optimizing modularity.

1. *Greedy algorithms:* The first attempt at optimising Q directly was through a a greedy optimization (hill climbing) approach introduced by Newman [106]. It is an agglomerative algorithm, where nodes are successively joined, and the change in Q as result of the merge is calculated. The algorithm keep the partition producing the largest change in Q, and the process is repated until a maximum value of Q is found. Another popular agglomerative greedy approach is the one introduced by Blondel et al. [31]. The algorithm consist of two main step, the main steps operates in a similar manner as the Newman algorithm. The second step consists of building a graph in which nodes are the communities that were found in the previous step, and two nodes of the new graph are connected if there is a least a link between nodes of the corresponding communities. The two steps of the algorithm are then repeated, until, modularity cannot increase any more.

2. *Simulated annealing method:* is a probabilistic approach for approximating the global optimum of a given function. First employed for modularity optimization by Guimerà *et al.* [105]. The algorithm performs an exploration of a space of possible states, looking for the global optimum of a function, modularity in this case. It relies on a temperature parameter τ, which decreases after each iteration by a factor c. At each time step, the algorithm randomly selects a configuration, which is accepted with probability 1 if Q increases after the change, otherwise with a probability $\exp(\frac{\Delta Q}{T})$, where ΔQ is the change on modularity. This small probability reduces the risk that the system gets trapped in local optima.

3. *Extremal Optimization (EO):* An heuristic search proposed by Boettcher and Percus [107] and used for modularity optimization by Duch and Arenas [29]. It consists of a divisive algorithm based on the optimization of a global variable by improving extremal local variables. In this case, the global variable to optimize is Q. Hence, the local variables should be related to the contribution of the individual nodes to the summation of Q. Starting from a random partition of the network into two groups with the same number of nodes. At each step, a local contribution to the total modularity for each node is calculated. The node with the lowest contribution is moved to the other partition. Each movement implies a change in the partition, and a recalculation of the fitness is computed. The process is repeated until the global modularity Q can no longer be improved. Then, each partition is considered as a graph on its own, and the procedure is repeated for each one, as long as Q increases with the new subdivisions. The algorithm represents a good tradeoff between accuracy and speed and will be employed and extended throughout this thesis.

1.5.3 Modularity's resolution limit

As described in the former section, the problem of community detection via modularity optimization is particularly tricky and has been the subject of discussion in various disciplines. Parallel to the constraints of the algorithmic strategies, the formulation of Q has an inherent limitation itself, which impedes to detect blocks that are smaller than $\sqrt{L}$. Intuitively, for the modularity function, this limit can be understood using a toy network formed by a set of cliques placed on a ring, where each pair of adjacent cliques is connected by a single inter-clique link, see Figure 1.7 (top left and middle left panels). This is the most modular connected network [102]. In this setting, one can show that the modularity has a scale detection problem. Even if the network has more cliques than $B \geq \sqrt{L}$, the modularity function will favor partitions where B blocks are detected. This somehow imposes a detection scale which can be intuitively understood by noticing that the expected number of edges between two blocks α and β is, approximately, $P_{\alpha\beta} = k_\alpha k_\beta/(2L)$, where $k_\alpha = \sum_{s\in i} k_s$ denotes the total degree of block α. When both k_α and k_β are of order $\sqrt{L}$ or smaller, $P_{\alpha\beta}$ becomes of order one or smaller, meaning that even a single link between blocks α and β is interpreted by the modularity function as a non-random connection, thereby favoring their merging into a single block [93].

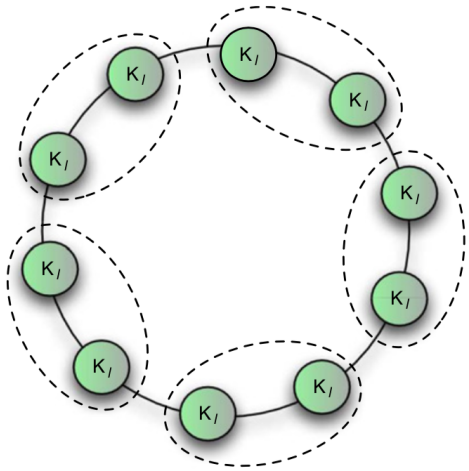


Figure 1.7: Representation of a ring of cliques (green circles) connected by a single link. If the number of cliques is larger than the modularity's intrinsic scale $\sqrt{L}$, modularity optimization would lead to a *wrong* partition, obtained by merging pairs of adjacent cliques (indicated by the dotted lines) [102].

In the maximally-modular network above, an alternative demonstration of the resolution limit can be obtained by comparing the modularity of the correct partition of the nodes into cliques, Q_{single}, against the modularity of the (wrong) partition obtained by merging pairs of adjacent cliques, Q_{pairs}. It turns out that $\Delta Q := Q_{single} - Q_{pairs} > 0$ if and only if $N < \sqrt{L}$. If we gradually increase N by adding new cliques, as soon as

N becomes larger than the modularity's intrinsic scale $\sqrt{L}$, the modularity of the wrong partition, Q_{pairs}, exceeds the modularity of the correct partition, Q_{single} ($\Delta Q < 0$). Alternative examples can be drawn to further prove the modularity's resolution limit in various scenarios [102].

1.5.4 Compound structures: Macroscale patterns at the mesoscale

Although the possibility of simultaneously looking for the presence of combined structural patterns in a network (e.g., communities with internal structures) has been previously suggested [21], the studies that take into account such type of structures are still scarce [51–53, 92, 108]. In the next sections, we will briefly introduce a couple of examples of such types of combined architectures.

1. Multiple Core-periphery: from the different works focusing on the characterization of core-periphery structure, the possibility that the networks may be better regarded as a collection of multiple core-periphery pairs has been suggested, but not explored in depth. Mainly because of the lack of the appropriate tools to perform such analysis. It was not until 2017, where Kojaku and Masuda [92] introduced a novel method to detect multiple groups of core-periphery structure networks (e.g., communities with core-periphery structures as in Fig 1.8). The authors extended the idealized core-periphery structure introduced in [26] to the case of multiple core-periphery pairs by defining an idealized matrix with N nodes:

$$B_{ij}(\mathbf{c},\mathbf{x}) = \begin{cases} \delta_{c_i,c_j} & x_i = 1 \text{ or } x_j = 1 \text{and } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{1.17}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, is a vector of length N, where $x_i = 0$ if node i is a peripheral node, and $x_i = 1$ if node i is a classified as a core node. $\mathbf{c} = (c_1, c_2, \ldots, c_N)$ is also a vector of length N, where $c_i \in \{1, 2, \ldots, C\}$ is the index of the core-periphery pair to which node i belongs, and C is the number of core-periphery pairs, δ is the Kronecker delta. Afterward, they looked for a $(\mathbf{c},\mathbf{x})$ that makes $\mathbf{B}(\mathbf{c},\mathbf{x})$ the closest to the adjacency matrix of the network by maximizing a modularity-like quality function.

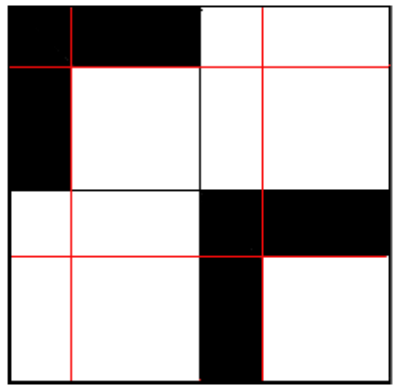


Figure 1.8: Example of an adjacency matrix with perfect Modular-Core-periphery [92].

Finally, they tested the performance of the methods on different types of synthetic and empirical networks by assessing the statistical of the significance of the detected structures. They started by analyzing a, is a political blog networks containing $N = 1222$ nodes and $L = 16714$ links. Each node in the networks corresponded to blogs for the U.S. presidential election of 2004, and two blogs were said to be connected if one blog cited the other blog on its front page. Each blog was labeled by their political leaning, liberal or conservative. Their method coul detect two core-periphery groups, each group comprising the blogs with the same political leaning Fig. 1.9 (left). Furthermore, they analyzed an airport network with $N = 2939$ nodes and $L = 15677$, each node represented an airport, and they were said to be connected if there was a direct commercial flight between them. Their method detected 10 geographically concentrated core-periphery pairs and the overall result indicated that hub metropolitan airports were not necessarily core airports Fig. 1.9 (right). Moreover, while this quality function is able to detect multiple core-periphery structures in a network, it inherits from the modularity function a similar resolution limit [109], which has motivated the introduction of a multiscale variant of the original algorithm [110], and the development of alternatives methods for the detection of multiple core-periphery pairs in real networks [111, 112].

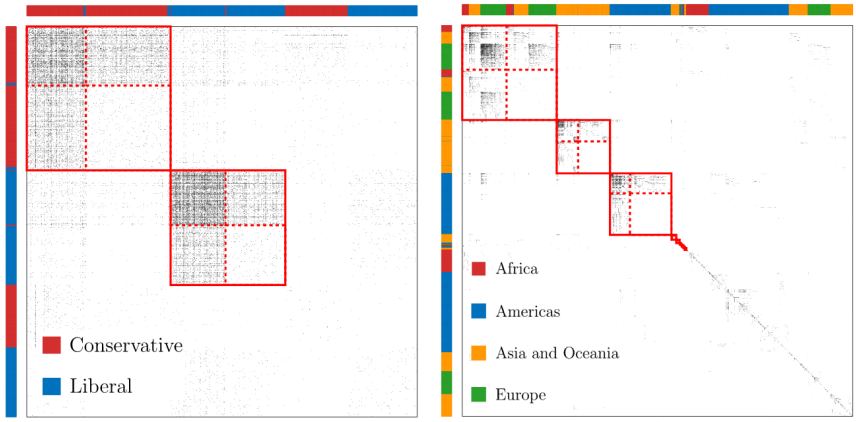


Figure 1.9: Empirical networks with Modular-Core-periphery structure. A political blog network (left) and an airport nework (right) [92].

2. Modular-Nested structure: Nestedness and modularity are emergent properties in many systems, but it is rare to find them in the same system. This apparent incompatibility has been noticed and it might be explained by different evolutive pressures: certain mechanisms favor the emergence of blocks, while others favor the emergence of nested patterns. Following this logic, if two such mechanisms are concurrent, then hybrid modular-nested or *in-block nested (IBN)* arrangements may appear. This type of hybrid or "compound" architectures were first described in Lewinsohn *e*t al.[21], see Fig. 1.10.

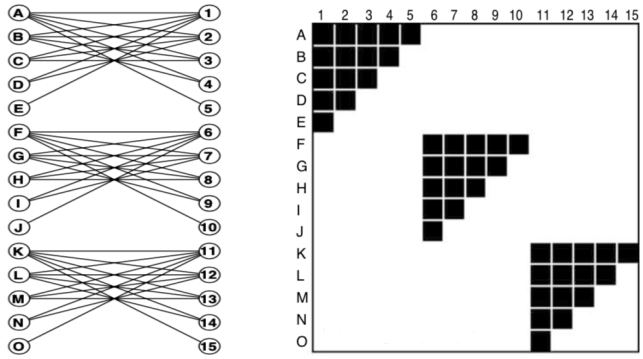


Figure 1.10: Graph (left) and adjacency matrix (right) representation of networks with in-block nested structure [21].

Subsequently, Flores *e*t al. [51] reported the presence of combined nested-modular structure in an infection network. In their work, the authors performed a multi-scale analysis of phage-bacteria infection networks composed of 286 bacteria strains and 215 phage strains with 1332 positive infection outcomes. They found that the infection network was significantly modular and that such modules had a

nested organization, see Fig. 1.11.

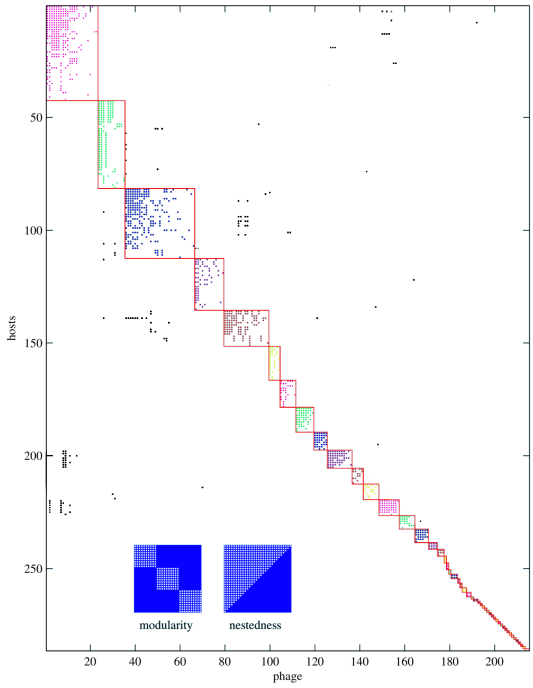


Figure 1.11: A phage-bacteria infection network with inblock-Nested structures analyzed in [51]. The authors detected 49 modules (red rectangles)[52].

The debate over the possible co-occurrence of nestedness and a modularity, in a single network has been covered for over a decade [113, 114], and the possibility of combination of both structures in the same network has been not left out of the debate [21, 50, 51, 53, 108]. The evidence from these works suggests that the emergent interactions, resulting from the dynamical processes in these networks, take place at different scales. Characterizing a network as purely modular or purely nested may be too simplistic; rather, it may integrate both properties reflecting that the system has evolved under different dynamical pressures. And yet, we are missing a systematic approach that tackles the plausible co-existence or combination of nested and modular patterns.

Among the objectives of this thesis, we aim to unravel the dynamical mechanisms that enable the emergence of in-block nested structures, and his role as bridging pattern in systems where a modular-to-nested topological transition have been reported [49]. Thereby, in the later section we will further expose some additional aspects regarding the identification in-block nested patterns.

1.6 Nestedness at the mesoscale: detection of in-block nested patterns

The above section discussed previous work on which the presence of communities with an internal nested organization was reported in empirical systems. The approach followed by the authors to detect such arrangements started first optimizing modularity and subsequently computing the level of nestedness within the detected communities. Although this sequential procedure may deliver good results in some situations (since, often, detected modules gather nodes with degree heterogeneity [98]), to correctly identify this type of pattern, the development of specialized tools is essential. The next sections aim at introducing what is, to the extent of our knowledge, the first methodological framework that jointly considers both patterns.

1.6.1 A new benchmark graph model

The use of a suitable benchmark is crucial to evaluate if measuring modularity and nestedness as two independent network properties, is the appropriate scheme to unveil the existence of IBN structures. In this regard, Solé-Ribalta *et* al. [53] introduced a probabilistic benchmark graph model that is able to generate structures that smoothly interpolate between purely nested, modular and in-block nested patterns. The model pivots on four parameters: the number of modules $B \in [1, \infty]$, noise regarding the existence of interactions outside species communities $\mu \in [0, 1]$, (inter-block noise), noise regarding interactions outside a perfect nested structure $p \in [0, 1]$ (intra-block noise), and a shape parameter for the generation of the nested structure $\xi \in [1, \infty]$ which controls the slimness of the nested structure. Although ξ affects the overall network connectance (total number of existing species interactions), it does not determine it: for example, for a network with a single block ($B = 1$) and $\xi = 1$, the matrix fill is 50%. For the same $\xi = 1$, with $B = 2$, the fill is 25%. On the other hand, p and μ do not alter the density of the network. A formal proof of this aspect is available at [53].

For a network with N nodes, the model allows generating networks with fractional communities. Starting off with B (a real-valued number) blocks, they built $\lfloor B \rfloor$ blocks of size $\lfloor N/B \rfloor$ and another with the remaining $N - \lfloor N/B \rfloor$ nodes. $\lfloor . \rfloor$ stands for the integer part function, forming a block diagonal matrix. In this way, the network communities that are produced have some level of heterogeneity.

Considering the described parameters, they derived the independent probability expressions for having an interaction between species i and j within community c as:

$$\begin{aligned} P(A_{ij}^{c}) = [(1 - p + pp_r)\Theta(jN - f_n(iN)) \\ + p_r(1 - \Theta(jN - f_n(iN)))](1 - p_\mu), \end{aligned} \tag{1.18}$$

where the term f_n corresponds to the p-norm ball curve, drawn for a given ξ value, and employed to generarte the perfect nested structure. Equation 1.18 implicitly models a two-step process:

- ($T1$) by which we first remove from the ideal nested structure the interactions that will be considered as noise, and then,
- ($T2$) these interactions are randomly distributed over the set of remaining non-existing interactions.

The term within square brackets differentiates between the probability of having an interaction within the nested part and outside the nested part inside the community. These two components are separated by the Heaviside function Θ. Having an interaction within the nested part implies either the interaction has not been removed in $T1$, $(1-p)$, or that have been removed in $T1$ then recovered in $T2$, pp_r. The probability of recovering the interaction p_r is proportional to number of interactions that have been removed in $T1$, pL, and inversely proportional to the number of non-existing interactions in the network, $N - L + pL$. That is, $p_r = pL(N - L + pL)^{-1}$, where L is the total number of interactions within the network. The rest of Eq. 1.18, p_r corresponds to the probability of having a link outside the initially nested part. Finally, the term $(1 - p_\mu)$ stands for the probability of not removing the link in the process of generating inter-block noise and $p_\mu = \mu(B-1)/B$. Finally, the probability of an inter-block link, between especies i and j belonging to different communities is given by

$$P(A^o_{ij}) = \frac{2Lp_\mu}{2(B-1)N^2} = \frac{\mu L}{N^2 B} \tag{1.19}$$

where the numerator corresponds to the number of removed interactions within communities in $T1$, and the denominator corresponds to the possible places where each of those links can be relocated in step $T2$.

Figure 1.12 shows some synthetic networks the model is able to generate. The first row of the figure shows perfectly nested networks generated with $B = 1$, varying values of ξ and fixing $p = \mu = 0$. The second row shows perfect in-block nested networks obtained with the same settings $B > 1$ and $p = \mu = 0$. Finally, the third row shows intermediate scenarios varying p and μ for a fixed ξ value. The left example shows an ideal modular network (in terms of modularity), i.e. with no links between communities (bottom-left); and the right example a purely random Erdős-Rényi networks, regardless of B (bottom-right).

1.6.2 An objective function for in-block nestedness

Inspired by the NODF and modularity optimization, Solé-Ribalta *et* al. [53] developed an objective function to quantify in-block nestedness, and proved that it overcomes the limitations of the previous approach, employing the benchmark model described above (we will deepen on this demonstration in the following section). The in-block nestedness quality function $\mathcal{I}$ naturally embeds a suitable null model to discount the expected overlap between pairs of nodes, that can be ascribed to randomness,

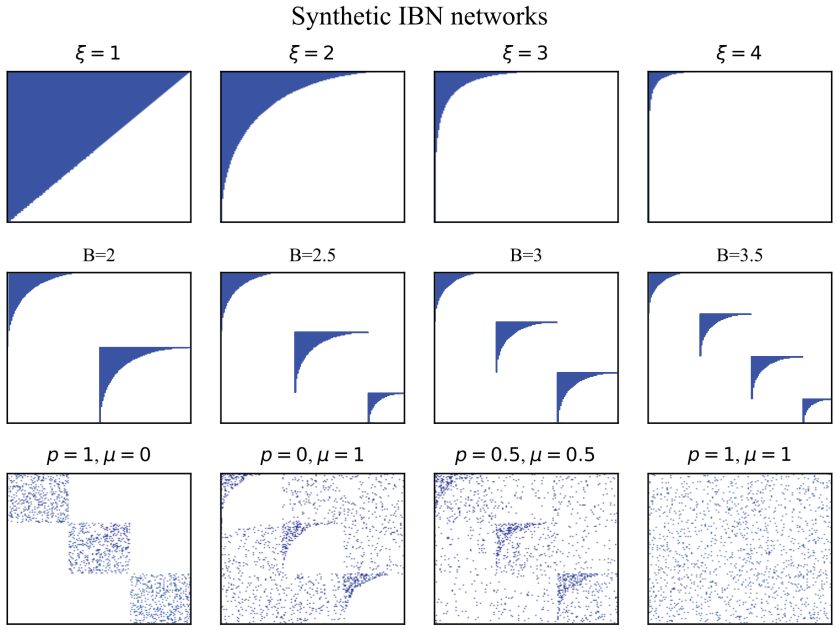


Figure 1.12: Examples of synthetic network generation with the model introduced in [53]. The top and middle rows show the effects of the shape parameter ξ and the number of blocks B, respectively, in a noiseless scenario ($p = \mu = 0$). The bottom row provides some examples of the effect of the noise parameters p and μ for a fixed ξ.

$$\mathcal{I} = \frac{2}{N+M}\left\{\sum_{i,j}^{N}\left[\frac{O_{i,j}-\langle O_{i,j}\rangle}{k_j(C_i-1)}\Theta(k_i-k_j)\delta(\alpha_i,\alpha_j)\right]+\sum_{l,m}^{M}\left[\frac{O_{l,m}-\langle O_{l,m}\rangle}{k_m(C_l-1)}\Theta(k_l-k_m)\delta(\alpha_l,\alpha_m)\right]\right\}, \tag{1.20}$$

where $\mathcal{O}_{i,j}$ and $\mathcal{O}_{l,m}$ measure the degree of overlap between node pairs i and j (or l and m) as in the NODF, k_i corresponds to the degree of the element i; $\Theta(\cdot)$ is the Heaviside step function that ensures that the only terms that contribute to the sum are those in which the outer index has larger degree than the inner. $\langle O_{i,j}\rangle$ represents the expected number of links between nodes i and j in the null model as in modularity. Finally, $\delta(\alpha_i,\alpha_j)$ is the Kronecker delta function that is equal to one when nodes i and j belong to the same module (i.e. they have the same label) and zero otherwise. The expression is valid for unipartite networks, if we consider that the two sets of nodes are identical. Note that, by definition, $\mathcal{I}$ reduces to $\mathcal{N}$ when the number of blocks is 1. As in the case of modularity, in-block nestedness detection is a hard computational problem and the use of heuristic algorithms is mandatory. Other methodologies that may detect communities with similar structural arrangements within modules, e.g. core-periphery [115], exist, but there is no guarantee that the detected communities have nested properties. In the remaining chapters of this thesis, we will perform the in-block nested characterization in networks through the use of the $\mathcal{I}$ objective function.

1.6.3 In-block nestedness and modularity in synthetic networks

By means of the benchmark graph model presented earlier on this chapter, the authors generated 3×10^4 noiseless unipartite networks ($p = \mu = 0$), with varying number of blocks and nested shapes ξ and tested the performance of a modularity optimization algorithm in reconstructing the planted IBN structures by measuring the normalized variation of information (NVI) [116] between the modules detected by modularity optimization and the planted blocks. They found that the modules detected by the Q-maximization were very different from the planted modules, see Fig. 1.13(b), especially in dense nested networks with few blocks, lower left region in panel (b), or in general, when the network is sparse, upper region Fig. 1.13(b). This result indicates that modularity optimization is only reliable in the limit of a large number of blocks and dense networks (lower-right corner of Fig. 1.13(b)). When looking at the values of Q, unsurprisingly, these values increase as the number of blocks increases Fig 1.13(a).

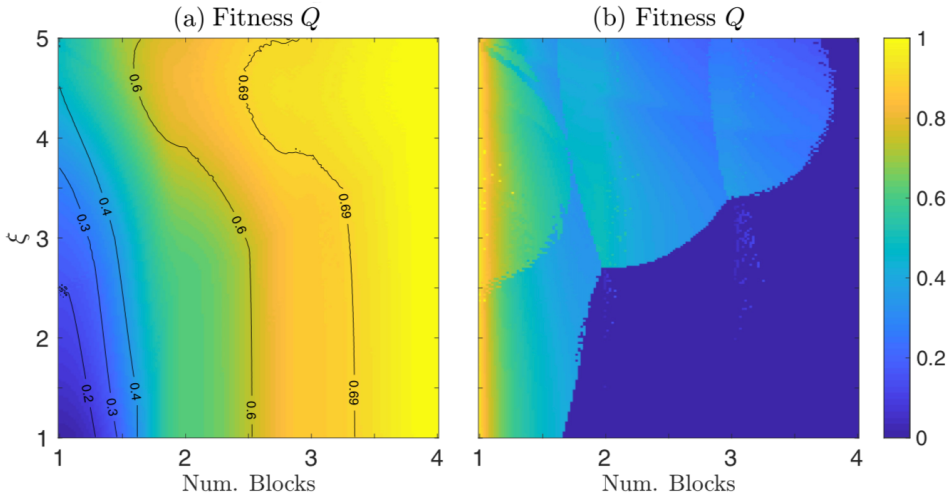


Figure 1.13: Behavior of modularity Q in noiseless IBN synthetic networks with varying values of B and ξ. Panel (a) illustrates how modularity Q increases with the number of blocks in a network. Panel (b) shows the normalized variation of information between the modules detected by modularity optimization and planted blocks [53].

Furthermore, the authors performed an exhaustive exploration of the (p, μ) parameter space for a couple of values of ξ and showed that under this scenario the limitations became more noticeable. They separately measured the NVI between the planted partitions versus the partitions obtained after maximizing $\mathcal{I}$ and Q, see Fig. 1.14. The found that modularity recovers the planted partition in the full range of p values as long as μ remained low. In contrast, they observed that $\mathcal{I}$ optimization allows us to unveil the planted partition for a region along the μ axis, as long as p remained low, showing that the Q-detected partitions were particularly unreliable when a clear internal nested structure with a significant number of interblock links is present. This weakness became

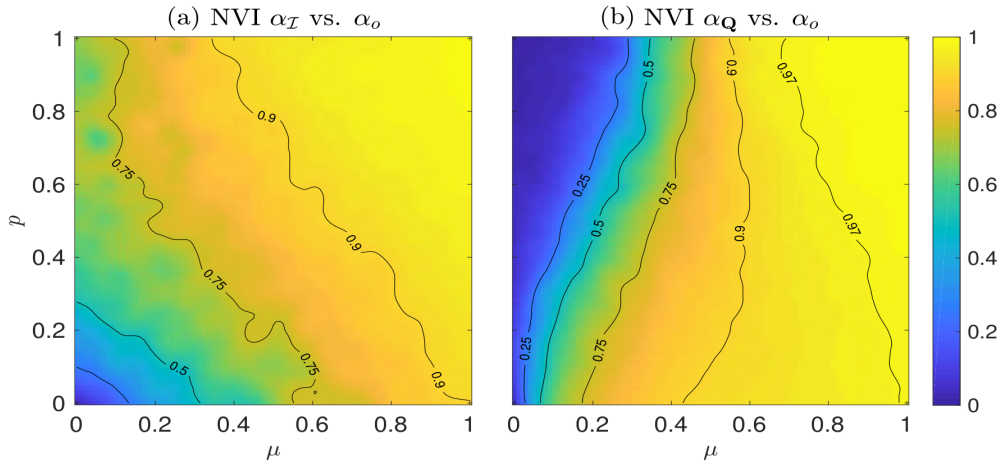


Figure 1.14: Results for synthetic IBN networks with varying values of p and μ. Panels (a) and (b) show the results of NVI between planted partitions α_o and the $\mathcal{I}$-optimized ($\alpha_{\mathcal{I}}$) and Q-optimized (α_Q) partitions, respectively [53].

more evident for highly stylized nested shapes (higher values of ξ), where the sparsity of the network itself, distorted even the almost perfectly modular partitions.

1.6.4 Detecting in-block nestedness in real networks

Besides demonstrating that computing modularity and nestedness independently is not the correct approach to unveil IBN structure for a large set of synthetic networks, the authors performed the Q and $\mathcal{I}$ analysis to more than 300 empirical networks. They considered 57 unipartite and 277 bipartite networks which are known to display some level of nested organization, coming from ecology [117], online platforms [49, 118] and social networks [119–121]. An example of the analysis for a pollination mutualistic network is shown in figure 1.15. Once again, their results pointed out that, even though modularity optimization may sometimes provide partitions with some amount of IBN organization, most of the time it fails to detect the IBN structure in most real-world networks, see Figure 1.15(bottom).

1.6.5 Limitations of in-block nestedness

Beyond assessing how common IBN arrangements appears in real systems, and demonstrating the robustness of the $\mathcal{I}$ objective function, it is important to analyze in-depth the possible shortcomings of such formulation. A relevant aspect to look at, is the possible existence of a resolution limit in $\mathcal{I}$, along the lines of the well-known resolution limit inherent to modularity [102] that was discussed in section 1.5.3. Although, in [53] the authors provided some preliminary numerical exploration on this matter, a deeper analysis accompanied by an analytical account of such resolution limit was left as an open problem. Tackling the latter subject is one of the focus of this thesis, and it will be

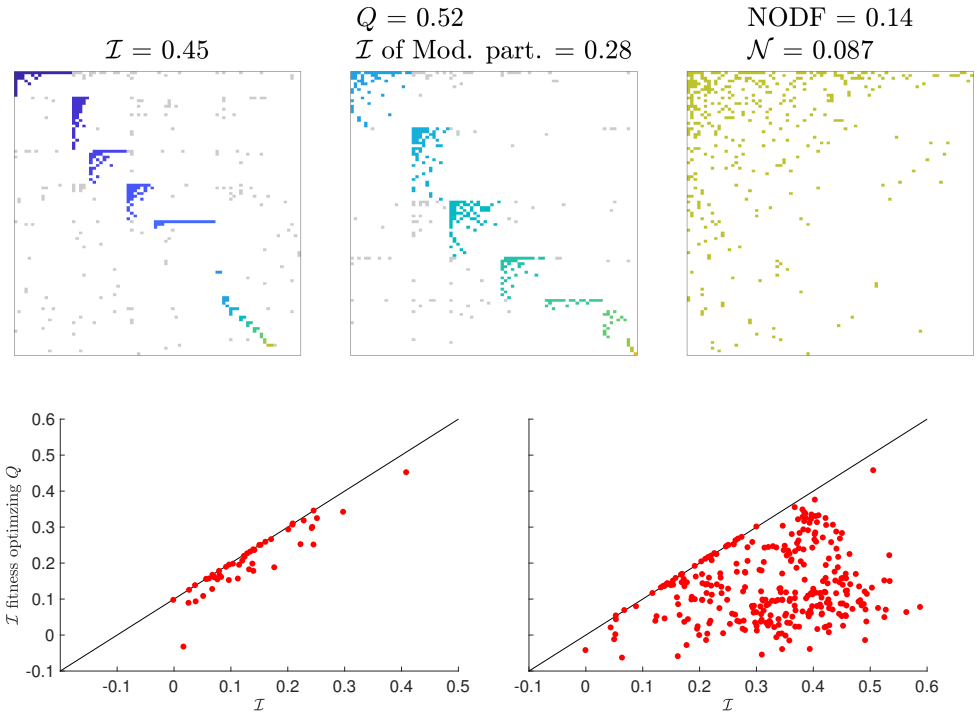


Figure 1.15: Modularity and in empirical networks. Top: Interaction matrix for a pollination mutualistic bipartite network in Cordón del Cepo, Chile. Rows and columns have been arranged to highlight the different arrangements. Bottom: Comparison of in-block nestedness value obtained optimising modularity versus the values obtained by optimizing $\mathcal{I}$, left panel corresponds to unipartite networks and right panel to bipartite networks [53].

explored in detail Chapter 2.

Part II

Structural properties of nestedness, modularity and in-block nestedness

This part of the thesis will be devoted to the examination of the structural properties of the in-block nestedness function, and how it interrelates with the nestedness and modularity, from a strictly analytical and numerical point of view. In particular, we will explore the possible existence of a resolution limit of the in-block nestedness quality function, and will examine the structural constraints that these three measures impose on each other.

Chapter 2

Absence of a resolution limit in in-block nestedness

In-block nestedness (IBN) has emerged as an interesting pattern in complex networks. Initially proposed merely as a hypothetical configuration [21], the idea of hybrid nested-modular structures has gained interest after empirical evidence has shown that such arrangements may play a prominent role in many systems, natural [50–52, 108] and artificial [118].

These findings have shifted the focus from the measurement of nestedness as a global property (*macro* level), to the detection of blocks (*meso* level) that internally exhibit a high degree of nestedness. This change of focus is supported by the existence of various kinds of constraints that operate in real-world systems, which naturally delimit the breadth of interactions [21]. And yet, the availability of methods to properly detect in-block nested partitions are still scarce. As explained in the previous Chapter, in most studies that focus on the identification of such compound structures a sequential approach is applied: after the identification of a network partition (usually in terms of modularity [33]), nestedness (usually in terms of NODF [73] or nestedness temperature [41]) is computed locally for each block.

Recently, the in-block nestedness quality function $\mathcal{I}$ has been proposed [53] and has proven to be a suitable method to correctly identify IBN patterns. Similarly to the popular Newman-Girvan's modularity Q, the optimization of the IBN quality function $\mathcal{I}$ is an NP problem. Nonetheless, while it is well known that Newman-Girvan's modularity, notoriously suffers from a resolution limit that impairs their ability to detect blocks smaller than a given scale [102], and that resolution limits can arise when optimizing a quality function different than modularity [122], the potential existence of such resolution limits for in-block nestedness remains unknown.

In this Chapter, we will examine whether the in-block nestedness function exhibits

a resolution limit, similar to the one found for the modularity function [102]–presented in Chapter 1, Section 1.5.3–. The existence of such a limit would imply the impossibility to detect interaction blocks smaller than a given scale [102], potentially making the interpretation of the detected nested blocks ambiguous [93]. After briefly providing some definitions (Section 2.1) that are relevant to our empirical and analytical explorations, we demonstrate empirically, analytically and numerically, the absence of a resolution limit for the in-block nestedness objective function.

2.1 Definitions: Weak and strong communities

An interesting property, that if often observed in real of networks at the mesoscale, is their heterogeneity: the distribution of its edges is not only globally, but also locally inhomogeneous, with high concentrations of edges *within* groups of nodes, and low concentrations *between* these groups [40].

Such feature of real networks –community structure– can be translated to a quantitative criterion. Radicchi *et al.* [35] propose the following: a block (also called *community*, *module*, *compartment*, or *cluster* depending on the research field [42]) constitutes a *weak community* if and only if its internal degree exceeds its external degree (i.e., the total degree of its nodes by only considering links with nodes that do not belong to the block). Conversely, a block constitutes a *strong community* if and only if, for each of its nodes, the node's internal degree is larger than the node's external degree.

2.2 Empirical insights: preliminary intuitions on Q and $\mathcal{I}$ resolution limit

To test whether there is a resolution limit for the in-block nestedness, we first perform, following the approach described in [102], an empirical exploration for some real networks. From this analysis, we should render some intuitions before the strictly formal approach that we will present in the following sections. Hopefully, by the end of this initial exercise we will get a glimpse on whether a resolution limit for in-block nestedness exists, or not, and how severe it is –if it does exist–, when compared to the resolution limit of modularity.

We collected a set of 82 real networks, from two different domains: ecological in most cases [117], with some collaboration networks taken from socio-technological systems [118, 123], and restricted the size of these networks in the range $[50, 10^3]$ nodes.

The empirical ecological networks analyzed here represent bipartite mutualistic and competitive systems, including macroscopic and microscopic environments. Network data can be downloaded from [117] in different formats, and can be filtered depending on the type of interaction of the system (e.g. plant-pollinator, host-parasite) and the type of data, e.g. binary or weighted. In this work, we have analyzed a total of 52 of these networks, all of them in their binary form. Thus, this kind of networks are represented as a rectangular $N \times M$ matrix, where rows and columns refer to interacting species. An

entry in the matrix $a_{ij} = 1$ if species i of one guild interacts with a species j of the other guild at least once, and 0 otherwise.

On the other hand, for the collaboration networks we collected data from open source software projects through GitHub [123], a social coding platform that provides source code management and collaboration features. Similar to the ecological networks described above, for each project (30 in total) we build a bipartite unweighted network as a rectangular $N \times M$ matrix, where rows and columns refer to the contributors and source files of each open source software project, respectively. An entry in the matrix $a_{ij} = 1$ if a contributor i have edited a file j at least once, and 0 otherwise. More details on this dataset can be found in [118] and in Chapter 4, Section 4.2.

For each network, and following the exercise proposed in [102], we proceed as follows: Initially, for a given network, the quality functions of interest–modularity Q and in-block nestedness $\mathcal{I}$– are maximized by means of an optimization strategy (extremal optimization [29], in our case). See Chapter 1, Section 1.5.2). In each case, a partition P_Q and $P_{\mathcal{I}}$ are obtained as a result. Then, all the links between the detected blocks are removed, and the optimization algorithms are applied again to the resulting network, obtaining now new partitions P'_Q and $P'_{\mathcal{I}}$. With two partitions for each optimization strategy, at hand, we compute the Jaccard index to measure how similar they are, J_{P_Q,P'_Q} and $J_{P_{\mathcal{I}},P'_{\mathcal{I}}}$, respectively. We iterate this procedure –remove links between communities and optimize the quality functions–, until the Jaccard index between consecutive partition vectors $J_{P,P'} \geqslant \tau$, i.e. the similarity between consecutive partitions is at least τ. Note that if we set $\tau = 1$, it implies that the algorithm is no longer able to split the current partition into one with higher score: both P and P' are identical. Along this process, we keep track of the number of iterative steps needed to reach $J_{P,P'} \geqslant \tau$.

If $\mathcal{I}$ suffers from a resolution limit like Q does, we hypothesize that the number of iterative steps needed to reach $J_{P,P'} \geqslant \tau$ will be similar for both $\mathcal{I}$ and Q. If the number of steps needed for $\mathcal{I}$ is larger than those needed for Q, this result would suggest that $\mathcal{I}$'s resolution limit is more severe than Q's. Finally, if the number of steps for $\mathcal{I}$ is smaller than Q's, the conclusion would be that $\mathcal{I}$ is mildly affected (or not at all) by a resolution limit. Note that, ideally, if $\mathcal{I}$ lacks a resolution limit (and assuming that the heuristics can reach the optimal partition), one should expect that after the initial optimization step, the algorithm should not be able to further split the detected blocks into smaller ones, i.e. $J_{P,P'} = 1$ after the first step.

The dataset with the OSS projects and the corresponding software codes for modularity and in-block nestedness optimization (for uni- and bipartite cases), can be downloaded from the web page http://cosin3.rdi.uoc.edu/, under the Resources section.

The result of this experiment is summarized in Figure 2.1. Setting $\tau = 1$, the left panel shows a scatter plot of the number of attempts needed to reach the stopping condition $J_{P,P'} \geqslant \tau$, after the maximization of Q, plotted against the corresponding number of attempts needed for in-block nestedness, for each network. The size of the points in

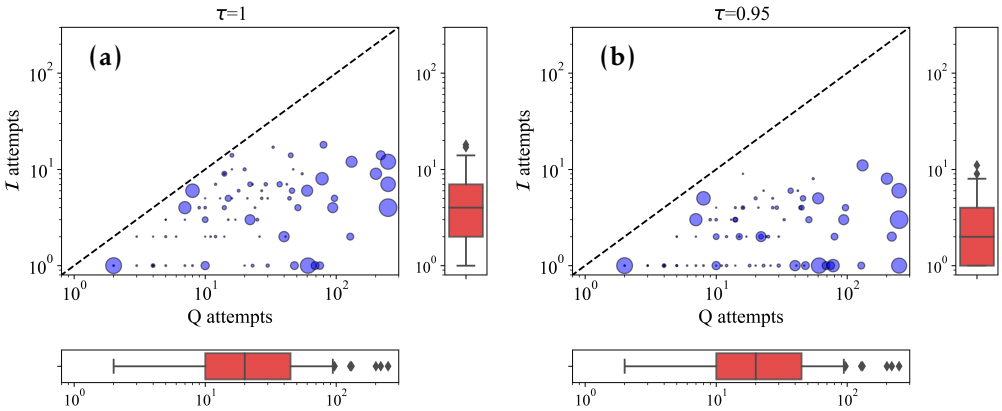


Figure 2.1: **Comparing the resolution limit of modularity and in-block nestedness in empirical data.** Scatter plots of the number of attempts needed to reach $\tau = 1$ (left) and $\tau = 0.95$ (right) for modularity and in-block nestedness. Marginal box plots show the distribution of the number of attempts needed for each network. The size of the points in the scatter plot is proportional to the total number of nodes of each network.

the scatter plot is proportional to the size of each network. Note that the plot is in log scale. To ease comparison, the number of attempts for Q and $\mathcal{I}$ have been plotted in the same scale (log-log), the function $y = x$ is plotted as a dashed black line as a visual aid. Marginal box plots show the distribution of the number of attempts needed for each network, for both Q and $\mathcal{I}$. Without exception, the number of iterations needed to reach the stopping condition is substantially larger for modularity.

Taken strictly, this result can be interpreted as informal evidence of a milder effect of the resolution limit for in-block nestedness (compared to Q). At the same time, this result is not a formal proof that the resolution limit is entirely absent: if the resolution limit is absent, the additional optimization steps could be due to the fact that the extremal optimization algorithm is unable to reach the optimal partition in each step. Relaxing the conditions for the stopping criteria, e.g. $J \geqslant \tau$, with $\tau \in [0.95, 0.99]$, strengthens this informal evidence: the number of attempts needed to reach $J \geqslant \tau$ for $\mathcal{I}$ drops to 1 for many networks, while the number of attempts for Q remains large in most cases: see Figure 2.1 (right panel), which shows this for $\tau = 0.95$.

2.3 Absence of resolution limit in $\mathcal{I}$: analytic approach

In this Section, we aim to provide an analytic explanation for the previous empirical intuitions, in an idealized family of synthetic networks. For the sake of analytic tractability, we consider a ring of interconnected blocks of equal size C, where each block has internally a stepwise structure. That is, the degrees of subsequent rows (columns) of the adjacency matrix differ by one (see Fig. 2.2, bottom-left panel). Additionally, contiguous blocks are interconnected by one single link, $\ell = 1$, that connects the two generalists[1] (or

[1] the node with the largest degree as to the block'shuborgeneralist

hubs) or specialists[2] of each block –in total, there are B inter-block links that connect the B generalists (specialists). Our strategy to perform the calculation is to first compute in-block nestedness $\mathcal{I}_0$ of a perfectly in-block nested network composed of B disconnected blocks, and then add to $\mathcal{I}_0$ the terms due to the interactions between the generalist (or specialists) nodes.

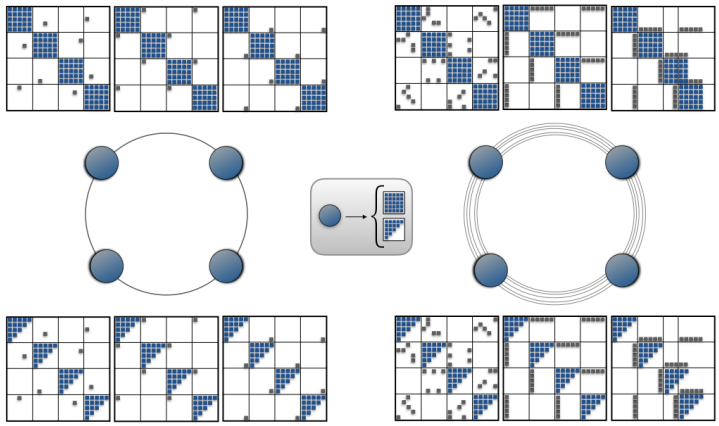


Figure 2.2: **Illustration of a ring of weakly-interconnected blocks.** Central row: Representation of a ring of sub-graphs (blue circles) connected by a single link (left) $\ell = 1$ and connected through several links (right) $\ell = 5$. The subgraphs, represented as blue circles, can take the form of identical cliques or perfect nested blocks. A matrix representation of these cases is shown in top (identical cliques) and bottom (identical perfect nested blocks) rows, respectively. Each adjacency matrix represents a different type of connectivity between communities: random, generalist oriented and specialist oriented.

2.3.1 Derivation of the in-block nestedness of a set of disconnected stepwise blocks

In order to compute $\mathcal{I}_0$, it is sufficient to derive the nestedness of a single stepwise block. Starting from the expression of the unipartite, and in the same way as in modularity, one can rewrite the expression for $\mathcal{I}$ as a sum over the network's blocks:

$$\mathcal{I} = \sum_{\alpha=1}^{B} \mathcal{N}_\alpha, \tag{2.1}$$

where B denotes the total number of blocks and

$$\mathcal{N}_\alpha = \frac{2}{N}\frac{1}{C_\alpha - 1}\sum_{s,t\in\alpha}\left(\frac{O_{st}}{k_t} - \frac{k_s}{N}\right)\Theta(k_s - k_t), \tag{2.2}$$

[2] a node with the smallest number of connections

can be interpreted as the level of block α's internal nestedness. Since each block has an internally nested structure, by definition, $O_{st} = k_t$ if $k_t < k_s$. Therefore, Eq. (2.2) becomes

$$\mathcal{N}_\alpha = \frac{2f(\boldsymbol{k}^{(\alpha)})}{N(C_\alpha - 1)}, \tag{2.3}$$

where $f(\boldsymbol{k}^{(\alpha)}) = \sum_{s\in\alpha}\left(1 - \frac{k_s}{N}\right)\sum_{t\in\alpha}\Theta(k_s - k_t)$.

In general, the function $f(\boldsymbol{k}^{(\alpha)})$ depends on the perfectly-nested block's internal shape or, equivalently, on the density of the perfectly-nested block. The factor $\sum_{t\in\alpha}\Theta(k_s - k_t)$ represents the number of nodes with degree strictly smaller than k_s. As we are considering stepwise perfectly nested networks, we have $\sum_{t\in\alpha}\Theta(k_s - k_t) = k_s - 1$. Hence, after rearranging some terms,

$$f(\boldsymbol{k}^{(\alpha)}) := \left(1 + \frac{1}{N}\right)\sum_{s\in\alpha} k_s - \frac{1}{N}\sum_{s\in\alpha} k_s^2 - C_\alpha. \tag{2.4}$$

Subsequently, for stepwise perfectly-nested networks, the following identities hold:

$$\begin{aligned}\sum_{s\in\alpha} k_s &= \sum_{s=1}^{C_\alpha} k_s = \sum_{s=1}^{C_\alpha} s = \frac{C_\alpha\,(C_\alpha + 1)}{2},\\ \sum_{s\in\alpha} k_s^2 &= \sum_{s=1}^{C_\alpha} k_s^2 = \sum_{s=1}^{C_\alpha} s^2 = \frac{C_\alpha\,(C_\alpha + 1)(2\,C_\alpha + 1)}{6}.\end{aligned} \tag{2.5}$$

By replacing (2.5) into (2.4), and after that into (2.3), and rearranging some terms, Eq. (2.2) becomes

$$\mathcal{N}_\alpha = \frac{C_\alpha}{N} - \frac{2}{3N^2}\,C_\alpha\,(C_\alpha + 1). \tag{2.6}$$

This represents the nestedness of a stepwise block α composed of C_α nodes.

Now the in-block nestedness, $\mathcal{I}_0$ of a (disconnected) network composed of a set of disconnected stepwise blocks is obtained by summing the contributions $\mathcal{N}_\alpha$ – given by Eq. (2.6) – over all the blocks that compose the network

$$\mathcal{I}_0 = 1 - \frac{2}{3N} - \frac{2}{3N^2}\sum_\alpha C_\alpha^2, \tag{2.7}$$

then for the case of equally-sized blocks, $C_\alpha = C = N/B$, $\mathcal{I}_0$ is equal to

$$\mathcal{I}_0 = 1 - \frac{2}{3N} - \frac{2}{3B}. \tag{2.8}$$

2.3.2 Derivation of the in-block nestedness of a ring of weakly-connected stepwise blocks

In order to prove the absence of a resolution limit, we need to calculate $\mathcal{I}_{single}$ and $\mathcal{I}_{pairs}$, and evaluate their difference. To calculate $\mathcal{I}_{single}$, we alter the perfectly in-block nested structure described above by connecting all the generalists (or specialists) nodes of the B blocks; each generalist (or specialist) will now be connected with two other generalists (or specialists), with B inter-block links, in total, see Fig. 2.2 for an illustration.

1. *Generalist-based strategy:* We start with the derivation of $\mathcal{I}_{single}$, altering the perfectly in-block nested structures by connecting the generalists nodes, i.e., *generalist-based strategy*. In this case, because of their links with the hubs of the two adjacent blocks, each generalist node have degree $C_\alpha + 2$. For simplicity, we assume that the blocks have the same size $C = N/B$; the hubs' degree is therefore $C + 2 = N/B + 2$, and the O_{st}/k_t term remains always equal to one if $k_s > k_t$ because internally, the blocks remain perfectly nested. The negative term receives now, for each block, an additional contribution given by the two extra link of each hub. Therefore, the in-block nestedness of the network, $\mathcal{I}_{single}$, can be expressed as $\mathcal{I}_{single} = \mathcal{I}_0 + \mathcal{I}_{int}$, where $\mathcal{I}_{int}$ is the "interaction" term that results from the edges that connect the hubs. Overall, this extra term is

$$\mathcal{I}_{int} = -\frac{2}{N}\sum_{\alpha=1}^{B}\frac{1}{C_\alpha - 1}\sum_{t\in\alpha}\frac{2}{N}\Theta(C+2-k_t) = -\frac{4B}{N^2}, \tag{2.9}$$

where we used the fact that there are $C_\alpha - 1$ nodes of degree smaller than $C + 2$ in each block (all the non-hub nodes, simply). Therefore, we obtain

$$\mathcal{I}_{single} = 1 - \frac{2}{3N} - \frac{2}{3B} - \frac{4B}{N^2}. \tag{2.10}$$

2. *Specialist-based strategy:* We calculate here the $\mathcal{I}^{Sp}_{\text{single}}$ for a network composed of specialist-connected stepwise blocks, and we assume $C \geq 4$. To compute $\mathcal{I}^{Sp}_{\text{single}}$, we evaluate the consequences of adding inter-block links among the specialists on the in-block nestedness $\mathcal{I}_0$ of a perfectly in-block nested network. There are two consequences: first, the degree of the specialist increases from 1 to 3, which affects the terms O_{st}/k_t for $k_s \geq 4$, which leads to a correction $\mathcal{I}^{(\geq 4)}_{int}$; Second, this increase in degree affects the relative degree of the nodes with $k_s \leq 3$, which affects the Heaviside functions and leads to a correction $\mathcal{I}^{(<4)}_{int}$. Overall, $\mathcal{I}^{Sp}_{\text{single}} = \mathcal{I}_0 + \mathcal{I}_{int}$, where $\mathcal{I}_{int} = \mathcal{I}^{(\geq 4)}_{int} + \mathcal{I}^{(<4)}_{int}$.

   To calculate $\mathcal{I}^{(\geq 4)}_{int}$, we remove the original terms that assumed that the specialist has degree equal to one, and add the corrected terms that assume that it has degree equal to three. This leads to two contributions:

$$\mathcal{I}^{(\geq 4)}_{int} = \frac{2B}{N(C-1)}\left(-\sum_{s=4}^{C}\left(1-\frac{s}{N}\right)+\sum_{s=4}^{C}\left(\frac{1}{3}-\frac{s}{N}\right)\right) \tag{2.11}$$

   To calculate $\mathcal{I}^{(<4)}_{int}$, we consider that the nodes with intra-block degree equal to 2 and 3 do not have anymore a larger degree than the specialist, and the specialist has a larger degree than the node with intra-block degree equal to 2. This leads to three contributions:

$$\mathcal{I}^{(<4)}_{int} = \frac{2B}{N(C-1)}\left(-\left(1-\frac{3}{N}\right)-\left(1-\frac{2}{N}\right)+\left(\frac{1}{2}-\frac{3}{N}\right)\right). \tag{2.12}$$

By performing the sums in (2.11) and re-grouping terms, we obtain

$$\mathcal{I}_{int} = \mathcal{I}_{int}^{(\geq 4)} + \mathcal{I}_{int}^{(<4)} = \frac{2B}{N(C-1)}\left(-\frac{2C}{3} + \frac{2}{N} + \frac{1}{2}\right). \tag{2.13}$$

By summing up expressions (2.8) and (2.13), one can calculate $\mathcal{I}_{\text{single}}^{Sp}$,

$$\mathcal{I}_{\text{single}}^{Sp} = 1 - \frac{2}{3N} - \frac{2}{3B} + \frac{2B}{N(C-1)}\left(-\frac{2C}{3} + \frac{2}{N} + \frac{1}{2}\right). \tag{2.14}$$

Note that, differently than the interaction terms for the generalist-based strategy $\mathcal{I}_{int}$ (Eq. 2.9), for the specialist-based strategy tends to a finite value for $N \to \infty$ with constant C: $\mathcal{I}_{int} \to 2(-2C/3 + 1/2)/(C(C-1))$. This limit value implies that connecting the blocks' specialists causes a bigger loss to the in-block nestedness than connecting the hubs, and the contribution to the in-block nestedness from specialist-based interactions is not negligible even in the thermodynamic limit.

2.3.3 Proving the absence of a resolution limit: generalist-based strategy

The in-block nestedness of a wrong partition, $\mathcal{I}_{pairs}$, where pairs of contiguous blocks, α_i and α_{i+1} ($i = 1,\ldots,B$), are merged, can be calculated by adding up the contributions from pairs of nodes that belong to the same block, and those from pairs of nodes that belong to a different blocks. The contributions from pairs of nodes that belong to the same in-block nested block $f(\boldsymbol{k}^{\alpha_1})$ are defined by Eq. (2.4). The contributions from all pairs of nodes that belong to the merged block α_{12}, denoted as $f(\boldsymbol{k}^{\alpha_{12}})$; and contributions from pairs of nodes that belong to the same merged block α_{12}, but different in-block nested blocks α_1 and α_2, $f_{12}(\boldsymbol{k}^{\alpha_{12}})$ are defined as follows

$$\begin{aligned} f(\boldsymbol{k}^{\alpha_{12}}) &= \sum_{s,t\in\alpha_{12}}\left(\frac{O_{st}}{k_t} - \frac{k_s}{N}\right)\Theta(k_s - k_t), \\ f_{12}(\boldsymbol{k}^{\alpha_{12}}) &= \sum_{s\in\alpha_1,t\in\alpha_2}\left(\frac{O_{st}}{k_t} - \frac{k_s}{N}\right)\Theta(k_s - k_t). \end{aligned} \tag{2.15}$$

Based on symmetry with respect to permutations of the blocks, we obtain:

$$\mathcal{I}_{pairs} = \frac{B}{2}\frac{2}{N}\frac{1}{2C-1} f(\boldsymbol{k}^{\alpha_{12}}). \tag{2.16}$$

Note that the block-size normalization factor is given by $1/(2C-1)$ and there is an overall factor $B/2$, which reflects the property that the partition comprises $B/2$ merged blocks which contain $2C$ nodes each. For symmetry with respect to permutation of α_1 and α_2, $f(\boldsymbol{k}^{\alpha_{12}})$ takes the form $f(\boldsymbol{k}^{\alpha_{12}}) = 2f_{11}(\boldsymbol{k}^{\alpha_1}) + 2f_{12}(\boldsymbol{k}^{\alpha_{12}})$, where $f_{12}(\boldsymbol{k}^{\alpha_{12}})$ is equal to

$$f_{12}(\boldsymbol{k}^{\alpha_{12}}) = \sum_{t=1}^{C-1}\frac{1}{t} - \frac{(C-1)(C+2)}{N} - \sum_{s=1}^{C-1}\frac{s(s-1)}{N}. \tag{2.17}$$

The first term on the r.h.s is the positive contribution that comes from the overlap between the hub of block α_1 and the $C-1$ non-hubs of block α_2. The second term is the negative contribution that comes from the expected overlap between the hub of block α_1 (with degree $C+2$)) and the $C-1$ non-hubs of block α_2. The third term is the negative contribution that comes from the expected overlap between the non-hubs of block α_1 and the non-hubs of block α_2; note that there is no overlap between the neighborhoods of the non-hubs of block α_1 and the non-hubs of block α_2. By using again the identities (2.5) and rearranging some terms, we obtain

$$f_{12}(\boldsymbol{k}^{\alpha_{12}}) = H_{C-1} - \frac{g(C)}{3N}, \tag{2.18}$$

where $H_{C-1} := \sum_{t=1}^{C-1} t^{-1}$ denotes the $C-1$th harmonic number, and we defined the polynomial function $g(C) := (C-1)(C^2+C+6)$. Note that the two terms in the r.h.s. represent the contribution of the observed and expected overlap between the nodes that belong to the two different original blocks that are joint together in the merged partition. By plugging Eq. (2.18) into Eq. (2.16), we obtain

$$\begin{aligned} \mathcal{I}_{pairs} &= \frac{1}{C(2C-1)}\left[2\, f_{11}(\boldsymbol{k}^{\alpha_1}) + 2\left(H_{C-1} - \frac{g(C)}{3N}\right)\right], \\ &= \frac{C-1}{2C-1}\mathcal{I}^{single} + \frac{2}{C(2C-1)}\left(H_{C-1} - \frac{g(C)}{3N}\right). \end{aligned} \tag{2.19}$$

Finally, by putting together Eqs. 2.10 and 2.19, we obtain

$$\Delta\mathcal{I} = \mathcal{I}_{single} - \mathcal{I}_{pairs} = \frac{C}{2C-1}\mathcal{I}_{single} - \frac{2}{C(2C-1)}\left(H_{C-1} - \frac{g(C)}{3N}\right). \tag{2.20}$$

Numerical results in Figure 2.3 show the perfect matching between the analytical insights in Eq. (2.20) (Fig. 2.3(a)), and Eqs. (2.10),(2.19) and (2.21) (Fig. 2.3(b)). For a fixed $C \gg 1$ value, in the limit $N \to \infty$ (or equivalently, $B \to \infty$), we obtain

$$\mathcal{I}^{pairs} \to \mathcal{I}^{single}/2, \tag{2.21}$$

confirming the numerical intutions in [53], and in accordance with Fig. 2.3(b). This implies that no matter how large the network is, the in-block nestedness of the partition with pairwise-merged blocks remains significantly smaller than the in-block nestedness of the partition with the original blocks. The same holds true for small values of C, because the second term in the r.h.s. of Eq. (2.20) tends to be substantially smaller than the first term. The reason is that the contribution from the null model is negligible compared to the penalty due to the merging of two blocks into a single one. Therefore, in this idealized example, the penalization for larger blocks in the in-block nestedness function prevents the resolution limit, allowing the in-block nestedness function of the partition composed of the individual blocks to stay always larger than the in-block nestedness of the partition composed of pairwise-merged blocks.

2.3.4 Proving the absence of a resolution: specialist-based strategy

To obtain $\mathcal{I}^{Sp}_{\text{pairs}}$ for the specialist-based strategy, we need to calculate $f_{12}(\boldsymbol{k}^{\alpha_{12}})$. Following 2.3.3, we consider a pair of adjacent blocks, (α_1, α_2). The only non-zero overlap between nodes in α_1 and nodes in α_2 (with the condition $k_s > k_t$) is the overlap between α_1's hub and α_2's specialist: both nodes are indeed connected with α_1's specialist, resulting in a contribution $f_{overlap} = 1/3$ to $f_{12}(\boldsymbol{k}^{\alpha_{12}})$. To calculate the null-model contribution to f_{12}, it is convenient to split the null-model contribution into the contribution $f^{(\geq 4)}_{null}$ from nodes $s \in \alpha_1$ with $k_s \geq 4$, and the contribution $f^{(<4)}_{null}$ from nodes with $s \in \alpha_1$ with $k_s < 4$. Overall, $f_{12}(\boldsymbol{k}^{\alpha_{12}}) = f_{overlap} + f^{(\geq 4)}_{null} + f^{(<4)}_{null}$. We obtain:

$$f^{(\geq 4)}_{null} = -\sum_{s=4}^{C} \frac{s(s-1)}{N} = -\frac{C^3 - C - 24}{3N}, \tag{2.22}$$

Among the nodes $s \in \alpha_1$ with $k_s < 4$, the only nodes that find a node in α_2 with a strictly smaller degree than them are: the node with intra-block degree equal to three and the specialist. Both of them provide a contribution $-3/N$ to f_{null}, resulting in $f^{(<4)}_{null} = -6/N$. Putting all together:

$$f_{12}(\boldsymbol{k}^{\alpha_{12}}) = \frac{1}{3} - \frac{f(C)}{3N} \tag{2.23}$$

where $f(C) = C^3 - C - 6$. Putting together Eqs. (2.14) and (2.23), we obtain:

$$\mathcal{I}^{Sp}_{\text{pairs}} = \frac{C-1}{2C-1}\mathcal{I}^{Sp}_{\text{single}} + \frac{2}{C(2C-1)}\left(\frac{1}{3} - \frac{f(C)}{3N}\right). \tag{2.24}$$

For a fixed $C \gg 1$ value, in the limit $N \to \infty$ (or equivalently, $B \to \infty$), we obtain

$$\mathcal{I}^{Sp}_{\text{pairs}} \simeq \frac{1}{2}\mathcal{I}^{Sp}_{\text{single}} + \frac{1}{3C^2} + \frac{C}{3N} \to \frac{1}{2}\mathcal{I}^{Sp}_{\text{single}} + \frac{1}{3C^2} \simeq \frac{1}{2}\mathcal{I}^{Sp}_{\text{single}}, \tag{2.25}$$

which proves again the absence of a resolution limit. In a similar manner as for the generalist-based strategy, Fig. 2.3(c) shows the agreement between analytical and numerical results for Eqs. (2.14) and (2.24). Gray symbols and dotted line in Fig. 2.3(d) confirm Eq. (2.25).

2.4 Generalizing the absence of resolution limit in $\mathcal{I}$: numerical approach on benchmark graphs

Supported by the excellent agreement between analytical and numerical results in Figure 2.3, we now carry out a numerical validation considering less idealized scenarios. We do so examining numerically whether the in-block nestedness function presents a resolution limit or not, in scenarios beyond $\ell = 1$ where modularity does. To this end, we analyze benchmark networks along the lines of Figure 2.2 (middle-right and bottom-right panels), that is, building unipartite synthetic networks, composed of a growing ring of blocks that internally exhibit a nested structure. We study a wide range of these

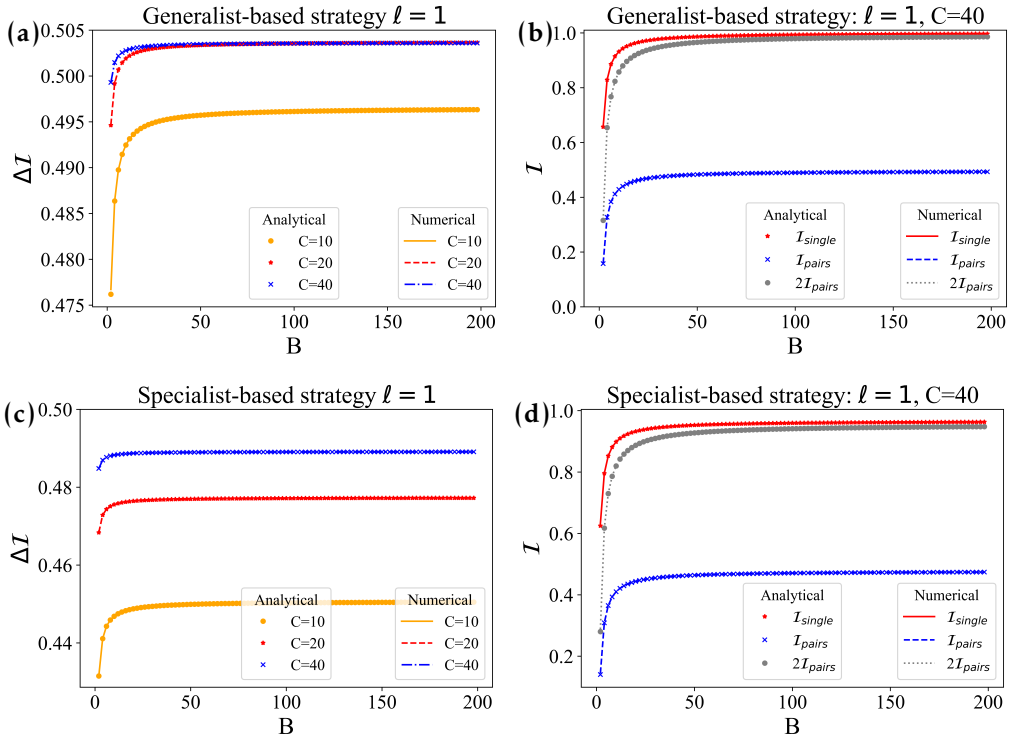


Figure 2.3: **Agreement between the analytical and numerical results.** Top and bottom left panels reports on the perfect match between of the analytical computation of $\Delta\mathcal{I} = \mathcal{I}_{\text{single}} - \mathcal{I}_{\text{pairs}}$ (symbols) and the actual calculation performed on synthetic graphs (lines). Similarly, the top and bottom right panels shows the agreement between analytical and numerical results for Eqs. (2.10) and (2.19) (top right), Eqs. (2.14) and (2.24) (bottom right). Gray symbols and dotted line in top and bottom right panels confirms Eq. (2.21) and Eq. (2.25), respetively.

networks, modifying the number of blocks B that conform the ring, and the number of inter-block links ℓ. We start with a network composed of $B = 3$ (perfectly nested) stepwise blocks connected as a ring, and then consider a growing number of blocks (up to $B = 200$). Regarding the inter-block connectivity ℓ, we start with $\ell = 1$, which corresponds to the analytical calculations above, up to $\ell = C(C-1)/2$ which corresponds to maximum possible connectivity between contiguous blocks. The internal nested structure of the blocks is generated following the approach developed by Solé-Ribalta *et al.* [53], described in Chapter 1, Section 1.6.1.

We carry out the numerical validation considering three strategies to generate inter-block connectivity: in one of them, we consider a *random strategy*, where the blocks are connected by adding a link between two randomly selected nodes from each block. For this case, we report results for an average over 25 realizations. In the other two, the addition of inter-block links ($\ell \geq 1$) is deterministic. The first builds upon connecting the most-generalist available nodes in each pair of adjacent communities (*generalist-based strategy*). Note that, strictly speaking, this strategy is the logical generalization of

our analytical results (where a single link was laid between adjacent blocks, connecting the generalist nodes in them). The second one builds upon connecting the most-specialist available nodes in each pair of adjacent communities (*specialist-based strategy*. See Fig. 2.2 for an illustration of the different linking strategies.

For all the strategies, we compare numerically the in-block nestedness of the ground-truth partition, $\mathcal{I}_{\text{single}}$, against the in-block nestedness of the wrong partition obtained by considering pairs of adjacent blocks as a single block, $\mathcal{I}_{\text{pairs}}$. If the in-block nestedness has a resolution limit beyond the scenario presented in the previous section, then for some value of B we would observe a crossover from $\Delta\mathcal{I} := \mathcal{I}_{\text{single}} - \mathcal{I}_{\text{pairs}} > 0$ to $\Delta\mathcal{I} < 0$, as indeed happens with Q. All these results are shown in Figure 2.4, where first to third rows present the results for $\Delta\mathcal{I}$ in the random, generalist and specialist strategies, respectively. The bottom row, conversely, corresponds to the results for ΔQ for the *random* strategy only, since ΔQ presents a qualitatively similar behavior across all strategies. For the sake of clarity, a black vertical line is drawn in each panel highlighting the weak community criterion. Beyond this limit, no recognizable block structure is compatible with the definition of community, and therefore it becomes irrelevant whether a given quality function identifies a "correct" block or not. Each column of the figure corresponds to different block sizes C.

For the random strategy, each point in the parameter space (B, ℓ) of the panels in Figure 5.6 reports the average value of $\Delta\mathcal{I}$ (top row), and ΔQ (bottom row), for 25 different realizations. There are at least three remarkable lessons from Figure 2.4, equally valid for all the adopted linking strategies (generalist-based, random, specialist-based). First, only Q shows the existence of a resolution limit consistently –no matter the number of inter-block links ℓ, we always find a large-enough number of blocks such that the resolution limit appears, i.e. Q_{pairs} is larger than Q_{single}. Second (a consequence of the first), the appearance of the resolution limit for Q is independent of the criterion of weak community: the crossover to $\Delta Q < 0$ can occur anywhere in the ℓ spectrum, and it depends on B only (i.e., on network size, in line with the analytic results in [102]). Of course, increasing ℓ reduces the amount of blocks B needed to reach the crossover (note the logarithmic scale on the B axis). Finally, the robustness of the single block as the best partitioning scheme for in-block nestedness (i.e. $\Delta\mathcal{I} > 0$) is remarkably high. Note that $\mathcal{I}_{single}$ remains systematically larger than $\mathcal{I}_{pairs}$ until ℓ has almost reached the weak community criterion. In other words, $\mathcal{I}$ identifies the correct block-by-block structure up to the point where such partition (or any other one) becomes unrecognizable.

The only relevant difference between the random (first row), the generalist (second row) and specialist (third row) linking strategies is related to ℓ. The area of the parameter space where the in-block nestedness cannot detect the correct block partition ($\Delta\mathcal{I} < 0$) is substantially smaller in the generalist strategy, compared to the same area under the random and the specialist strategies. This indicates that when inter-block connections are preferentially established by local hubs (or generalists), in-block nest-

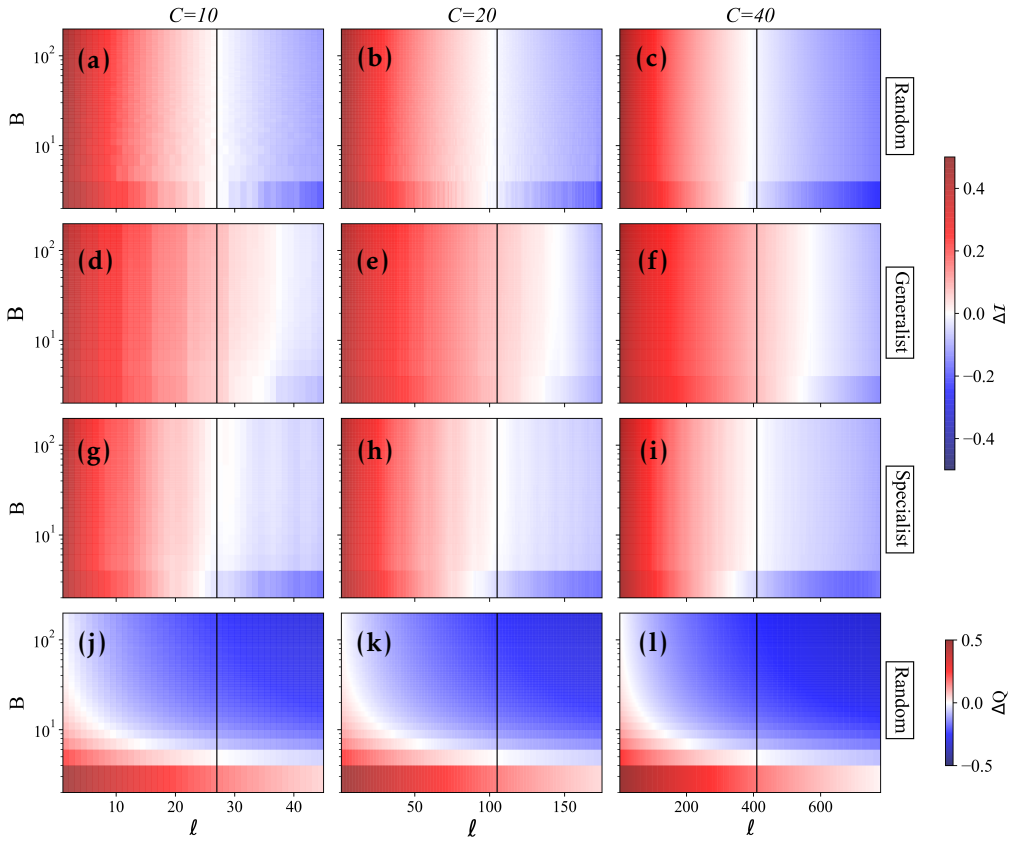


Figure 2.4: **Resolution limit in random- and generalist-connected rings of nested blocks.** The panels represent three-dimensional plots in the parameter space (B, ℓ) showing, in the z-axis (color code), the values of $\Delta\mathcal{I}$ (first to third row panels) and ΔQ (fourth row panel). In the random linking strategy (first and last rows), results are averaged over 25 different realizations. The solid black line indicates the transition from weak communities to no communities, as defined by Radicchi *et* al. [35].

edness can detect blocks of locally nested interactions even when these blocks are not communities in the traditional sense. The specialist linking strategy performs similarly to the random strategy, and in-block nested communities can be detected up to the point where communities diffuse with the rest of the network (vertical black line on the plot). Other than these remarks, the previous conclusion holds: $\mathcal{I}$ does not show a dependency on B (and thus on N) by which its ability to detect the right partition is affected, and thus $\mathcal{I}$ appears to lack a size-related resolution limit in the traditional sense.

2.5 Summary

In this Chapter, we have verified whether the in-block nestedness function exhibits a modularity-like resolution limit, i.e., the inability to identify blocks smaller than a certain scale. We have approached the question of in-block nestedness' resolution limit as a three-step process. First, we have performed an informal test on empirical networks,

to assess the extent to which a network can be recursively split into smaller blocks, which may indicate the existence of a resolution limit [102]. From there, upon the intuition that in-block nestedness lacks a resolution limit (or, at least, it is less severe than Q's), we provide a formal proof that $\mathcal{I}$ does not have a resolution limit, at least in a specific setting –that in which different blocks are connected by a single link. Finally, we have numerically generalized and confirmed the analytical argument, exhaustively studying a large parameter space with varying network size and inter-block connectivity. Thus, we have showed that our capacity to detect correct IBN partitions in networks via its maximization may depends solely on the accuracy of the optimization algorithms.

Chapter 3

Macro and mesoscale pattern interdependencies in complex networks

The detection and identification of emergent structural patterns constitute one of the main focus in the development of modern network theory, with many efforts devoted to the technical aspect, i.e., the development of appropriate metrics for its characterization along with the exploration of the inherent limitations of such metrics. Following this tradition, in the previous Chapter, we have shown that the in-block nestedness function differs from the traditional community detection methods, and does not suffer from a resolution limit.

In light of these differences, in this Chapter, we will further explore the structural properties of the in-block nested function, and how it relates with the nestedness modularity functions, for both uni- and bipartite settings. The study of nestedness and modularity has, by far, concentrated most of the attention of scholars. Such interest is not surprising, because these arrangements had been observed in a wide variety of systems and had shown to have important implications in its different dynamical properties. On one side, modularity is a near ubiquitous mesoscale configuration [28–34, 36–40], that appears to be a crucial actor in the stability of ecological systems [86, 88, 89], and in the diffusion dynamics of social systems [20, 90]. On the other, nestedness stands as a frequent macroscale pattern which has been observed prominently in ecology [41, 45], but also in economy [46–48] and social systems [49], and that seems to play a key role in the persistence and stability of mutualistic ecological systems [64–67, 124].

Overall, the knowledge acquired in the last 40 years has unveiled some of the implications each of these individual organizational patterns have on a system's dynamics. However, we have limited knowledge on how different structural signatures may interlace, or how –if ever– they affect and limit each other. A clear example, is the enduring debate over the possible co-occurrence of nestedness and modularity in a single net-

work [113, 114], or regarding the presence of intermediate nested-modular regimes, in the form of in-block nested structures [50–53].

We will show experimentally and analytically that nestedness imposes bounds to modularity, with exact analytical results in idealised scenarios. Furthermore, we analytically evidence that in-block nestedness provides a natural combination between nested and modular networks, taking structural properties of both. Far from a mere theoretical exercise, understanding the boundaries that discriminate each architecture is fundamental, to the extent that modularity and nestedness are known to place heavy dynamical effects on processes, such as species abundances and stability in Ecology.

3.1 Structural analysis on synthetic networks

We start the analysis of the trade-offs between nestedness $\mathcal{N}$, modularity Q and in-block nestedness $\mathcal{I}$ by exploring a controlled synthetic setting. To this aim, we have extended the network generative model described in Chapter 1, Section 1.6.1 [53], to generate networks with a fixed block size and increasing number of blocks (hence, increasing network size), instead of networks with fixed size. As a reminder, the model assumes statistical independence between the existence of species interactions and pivots on four parameters: the number of communities $B \in [1, \infty)$, noise regarding the existence of interactions outside species communities $\mu \in [0, 1]$, noise regarding interactions outside a perfectly nested structure $p \in [0, 1]$ and the shape parameter that defines the slimness of the nested structure $\xi \in [1, \infty]$.

We generated a set of 2×10^5 unipartite networks with varying parameters, covering the following ranges: $B \in [1, 9]$; $\xi \in [1.5, 7]$; $p \in [0, 0.6]$; and $\mu \in [0, 0.6]$. We restrict p and μ to 0.6 to guarantee that still some identifiable pattern is still present, e.g., maintaining the requirements of weak community structure as defined in [35], while avoiding spurious outcomes [105]. See Appendix A, Section A.2 for detailed information. We have assumed a fixed community size of $N_B = 50$. Thus, as we add communities, network size increases proportionally to the B parameter. The alternative process of fixing N and reducing the size of the communities as we increase B produces equivalent results, but difficult the analytical approach of the following sections. For modularity and in-block nestedness maximisation, we have used the extremal optimisation algorithm [29], adapted to the corresponding objective functions. The corresponding software codes, both for uni- and bipartite cases, can be downloaded from the web page of the group http://cosin3.rdi.uoc.edu/, under the Resources section.

Figure 3.1 and Figure 3.2 present the results over four ternary heat-map plots. This is a convenient diagram, since it allows the joint assessment of the mutual relationships between the three different structural patterns under consideration. Figure 3.1 (a) shows a density plot, showing the structural properties of the generated networks. The colour indicates the amount of networks in each bin of the ternary plot. As we see, the network generation model does not produce a sampling homogeneously distributed over

all the domain. It is apparent that the predominant architecture is modular. This is expected, since any parameter configuration with $B > 1$ (more than 95% of the generated benchmark) presents some sort of community organisation. Furthermore, modularity is the most favoured arrangement among the three under discussion: any departure from $B = 1$, and any departure from $p = 0$, decreases nestedness and in-block nestedness, but leaves Q unmodified. In other words, only parameter μ affects modularity in a negative way. However, this bias in the generation process does not affect our conclusions since we ensured that expected values for each hexagonal bin presented on the rest of the paper contains a with a minimum sample size of 20 networks. An alternative way to overcome this sampling problem in the non-homogeneous space may be to apply stratified sampling techniques.

Panels (b), (c), and (d) of the same figure reports the average value of $\mathcal{N}$, $\mathcal{I}$, and Q in each hexagonal bin of the ternary. Similarly, for panels (a)-(d) in Figure 3.2 each hexagonal bin of the ternary reports the average value of the parameters of the probabilistic network generation model. A preliminary visual analysis from Fig. 3.1 shows that the highest values of $\mathcal{N}$ and Q never overlap (red areas in panel (b) and (d)). In contrast, $\mathcal{I}$ is able to maintain high values for networks that are either modular or nested. These are valuable insights for the analytical results in the remainder of the Chapter: they numerically confirm that networks cannot acquire properties of nestedness and modularity simultaneously, whereas in-block nested networks might.

Another remarkable feature of results in Fig. 3.1–and Figure. 3.2– is the existence of sharp boundaries in the ternary plots, conveniently marked in dashed pink lines. The first boundary, F_1, results from the definition of $\mathcal{I}$, which generalises $\mathcal{N}$. By definition $\mathcal{I}$ reduces to $\mathcal{N}$ when $B = 1$. Translated to coordinates on the ternary, F_1 simply reflects that the contribution of $\mathcal{N}$ is always equal or smaller than the contribution of $\mathcal{I}$. Thus, this holds also in fractions $f_{\mathcal{N}} \leq f_{\mathcal{I}}$. More interesting, however, is the existence of F_2, which suggests that there is an inherent limit that constrains in-block nestedness to dominate over Q. On close inspection to Figure 3.2, networks which map onto F_2 exhibit high values of ξ, and very low values of p and μ, panels (b)-(d). We build on this observation to elaborate on our analytical exploration below.

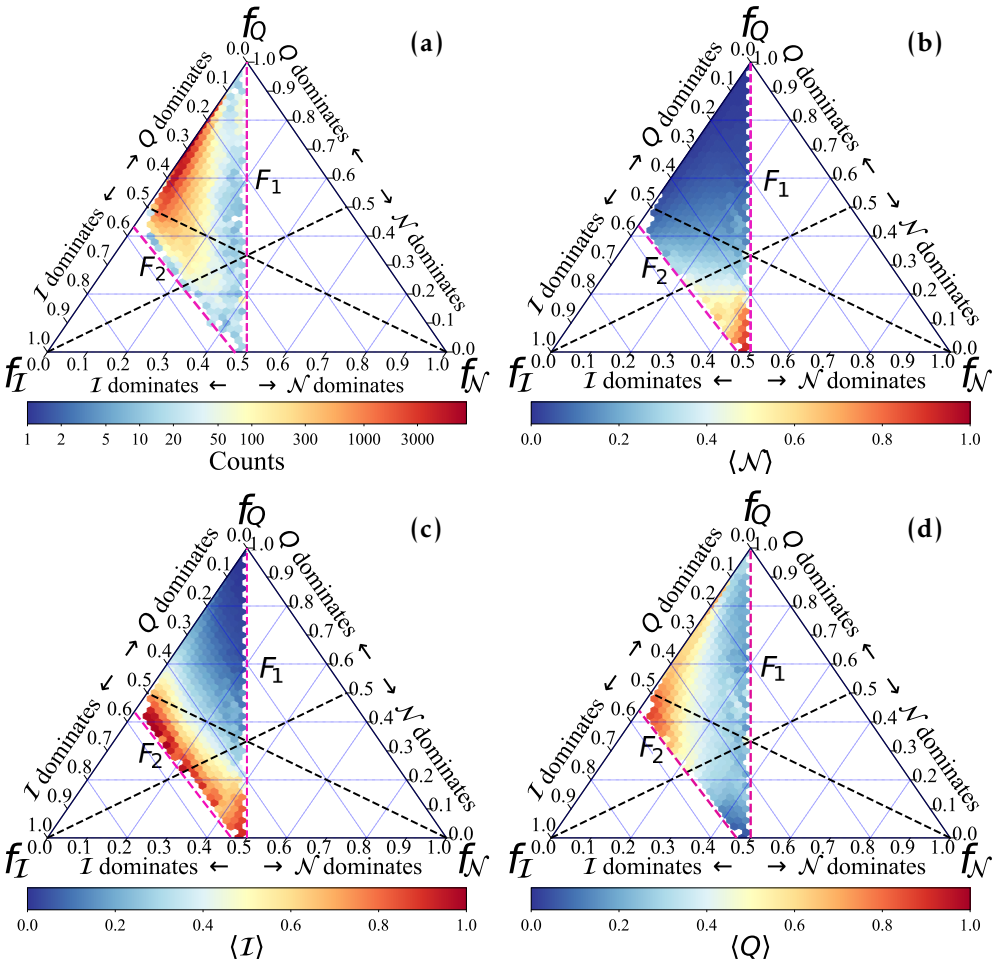


Figure 3.1: **Ternary plots showing the joint influence of the different structural patterns analysed within the paper**. Each axis corresponds to the fractional values of the three structural patterns, i.e. $f_{\mathcal{N}} = \frac{\mathcal{N}}{\mathcal{N}+Q+\mathcal{I}}$, $f_Q = \frac{Q}{\mathcal{N}+Q+\mathcal{I}}$ and $f_{\mathcal{I}} = \frac{\mathcal{I}}{\mathcal{N}+Q+\mathcal{I}}$. The bottom axis represents $\mathcal{N}$, and its right vertex corresponds to perfectly nested networks ($f_{\mathcal{N}} = 1$). Other values of $f_{\mathcal{N}}$ are indicated by the dashed blue lines in direction ↗ of the triangle. The right axis represents f_Q, and the top vertex thus corresponds to purely modular networks ($f_Q = 1$). Other f_Q values are indicated by horizontal dashed blue lines. Finally, the left axis represents $f_{\mathcal{I}}$, and the left vertex corresponds to networks that are purely in-block nested ($f_{\mathcal{I}} = 1$). Other $f_{\mathcal{I}}$ values are indicated by lines in direction ↘ of the triangle. Additionally, the black dashed lines delimit dominance regions. Each dominance region indicates (by pairs) which is the dominating structural pattern. For the sake of clarity, the dominant structure is also indicated close to the plot axis. For further details on the construction and interpretation of ternary plots, see Appendix A. Panel (a) shows the distribution of the generated networks over the ternary plot. The colour bar indicates the amounts of networks in each bin. Panels (b), (c) and (d) show the average values of $\mathcal{N}$, $\mathcal{I}$ and Q, respectively.

3.2 Structural analysis for a ring of star graphs

In this section we derive analytically the expressions for $\mathcal{N}$, $\mathcal{I}$ and Q at the boundary F_2 represented in Figs. 3.1 and 3.2. This is possible because, as mentioned, networks that map onto that boundary have common and very specific features: an extreme fill parameter, $\xi \to \infty$ (i.e. very sparse network), a perfectly nested intra-block structure ($p = 0$), and minimum inter-block connectivity ($\mu \approx 0$); see Fig. 3.2 panels (b)-(d). Such organisation corresponds exactly to a well-defined family of network configurations: a ring of star graphs, $G^\star$ hereafter, that reduces to a single star when $B = 1$, and resembles a set of stars connected with a single link through their central nodes when $B > 1$ (so as to guarantee a single giant component), see Fig. 3.3.

The exact expressions for $\mathcal{N}$, $\mathcal{I}$ and Q for $G^\star$ are the key to understand the mutual constraints that the different network arrangements impose to each other, strictly for such idealised case, and more loosely in general. In the following, we consider a ring of star graphs with B communities and N_B nodes per community–N_{B_c} andN_{B_r} for bipartite networks–. For such given graph, we find the exact values $\mathcal{N}_{G^\star}$, $\mathcal{I}_{G^\star}$ and $Q_{G^\star}$.

3.2.1 Nestedness.

We derive the analytical expression for $\mathcal{N}_{G^\star}$ from the expression in Eq. 1.9 and its unipartite counterpart. We start deriving the corresponding expression for the unipartite case. Here, the pair overlap of a generalist node (the centre of each star subgraph), g, with a specialist node (periphery of a star), s, is $O_{gs}/k_s = 1$ if g and s belong to the same star (and 0 otherwise). For all those pairs (regardless of the star they belong to), the null model contribution is $\langle O_{gs}/k_s \rangle = (N_B + 1)/BN_B$. We can obtain in a similar way the terms for the the generalist-generalist pairs between stars. Summing up all the contributions, the final expression for $\mathcal{N}_{G^\star}$ is:

$$\mathcal{N}_{G^\star} = \frac{BN_B^3 - BN_B^2 - 3BN_B + B + 2N_B + 2}{BN_B\left(BN_B^2 + BN_B - N_B - 1\right)}. \tag{3.1}$$

The corresponding expression for bipartite networks can be obtained following a similar logic, but taking into consideration the contributions of rows and columns separately (N_{B_r} and N_{B_c}) to obtain:

$$\mathcal{N}_{G^\star} = \frac{\mathcal{X}^c + \mathcal{X}^r}{BN_{B_c}N_{B_r}(N_{B_c}+2)(N_{B_c}+N_{B_r})(N_{B_r}+2)(BN_{B_c}-1)(BN_{B_r}-1)}, \tag{3.2}$$

where the first contributor term in the numerator $\mathcal{X}^c$, takes the form:
$\mathcal{X}^c = N_{B_r}(N_{B_r}+2)(BN_{B_c}-1)\Big\{B\Big[(N_{B_r}^2+N_{B_r}-3)N_{B_c}^2+N_c(N_{B_r}^2+1)-2N_{B_r}(N_{B_r}+2)+2\Big]+(N_{B_c}+2)(N_{B_c}+N_{B_r}+1)\Big\}$. The second contributor term in the numerator $\mathcal{X}^r$, takes an equivalent form, but one needs to interchange the terms N_{B_r} and N_{B_c}, accordingly.

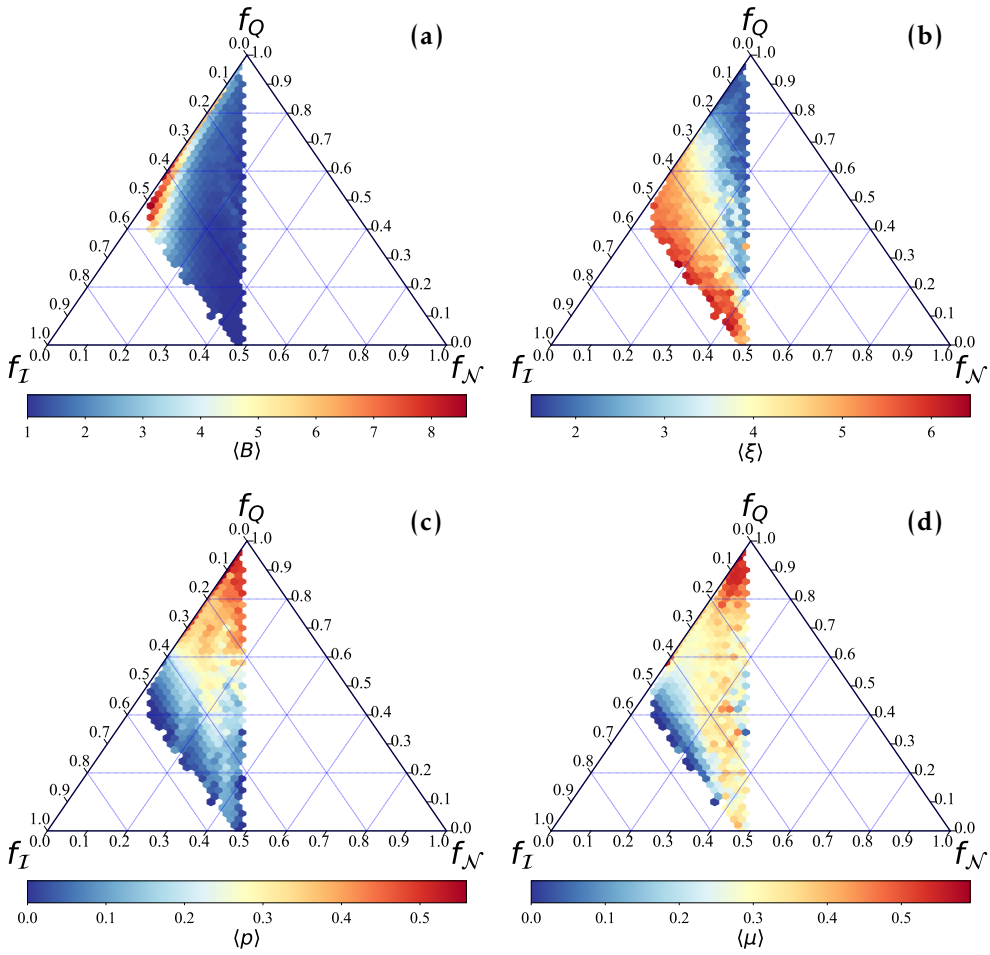


Figure 3.2: **Ternary plots showing the effect of the parameters of the probabilistic network generation model**. In this case, color in each bin of the simplex indicates the average number of blocks B (a); average shape parameter ξ (b); average intra-block noise p (c); and finally average inter-block noise μ (d).

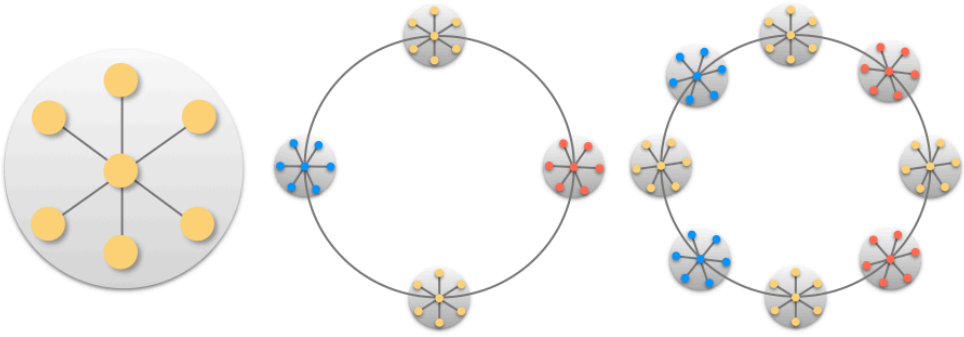


Figure 3.3: **Design of a ring of star graphs**. The star graphs are connected through their central nodes.

3.2.2 Modularity.

In general the optimal partition for an arbitrary network cannot be easily obtained, except for very idealised cases such as $G^\star$, where each star in the ring forms a community (note that $G^\star$ does not suffer, like a ring of cliques, from the well-known Q's resolution limit [102]). On such setting, we can easily derive the contribution of each star to the total Q following Eqs. 1.13 and 1.16. In unipartite settings, the total number of links within communities is $l_c = N_B - 1$ and the amount of links of the network, including links within and between communities, is $L = B(N_B - 1) + B = BN_B$. The last term, the sum of the degrees of all the nodes in community c, corresponds to $d_c = 2N_B$. Assembling these, we obtain the modularity of $G^\star$ as

$$Q_{G^\star} = B\left[\frac{N_B - 1}{BN_B} - \left(\frac{2N_B}{2BN_B}\right)^2\right] = 1 - \frac{1}{N_B} - \frac{1}{B}, \tag{3.3}$$

which is equivalent to the general expression derived in [102].

In the bipartite counterpart, the total number of links in the networks is equal to $B(N_{B_r} + N_{B_c} + 1)$, and the number of links per community is equal to $N_{B_r} + N_{B_c} - 1$. Putting all this together, we have the maximum bipartite modularity expressed as

$$Q_{G^\star} = 1 - \frac{2}{N_{B_r} + N_{B_c} + 1} - \frac{1}{B}. \tag{3.4}$$

3.2.3 In-block nestedness.

The derivation of $\mathcal{I}_{G^\star}$ resembles that of $\mathcal{N}_{G^\star}$, with the difference that only nodes within the same community contribute; thus, all stars have the same contribution. Focusing now on each star, we have only two contributing terms to the sum: the pair overlap between specialist nodes, s, and the pair overlap of the generalist node, g, with the specialists. In both cases, the contribution is 1. For unipartite settings, the null model corrections are $\langle O_{gs}\rangle = k_g k_s / BN_B = (N_B + 1)/BN_B$ and $\langle O_{ss}\rangle = k_s k_s / BN_B = 1/BN_B$. Finally, the size of the communities is $C_g = C_s = N_B$. After taking Eq. 1.20 in its unipartite form, and replacing all the contributions above, we obtain

$$\mathcal{I}_{G^\star} = 1 - \frac{3}{BN_B} - \frac{2}{N_B}. \tag{3.5}$$

For the bipartite networks (Eq. 1.20), the null modell corrections are equal to $\langle O_{g,s}\rangle = k_g k_s / BN_{r,c} = (N_{B_c} + 2)/BN_{B_c} = (N_{B_r} + 2)/N_{B_r}$ and $\langle O_{s,s}\rangle = k_s k_s / BN_{r,c} = 1/BN_{B_c} = 1/BN_{B_r}$, respectively. Finally, replacing all the contributions for the bipartite case, we obtain

$$\mathcal{I}_{G^\star} = 1 - \frac{1}{BN_{B_c}} - \frac{1}{BN_{B_r}} - \frac{2}{B(N_{B_r} + N_{B_c})} - \frac{2}{BN_{N_r} N_{B_c}}. \tag{3.6}$$

In all the expressions presented above, we consider a closed ring, on which the number of inter-community links is B. For the cases $B = 1$ and $B = 2$, the number of inter-community links is $B - 1$ and the degree of the generalist nodes have to be modified accordingly. Thus, the expressions for these particular settings demand a specific treatment, see Appendix A, Section A.3 for details on these cases.

3.3 Exact constraints between $\mathcal{N}_{G^\star}$ and $Q_{G^\star}$

The interdependences of Eqs. 3.1-3.6 become apparent when the number of blocks, B, or the size of the blocks, N_B, are large. For the case where $N_B \to \infty$, in both unipartite and bipartite cases, Eqs. 3.1-3.6 reduce to

$$\lim_{N_B\to\infty} \mathcal{N}_{G^\star} = \frac{1}{B}, \qquad \lim_{N_B\to\infty} Q_{G^\star} = 1 - \frac{1}{B} = 1 - \mathcal{N}_{G^\star}, \qquad \lim_{N_B\to\infty} \mathcal{I}_{G^\star} = 1, \tag{3.7}$$

As we see, for large $G^\star$ networks, nestedness and modularity are complementary – corroborating the empirical observations in [49]. This result shows analytically that, in large systems with low fill, the ecosystem needs to choose (with the dynamical consequences it may bear) between maximising community structure, or maximising nested arrangements, but not both at the same time.

With respect to the case $B \to \infty$, Eqs. 3.1-3.6, in the unipartite case reduce to

$$\lim_{B\to\infty} \mathcal{N}_{G^\star} = 0, \qquad \lim_{B\to\infty} Q_{G^\star} = \frac{N_B - 1}{N_B} \approx 1, \qquad \lim_{B\to\infty} \mathcal{I}_{G^\star} = \frac{N_B - 2}{N_B} \approx 1, \tag{3.8}$$

as for the bipartite case they reduce to

$$\lim_{B\to\infty} \mathcal{N}_{G^\star} = 0, \qquad \lim_{B\to\infty} Q_{G^\star} = \frac{N_{B_r} + N_{B_c} - 1}{N_{B_r} + N_{B_c} + 1} \approx 1 \qquad \lim_{B\to\infty} \mathcal{I}_{G^\star} = 1. \tag{3.9}$$

In this case, the existence of many communities implies the impossibility to develop a purely nested pattern. Indeed, the mutual bounds that $\mathcal{N}_{G^\star}$ and $Q_{G^\star}$ impose on each other are evident, displaying a perfect anti-correlated behaviour. A plausible way to preserve both nested arrangements and community structure is under the form of in-block nestedness, which yields to the maximum possible value in both limits. Importantly, this suggests that $\mathcal{I}$ doesn't show any incompatibility with either $\mathcal{N}$ or Q.

Figure 3.4(a) illustrates these results, comparing the analytical estimation of $\mathcal{N}_{G^\star}$, $Q_{G^\star}$ and $\mathcal{I}_{G^\star}$ (Eqs. 3.1,3.3 and 3.5) against B, and the numerical results for networks generated from different parameters. As the generated networks deviate from the ring of stars $G^\star$ (i.e. $p > 0$ and $\mu > 0$), results show a worse fit to the analytical prediction, but the overall anti-correlated pattern clearly remains. Finally, we observe that as networks transition from a nested ($B = 1$) to a modular ($B > 1$) architecture, the values of in-block nestedness remain very high (close to one as Eqs. 3.7, 3.8 and 3.8 indicate) and almost constant. Fig. 3.4(b) tests the same evolution for a much denser network (50% of matrix fill when $B = 1$, clearly far above most real networks). The anti-correlated behaviour of $\mathcal{N}$ and Q is preserved, but the effects of the null model term are notable: the maximum value that nestedness can take (at $B = 1$) is $\mathcal{N} \approx 0.3$.

3.4 Approximate constraints $\mathcal{N}$ and Q (general case)

Results in Section 3.3 obtained for idealised settings ($G^\star$) point at a more general question: can the exact constraints in Eqs. 3.7, 3.8 and 3.8 be used to understand the co-occurrence of macro- and mesoscale patterns for the general case? Can we exploit

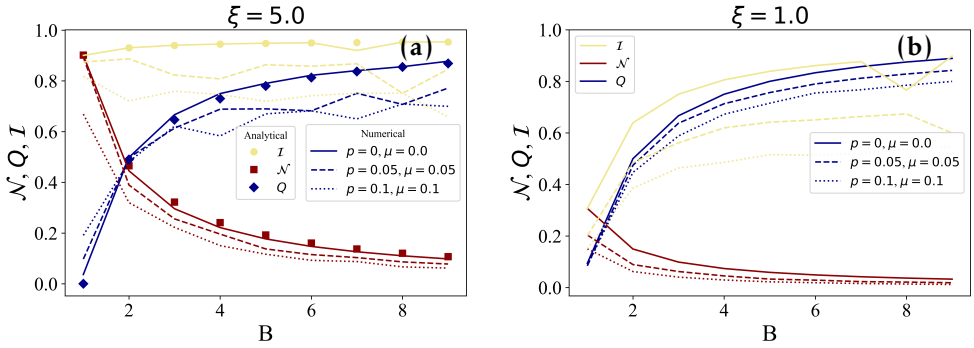


Figure 3.4: **Analytical and numerical agreement.** (a) Comparison of the analytical (symbols; Eqs. 3.1-3.5) and numerical (lines) values of $\mathcal{N}$, Q, $\mathcal{I}$ with respect to B. All the calculations were performed by taking $N_B = 50$ and $\xi = 5$. The values for p and μ parameters, which increasingly depart from the ideal configuration $G^\star$, are indicated in the plot legend. (b) Same exercise as panel (a), for a very dense network ($\xi = 1$). The overall behaviour of $\mathcal{N}$ and Q is preserved, with monotonic decrease and increase, respectively. $\mathcal{I}$ is closer to $\mathcal{N}$ initially, but quickly converges to Q thereon. Notably, the analytical $\mathcal{N}$-Q antagonism does not hold anymore for low B, as these networks deviate strongly from $G^\star$.

the complementarity between $\mathcal{N}$ and Q beyond the strict conditions of $G^\star$? This and next Section target these questions, proposing soft bounds for Q (and for $\mathcal{I}$) in terms of $\mathcal{N}$ when networks deviate from idealised scenarios. We stress the importance of this attempt since $\mathcal{N}$ can be obtained for any network in polynomial time, $\mathcal{O}(N_T^3)$, while the maximization of Q and $\mathcal{I}$ are NP problems. In this situation, these bounds offer a valuable *a* priori intuition of the mesoscale organization of a network. The derivation of these bounds for an arbitrary network G is presented below.

3.4.1 Upper bound.

The calculation of $\mathcal{N}$ is computationally cheap even for very large networks. Thus, given $\mathcal{N}$ for a graph G, to obtain the maximum Q value compatible with such level of nestedness, we assume G can be approximated to $G^\star$ with the same number of nodes, N_T, and nestedness $\mathcal{N}$. That is, G is assumed to have a relatively large ξ. The rationale behind this mapping (G to $G^\star$) responds to the fact that, for any network with the given $\mathcal{N}$, the largest possible modularity value corresponds to a network lying on F_2, i.e. the $G^\star$ graph, see Fig. 3.1(d). With this approximation, the upper bound reduces to computing $Q_{G^\star}$ (Eq. 3.3) and $\mathcal{I}_{G^\star}$ (Eq. 3.5) for a $G^\star$ network compatible with the observed values of $\mathcal{N}_{G^\star}$ (Eqs. 3.1). To attain these, the only missing information is the number of communities, B , which, for the case of $G^\star$ with equally-sized modules (that is, $N_T = BN_B$), can be obtained exploiting Eq. 3.1:

$$\mathcal{N}_{G^\star}(N_T, B) = \frac{(B^3 + B^2(2 - 3N_T) - B(N_T - 2)N_T + N_T^3)}{B(N_T - 1)N_T(B + N_T)}. \tag{3.10}$$

This polynomial equation has three possible roots, two of them being in the imaginary domain. The upper bounds for Q and $\mathcal{I}$ are thus readily available, applying B to Eqs. 3.3 and 3.5. We remark that this is a heuristic approximation to Q upper bound. The consideration of a more nuanced estimation, which should consider the density of G and a non-homogeneous communities, is beyond the purpose of this work. Additionally, we can obtain the fractional contributions of Q, $\mathcal{I}$ and $\mathcal{N}$ over the F_2 boundary, $f_Q^{F_2}$, $f_{\mathcal{I}}^{F_2}$, $f_{\mathcal{N}}^{F_2}$ (see Fig. 3.1). In particular, $f_Q^{F_2}$ will be required to estimate the lower bound.

3.4.2 Lower bound.

We now turn our attention to the minimum value that Q can attain, which is obtained at the boundary F_1, see Fig. 3.1(d). Heuristically, this makes sense because networks which belong to the region along F_1 are those with $B = 1$, see Fig. 3.2(a). To obtain the lower bounds for Q we require to assume that $\mathcal{N}$ values are approximately constant with respect to the contributions f_Q. This is not a strict fact, but an observation from Fig. 3.1(b). Additionally, at the boundary F_1, we know that $\mathcal{N} = \mathcal{I}$. Thus, with the actual measure of $\mathcal{N}$, and $f_Q^{F_2}$ obtained though the upper bound estimation, we can obtain Q as

$$f_Q^{F_1} \approx f_Q^{F_2} = \frac{Q}{Q + \mathcal{I} + \mathcal{N}} = \frac{Q}{Q + 2\mathcal{N}}. \tag{3.11}$$

The lower bound for $\mathcal{I}$ doesn't need a heuristic estimation because, as mentioned, its definition implies a hard lower limit when $B = 1$, i.e. $\mathcal{I} = \mathcal{N}$.

Figure 3.5(a) shows the values of Q as a function of $\mathcal{N}$ for the previous synthetic ensemble ($\sim 2 \times 10^5$ networks). Q values, as obtained with the optimisation algorithm, are plotted in grey or yellow, and the values of the theoretical upper and lower bounds are plotted in black. The red line indicates the average values of Q for a fixed $\mathcal{N}$, and the overlaid error area represents one standard deviation above and below that average. Our approximation of Q bounds is in good agreement with actual values obtained after optimisation: most of the optimised Q values lie within the estimated soft bounds. Despite the wide range of parameters ξ, B, p and μ –far from limiting cases in most cases–, estimated upper bounds behave like $Q = 1 - \mathcal{N}$ almost perfectly, in accordance with our analytical insights. While these bounds are trivial when $\mathcal{N} \approx 0$, we observe that, for intermediate-to-high values of nestedness, these provide relevant information about the possible mesoscale organisation of the network.

Modularity values, Q, above the upper bound correspond to networks with a single community $B = 1$ and perfectly nested structure, $p = 0$ (see App. A, Fig. A.3). These networks, coloured as yellow in Fig. 3.5(a) and less than 0.1% of the total, are dense enough to allow a partition with $B > 1$ where the nodes of higher degree are gathered in a block, resulting in values of Q larger than expected [53]. The small fraction of yellow values below the lower bound approximation also corresponds to networks with $B = 1$,

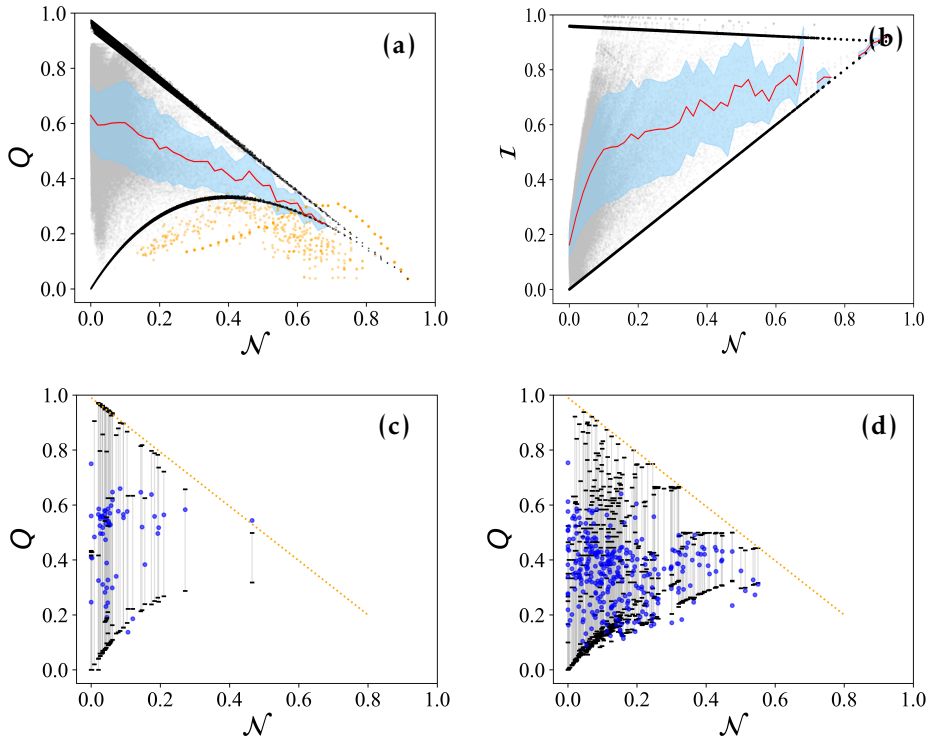


Figure 3.5: **Estimated upper and lower bounds for modularity.** Panel (a) shows the values of Q obtained after optimisation (grey and yellow dots), plotted against $\mathcal{N}$ for over 2×10^5 generated networks. Yellow points correspond to networks with $B = 1$ (network with a single block), for which modularity optimisation algorithms detect more than one block, although only one planted block exists in the network. The red line in panel (a) indicates the average values of Q for a given $\mathcal{N}$, the error shaded area represents one standard deviation above and below that average; networks with $B = 1$ are excluded in this computation. Panel (b) shows the values of $\mathcal{I}$ obtained after optimization (grey dots), plotted against $\mathcal{N}$ for the generated networks. Panel (c) shows results obtained for the set of 57 unipartite social networks analysed in [53]; and panel (d) for the set of bipartite social and ecological networks [117, 118]. In these two panels, dark blue dots represent the real Q value after optimisation, and bars represent the corresponding estimated upper and lower bounds for the same network. In all scenarios, the upper and lower bounds of Q are marked by black dots.

but with different (ξ, p) parameters. In the same spirit, upper and lower bounds for $\mathcal{I}$ can be as well approximated from the actual value of $\mathcal{N}$, see Fig. 3.5(b). Remarkably, none of the optimized values of $\mathcal{I}$ violates such bounds. There is no surprise with respect to lower bounds, since the lower bound simply represents the hard limit $\mathcal{I} = \mathcal{N}$. But even the upper bounds, which represent an estimation, are in excellent agreement with respect to the optimized values of $\mathcal{I}$. For the sake of completeness, Q-$\mathcal{I}$ scatter plots are shown in Fig. A.4 of Appendix A, where we reconfirm that $\mathcal{I}$ and Q can coexist, i.e. there is no clear map or mutually imposed constraints from one to the other.

The corresponding software codes to obtain the upper and lower bounds for Q and $\mathcal{I}$ are included in the package that can be downloaded from the web page of the group (http://cosin3.rdi.uoc.edu/), under the Resources section.

3.5 Application to real networks

For the conducted experiments with synthetic networks, we have seen that $\mathcal{N}$ provides informative bounds to the mesoscale organisation. However, real networks differ from idealised synthetic networks, e.g. the assumption of homogeneous sizes of communities, or uncorrelated noise. To assess the accuracy that our development has in real scenarios, we perform experiments on 347 real networks, covering several domains: 57 real unipartite networks [53] (mostly social and economic networks), and 290 bipartite networks (ecological in most cases [117], with some social networks [118] as well).

Remarkably, for these real networks, see Figure 3.5(c) and (d), our bound estimations also hold quite accurately. In general, we observe that for both uni- and bipartite real networks, the limits and bounds for the bipartite case behave as expected. In general, the larger $\mathcal{N}$, the tighter are the bounds of Q, and smaller the maximum value of Q. In 45 of these networks the bounds fail: the obtained modularity is either above or below than the upper or lower bound respectively. To ease visualization, Fig. 3.6 presents the same results, sorted by the difference between the upper and lower bounds. Results show that in general we have a good estimation of the bounds. For the bipartite networks results are clearer, since higher values of nestedness produce tighter bounds. For the unipartite case, the low values of nestedness of these networks derive wider bounds.

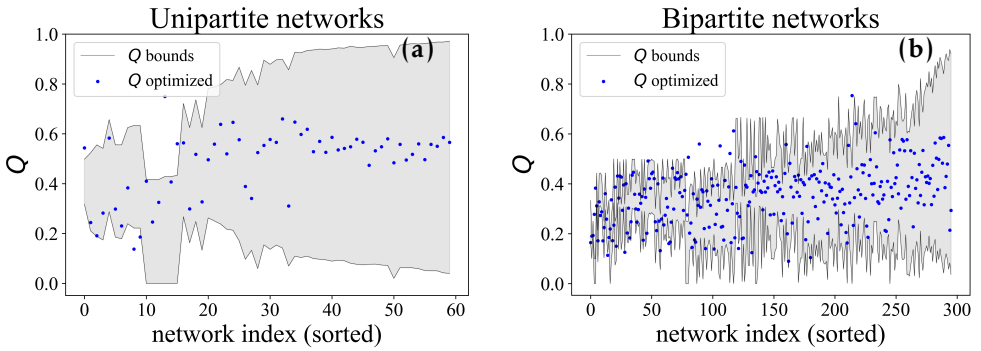


Figure 3.6: **Assessment of the accuracy of the bounds for real networks.** Panel (a) corresponds to 57 unipartite networks (see Fig. 3.5(b)), and panel (b) to 290 bipartite networks (see Fig. 3.5(c)). In both cases networks are sorted by the difference between the upper and lower bound.

3.6 Summary

In this chapter, we have explored the relationship between three different organizational patterns: two at the mesoscale (modularity and in-block nestedness); and one at the macroscale (nestedness). We have quantified numerically and analytically, the

interference between these three different structural organizations, in both uni- and bipartite settings. We show that modularity and nestedness are antagonistic architectures, the growth of one implies the decline of the other, and this antagonism can be used to estimate mutual bounds in synthetic and real settings. The need to preserve ingredients from both nested and modular arrangements points at the possibility of intermediate structures, which are indeed plausible with in-block nested structures. Our results stand as a theoretical and numerical step forward to better understand past empirical evidence, which pointed at the harsh (but not impossible) coexistence of nestedness and modularity, in Ecology and elsewhere. Notwithstanding, it is worth highlighting that our approach takes into account solely the structural aspect of the problem, without considering a plausible dynamic co-emergence of both patterns [52, 125], which we foresee as the next relevant problem.

Part III

Empirical applications and implications to systems' dynamics

For this part of the thesis, instead of adopting a purely structural approach, i.e., solely employing an analytical and numerical perspective, towards the study of nestedness, modularity, and in-block nestedness. Our goal here is to further explore which are some of the mechanisms that facilitate the emergence of in-block nested arrangements. To this aim, we will provide a combination of empirical work and theoretical modeling with a focus on online social and socio-technological environments.

Chapter 4

Online division of labour: emergent structures in Open Source Software

We start part III of the thesis with a purely empirical study. Specifically, we will focus on the analysis of a set of open source software (OSS) projects from public the repository platform Github. Our interest on the structural features of OSS projects departs from some obvious, but worth highlighting, observations. In first place, open Source Software projects strongly depend on the participation and commitment of volunteer developers to progress on a particular task, yet, little is known on how these diverse groups of developers self-organise to work together. Second, public repositories provide a virtually unlimited development framework: any number of actors can potentially join to contribute in a self-organized, decentralised, distributed, remote, and asynchronous manner. However, it seems reasonable that some sort of hierarchy and division of labour must be in place to meet human biological and cognitive limits, and also to achieve some level of efficiency. Based on past evidence, we think that these latter features (hierarchy and division of labour) should translate into detectable structural arrangements, such like nestedness, modularity and in-block nestedness, when projects are represented as developer-file bipartite networks.

4.1 Background in Open Source Software

Open Source Software (OSS) is a key actor in the current software market, and a major factor in the consistent growth of the software economy. The promise of OSS is better quality, higher reliability, more flexibility, lower cost, and an end to predatory vendor lock-in, according to the Open Source initiative [126]. These goals are achieved thanks to the active participation of the community [127]: indeed, OSS projects depend on the participation and commitment of volunteer developers to progress [128, 129].

The emergence of GitHub and other platforms as prominent public repositories, to-

gether with the availability of APIs to access comprehensive datasets on most projects' history, has opened up the opportunities for more systematic and inclusive analyses of how OSS communities operate. In the last years, research on OSS has left behind a rich trace of facts. For example, we now know that the majority of code contributions are highly skewed towards a small subset of projects [130, 131], with many projects quickly losing community interest and being abandoned at very early stages [132]. Moreover, most projects have a low *truck factor*, meaning that a small group of developers is responsible for a large set of code contributions [133–135]. This pushes projects to depend more and more on their ability to attract and retain occasional contributors (also known as "drive-by" commits [136]) that can complement the few core developers and help them to move the project forward. Along these lines, several works have focused on strategies to increase the on-boarding and engagement of such contributors (e.g., by using simple contribution processes [137], extensive documentation [138], gamification techniques [139] or *ad hoc* on-boarding portals [140], among others [141]). Other social, economic, and geographical factors affecting the development of OSS have been scrutinised as well, see Cosentino *et al.* [142] for a thorough review.

Parallel to these macroscopic observations and statistical analyses, social scientists and complex networks researchers have focused, in relatively much fewer papers, on analysing how a diverse group of (distributed) contributors work together, i.e. the structural features of projects. Most often, these works pivot on the interactions between developers, building explicit or implicit collaborative networks, e.g. email exchanges[143, 144] and unipartite projections from the contributors-files bipartite networks [145], respectively. These developer social networks have been analysed to better understand the hierarchies that emerge among contributors, as well as to identify topical clusters, i.e. cohesive subgroups that manifest strongly in technical discussions. However, the behaviour of OSS communities cannot be fully understood only accounting for the relations between project contributors, since their interactions are mostly mediated through the edition of project files (no direct communication is present between group members). To overcome this limitation, here we focus on studying the structural organisation of OSS projects as contributor-file bipartite graphs. On top of technical and methodological adaptations, the consideration of these two elements composing the OSS system allows retaining valuable information (as opposed to collapsing it on a unipartite network) and, above all, recognising both classes as co-evolutionary units that place mutual constraints on each other.

As it was briefly mentioned above, public collaborative repositories place no limits, in principle, to the number of developers (and files) that a project should host. In this sense, platforms like GitHub resemble online social networks (e.g. Twitter or Facebook), in which the number of allowed connections is virtually unbounded. However, we know that other factors –biological, cognitive– set well-defined limits to the amount of active social connections an individual can have [146], also online [147]. But, do these limits

apply in collaborative networks, where contributors work remotely and asynchronously? Does a division of labour arise, even when interaction among developers is mostly indirect (that is, via the files that they edit in common)? And, even if specialised subgroups emerge (as some evidence already suggests, at least in developer social networks [145]), do these exhibit some sort of internal organisation?

We aim to answer these questions, by placing the accent on the analysis on of nestedness modularity and in-block nestedness as key structural signatures. The first one, nestedness, is a suitable measure to quantify and visualise how the mentioned low truck factor, and the existence of core/drive-by developers [115], translates into a project's network structure. As for modularity, it provides a natural way to check whether OSS projects split in identifiable compartments, suggesting specialisation, and whether such compartments are subject to size limitations, along the mentioned bio-cognitive limits. Finally, since modularity and nestedness are, to some extent, incompatible in the same network –as shown in Chapter 3–, in-block nestedness (or the lack of it) can help to determine how projects solve the tension between the emergence of nested (hierarchy, asymmetry) and modular (specialisation, division of labour, bounds to social connections) patterns.

4.2 Data and methods

Our open source projects dataset was collected from GitHub [123], a social coding platform which provides source code management and collaboration features such as bug tracking, feature requests, tasks management and wiki for every project. Given that GitHub users can star a project (to show interest in its development and follow its advances), we chose to measure the popularity of a GitHub project in terms of its number of stars (i.e. the more stars the more popular the project is considered) and selected the 100 most popular projects. This criterium mainly responds to two arguments: maturity and success. That is, here we purposefully pay attention to projects which have reached a reasonable degree of evolution, regardless of the absence (or presence) of any given structural organisation at the initial stages. Other possible criteria –number of forks, open issues, watchers, commits and branches– are positively correlated with stars [142], and so our proxy to mature, successful and active projects probably overlaps with other sampling procedures.

4.2.1 Collection and pruning

The collection and cleaning of the dataset involved three phases, namely: (1) cloning, (2) import, and (3) enrichment.

1. **Cloning and import.** After collecting the list of 100 most popular projects in GitHub (at the moment of collecting the data) via its API [148], we cloned them to collect 100 Git repositories. We analysed the cloned repositories and discarded those ones not involving the development of a software artifact (e.g. collection of

links or questions), rejecting 15 projects out of the initial 100. We then imported the remaining Git repositories into a relational database using the Gitana [149] tool to facilitate the query and exploration of the projects for further analysis. In the Gitana database, Git repositories are represented in terms of users (i.e. contributors with a name and an email); files; commits (i.e. changes performed to the files); references (i.e. branches and tags); and file modifications. For two projects, the import process failed to complete due missing or corrupted information in the source GitHub repository.

2. **Enrichment.** Our analysis needs a clear identification of the author of each commit so that we can properly link contributors and files they have modified. Unfortunately, Git does not control the name and email contributors indicate when pushing commits resulting on clashing and duplicate problems in the data. Clashing appears when two or more contributors have set the same name value (in Git the contributor name is manually configured), resulting in commits actually coming from different contributors appearing with the same commit name (e.g., often when using common names such as "mike"). In addition, duplicity appears when a contributor has several emails, thus there are commits that come from the same person, but are linked to different emails suggesting different contributors. We found that, on average, around 60% of the commits in each project were modified by contributors that involved a clashing/duplicity problem (and affecting a similar number of files). To address this problem, we relied on data provided by GitHub for each project (in particular, GitHub usernames, which are unique). By linking commits to unique usernames, we could disambiguate the contributors behind the commits. Thus, we enriched our repository data by querying GitHub API to discover the actual username for each commit in our repository, and relied on those instead on the information provided as part of the Git commit metadata. This method only failed for commits without a GitHub username associated (e.g. when the user that made that commit was no longer existing in GitHub). In those cases we stick to the email in Git commit as contributor identifier. We reduced considerably the clashing/duplicity problem in our dataset. The percentage of commits modified by contributors that may involve a clashing/duplicity problem was reduced to 0.004% on average ($\sigma = 0.011$), and the percentage of files affected was reduced to 0.020% ($\sigma = 0.042$).

At the end of this process, we had successfully collected a total number of 83 projects, adding up to 48,015 contributors, 668,283 files and 912,766 commits. 18 more projects were rejected due to other limitations. On one hand, after exploring the relationship between the number of files and contributors within the projects –Pearson coefficient $r = 0.34$–, we discarded some projects that presented very strong divergence between the two sets, e.g. projects with a very large number of files but very few contributors. In

these cases, although $\mathcal{N}$, Q, and $\mathcal{I}$ can be quantified, the outcome is hardly interpretable. An example of this is the project *material-designs-icons*, with 15 contributors involved in the development of 12,651 files. Finally, we considered only projects with a bipartite network size within the range $10^1 \leq S \leq 10^4$, as the computational costs to optimise in-block nestedness and modularity for larger sizes were too severe. After this, we ended up retaining 65 projects to perform the structural analysis. Nonetheless, as can be seen in Table 5.1, we have a sufficiently broad distribution of project sizes and age. Note also that popularity (number of stars) is not necessarily related to their size (Pearson coefficient $r = -0.03$) nor age ($r = -0.1$). The complete dataset with the final 65 projects is available at http://cosin3.rdi.uoc.edu, under the Resources section.

	contributors	files	commits	stars	Project age
Largest project	1,061	12,321	75,757	27,500	4 years 11 months
Smallest project	55	27	444	36,900	5 years 6 months
Average	422	3,247	33,936	46,334	4 years 9 months
Most popular project	516	2,833	34,666	293,000	2 years 10 months
Least popular project	117	103	4,057	21,700	5 years 5 months
Oldest project	1,434	10,413	174,452	35,000	11 years 3 months
Youngest project	43	51	210	31,600	0 years 3 months

Table 4.1: Statistics of our dataset.

4.2.2 Matrix generation.

We build a bipartite unweighted network as a rectangular $N \times M$ matrix, where rows and columns refer to contributors and source files of an OSS project, respectively, and total size $S = N + M$. Cells therefore represent links in the bipartite network, i.e. if the cell a_{ij} has a value of 1, it represents that the contributor i has modified the file j at least once, otherwise a_{ij} is set to 0.

We are aware that an unweighted scheme may be discarding important information, i.e. the heterogeneity of time and effort that developers devote to files. We stress that including weights in our analysis can introduce ambiguities in our results. In the Github environment, the size of a contribution could be regarded either as the number of times a developer commits to a file, or as the number of lines of code (LOC) that a developer modified when updating the file. Indeed, both could represent additional dimensions to our study. Furthermore, at least for the first (number of commits), it is readily available from the data collection methods. However, weighting the links of the network by the number of commits is risky. Consider for example a contributor who, after hours or days of coding and testing, performs a commit that substantially changes a file in a project. On the other side, consider a contributor who is simply documenting some code, thus committing many times small comments to an existing software –without changing the internal logic of it. There is no simple way to distinguish these cases.

The consideration of the second item (number of LOC modified) could be a proxy to

such distinction, but this is information is not realistically accessible given the current limitations to data collection. Getting a precise number of LOCs requires a deeper analysis of the Git repository associated to the GitHub project, parsing the commit change information one by one –an unfeasible task if we aim at analysing a large set of projects. The same scalability issue would appear if we rely on the GitHub API to get this information, which additionally would involve quota problems with such API.

One might consider even a third argument: not every programming language "weighs" contributions in the same way. Many lines of HTML code may have a small effect on the actual advancement of a project, while two brief lines in C may completely change a whole algorithm. In conclusion, we believe there is no generic solution that allows to assess the importance of a LOC variation in a contribution. This will depend first on the kind of file, then on the programming style of each project and finally on an individual analysis of each change. Thus, adding informative and reliable weights to the network is semantically unclear (how should we interpret those weights?) and operationally out of reach.

4.3 Preliminary observations: Developers implicit degree

Before we focus on the structural arrangements of interest (nestedness, modularity, in-block nestedness), we explore whether a potentially unbounded interaction capability is mirrored in actual OSS projects across 4 orders of magnitude in size. To do so, we work on the projected contributor-contributor network, to measure the developer's implicit average degree $\langle k \rangle$, i.e. the average amount of contributors with whom an individual shares at least one file. Figure 4.1(a) shows a scatter plot of $\langle k \rangle$ against S (note the semi-log scaling). Panels (b) and (c) in Fig. 4.1 shows the scatter plots of $\langle k \rangle$ against N and M, respectively. Despite the changes in the $x-$axis scale (which affects the order in which projects are represented), there are no significant differences in the results. Such results indicate that, besides the initial fluctuating pattern, $\langle k \rangle$ presents an almost flat trajectory suggesting that, on average, a contributor indirectly interacts with ~ 70 peers, regardless of the size of the project. As visual aid, we have added a vertical red line in panel (b) at $N = 70$, to differentiate those networks with $N < 70$ and for which it is not possible to exhibit $\langle k \rangle \approx 70$. The stable behaviour of this average was statistically validated with the Augmented Dickey-Fuller (ADF) test for stationarity of time series [150]. The idea of stationarity on a time series implies that summary statistics of the data, like the mean or variance, are approximately constant when measured from any two starting points in series (different project sizes in our case). Typically, statistical stationarity tests are done by checking for the presence (or absence) of a unit root on the time series[1] (null hypothesis). The results of the analysis indicate that, since the test statistic is less than the critical value at 5% significance level, then, the null hypothesis is rejected, and we

[1] A time series is said to have a unit root if we can write it as $y_t = a^n y_{t-n} + \sum_i \epsilon_{t-i} a^i$, where $a = 1$ and ϵ is an error term.

can conclude that the data series is stationary.

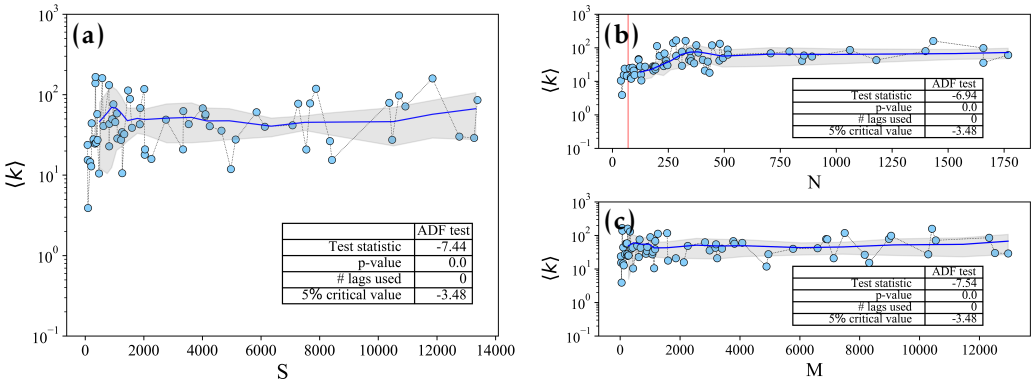


Figure 4.1: **Contributor-contributor network.** Scatter plots of the developers implicit average degree $\langle k \rangle$ against project size $S = N + M$ (panel (a)), number of contributors (panel (b)) and number of files (panel (c)). The shadowed grey area represents one standard deviation above and below the average, while circles represent each individual project. The red line in panel (b) indicates $N = 70$ contributors. Inset tables in all panels show the results of the ADF stationarity test. All plots are presented in semi-log axes.

The stationary pattern for the developers implicit average degree in Figure 4.1 is interesting in two aspects. First, it points to an inherent limitation to the number of connections (even indirect ones) that a contributor to a project can sustain. Notably, such limitation is below (but not far) from the Dunbar number (somewhere between 100 and 300), which is echoed as well in digital environments [147]. Second, the result is an indication of the existence of some sort of mesoscale organisation in the projects. In Bird *et* al. [144], the authors find that developers in the same community have more files in common than pairs of developers from different communities. Reversing the argument, one may say that relatively small contributor neighbourhoods are indicative, though not a guarantee, of the presence of well-defined subgroups in OSS projects.

4.4 Strcutural analysis: mesoscale patterns

From the previous encouraging result, we move on to the analysis of a comprehensive view of projects. The specificities of the methods to calculate nestedness $\mathcal{N}$, and to optimise modularity Q and in-block nestedness $\mathcal{I}$ are detailed in the Materials and Methods section. For the sake of illustration, Figure 4.2 (top row) shows adjacency matrices of three projects with high values of each structural measure. In this Figure, rows and columns have been rearranged to highlight the different properties.

We start out with a general overview of the results for the three measures of interest. Figure 4.3 plots the obtained values for $\mathcal{N}$, Q, and $\mathcal{I}$ over all the projects considered in this work. To ease visualisation, and considering that nestedness and modularity are antagonistic organisations [151], projects are sorted to maximise the difference between

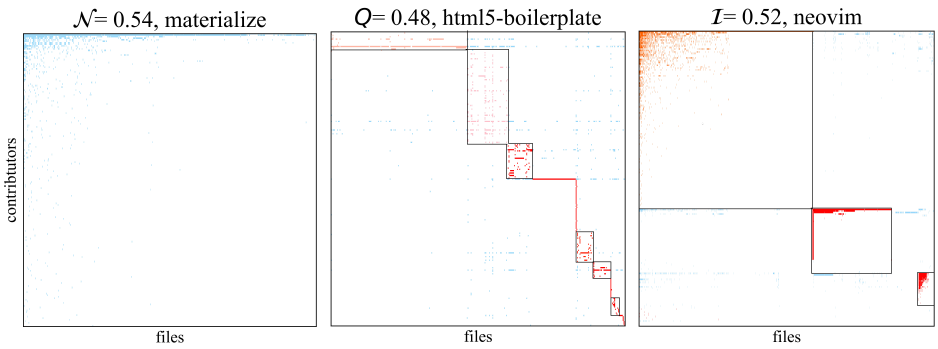


Figure 4.2: **Interaction matrices for three projects with high values for each one the structural patterns of interest.** Left: Nestedness $\mathcal{N}$, middle: Modularity Q, right: In-block nestedness $\mathcal{I}$.

$\mathcal{N}$ and Q. In general, nestedness is the lowest of the three values at stake, and in-block nestedness is, more often than not, the highest. It can be safely said, thus, that a tendency to self-organise as a block structure is present: 90% of the projects exhibit either Q or $\mathcal{I}$ above 0.4, and values beyond 0.5 are not rare. This evidence is compatible with previous results regarding the division of labour: indeed, be them modular or in-block nested, most projects can be splitted into communities of developers and files, forming subgroups around product-related activities [144].

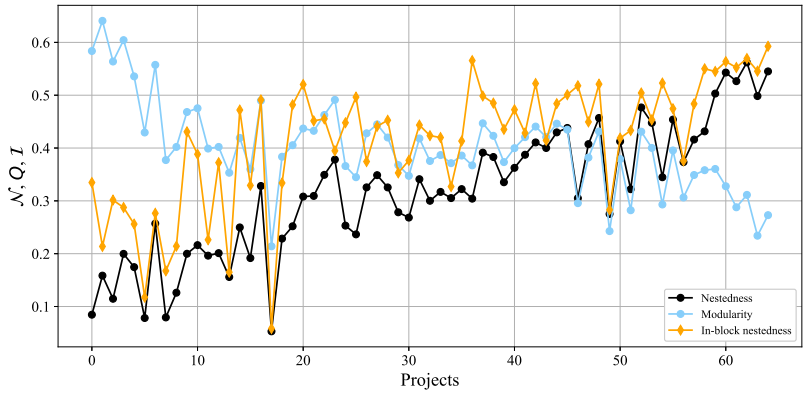


Figure 4.3: $\mathcal{N}$, Q, **and** $\mathcal{I}$ **obtained values, for each project of our dataset.** The projects were sorted to maximise the difference between $\mathcal{N}$ and Q.

Just like there is virtually no technical limit to the overall size of a project, there is not either an explicit bound to the size that a sub-group should have. And yet, previous theory and evidence suggests that larger communities come at an efficiency cost: the dynamics of a group change fundamentally when they exceed the Dunbar number, which is estimated around 150. While most often the number refers to personal acquaintances, it has been (and still is) applied in the industrial sphere [152]. Applied to the OSS environment, exceedingly small working sub-groups might hamper a project's advance;

while too many contributors may not allow the group to converge towards a solution [153, 154]. We explore whether, indeed, size limitations arise in developers sub-groups, as they emerge from either Q or $\mathcal{I}$ optimization procedures. Although partitions are hybrid, i.e. a community has both developers and files, in the following, we will report the community sizes in terms of developers.

Figure 4.4 provides a global overview of the 65 projects studied here, with the distribution of their largest subgroup sizes as they are identified via Q (panel (a)) or $\mathcal{I}$ (panel (b)). In both cases the average (dashed orange vertical line) is below 200, and the histogram is evenly distributed around 100: most communities belong in the range from 80 to 200. Given the obvious similarity between both distributions, we perform a Mann-Whitney U test, so as to find out whether these two distributions are actually compatible (the null hypothesis cannot be rejected, p-value = 0.3). In other words, block sizes are independent from the optimization strategy adopted. Indeed, the test indicates that both size distributions can be regarded as drawn from populations having the same distribution, and the combined distribution is shown in panel (c). The solid red line represents a log-normal fit (notice the logarithmic scale in the x-axis), and the insets in all panels show the QQ plots, to compare both theoretical and empirical distributions revealing that the fit is accurate.

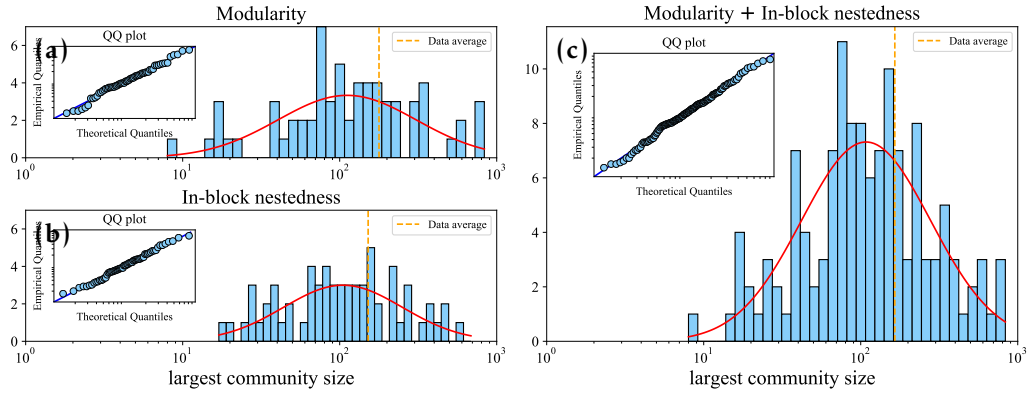


Figure 4.4: **Frequency distribution of the largest community size for each project.** Values obtained after optimization of modularity (panel a) and in-block nestedness (panel b). The distribution of largest community size when combining both opmtimization strategies is shown in panel c. In the three panels, the solid red line corresponds to the log-normal fit performed to each distribution, which are centred around 100. The dashed orange line indicates the average values of our dataset, and inset panels show the QQ plots of the empirical versus theoretical quantiles from the log-normal distribution fit.

Although Figure 4.4(c) evidences, on average, a well-defined maximum community size (at 169.7 users, and 95% confidence interval [139.4, 206.7] as measured for log-normal distributions [155]), we must ensure that the size of the largest communities detected for each project is independent of the size of the project, in order to validate

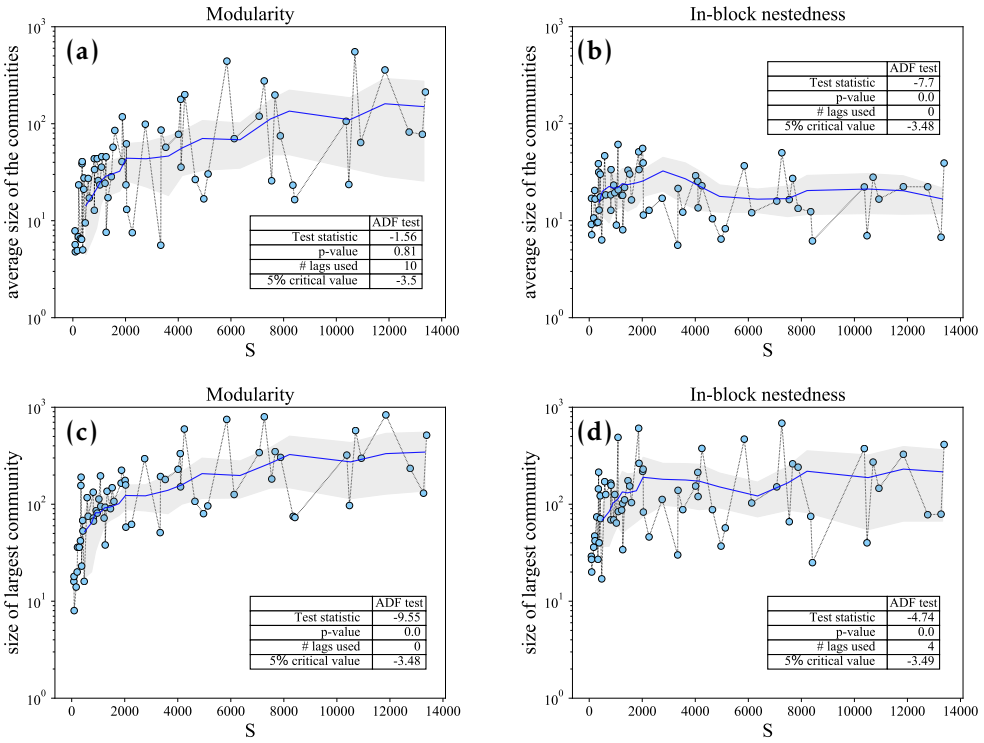


Figure 4.5: **Evolution of the average community size as a function of S:** for Q- and $\mathcal{I}$-optimised partitions (panels (a) and (b), respectively). Regarding the size, average Q-communities are in general larger than $\mathcal{I}$-communities. Furthermore, the scaling behaviour is also different: an average community size for Q-optimised partitions moderately grows with S, while it remains almost constant for $\mathcal{I}$ beyond $S > 2000$. Turning from average to maximum community size, Q- and $\mathcal{I}$-optimised partitions (panels (c) and (d), respectively) present very similar bounds, from 30 to 300 contributors. Again, the largest Q-community slightly tends to grow with S, while this size stabilises around 100 for the case of $\mathcal{I}$. Inset tables in all panels show the results of the ADF stationarity test, confirming the presence of bounded values for the maximum subgroup sizes (panels (c) and (d), respectively). Note semi-log scaling.

such organizational limit. That is, we need to test that the largest blocks (far right in panel (c)) do not necessarily correspond to the largest projects. To do so, we go down to the project level. Figure 4.5 reports average (panels (a) and (b)) and maximum (panels (c) and (d)) subgroup sizes for both community identification strategies, as a function of the project size S. In general, results point at the existence of upper bounds to community size. This impression is confirmed statistically, as the ADF test for stationarity indicates (see p-values in insets) that subgroup sizes, after a fluctuating behaviour when $S < 2000$, remain stable across S in panels (b) to (d). This is not so in panel (a): average Q-communities exhibit a subtle growth with respect to project size, and the ADF test signals such non-stationarity. Nonetheless, it is apparent that in all cases –even in panel (a), despite its increasing trend– the size of communities is compatible with the

limits described by Dunbar's number: in panel(c), largest community size is slightly above 200. In panel (d), even the largest projects reflect that the maximum size of a community is between 100 and 200.

These results are surprising, since such trend towards the compartmentalization of the workload is not only decentralised, in the sense that it does not emerge from a predefined plan, but also implicit, because the interaction between developers is most often indirect.

4.5 Co-existing architectures and project maturity

As it has been suggested [151], empirical evidence indicates that more than one structural pattern may concur within a network, each evincing different properties of the system. We take the same stance here: a network is not regarded, for example, as *c*ompletely modular or *c*ompletely nested; rather, it may combine structural features that reflect the evolutionary history of the system, or the fact that the system evolves under different dynamical pressures that favour competing arrangements.

A convenient way to grasp this mixture is a ternary plot (or simplex), see Figure 4.6. In the ternary plot, each project is located with three coordinates $f_{\mathcal{N}}$, f_{Q} and $f_{\mathcal{I}}$, which are simply calculated from the original scores, e.g. $f_{\mathcal{N}} = \mathcal{N}/(\mathcal{N}+Q+\mathcal{I})$ (note that the three quantities are, by definition, in the $[0,1)$ range). The simplex can be partitioned according to "dominance regions", bounded by the three angle bisectors. These regions intuitively tell us which of the three patterns is more prominent for any given project.

Figure 4.6 reveals that most projects lie in the nested regions, while the predominantly modular region is relatively empty. Note that certain areas of the simplex (in grey in Figure 4.6) are necessarily empty. In particular, the right half of the ternary, i.e., $f_{\mathcal{I}} \geq f_{\mathcal{N}}$, is empty since, by definition, $\mathcal{I}$ reduces to $\mathcal{N}$ when the number of blocks is 1, hence, the contribution of $\mathcal{N}$ is always equal or smaller than the contribution of $\mathcal{I}$. On the other hand, as explained in Chapter 3, an in-block nested structure exhibits necessarily some level of modularity, but not the other way around. This explains why the lower-left area of the simplex in Figure 4.6 is empty as well.

Together with their dominant architecture, points in Figure 4.6 are colour-coded according to the total number of commits that each project has received. We take this number as a proxy to the level of development or maturity of the project (note that a project's age may be misleading due to periods of inactivity). The distribution of colour on the simplex suggests that more mature projects tend to exhibit nested or in-block nested structures, whereas predominantly modular projects appear to be relatively immature (with exceptions, admittedly). Such result is resonant to the fact that topical conversations in online social networks ("information ecosystems") evolve through different stages –modular when the discussion is still brewing in a scattered way; nested when the discussion becomes mainstream to the group of interest [49]. More relevant to OSS development, Figure 4.6 reconciles the idea of workload compartmentalization

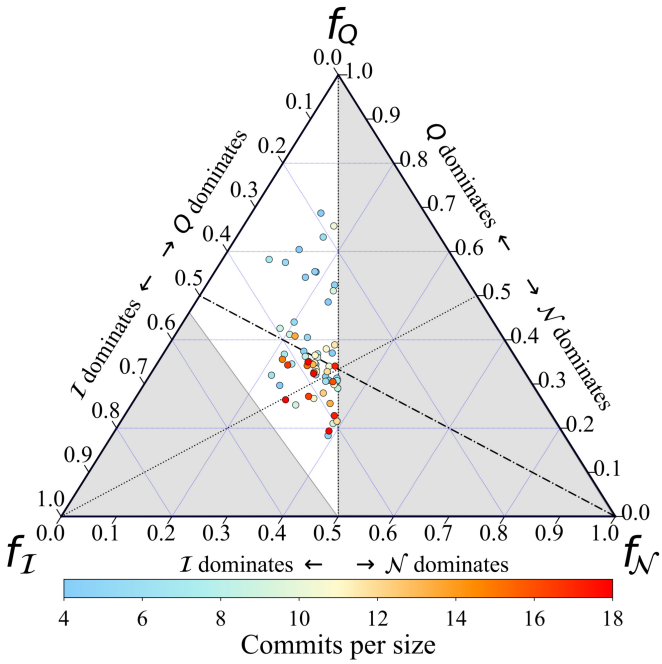


Figure 4.6: **Distribution of the three architectural patterns for each the projects across a ternary plot.** The colourbar indicates the number of commits received by each project, normalized by the size of it.

(sub-communities forming around product-related activities) [144], and the emergence of hierarchies [133] or a rich club [143] of developers, at least in well-developed projects. This partial picture is however complemented by the fact that hierarchies emerge as well on the code class: the presence of generalists and specialists applies to both developers and files in a nested or in-block nested scenario.

4.6 Summary

In this Chapter we have analysed a set of popular open source projects from GitHub, placing the accent on nestedness, modularity and in-block nestedness –which typify the emergence of heterogeneities among contributors, the emergence of communities of developers working on specific subgroups of files, and a mixture of the two previous, respectively. Our analyses have shown that indeed projects evolve into a relatively narrow set of structural arrangements. At the mesoscale, we have observed that projects tend to form blocks, a fact that can be related to the need of contributors to distribute coding efforts, allowing a project to develop steadily and in a balanced way. Focusing on the file class, the emergence of blocks is interesting as well, since a modular architecture (understood now as a software design principle) is a desired feature in any complex software project. Furthermore, those blocks or subgroups have a relatively stable size no matter how large a project is. Remarkably, such size is compatible with the Dunbar number.

Our results create a link between bio-cognitive constraints, group formation and online working environments, opening up a rich scenario for future research on (online) work team assembly (e.g. size, composition, and formation). From a complex network perspective, our results pave the way for the study of time-resolved datasets, and the design of suitable models that can mimic the growth and evolution of OSS projects.

Chapter 5

Structural Elasticity in Online Communication Networks: an ecological approach

In previous Chapters, we have shown that different structural architectures can coexist in a system, and had introduced a discussion on how the presence of combined patterns or transitional states within a system could appear as the system's response when facing two driving forces at the same. It's is not surprising then that one of the most repeated mantras in the study of complex systems refers to the intertwined nature of structure and dynamics: different structural arrangements place distinct constraints to the dynamics on that structure, while the dynamics modify the structure that supports it. With this in mind, the design of mathematical models that mimic the observed changes in network topology seems like a suitable approach to bring to light the underlying mechanisms behind their emergence.

This Chapter is focused to the development of a dynamical model to study these type of structural connectivity patterns, which have been observed in online communication networks. First, we report evidence that these information networks, in resemblance to natural ecosystems, show remarkable structural resilience to environmental changes and, secondly, we provide an ecology-inspired theoretical model that explains the dynamical reorganization observed in the data. We adopt a novel perspective on the dynamics of information networks, in which co-adaptation and surrounding conditions are naturally inserted. Our proposal builds on the idea that the network structure between users and memes is the result of a local optimization process [66], i.e. the individual maximization of visibility, and that the nature of the interactions is mutually beneficial,i.e., mutualistic.

5.1 An ecological approach to model information ecosystems

Perceptual and cognitive human capabilities are limited resources [156–158]. However, their finiteness had not generally emerged in day-to-day communication processes: not in the pre-industrial era, where a physical (face-to-face) or low-bandwidth interaction governed the slow change in public opinion; nor during the dominance of the media when exposure to an oligopolistic media environment puts little pressure on the audience's attention resources. In both cases, the public sphere was hierarchically structured and framed by the operations of few actors on a rather slow time scale. On the contrary, the paradigm of online communication is characterized by the fragmentation of the public sphere [159], in which elite and non-elite actors behave as sources and receivers of information on the virtual stage. Only in this new scenario, attention, memory and processing time suddenly become critical assets to compete for [147, 160–162]: their scarcity has been exposed.

Complementary to direct competition (among actors), interaction with other units in the system is often mutualistic. For the same reason that two actors compete with each other, they establish cooperative relationships with the memes (keywords, hashtags). These "information chunks" may –if correctly chosen– optimally spread information and consolidate the visibility they strive for. Hence, for example, the (ab)use of hyper-emotional language that we suffer in nowadays politics, as an arms race to impact optimization.

Of course, the choice of a meme is context-dependent ("past performance is no guarantee of future results"), and thus the interactions between actors and memes are adaptative and extremely sensitive to changes in the communication environment –breaking news, fads and rumours, celebrity gatherings, etc–. In turn, changes in the surrounding conditions tend to be ephemeral although frequent, in the more open and fluid access to many digital sources.

Under the light of these four drivers –competition, mutualism, adaptation, and environment–, online communication systems and natural mutualistic assemblages become special cases of a broader class of mutualistic bipartite systems, i.e. those dominated by intra-class competition and inter-class mutually beneficial interactions, although clearly functioning at very different spatial and temporal scales. Our failure to realize this in the past is due to several factors. Previous approaches to an "info-ecological" understanding of online communication dynamics typically focused on one of the dimensions of the problem actors [147, 160–162] or memes [163]), missing the co-evolutionary interplay of topologies and states in the network [164–167]. This picture changes dramatically if the focus is shifted from the relatively stable peer-to-peer network to the fluid information bipartite network, that is, *a*d hoc groups of users, which loosely gather around and engage in shared memes [168], operating in a hyper-competitive environment [169]. Other approaches, which did include the bipartite perspective, were limited to a qualitative discussion as a result of empirical observa-

tions on a single dataset [49], failing to identify the mechanisms that drive the whole system.

A picture that embraces the mentioned ingredients opens new promising possibilities to analyse and model online social networks, if we consider that Ecology is rich in theoretical frameworks where the co-evolutive coupling between structure and dynamics is studied [66, 125, 170]. Moreover, while testing these theories empirically in natural ecosystems is difficult –mainly because of the resource-intensive demands to collect data [171]–, digital streams from social interactions are abundant on several spatial and temporal scales, and precise knowledge about the environmental (external) conditions –related to specific information flows– can also be collected.

The first problem to address under this *i*nformation ecosystems framework is the network's structural volatility, which is coupled to the fluctuating nature of the environment. Online communication is heavily driven by the events surrounding it, which constantly trigger attention shifts that modify the behaviour of otherwise loosely linked assemblages of individuals and groups [169]. It is precisely this hectic, information-dense environment that dictates the emergence and fall of ephemeral synchronized attention episodes, which translate in fast structural changes.

Here, we provide evidence that information ecosystems exhibit a remarkable structural elasticity to environmental changes , recovering its original architecture in the aftermath of an external event affecting it. To do so, we first report on theory-free, empirical observations of the characteristic dynamical re-organisation in communication networks, as they react to environmental "shocks". Analyzing the response of the Twitter ecosystem to different types of external events, we quantify how collective attention episodes reshape the user-hashtag information network, from a modular [33, 40] to a nested [45, 72] architecture, and back. The emergence of these structural signatures is, remarkably, consistent across different topics and time scales. Next, we propose a theoretical framework that explains the emergence of the patterns observed in real data streams, as a result of an adaptive mechanism. The model builds on the idea that the user-meme network structure is effectively driven by an optimisation process [66], aiming at the maximization of visibility, and that the nature of the user-meme interactions is mutualistic, i.e., beneficial for both. Furthermore, through our modeling framework we predict that the users' struggle for visibility in any context facilitates the emergence of nested arrangements at multiple scales: either mesoscale (in-block) nestedness [21, 53] during the compartmentalized stages, or macroscale nestedness in exceptional global attention episodes. These predictions are supported by the data. Finally, we present some results that link our observations with the model at the microscale, which suggest that environmental shocks may leave a trace, if not at the structural level, at the dynamical one.

5.2 Data and Methods

5.2.1 Datasets

Biased as it may be [172], Twitter is without a doubt a sensitive platform that mirrors, practically without delay, exogenous events occurring in offline environments. In this sense, Twitter data constitute a rich stream, providing a public and machine-readable reflection of the real world. Therefore, the empirical data employed in this work was collected from the online platform www.twitter.com.

Following the analogy with interactions in ecological systems, two types of species are considered: users and hashtags (memes). For each tweet on the different datasets, we only extracted the user's name, the hashtags in the tweets, and the time at which they were posted.

We considered two events of different nature: the Spanish general elections april 2019 (28A), the 2015 Nepal earthquakes. The dataset corresponding to the 2015 Nepal earthquakes was collected in [173].

1. **Spanish general Elections (April 2019):** The April 2019 Spanish general elections were held on Sunday, 28 April 2019, to elect the 13th bicameral legislative chambers of the Kingdom of Spain, the 350 seats in the Congress of Deputies and 208 out of 266 seats in the Senate. The observation period started at the beginning of the electoral campaign, on the 12th of April and lasted until the 6th of May, a few days before the beginning of the electoral campaign for the election of the 54 Spanish members of the European Parliament. Hence, the observation period was marked by intense political activity. For this event, we collected a dataset composed of $3,0107,629$ unique tweets containing at least one hashtag, with a total $124,062$ unique hashtags and $1,883,468$ users. The dataset was collected by selecting all the tweets containing at least one of a total set composed of 300 relevant keywords that could be either user names or hashtags related to the electoral process, i.e, names of candidates, electoral actvities (debates, meetings) and name of the parties involved, etc.

2. **Nepal Earthquake (April-May 2015):** The next dataset taken into consideration for analysis corresponds to an unexpected event, specifically, a series of earthquakes registered in Nepal in 2015. The first earthquake occurred on the 25 of April 2015, registering around 9000 casualties. This event was followed by several continued aftershocks, with a major aftershock of similar magnitude of the first quake, registered on May 12th. Given the unpredictable nature of this type of event, we have focused on the study of the second major earthquake. The observation period covers a total of six days, from 8 to 14 of May, a few days after the second aftershock. The dataset contains $1,918,045$ unique tweets containing at least one hashtag, with a total $35,795$ unique hashtags and $810,744$ users. The

dataset was collected by selecting all the tweets containing at least one of the following hashtags or keywords: *nepal*, *earthquake*, *#nepalearthquake*.

Table 5.1: Summary of the datasets

Dataset	Data length	Total days	Tweets	Users	Hashtags
2019 Spanish general elections	April 12 - May 6	24	30,107,629	1,883,468	124,062
2015 Nepal Earthquake	May 8-14	6	1,918,045	810,744	35,795

5.2.2 Matrix construction

Prior the construction of the interaction matrices, we performed a selection criteria that allowed us to capture the structural changes of the data in a smooth way, and, at the same time, reducing the computational cost.

For each dataset, we split the timestream into chunks according to non-overlapping time windows with three hours of duration $\omega = 3$h, Fig. 5.1 top row. For each chunk, we built matrices $a_{uh}^{(t)}$ containing the 2000 most active unique users and a variable number of hashtags, depending on the amount produced by those 2000 users [49]. Each cell in the matrices $a_{uh}^{(t)}$ is equal to 1 if user u has posted a message containing the hashtag h at least once, and 0 otherwise. Note that each matrix will have a different duration, spanning from a few minutes during the times of high activity (when an event is taking place), to the total duration of the time window. For each one of these 3-hour chunks, we select the matrices that are closer to the middle of the time window, e.g., $a^{(t\approx\omega/2)}$ to perform the structural analysis. Around the periods of high activity –on the onset of the events– the procedure is repeated considering time windows of 15 minutes of duration.

It is also important to highlight that the $a_{uh}^{(t)}$ matrices may not contain the same nodes across t: as time advances, users join (disappear) as they start (cease) to show activity; the same applies for hashtags, which might or might not be in the focus of attention of users. This volatile situation is quite normal in time-resolved ecology field studies [174–176], where the accent is placed on the system's dynamics –rather than individual species.

5.2.3 Structural measures.

We have explored the structural evolution of the network by means of three arrangements, one at the macroscale (nestedness [41, 43]), and two at the mesoscale (modularity [33], in-block nestedness [21, 53, 151]). We focus our attention on modular, nested and in-block nested patterns since all of them have been observed prominently in ecology [50, 65, 113, 114] and in information systems [49, 53]. We quantify the amount of nestedness by means of a global nestedness fitness $\mathcal{N}$, introduced by Solé-Ribalta *et* al. [53], an overlap measure [73] that includes a suitable null model. We follow this work as

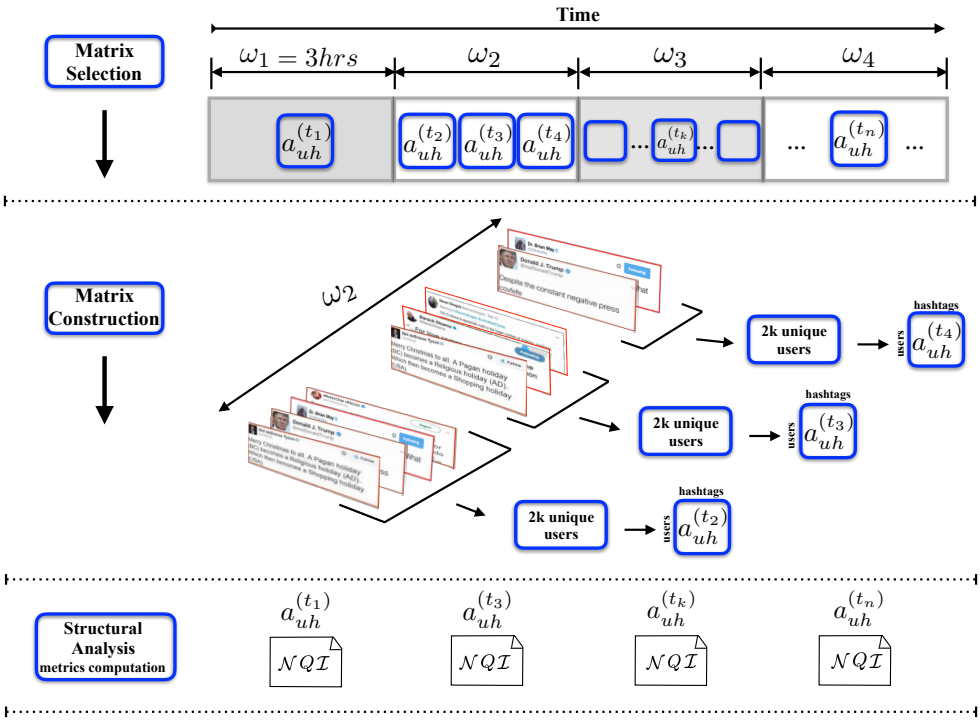


Figure 5.1: **Schematic representation of the implemented methodology in the analysis of empirical data.** The applied methodology comprises three steps: Selection and construction of the adjacency matrices (top and middle rows) , structural analysis of the selected matrices by means of nestedness, modularity and in-block nestedness (bottom row).

well for the definition and optimization strategy of in-block nestedness $\mathcal{I}$. With respect to community analysis, we apply a variant of the extremal optimisation algorithm [29], adapted [151] to maximize Barber's bipartite modularity [30].

5.3 Structural elasticity in information systems

Despite the highly fluctuating nature of the timestream datasets, some reliable patterns emerge from its apparently hectic activity. For now, we focus on two of them: modularity [33, 40] (Q) and nestedness [41, 43, 45] ($\mathcal{N}$). High levels of modularity correspond to a fragmented attention scenario, and can be considered as the *r*esting state of the system. In this stage, users mostly focus on their own topics of interest, *i*.e. a certain subset of memes, facilitating the emergence of identifiable blocks. High values of nestedness, on the other hand, reflect an extraordinary (and, thus, ephemeral) stage in which the system self-organizes to attend one or few topics. In these cases, the discussion revolves around a small set of generalist memes (hashtags used virtually by everybody) and users (highly active individuals participating in many facets of the discussion).

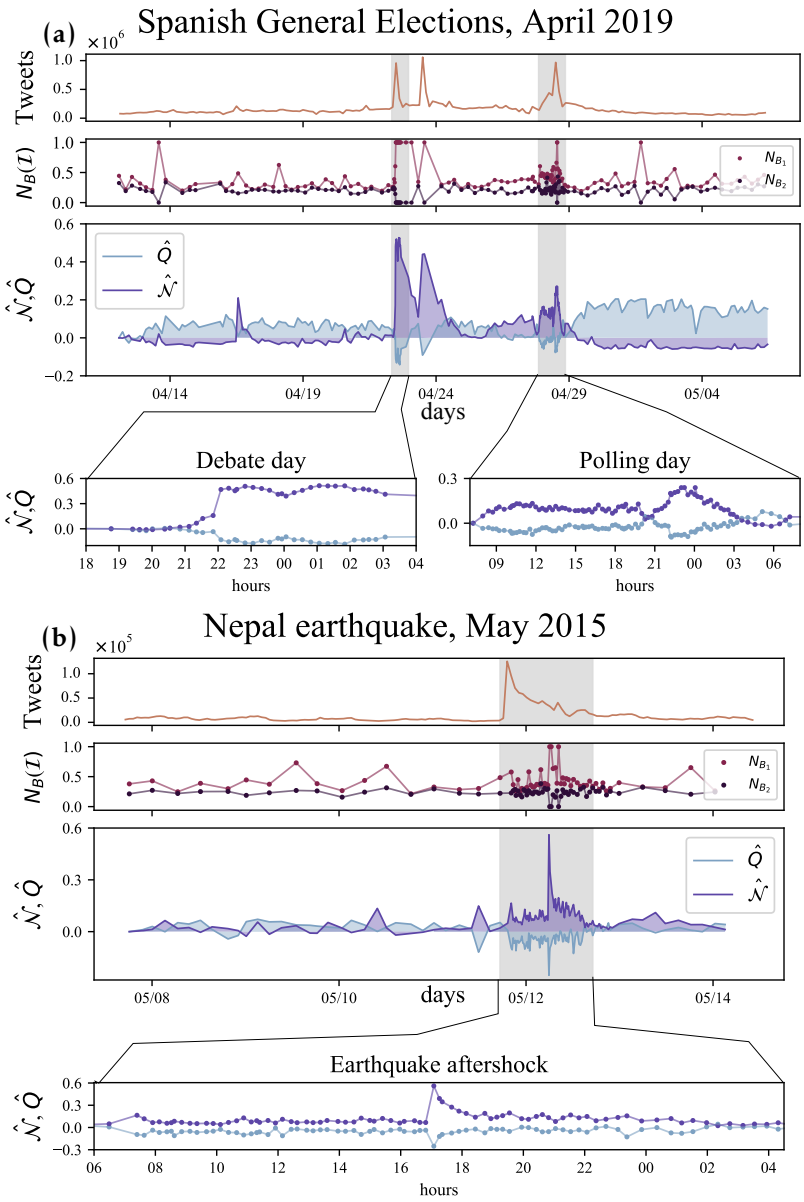


Figure 5.2: **Structural measures over time for two datasets:** Twitter streams covering two different topics, i.e. Spanish general election of 2019 (panel (a)) and 2015 Nepal earthquake (panel (b)). Spanning different time ranges and attracting varying levels of attention (see tweet volume in top panels), the information ecosystems self-organise in similar ways: a block organization dominates the system (positive modularity $\hat{Q}$), reflecting the separate interests of users, until external events induce large-scale attention shifts, which rearrange completely the network connectivity towards a nested architecture (high $\hat{\mathcal{N}}$). For a closer view, we highlight specific time windows in each dataset with some identifiable events happening in them (lower panels). In each plot, measures of modularity and nestedness are shifted from their initial values (Q_0 and $\mathcal{N}_0$, respectively). The panels corresponding to $N_B(\mathcal{I})$ highlight the nested self-similar arrangements at different scales, which is discussed later on.

Figure 5.2 presents the evolution of Q and $\mathcal{N}$ for the two datasets described above (several more are shown in Appendix B, Section B.2, with similar insights). For example, Fig. 5.2 (a) corresponds to a period of over 20 days around the local elections in Spain (April-May 2019). For this dataset, the evolution of Q and $\mathcal{N}$ shows a remarkable anti-correlated behaviour, that was statistically validated by computing the Peason coefficient, see Section B.5, Table S2 in Appendix B. Such behaviour can be explained by the mutual structural constraints that these two arrangements impose on each other [151]. Remarkably, however, the significant growth of nestedness is not caused by a decrease in modularity of the system, but, on the contrary, tightly linked to external events: see for instance the sudden changes in the structure on specific dates, shadowed in grey in the figure (debate and polling day, respectively). These extraordinary events are accompanied, unsurprisingly, by an increased volume of messages (top panel) and connectance. Despite previous research [177], neither volume nor connectance can explain, *p*er se, the rapid surge of nestedness. We discuss this aspect in more detail in Appendix B, Section B.7. The figure, at the scale of days, is complemented with high-resolution monitoring of portions of these exceptional events (bottom panel). Finally, the most outstanding feature highlighted by the figure is the elasticity of the network: no matter how abrupt and large the excursion to a nested arrangement is, the system bounces back to its "ground" –predominantly modular– state soon after, when the interest in the breaking news fades out. The observed elasticity can be considered as an aspect of the network's structural resilience. System resilience or stability is defined in different ways in ecology and environmental science [178–180], but can generally be thought as the ability of the system to recover the original system's state after a perturbation of the model state variables [181, 182] or parameters [183, 184]. Specifically, in the case of structural elasticity, the system state is not given by the nodes' configuration (*e*.g. the abundance of each species), but by the overall network architecture (*i*.e. modular, nested), which is perturbed by the external event.

This behaviour is stable across different types of event. Figure 5.2b shows an equivalent behaviour for the reaction after the Nepal earthquake in 2015 [173]. Unlike a political debate or an election date, this example is inherently unexpected and unpredictable –an important fact, attending the taxonomy of collective attention described in Lehmann *e*t al. [185]. As in Fig. 5.2 (a), the coarse grain scale of days and weeks in Fig. 5.2 (b) is complemented with high-resolution monitoring a portion of exceptional events.

These analyses suggest that there is a tight logic underlying the structural fluctuations of the information network: the level of fragmentation of collective attention maps onto specific network arrangements, and is independent of the particular contents of the data stream. Online activity on different topics translates to comparable changes in the resulting patterns, no matter the semantics of the underlying discussion. The observed differences in the emergence, magnitude and persistence of structural changes

are directly related to the predictability, intensity and duration of the exogenous events (*i*.e. related to the environmental conditions), and therefore cannot be explained as intrinsic to the communication system itself. The question remains, however, how a networked system can fluctuate so fast between two states which have often been considered incompatible [65, 114, 151]. The key to this puzzle is in-block nestedness, a hybrid modular-nested architecture that bridges the apparent antagonism between nestedness and modularity [151].

5.4 Theoretical Framework

To understand the mechanisms that govern the observed elasticity, and, at the same time, to solve the puzzle around the network's nested-modular oscillations, we propose a model founded on the ecological drivers introduced above: competition, mutualism, adaptation and environment. The model builds on the simple idea that the network architecture between users and memes is the result of several local optimization processes, *i*.e. each individual's maximisation of visibility, and that such process operates on top of attentional dynamics. To do so, we generalize the ecological adaptive modeling proposed by Suweis *e*t al. [66, 125], in which the system's actors (plant and pollinator species) strive for larger individual abundance, rewiring their interactions accordingly.

5.4.1 Niche Model.

The synthetic information network model is developed for a bipartite network that comprises a total of N interacting "species" or nodes (N_U users and N_H hashtags or memes). Each species i has an associated niche [186] which, in the context of an information ecosystem, represents their topical domain (*i*.e. the topic to which a user attends preferentially, and, conversely, the semantic space where a meme belongs to). For the sake of simplicity, each species' niche is represented as a Gaussian distribution $G_i(s)$ with a given standard deviation σ_i [125]. Both users and memes niches center positions $\bar{s}_i$ are anchored around T different points in the range $[0,1]$, to express different topic preferences (users), and semantic domain (memes). To model the inherent diversity of users and memes within their topic, their position over the line is perturbed by a small amount, randomly sampled from a uniform distribution.

Competition occurs between species of the same class (or guild), whereas mutualistic interactions couple the dynamics of abundance of users and memes. Following the proposal of Cai *e*t al. [125], the strength of the competitive interactions between a pair of users (memes) is tuned by a fixed parameter (Ω_c) scaled by a quantity that depends on the niche overlap G_{ij} between them. Similarly, the strength of the mutualistic interactions between a pair user-meme results from a fixed parameter (Ω_m) scaled by the niche overlap between the pair user-meme –*i*.e. the similarity between the user's topic preference and the adequacy of the meme within this topic–, and constrained to the existence of a link between them. Then, we define the niche overlap G_{ij} of a pair of nodes i and j as:

$$G_{ij}^{gg'} = \int G_i^g(s) G_j^{g'}(s) ds, \tag{5.1}$$

with g and g' denoting the guild of the considered species, either users or hashtags.

Following this, we define the mutualistic interaction matrix as:

$$\text{Mutualism: } \gamma_{ik}^{UH} = \Omega_m \cdot \theta_{ik} \cdot G_{ik}^{UH}, \tag{5.2}$$

where θ_{ik} is the adjacency matrix, with entries equal to 1 if i and k interact, and 0 otherwise.

With respect to the competitive interactions, we distinguish two levels of competition. At the local level, users (hashtags) in the same topic compete to gain visibility among those with related interests (meaning). At the aggregate level, a given topic strives to prevail among other topics. In order to capture this double competition as a trade off between both tendencies in our model, we define the competitive interaction matrix as:

$$\text{Competition: } \beta_{ij}^U = \begin{cases} 1 & \text{if } i = j \\ \Omega_c \left[\lambda(1 - G_{ij}^{UU}) + (1-\lambda) G_{ij}^{UU} \right], & \text{otherwise,} \end{cases} \tag{5.3}$$

where $\lambda \in [0,1]$ is the inter-intra topic competition parameter, the same definition applies to the competitive interactions among hashtags. For the case $\lambda = 1$, the competitive matrix neglects the competition among users belonging to the same topic. The case when $\lambda = 0$ corresponds to the original formulation [125].

Figure 5.3a summarises the ingredients of the model. We note that, in contrast to natural ecosystems, memes are an infinite resource –which explains why user-user competition does not grow with the amount of shared memes.

5.4.2 Population dynamics and optimization process

On the dynamical side, the species abundances evolve according to a set of Lotka-Volterra equations with Holling-Type II mutualistic functional response with handling time h:

$$\begin{aligned} \frac{dn_i^U}{dt} &= n_i^U \left(\rho_i^U - \sum_j \beta_{ij}^U n_j^U + \frac{\sum_k \gamma_{ik}^{UH} n_k^H}{1 + h \sum_k \theta_{ik}^{UH} n_k^H} \right) \\ \frac{dn_i^H}{dt} &= n_i^H \left(\rho_i^H - \sum_j \beta_{ij}^H n_j^H + \frac{\sum_k \gamma_{ik}^{HU} n_k^U}{1 + h \sum_k \theta_{ik}^{HU} n_k^U} \right). \end{aligned} \tag{5.4}$$

Within the information ecosystem context, these equations represent a phenomenological way to describe the evolution of the nodes visibility as a function of their interaction. In particular, n_i^U may represent the number of instances in which user i is present in other users' screens, while n_j^H may quantify the popularity of a given hashtag j. Assuming that preferential attachment mechanisms of various type affect the

nodes visibility, ρ_i^U and ρ_i^H model the associated exponential growth (if they are positive). The handling time h effectively models the constraint that users cannot interact with a very large number of hashtags due both to time and character constraints. Due to these limitations, the benefit obtained through mutualistic interactions does not grow monotonically with the number of partners.

Next, we cintroduce in the model a rewiring adaptation process that follows the approaches in [66, 125]. Each user attempts to change its mutualistic partners (memes) in order to maximize the benefit obtained from their use in the following way:

1. **Rewiring:** At each time step $t = mT$ (m is a positive integer and T is the integration time), a random species U, with a least one link, is selected and rewired to a randomly selected species H', removing one of its previous links H, with probability $p_{UH} \propto 1 - k_H^{-1}$. The rewiring probability is defined in such a way that the larger the species' degree is, the more prone to losing links. Once the rewiring is completed, we recalculate the mutualistic interaction factor of the new pair of nodes $\gamma_{ij'} = \Omega_m \cdot \theta_{ij'} G_{ij'}$ and integrate the dynamics according to Eq. 5.4, until the abundances of all species reach an equilibrium (integration time T is set sufficiently large).

2. **Link recovery:** At the end of each time step t, we compare the actual abundance of species U with its previous value. If the current abundance is greater than the previous value, the current (new) link is kept; otherwise, the previous one is recovered. Note that, in the case of abundance loss, only the connections are rolled back to the situation in $t - 1$; however, the vectors of abundances continue from their current state, $\vec{n}^U(t)$ and $\vec{n}^H(t)$.

This optimization principle may then be interpreted within an adaptive framework, in which users incrementally enhance their visibility by choosing the appropriate memes, and memes are created so as to maximize their diffusive capacity, see Figure 5.3b. In summary, both classes optimize the efficiency of resource usage, decreasing their chances of becoming extinct due to stochastic perturbations [161]. Within the model, this translates into reiterative rewiring interactions of randomly drawn users so as to increase their visibility –"abundance" in the ecological jargon.

5.4.3 Introduction of external events

At last, since our primary objective is to reproduce structural changes under the irruption of external events, the dynamical model includes as well a mechanism to introduce exogenous events in the environment. These can be understood as transitory shifts in the users' attentional niches, which are tantamount to (typically short-lived) changes in their interests (Figure 5.3c). In this altered environment, users temporarily engage with new kinds of hashtags, different from those they usually interact with.

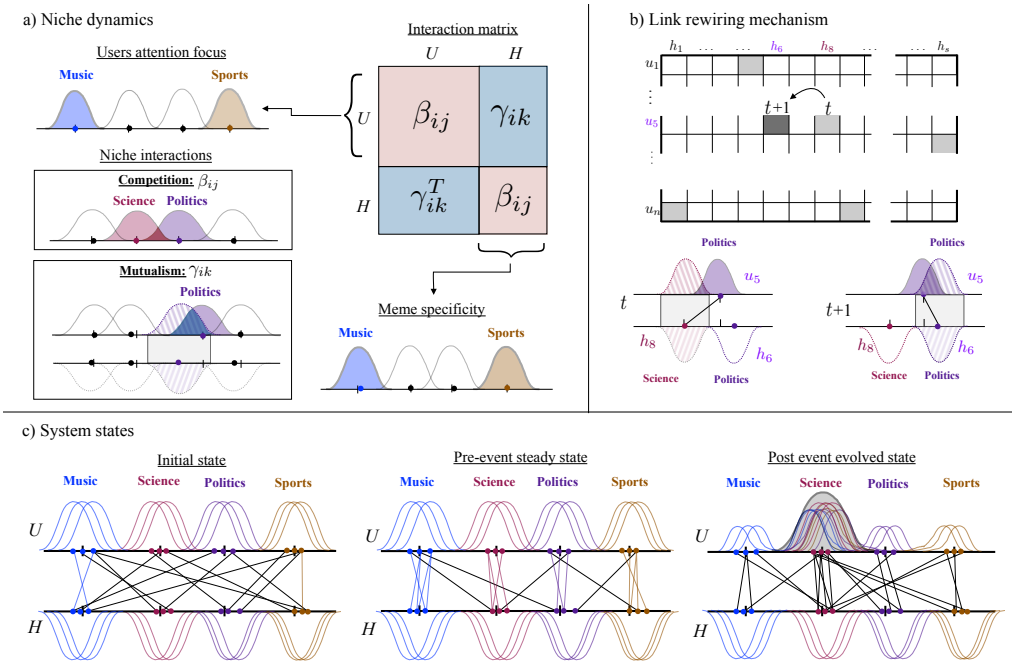


Figure 5.3: **S**chematic representation of the visibility optimization model. **(a)** Users and memes are represented as points in the range $[0,1]$ in a niche axis. We modeled each niche as a Gaussian curve with standard deviation σ. Topics are modelled as clusters of users (memes), *i*.e. $T = 4$. The coupling matrices β and γ, that define the competitive (within guilds) and mutualistic interactions, are define to be proportional to the niche overlap between pair of species. **(b)** At each time-step, species rewire their connections trying to optimise their abundance (popularity). If the rewiring leads to a larger popularity the connection is kept, otherwise the change is reverted. **(c)** Initially, the interactions are laid at random, and the rewiring proccess takes place. When the system reaches an evolved steady state, an external event enters the system. Users' niches are temporarily focused on a single common topic and the rewiring proccess continues while the effect of the event decays over time. As the event fades out, all species return to their original niche.

We modelled this situation as the change of every user's niche center towards a single common topic for a limited period of time. After that period of time, users were slowly moved back to their original niche centers, i.e. back to their respective topics.

An event modifies each users' niche in the following way:

$$G_i^E(s) = [1 - f(t_E)]G_i(s) + f(t_E)G^{E'}(s), \tag{5.5}$$

That is, a user's niche is now the composition of two Gaussian niches: one corresponding to the general event E (defined as a new niche profile $G^{E'}(s)$ centered at $\bar{s}_E$ and width σ_E), and the original one corresponding to the user's intrinsic interests $G_i(s)$. In this formulation, $f(t_E)$ is the function that governs the growth and decay of the external event, depending on the time t_E since its onset. We modelled two profiles, see Fig. 5.4, along the lines of Lehmann *e*t al. [185].

1. **Sudden event** The first event considered for study was modelled as a sudden and unexpected one. In this case $f(t_E)$ takes the form

$$f(t_E) = e^{-\alpha t_E}, \tag{5.6}$$

where α is the decay constant. Note that, at the onset of the event ($t_E = 0$), all users are focused on the same topic, and their niche overlap will be maximum. For sufficiently large t_E, namely $t_E \gg \alpha^{-1}$, the influence of the event becomes negligible.

2. **Expected event** In second place, we considered an expected event. In this case, the attention of users will slowly moves towards the one of the event, that is expected to happen at a specific time, in which the user's attention will be maximal. Here, $f(t_E)$ has the form

$$f(t_E) = \frac{1}{\left(1 + \left(\frac{t_E - t_o}{\alpha}\right)^{2a}\right)}, \tag{5.7}$$

where a and α are the parameters that regulate the width of the function and the duration of the plateu, while t_o specifies the location of the function peak, note that at $t_E = t_o$ the users niche overlap will be maximum. Again, we modelled the event such that for sufficiently large t_E, the influence of the event becomes negligible.

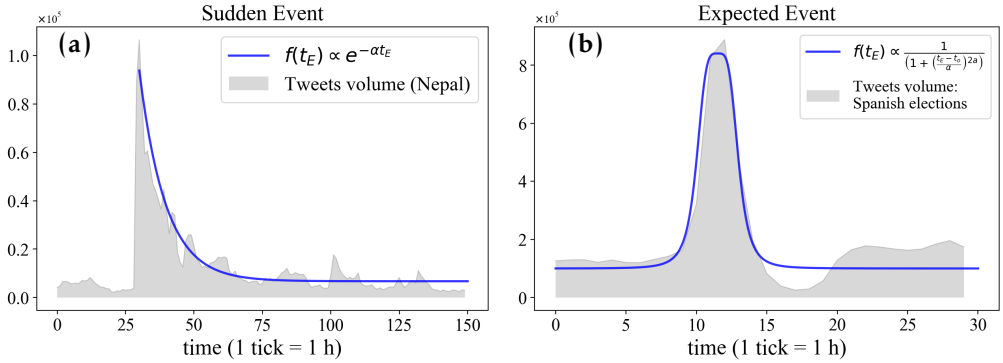


Figure 5.4: **Representation of the two different type of events included in our model**. Panel (a) shows the results a sudden and unexpected event, while panel (b) correspond to an expected event. To ease comparison with real scenarios, in both cases, $f(t_E)$ was shifted in order to align with the maximum and baseline activity of two empirical datasets (shadowed gray areas).

5.5 Numerical results

To avoid excessive computational costs, we consider small synthetic networks of $N_U = 100$ users and $N_H = 100$ hashtags with random connections across guilds, and density (connectance) $C \sim 10^{-2}$. We do so to match the same order of magnitude of empirical networks when we take $N_U = 100$, see Appendix B, Section B.3. We assign the same initial abundance $n_0 = 0.2$ to all the users and hashtags, the same intrinsic growth

rates $\rho_U = \rho_H = 1$ and handling time equals to $h = 0.1$. All the results presented below correspond to an average over 10 different realizations.

5.5.1 Species survival

Before exploring whether the observed structural elasticity can be reproduced within theoretical framework, we want to pay attention to the abundances of individual species over the mutualistic-competitive parameter space (Ω_m, Ω_c). Particularly, we want to characterize the regions over the (Ω_m, Ω_c) space where extinctions may occur, in order to guarentee the maximal survival species in the system prior to the introduction of the events. To this aim, we perform controlled numerical experiments at the stable stationary state on the (Ω_m, Ω_c) parameter space, for different values of the inter-intra competition parameter λ. We set both Ω_m and Ω_c in the interval $[0.1, 0.4]$ and perform simulations for 1200 different combinations of these parameters, for each value of λ. We consider that a species goes extinct if its abundance falls below 10^{-4}.

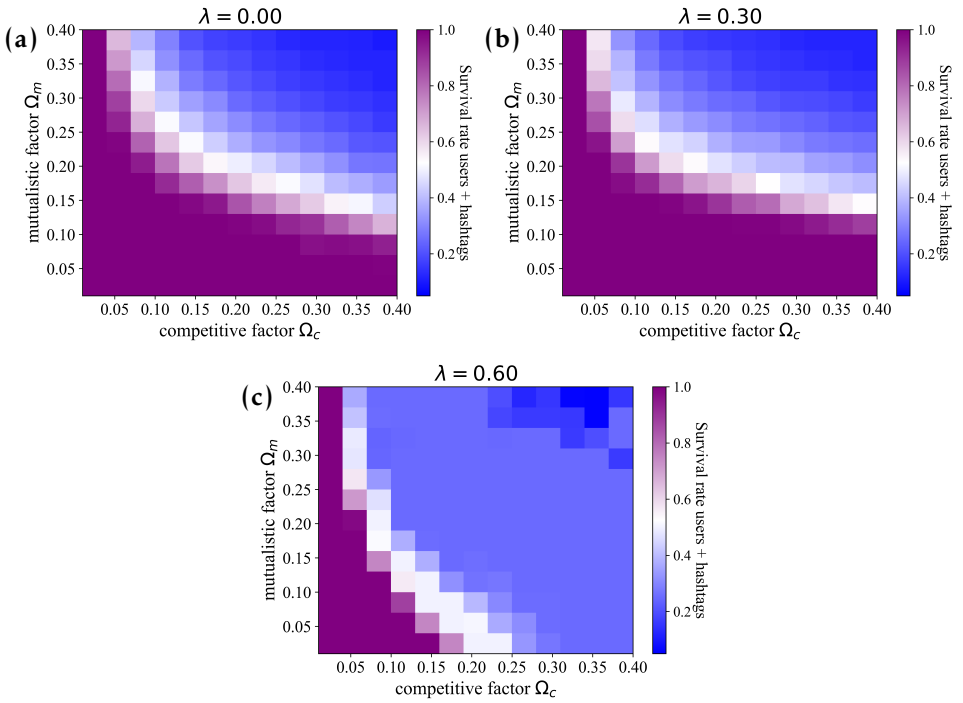


Figure 5.5: **Survival rate at the pre-event steady state:** two-dimensional plots in the $\Omega_m - \Omega_c$ parameter space showing steady survival rate of the species for different values of the inter-intra competition parameter.

Fig. 5.5 shows the fraction of especies survival in the two dimensional plot in the $\Omega_m - \Omega_c$ parameter space. For all the cases, we observe that as Ω_m and Ω_c increase, extinctions start to occur, even for favorable configurations of the system in which $\Omega_m > \Omega_c$. As expected, for low values of the inter-intra competition parameter $\lambda = 0$ and

$\lambda = 0.3$ the region in which extinction do not occur is wider. Under this configuration, the species compete more strongly within their topics, which correspond to just a fraction of all the species on the system. On the contrary, as we increase λ, the region of extinctions may increase, since now each specie starts to compete with a higher fraction of the system, making the system more susceptible to the values Ω_m, Ω_c. To guarantee the maximal survival of species before introducing the events, we will restrict the rest of our exploration on the $\Omega_m - \Omega_c$ space to the interval $[0.01, 0.1]$. On the other hand, since we know that the inter-intra competition λ parameter helps to balance the competition between the species, once the introduction of event take place, see Appendix B, Section B.4.1, from now on we will keep a fixed value of the inter-intra competition parameter. We fix the value of the inter-intra parameter $\lambda = 0.6$ since it offers a better trade off between the two competitive tendencies of our model –inter and intra topics competion–, maximizing the survival of the species (Figure B.3, in Appendix B).

5.5.2 Structural evolution.

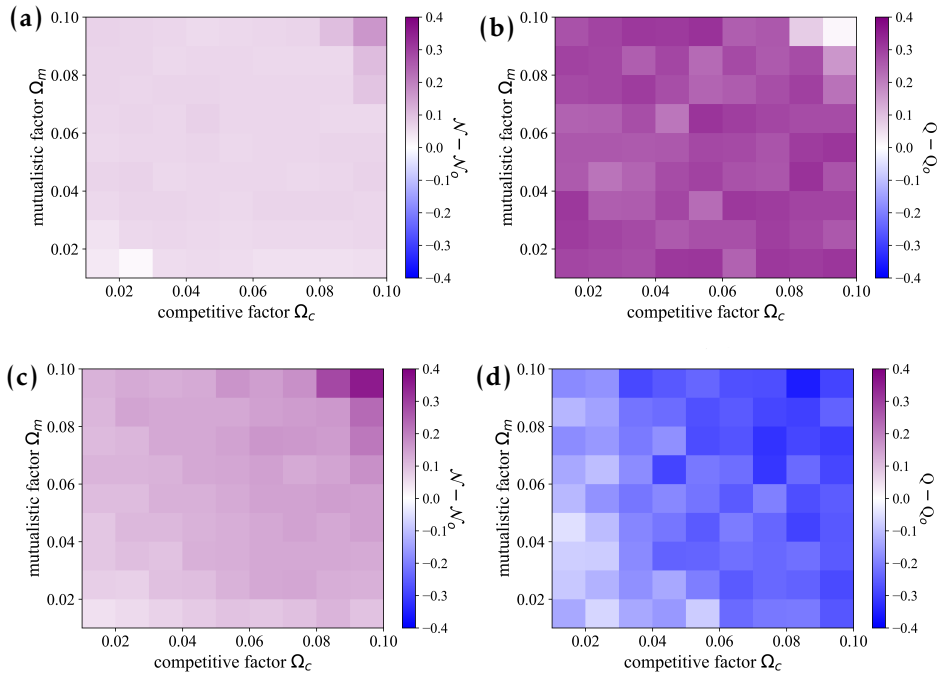


Figure 5.6: **Structural measures before and after the introduction of a sudden external event:** two-dimensional plots in the $\Omega_m - \Omega_c$ parameter space showing the evolution $\mathcal{N}$ and Q before (panels (a) and (b)) and after (panels (c) and (d)) the external event, for $\lambda = 0.6$. $\mathcal{N}_o$ and Q_o, correspond to the values at the beginning of the simulation.

In an unperturbed simulated environment, the observed structural arrangement mimics the prescribed organisation of niches in topical blocks. A modular architecture arises from the random initial one, while nestedness remain low, for a wide region of

the $\Omega_m - \Omega_c$ space, see Figure 5.6(a) and (b). Note that each point within the plots is shifted by Q_0, *i*.e. modularity value once the system has stabilized its architecture. This is in line with the *resting state* observed in the datasets (Figure 5.2), where users are focused on their own topics of interest. It is important to underline that the emergence of a modular architecture is not an artefact of the model: users (memes) do not rewire because of similarity reasons; it is the search for an improvement in their individual visibility that naturally drives to the consolidation of those new connections. Also note that, in empirical settings, the random initial stage is impossible to observe since the network already has a modular organization from the very beginning.

A change in the environment –*e*.g. breaking news– alters this scenario. The systems reacts with a decrease in Q, and an increase in the amount of $\mathcal{N}$ in the system, Figure 5.6 (c) and (d). To better visualize such transitions, Figure 5.7 shows the evolution of Q and $\mathcal{N}$ over the entire simulation time for a fixed combination of the (Ω_m,Ω_c) parameters $\Omega_c = \Omega_c = 0.07$. Panel (a) in Figure 5.7 models the increase, sustainment and decay of attention in programmed events (e.g. election day), while panel (b), mimics the arrival of an unexpected event. Note that if the simulation refers to an abrupt event (Fig. 5.7 (b)), the decrease in Q is sharp and almost immediate. If the simulation refers to a predictable event, (Fig. 5.7 (a)), the collapse of Q is smoother, and the emergence of $\mathcal{N}$ is slightly delayed. Indeed, in this situation we recover the results in Suweis *e*t al. [66] –the emergence of global nestedness–, because the existence of attentional niches becomes irrelevant when all niches are equally centred, at least on the users' side. In this sense, our niche-based population dynamics is a generalisation of Suweis and co-authors' model. As the environmental shock fades out, the network architecture tends to recover the general layout presented before the event was introduced. The elasticity of empirical information ecosystems is thus replicated here, and explained as a consequence of the adaptation to contextual changes –while the species' local strategies remain constant.

5.5.3 Nestedness reframed: meso- and macroscale analysis.

Beyond examining the evolution of Q and $\mathcal{N}$, we now take a look at the intra-modular organization of connections during the fragmentary stage of the system ($t < 3\times10^4$). For visualization purposes, the rows and columns of the adjacency matrices in the top-left part of Figures 5.7 (a) and 5.7 (b) have been arranged to highlight the block structure resulting from the modularity optimization. Additionally, rows and columns inside modules were sorted, in the bottom-left part, to highlight the possible nested structure within them [50, 51]. Clear to the naked eye, each compartment presents an internal nested architecture. This is a natural consequence of the node-level visibility-maximization strategy as it adapts to system-wide environmental conditions: as long as these conditions are stable around weakly connected topics, nestedness emerges in those relatively isolated subsystems. As soon as the boundaries across subsystems are blurred ($t > 3\times10^4$, top-right of Figures 5.7 (a) and 5.7(b), global nestedness prevails.

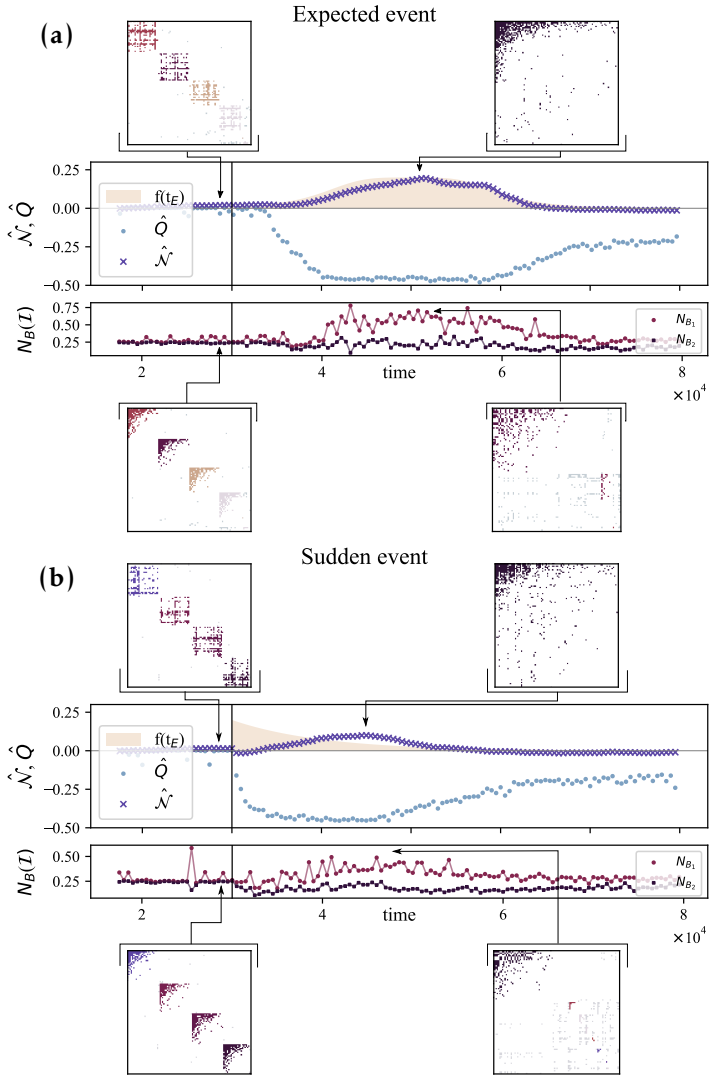


Figure 5.7: **Structural evolution in the visibility optimization model:** numerical experiments with fixed number of species $N_U = N_H = 100$, connectance $C \sim 10^{-2}$, initial abundance $n_0 = 0.2$, intrinsic growth rates $\rho_U = \rho_H = 1$, inter-intra competition parameter $\lambda = 0.6$, number of topics $T = 4$ (as in Figure 5.3), and mutualistic and competitive interaction factors $\Omega_c = \Omega_c = 0.07$. Initially, links between users and hashtags are laid at random. Panel (a) models the increase, sustainment and decay of attention in programmed events (yellow shade). Panel (b), mimics the arrival of an unexpected event (yellow shade). In the absence of external events, the system organises in a clear block structure. Once the external events enters, the system evolves from a modular towards a hierarchical, nested configuration. After the effects of the shock fade, the network slowly recovers its baseline modular configuration. The adjacency matrices surrounding the plots show the block and in-block nested structure of the bipartite network immediately before (top- and bottom-left panels, respectively) the onset of the perturbation, and the nested and in-block nested arrangement some time after (top- and bottom-right panels, respectively).

This subtle insight, which stems from the model, reframes the empirical findings presented above. Indeed, the information network is not swapping between two radically different architectures –often even antagonistic [65, 151]–, but rather fluctuating from mesoscale to macroscale nested arrangements. To quantify them, $\mathcal{N}$ is not a suitable tool, because it is designed to capture nestedness at the global scale only. For this reason, we resort to in-block nestedness $\mathcal{I}$ [53, 118, 151], which generalizes $\mathcal{N}$. On the one hand, when nestedness emerges at the global scale (one block, $B = 1$), then we have that $\mathcal{I} = \mathcal{N}$. On the other hand, when the network presents several blocks ($B > 1$), each one arranged in a nested manner, then $\mathcal{I} > \mathcal{N}$.

It makes sense now to revisit the previous numerical and empirical results, now through the lens of in-block nestedness. Figure 5.2 (second panels in (a) and (b)), and Figure 5.7 (bottom panels in (a) and (b)) monitor the relative size of the largest (N_{B_1}/N) and second largest (N_{B_2}/N) nested blocks. In both empirical and numerical cases, we observe that nearly-perfect consensus is reached at different moments ($N_{B_1}/N \approx 1$), while a fragmented public sphere dominates most of the time. The relative size of the second largest nested block (N_{B_2}/N) allows for an easier interpretation of the level of consensus reached at each time. The general character of this fluctuating nested multiscale organization over the $\Omega_m - \Omega_c$ is confirmed in Appendix B, Section B.4.2.

Our framework allows to explain the puzzling transition between partial and global consensus. A fast re-organization from modular (nested) to nested (modular) architectures seems paradoxical and hard to achieve. Nevertheless, the system can swiftly adapt to any state of collective attention through an intermediary arrangement that combines the structural signature of visibility maximization with the existence of a fragmented public sphere.

5.6 Effects at the microscopic level

Untill now, we have shown that the proposed model is able to reproduce the structural macro- and mesoscale fluctuations observed in the empirical data. In this section, we want to connect our empirical observations with the model at the microscopic level. Specifically, we attempt to perform a comparison –even if qualitative– between the model and the data, by exploiting the concept of abundance. As mentioned above, the translation of the concept of abundance to the online communication context can be thought of as the number of times an item is present on screens. With a language abuse, this is tantamount to the number of individuals (e.g. hashtag instances) that build up the species (e.g. *t*he hashtag). Following this line of reasoning, for the empirical data on the hashtags' side, we can track the hashtag usage frequency over time, as a proxy for hashtag abundance from the model.

We compare the evolution of such abundance in the model and the data (Spanish and Nepal datasets). Top-left plot of Figure 5.8 shows, from our numerical simulations, the changes in abundances of the hashtags over time, identifying with a colour the topic they

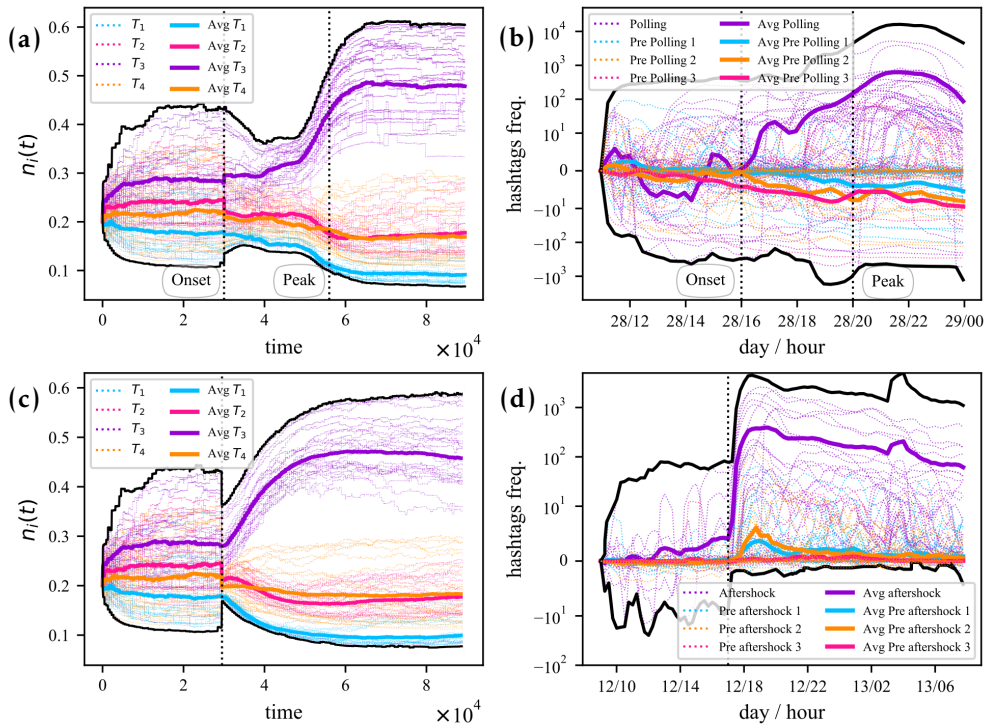


Figure 5.8: **Evolution of abundances for a 4-topic information ecosystem.** Plots in the top row correspond to synthetic (a) and empirical (b) expected events. For the numerical simulations, hashtags in T_3 (purple) begin a smooth increase in abundance at the event onset, which becomes steeper as the peak of the event approaches. Hashtags in other topics (blue, orange, fuchsia) experiment a slow decline. The same happens for the Spanish election day to the right, although admittedly with fluctuations. In this case, each colour corresponds to different communities, as detected from the networks maximising Q. Plots in the bottom row correspond to synthetic (c) and empirical (d) unexpected events. Except for the abruptness in the increase of the purple hashtags in the Nepal dataset (much faster than its synthetic counterpart), the similarities are clear to the naked eye. Remarkably, all four panels evidence that, at the microscale, a sufficiently strong perturbation impedes the system to recover the pre-event state, i.e. the system has achieved a new stable state. This result contrasts with the structural elasticity observed at the meso- and macroscale, in which the system remains within a narrow set of possible arrangements.

are ascribed to. We observe that, prior to the event, the abundances of the hashtags are distributed rather uniformly within a narrow range. After the onset of the expected event, however, the abundance of the hashtags in topic $T3$ (the one to which users' attention is shifted to) begins to increase. In the time range $3 \times 10^4 < t < 5.5 \times 10^4$ we observe a clear separation between the hashtags from $T3$ with respect to the ones from the other topics. In the simulations mimicking expected events, the artificial shock peaks at $t = 5.5 \times 10^4$. Slightly before that time, hashtags in $T3$ witness an even stronger increase up to $t = 6 \times 10^4$ (that is, beyond the peak time). After that, the system stabilises

and appears to be unable to bounce back to the original, quite uniform abundances.

The top-right plot of Fig. 5.8 shows the usage frequency of actual hashtags over time in the Spanish dataset, where events are known in advance (in this case, election day on April 28). Adapting the logic of the model to empirical data, we show the trajectories of a group of hashtags which belong to 4 different communities, the largest ones shortly before (light blue, orange, fuchsia) and at the time the ballots were closed (violet). As in its model counterpart, the vertical lines show the buildup of conversations ahead of the results (around 4pm, "event onset" tag), and the electoral schools closing time (8pm, "event peak" tag). Overall, we observe a striking qualitative agreement between the simulated hashtags abundance (model) and the hashtag frequencies (data). Until 4pm, all 4 communities present a rather flat and uniform activity (note the logarithmic scale: apparently large fluctuations, e.g. between 12pm and 2pm, imply frequency changes below 10). In the period 4pm-6pm, the behaviour of the violet subset of hashtags resembles that of the hashtags of T_3 when the event occurs (slow but steady separation from the other hashtags, with a frequency increase between 10^1 and 10^2); and also a more pronounced boost in the period 6pm-10pm (i.e. 2 hours before and after the event peak). The violet subset of hashtags clearly dominates the scenario even at midnight, and starts an expected decline as conversations mostly halt during the late night period. On the other hand, the subset of hashtags from the pre-debate stage (following T_1, T_2, T_4 in the model) present moderate decreases before 4pm, and losses are stronger after that time (especially light blue and orange topics).

For a complete picture we study as well an unexpected event. The bottom panels of Fig. 5.8 represent the evolution of abundances in an artificial setting with an unexpected event happening at $t = 3 \times 10^4$ (left); and the evolution of hashtag frequencies around the time of Nepal's earthquake main aftershock (May 12, around 5pm). Similar to its "expected" counterpart, our numerical experiments on the left show a separation of the violet hashtags in T_3, with slight decreases of the other topics T_1, T_2 and T_4. The system also appears to be unable to return to the pre-event stage, and so the only obvious difference is that the separation occurs in an abrupt way. On the right, we see the evolution of the frequencies of hashtags that belong to four of the largest communities detected in the data, slightly before (light blue, orange, fuchsia) and right after the aftershock (violet). Clearly, hashtags in the violet community present a sudden increase, followed by a very slow decrease resembling the one observed for $t > 6 \times 10^4$ in the left panel. Given the international impact of the earthquake in Nepal, there is not a decay during the night period.

These two examples extend the meso- and macroscale connections between data and model to the microscale. Furthermore, they provide a different perspective of our approach with regard to the memory of the system, and the trace that exceptional events leave behind. From the meso- and macroscale, it is still valid to say that the system is trapped in a narrow set of structural configurations (namely, nested arrangements

with only one or several blocks): this explains our use of the term "elastic". And yet, structural elasticity does not imply that the dynamical states of the system remain the same. Strong enough perturbations push the system away from its present stable state towards a new one. This apparent contradiction –structural persistence against dynamical variation– is intriguing, and demands further investigation.

5.7 Summary

The transit from a secular hierarchical management of public information to a decentralised and fragmentary scenario calls for a new vision in which the relevant drivers are identified: competition for cognitive resources, mutualistic exploitation of content, co-adaptation of users' and memes' visibility, and environmental conditions. So far, incursions in such ecological mindset have been sparse [49, 160, 161, 163]. In this Chapter, going beyond a simple metaphoric interpretation, we prove that an ecological framework –with explicit use of competitive and mutualistic interactions as drivers of information dynamics– is a powerful tool to describe the evolution of information ecosystems. Indeed, although simple neutral models may account for emergent patterns in the popularity distribution [160, 161], we show that our non-neutral, niche-based population dynamics model can successfully explain the complex interplay between users-memes interactions, attentional niches and environmental shocks. In particular, we show in spite all this complexity, the underlying architecture of the users-memes interaction in information ecosystems, apparently frenetic and noisy, actually evolves towards emergent patterns, reminiscent of those found in natural ecosystems [86, 187]. In addition, we show that the such systems are strucurally elastic, i.e., fluctuating from modular to nested architecture as a response to environmental perturbations (*e*.g. extraordinary events) [49]. Furthermore, our model predicts –and the data confirm– that the users' struggle for visibility induces a re-equilibration of the network towards a very constrained organization: the emergence of meso- and macroscale nested arrangements. Finally, we provide some results connecting the empirical and numerical observations at the microscale, that suggests that environmental shocks may leave a trace on the dynamical states of the system.

Part IV

Conclusion and future directions

Chapter 6

Conclusions and future work

The thesis provides a combination of empirical work, analytical development, and mathematical modeling to shed light on the intertwined nature between structure and dynamics in complex networks. The main research focus was to perform structural network analysis to a variety of systems, at different scales of organization. Our goal was to explore the relationship between multiple architectural arrangements, to unravel which are the mechanisms that allow the emergence of certain architectural patterns, and to explore how the transitions between these structural configurations occur. Concretely, we performed a mixture of empirical, numerical, and analytical work, along with theoretical modeling, to the study of nestedness, modularity, and in-block nestedness.

6.1 Summary of contributions

We can summarize the results presented in this thesis by grouping them according to the two central parts that composed this work, and that we employed to address our research objectives:

1. The second part of the thesis was dedicated to the development of a coherent methodology that tackles the plausible co-existence or combination of nested and modular patterns within the same system, from a purely structural point of view, i.e., by adopting a strictly analytical and numerical approach. After the recent introduction of a dedicated measure for the detection of in-block nested patterns, we first performed an in-depth examination of the inherent limitations of this measure, before disentangling how it relates with nestedness and modularity measures. We started our work in Chapter 2, by providing empirical, analytical, and numerical evidence that the in-block nestedness function lacks a modularity-like resolution limit. We have performed an empirical exploration, following a similar approach to the one in [102], that allowed us to assess to what extent the networks could be recursively split into smaller and smaller blocks. The results from

this exploration served as informal evidence that if the in-block nestedness function exhibits a resolution limit, the effect is milder when compared to modularity. Next, we provided analytical proof that, at least in an idealized setting in which the different blocks are connected through a single link, the in-block nestedness function does not have a modularity-like resolution limit. We concluded the Chapter with a numerical study over a large parameter space with varying network size and inter-block connectivity, that generalizes and confirms our analytical insights.

Moving on, in Chapter 3 we have quantified, empirically numerically and analytically, the relationship between nestedness (at the macroscale), and modularity and in-block nestedness (at the mesoscale) structural organizations. We first performed extensive numerical experiments over a rich ensemble of synthetic unipartite networks covering a wide range of parameters from our network generation model. Our results showed that high values of modularity and nestedness never overlap. In fact, we observe that, as one growth the other one declines, while in-block nestedness is able to maintain high values for networks that either highly modular or highly nested. Afterward, we demonstrated analytically that nestedness imposes bounds on modularity, with exact results in idealized scenarios, in both uni- and bipartite configurations. Specifically, we showed that nestedness and modularity are antagonistic architectures in certain settings. Furthermore, we analytically proved that in-block nestedness provides a natural combination between nested and modular networks, taking structural properties of both. This results offer an explanation to the inconclusiveness of past empirical studies regarding the coexistence of nestedness and modularity within a single network. Our findings pave the way to future research that aims to clarify, from a richer perspective, the role of one or more structural patterns in the assembly and evolution of networked systems.

2. In the third part of the thesis, on the other hand, we shifted our focus towards the investigation of some of the possible mechanisms that enable the emergence of in-block nested patterns, and the transitions from modular to nested arrangements, observed in multiple real systems, by performing extensive empirical analysis and dynamical modeling. We began this examination in Chapter 4 by analyzing a set of popular open-source projects from GitHub, through the characterization of nestedness, that allowed us to quantify and visualize the emergence of hierarchy among contributors; modularity that provided us with a way to verify to what extent a division of labor arise on the projects, and in-block nestedness that can help us to determine how projects solve the tension between these two driving forces. Our analyses have unveiled that, in general, mature OSS projects evolve evolve into internally organised blocks. Thus, the presence of workload compartmentalisation is compatible with the emergence of hierarchies, with generalists and spe-

cialists throughout a project. Furthermore, we found that the distribution of sizes of such blocks is bounded, connecting our results to the celebrated Dunbar number both in off- and on-line environments. Notwithstanding, a more evolved and structured architecture does not imply better overall performance, the nested arrangement inside blocks can hamper a project's progress, since the occasional and least committed contributors (those acting upon a small part of the code) tend to edit precisely the most generalist files, neglecting the least developed ones –a fact that has been observed from very different methodologies. Our results contribute to an understanding of how successful projects self-organise towards a modular architecture: large and complex tasks, involving hundreds (and even thousands) of files appear to be broken down, presumably for the sake of efficiency and task specialization (division of labour). Within this compartmentalization, mature projects exhibit even further organization, arranging the internal structure of subgroups in a nested way –something that is not grasped by modularity optimization only.

Finally, in Chapter 5 we showed that the architecture of the user-meme interactions in information ecosystems evolves towards a set of structural patterns that are similar those found in natural ecosystems. Particularly, we showed, through the analysis of empirical Twitter data streams, that communication networks are structurally elastic, i.e. they transition from a modular to a nested architecture, and back, as a response to environmental shocks. We then introduced an ecology-inspired modeling framework, bringing to light the precise mechanisms causing the observed dynamical reorganization. The model is founded on four ecological drivers: competition, mutualism, adaptation and enviromental conditions. We generalized an ecological optimization process in which the system's actor aim at maximizing their individual visibility, by rewiring their interactions accordingly, and included a mechanism to introduce exogenous events in the enviroment. Furthermore, our modeling framework predicts that, as a consequence of the users' struggle for visibility, the information network fluctuates across nested arrangements at different scales. The system is not oscillating between two antagonistic architectures, but rather adapting to different states of collective attention through an intermediary arrangement that combines, the existence of a fragmented attention scenario with the individual visibility maximization strategy that is characterized by the emergence of a global nested architecture. Finally, we performed a qualitative comparison between model and data ath the microscale level, by exploiting the concept of abundance, and our results suggests that exceptional event may leave a trace behind, that affects the dynamical states of the system.

6.2 Perspectives for future work

In this thesis, we have numerically and empirically studied, from macro and mesoscales perspectives, some relevant structural patterns that emerge in complex

networks. From these studies, here we identify several interesting research open questions that deserve to be addressed. Once again, we group this contributions according to the two main perspectives or objectives that were covered along this work:

1. From the first part, in which we undertook our analyses from a purely structural point of view, our results from the study of the pattern interdependencies in complex networks performed Chapter 3 open the door for two direct lines of development: one, aiming at analytical results for a more general family of networks, particularly for networks with higher connectance (lower shape parameter ξ of the probabilistic network generation model) and the inclusion of more realistic settings, such as heterogeneous community size distributions. Secondly, in-depth evaluation of the dynamical properties of in-block nested structures is needed, following the trail of works that have studied ecologically relevant processes, e.g. feasibility and local stability, which are two fundamental properties behind the persistence of an ecosystem. As mentioned previously in this document, much interest was devoted to investigating the effects of the network architecture on local stability in ecological systems [65, 67, 68, 86, 88, 89, 124]. Specifically, there is some evidence pointing at the fact that nested networks are less likely to be stable [67, 68]. Contrary to the stabilizing effects, that have been associated with modular configurations [86, 88, 89]. Notwithstanding, the effects of the network structure on feasibility have been largely overlooked. To the best of our knowledge, there only one study that suggests that nestedness is positively correlated with feasibility, but it refers to communities with a small number of species [188]. This result highlights that the simultaneous fulfillment of stability and feasibility is deeply linked to the interplay between nestedness and modularity, and points out at in-block nested structures as the optimal pattern to ensure the persistence of ecosystems.

2. Regarding part three of the thesis, our future research will expand across two dimensions: (1) to encompass richer types of empirical data, and (2) to expand our modeling framework to fully understand the conditions for the emergence in-block nested patterns in a broader type of systems. Precisely, the analysis from Chapter 4, could be complemented with weighted information. Initially, this is within reach –one should just adapt the techniques and measurements to a weighted scenario. However, the problem is not so much methodological, but semantic: the number of times that a contributor interacts with a file is not necessarily an accurate measure of the amount of information allocated in the file. Further, future research should tackle a larger and more heterogeneous set of projects, and even across different platforms. Other sampling criteria should be discussed and considered in the future, to ensure richer and more diverse project collection. Finally, two obvious lines of research are related to time-resolved

datasets, and the design of a mathematical model that can mimic the growth and evolution of the OSS projects. Regarding a temporal account of OSS projects, some challenges emerge due to the bursty development of projects in git-like environments. For example, a fixed sliding-window scheme would probably harm, rather than improve, possible insights into software development. On the modeling side, further empirical knowledge is needed to better grasp the cooperative-competitive interactions within these type of projects, which in turn determine the dynamical rules for both contributors and files.

With respect to the modeling framework introduce in Chapter 5, our findings open up an ambitious research alley along the lines of computational human ecology. In the shorter term, future efforts should attempt to better reproduce – at a more quantitave level– the microscopic dynamics of users and memes abundances before and after breaking events. These cannot be explained without including death-birth and invasion processes, which are in turn necessary to understand how influential users and viral contents emerge. Similarly, this initial proposal rules out “cultural drift"–the slower changes in the users’ topical preferences–, which leads to persistent structures and shapes communication flows. Reaching further, the tradition in theoretical ecology aimed at understanding and preventing the collapse of ecosystems can be adopted to decipher how social media and information bubbles shape our thinking, or, in the opposite direction to disrupt and break misinformation dynamics and polarization. Related to this, we foresee as well a connection between the extensive research on stability and resilience in natural ecosystems, and their informational counterparts. In this sense, we are convinced that such interchange of techniques and models could be beneficial for theoretical ecology too as it will allow to test theories and methodologies in a more controlled, data-rich environment with faster time scale at play.

Appendix A

Additional results: Macro- and mesoscale pattern interdependencies in complex networks

The appendix provides a detailed explanation regarding the construction and reading of the ternary plots, a complementary formulation for the two particular cases of the ring of star graphs $G^{\star}$ when $B = 1$ and $B = 2$ along with some additional results exploring the effect of noise parameters of the probabilistic network generation model, and some complementary plots that help to strengthen the main conclusions obtained from the results presented in Section 3.4 of Chapter 3.

A.1 Ternary plot: Dominance regions

A ternary plot is a three-variable diagram on which each point represents the proportions between three variables. Given the values of the variables, $\mathcal{N}$, $\mathcal{I}$ and Q, the proportions that are eventually represented in the plot are obtained as $f_{\mathcal{N}} = \mathcal{N}^{-1}(\mathcal{N} + \mathcal{I} + Q)$, $f_{\mathcal{I}} = \mathcal{I}^{-1}(\mathcal{N} + \mathcal{I} + Q)$ and $f_Q = Q^{-1}(\mathcal{N} + \mathcal{I} + Q)$. In Fig. A.1 the bottom axis represents $\mathcal{N}$ and its rigth vertex perfectly nested networks ($f_{\mathcal{N}} = 1$). Other values of $f_{\mathcal{N}}$ are indicated by the dashed blue lines in direction $\nearrow$ of the triangle. Rigth axis represents f_Q and the top vertex purely modular networks ($f_Q = 1$). Other f_Q values correspond to horizontal dashed blue lines. Finally, the left axis represents $f_{\mathcal{I}}$ and the left vertex networks that are purely nested ($f_{\mathcal{I}} = 1$). Other $f_{\mathcal{N}}$ values are indicated by lines in direction $\searrow$ of the triangle. Additionally, the black dashed lines delimit dominance regions, which are highlighted in different grey tones for variable pairs in panel a-c and in triads for in panel d. Each dominance region spots (by pairs) which is the dominating structural pattern. For ease of identification the dominant structure is also indicated close to the plot axis. Points over the line of dominance equilibrium in panels a to c correspond to

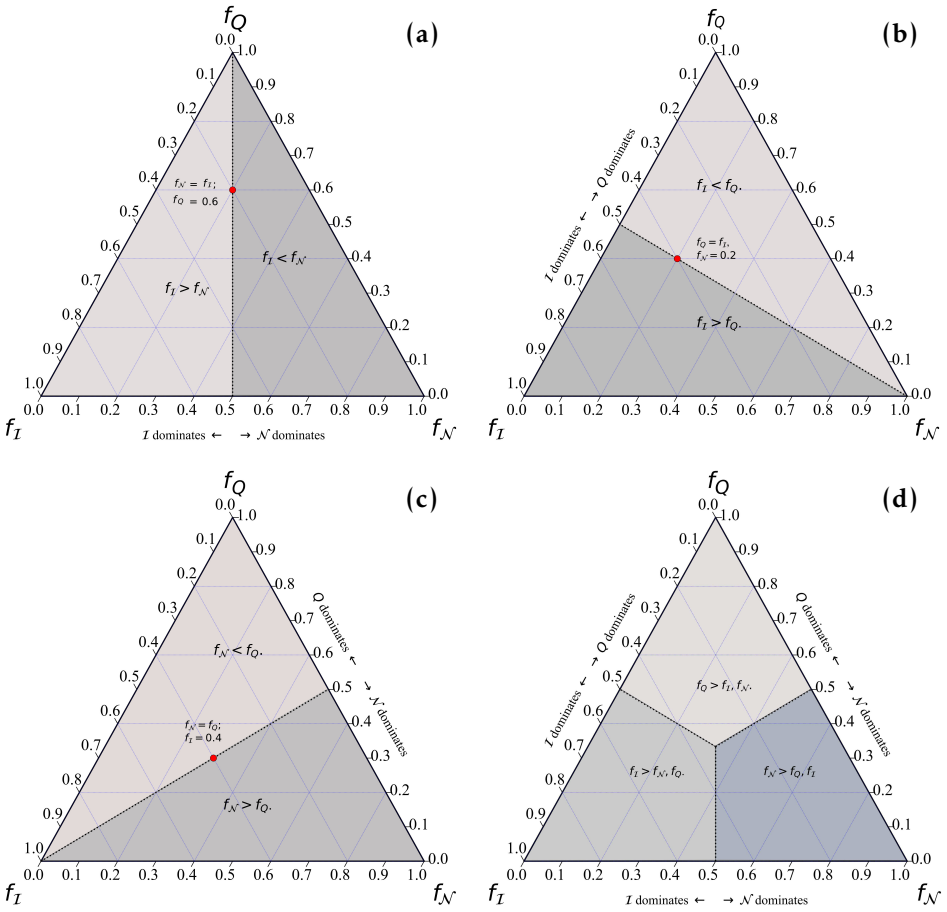


Figure A.1: Representation of the three variables in the ternary plot showing exemplar points with different proportions and evincing the dominance regions. Panel (a) to (c) delimit the dominance regions (in pairs) between nestedness, modularity and in-block nestedness structures. Panel (d) delimits these regions jointly considering all the variables.

points where the contribution of the two contrasted variables is equivalent.

A.2 Noise sensitivy test for weak communities

Figure A.2 explains the decision, stated in Section 3.1 of Chapter 3, to restrict the levels of intra- and inter-block noise (i.e. $p \leq 0.6$, $\mu \leq 0.6$) in the synthetic benchmark. We wanted to introduce a considerable level of noise while guaranteeing that some identifiable pattern was still present. To this aim, we have followed the concept of weak modularity as introduced in Radicchi *et* al. (ref. [15] of the main text). In that work, authors define weak modularity as a network partition in which modules have more internal than external links, but that is not true for each and every node in those modules. This informal notion guided our limitation of μ noise: we stop generating networks beyond $\mu = 0.6$ because the imposed community structure does not even comply with the

weak modularity condition (and thus an algorithm could hardly detect it). To prove this point, we show in Fig. A.2 that, for $\mu = 0.55$, still $\sim 65\%$ of the generated networks are weakly modular; by the time $\mu = 0.6$, only 40% of them fulfils the condition.

It should also be noted that, even completely random networks exhibit a remarkable level of modularity Q [105]. Since, by definition, $\mathcal{I} = \mathcal{N} = 0$ in such situation, any completely random realisation of the benchmark will induce a (false) modular-dominant network in the ternary plot –as, indeed, it already happens for high levels of p and μ, see the corresponding panels in Fig. 3.2.

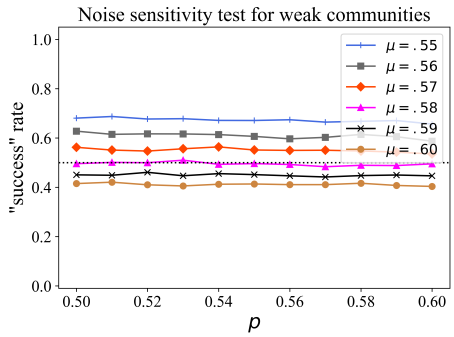


Figure A.2: Noise sensitivity test: Rate of networks that fulfill the condition for weak communities for different combinations of the noise parameters p and μ.

A.3 Analytic expression for $\mathcal{N}_{G^\star}$,$Q_{G^\star}$, $\mathcal{I}_{G^\star}$ along F_2 for the cases $B = 1$ and $B = 2$

A.3.1 Nestedness: $\mathcal{N}_{G^\star}$ for $B = 1$ and $B = 2$

The computation of the pair overlap for the evaluation of nestedness when $B = 1$ requires only the following terms: the pair overlap of a generalist node (the center of each star subgraph), g, with the specialist nodes s which is $O_{gs}/k_s = 0$; and the pair overlap between all the specialists nodes $O_{ss}/k_s = 1$, the degree of the generalist node is $k_g = N_B - 1$ and the null model corrections $\langle O_{gs}\rangle = k_g k_s/BN_B = (N_B - 1)/BN_B$ and $\langle O_{ss}\rangle = k_s k_s/BN_B = 1/BN_B$.

Furthermore, for the case $B = 2$ we have to take into account the change on the degree of the generalist node $k_g = N_B$ and additional terms such as: the pair overlap between the two generalists O_{gg}/k_g, the pair overlap between a generalist with the specialist from the other community $O_{gs_{out}}/k_s$, and the pair overlap between a specialist with the specialists from the other community$O_{ss_{out}}/k_s$.

B = 1	**B = 2**
$\mathcal{N}_{G^\star} = \frac{(N_B-2)(N_B-1)}{N_B^2} - \frac{2(N_B-1)^2}{N_B^2(N_B-1)}$ (A.1)	$\mathcal{N}_{G^\star} = \frac{BN_B^3 - BN_B^2 - 6N_B^2 + 8N_B - 3}{BN_B^2(BN_B-1)}$ (A.2)

A.3.2 Modularity: $Q_{G^\star}$ for $B = 1$ and $B = 2$

Starting from the equation for modularity expressed as sum over the communities (Eq. 1.13), we obtain the total number of links in the network and the number of links per community for a single star graph $N_B - 1$, and the sum of the degrees of the nodes in the community $d_c = 2(N_B - 1)$.

Moreover, when $B = 2$ the total number of links changes to $L = B(N_B - 1) + 1$ and the sum of the degrees of the nodes in the community is $d_c = 2(N_B - 1) + 1$. So we obtain

B = 1	**B = 2**
$Q_{G^\star} = B\left[\frac{(N_B-1)}{(N_B-1)} - \left(\frac{2(N_B-1)}{2(N_B-1)}\right)^2\right] = 0,$ (A.3)	$Q_{G^\star} = \left[\frac{(2N_B-2)}{2N_B-1} - \frac{1}{2}\right].$ (A.4)

A.3.3 In-block nestedness: $\mathcal{I}_{G^\star}$ for $B = 1$ and $B = 2$

Once again, we know that for $B = 1$, we will have only two contributing terms to our sum; the pair overlap between specialists (s) nodes and the pair overlap of the generalist (g) node with the specialists. Additionally, we know that for this case the degree of the generalist node is $k_G = N_B - 1$ and the rest of the terms are: the number of specialists nodes $N_s = (N_B - 1)$, the null model corrections $\langle O_{g,s}\rangle = k_g k_g / BN_B = (N_B - 1)/BN_B$ and $\langle O_{s,s}\rangle = k_s k_s / BN_B = 1/BN_B$, an the size of the communities is $C = N_B$.

Finally, for $B = 2$ we have that the degree of the generalist node is $k_G = N_B$. Substituting this term for each specific case, we obtain

B = 1	**B = 2**
$\mathcal{I}_{G^\star} = \frac{2}{N_B}\left\{\left[-\frac{(N_B-1)}{N_B}\right] + \left[\frac{(N_B-1)(N_B-2)}{2N_B}\right]\right\},$ (A.5)	$\mathcal{I}_{G^\star} = \frac{2}{N_B}\left\{\left[-\frac{1}{N_B}\right] + \left[\frac{(2N_B-1)(N_B-2)}{4N_B}\right]\right\}.$ (A.6)

A.4 Complementary figures: Approximate constraints $\mathcal{N}$, Q and $\mathcal{I}$

Figure A.3 shows the values of Q plotted against $\mathcal{N}$, for the all the generated synthetic networks ($\sim 2 \times 10^5$). The corresponding upper and lower bounds were plotted on top. The color bar, in each case, indicates the values of the respective parameters of the probabilistic network generation model (number of blocks (B), shape parameter (ξ), intra-block (p) and inter-block noise (μ), respectively). We observe that the modularity values, Q, above the upper bound correspond to networks with a single community $B = 1$ and perfectly nested structure, $p = 0$.

Fig. A.4 shows the values of Q against $\mathcal{I}$. As stated in Chapter 3, Section 3.4 the results from this figure corroborate that Q and $\mathcal{I}$ can coexist, i.e., there is no clear map between them.

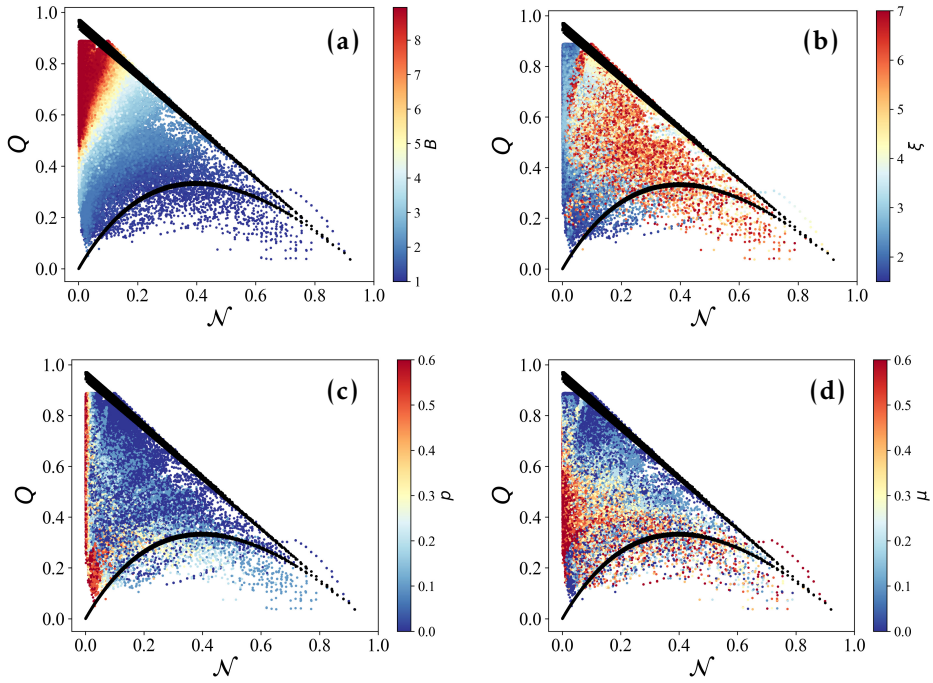


Figure A.3: Optimized values of Q plotted against $\mathcal{N}$, for the generated networks. The values of the corresponding upper and lower bounds were plotted on top (black dots). The color bar indicates the value of the respective parameters of the probabilistic network generation model (number of blocks (B), shape parameter (ξ), intra-block (p) and inter-block noise (μ), respectively).

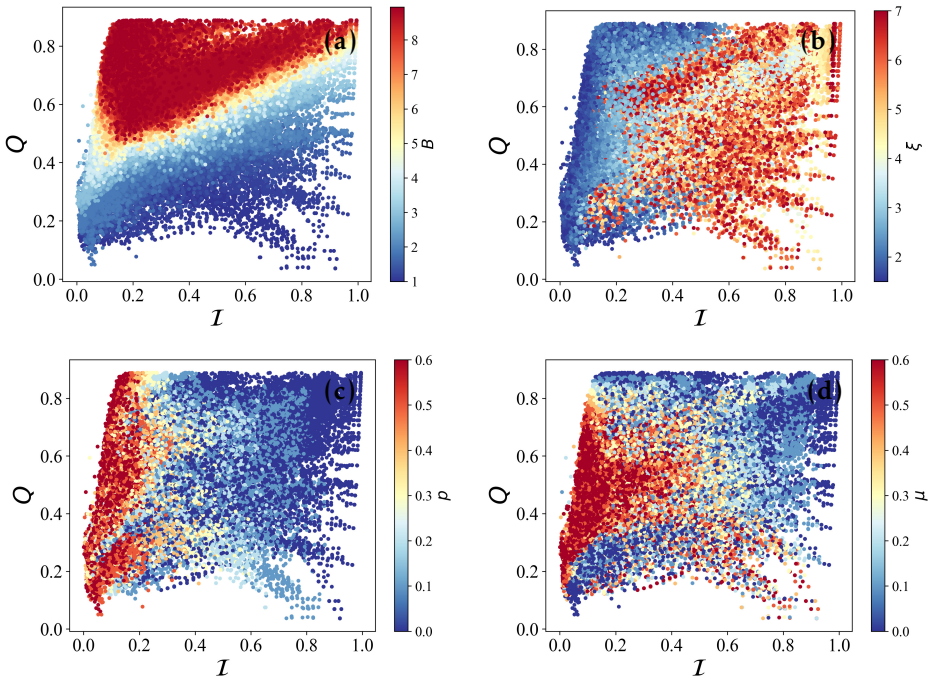


Figure A.4: Optimized values of Q plotted against the optimized values of $\mathcal{I}$, for the generated networks. The color bar indicates the value of the respective parameters of the probabilistic network generation model. Panel (a) shows the results with respect to the number of blocks. Panel (b) corresponds to the shape parameter ξ. Panels (c) and (d) corresponds to the noise parameters p and μ, respectively.

Appendix B

Additional results. Structural Elasticity in Online Communication Networks: an ecological approach

In this Appendix we provide additional details and results for other four Twitter datasets, along with some complementary results exploring the link density of the empirical data and extra numerical results.

B.1 Additional datasets

We considered four additional events of different nature: the 2012 UEFA European Football Championship, the 2014 Catalan self-determination referendum, the 2015 Charlie Hebdo Shooting and the 2014 Hong Kong streets protests. All the datasets excepting the ones from the Spanish general elections, presented in Chapter 5, and the Catalan self-determination referendum were collected by Zubiaga A. in [173].

1. **Catalan self-determination referendum (Nov 2014):** This dataset corresponds to the Citizen's Participation Process on the Political Future of Catalonia, a popular consultation about the process of independence of Catalonia from the Spanish Kingdom. The consultation was held on Sunday, 9 November 2014, after the approval decree was signed by the president of Catalonia on September 27 of the same year. The dataset contains 220,364 unique tweets containing at least one hashtag, with a total 18,116 unique hashtags and 78,270 users, ranging from September 1st to November 13 of 2014. Similarly to the Spanish election dataset, this dataset was collected by selecting all the tweets containing at least one of a preselected set of ≈ 70 hashtags and ≈ 50 Twitter accounts related to the referendum process and the Catalan independence movement.

2. **European football championship (2012):** Afterward, we considered the 2012

UEFA Football Championship, an European championship for men's national football teams. The tournament was held between 2 June and 1 July of 2012, and co-hosted by Poland and Ukraine. The observation period started a day before of the quarter-finals, on 19 June and lasted until the 4th of July, right after the final game. It contains 3,907,418 unique tweets containing at least one hashtag, with a total 147,646 unique hashtags and 1,325,631 users. This dataset was collected by selecting all the tweets containing the hashtag *#euro2012*.

3. **Hong Kong protests (Sept-Oct 2014):** Another dataset considered in our stude corresponds to a series of streets protests that took place in Hong Kong from September to December 2014. The protests, are often referred to as the Umbrella Movement or Occupy movement. The protests were initiated after a proposal from the Standing Committee of the National People's Congress to reform the electoral law. The dataset contains 826,194 unique tweets containing at least one hashtag, with a total 30,105 unique hashtags and 239,432 users. The observation period started on 27th of September, right after the protests escalated, resulting in several people detained, until October 10. The dataset was collected by selecting all the tweets containing at least one of the following hashtags or keywords: *#hongkong*, *#umbrellamovement*, *#occupycentral*, *#hongkongprotests*, *#occupyhongkong*.

4. **Charlie Hebdo Shooting (Jan 2015):** The last dataset considered for analysis also corresponds to an unexpected event, specifically, the shooting perpetrated at the offices of the french magazine Charlie Hebdo, on January 7 2015. On the morning of January 7 of 2015, two heavily armed brothers forced their entry into the magazine offices, killing 12 people and injuring 11 more. The dataset contains 6,002,087 unique tweets containing at least one hashtag, with a total of 102,799 unique hashtags and 2,001,826 users. The observation period started on the 8th of January, right after the shooting took place and lasted until the 10th of January, after the two main suspects were killed. The dataset was collected by selecting all the tweets containing at least one of the following hashtags or keywords: *#jesuischarlie*, *#charliehebdo*, *charlie hebdo paris*.

Table B.1: Summary of the additional datasets analyzed.

Dataset	Data length	Total days	Tweets	Users	Hashtags
2014 Catalan referendum	Sep 2 - Nov 12	10	220,364	78,270	18,116
2012 UEFA championship	Jun 19 - July 4	15	3,907,418	1,325,631	147,646
2014 Hong Kong protests	Sep 27 - Oct 7	10	826,194	239,432	30,105
2015 Charlie Hebdo shooting	Jan 8-9	2	6,002,087	2,001,826	102,799

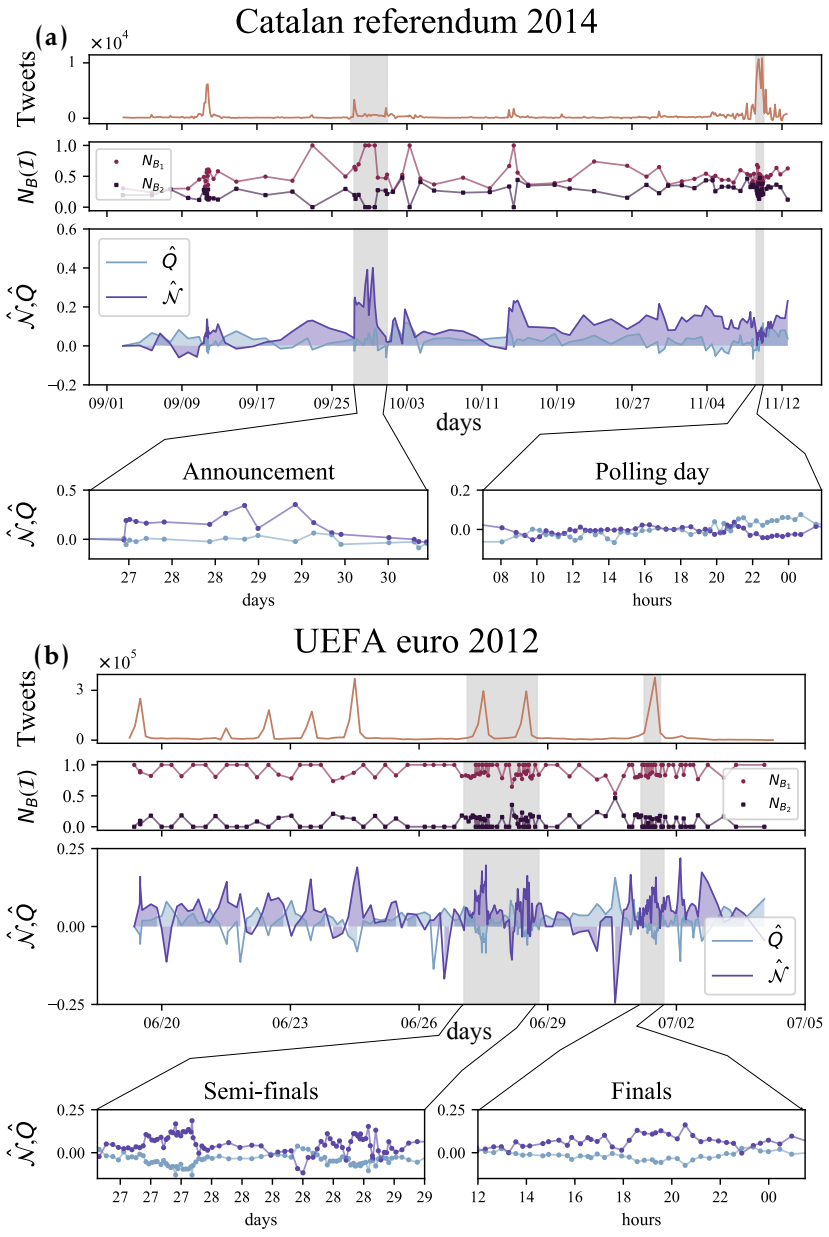
B.2 Complementary empirical results

In this section we present the results of the temporal structural analysis of the four additional additional empirical datasets. For the sake of consistency, once again, we analyse these datasets monitoring the system's modularity [33] (Q), nestedness [41, 43, 45] ($\mathcal{N}$) and in-block nestedness ($\mathcal{I}$) [53].

Figure B.1 shows the evolution of Q and $\mathcal{N}$ for the Catalan self-determination referendum from 2014, the 2012 European football championship, the 2014 Hong Kong street protests and the 2015 Charlie Hebdo Magazine shooting. The duration of the snapshot was adjusted to provide a better visualization of the structural transitions during the different events, and highlighted the location of the events in the main panels of each plot. Some of these events are pointed out in the pair of insets in each figure.

Overall, we observe that for all different datasets the behaviour is in qualitative agreement with the ones presented in Chapter 5. First, the anticorrelated behaviour between global nestedness and modularity is preserved. Further, in each case, regardless the nature of the different datasets, we observe a smooth transition into self-similar nested arrangements, which develop in accordance to the level of fragmentation of the surrounding conditions, i.e this transition is linked to external events (second row in all panels). The different datasets, regardless of their nature, lie along the lines of the different classes of collective attention described in Lehmann *et* al. [185]. The highly fluctuating pattern in Fig. B.1(b), corresponding to the UEFA championship, is due to the periodicity in which football games happen throughout the competition, with a slowdown by the end of the period when only the semifinal and final game are left.

Although mentioned in Section B.1, it worth highlighting that different data acquisition procedures employed to build the analysed datasets. The Spanish elections and Catalan referendum datasets were collected from a rich collection of hashtags and keywords that were manually chosen following the evolution of the event, even introducing new hashtags –or keywords– as the event unfolded. In contrast, the rest of the datasets were collected from a small set of hashtags (often just one) [173], resulting in the presence of "super"-generalist memes during all the stages of the discussion. Regardless of the possible biases induced by the presence of these "super"-generalist memes in some of the datasets, many (possibly most) important hashtags emerging at later stages are captured as well, since they tend to co-occur with the original chosen keyword. Thus, we were able to capture the different states of collective attention, from fragmented to global stages of public consensus.


Catalan referendum 2014
(a)
Tweets
$N_B(I)$
$\hat{\mathcal{N}},\hat{Q}$
N_{B_1}
N_{B_2}
$\hat{Q}$
$\hat{\mathcal{N}}$
09/01
09/09
09/17
09/25
10/03
10/11
10/19
10/27
11/04
11/12
days
Announcement
Polling day
days
hours
UEFA euro 2012
(b)
Tweets
$N_B(I)$
$\hat{\mathcal{N}},\hat{Q}$
06/20
06/23
06/26
06/29
07/02
07/05
days
Semi-finals
Finals
days
hours

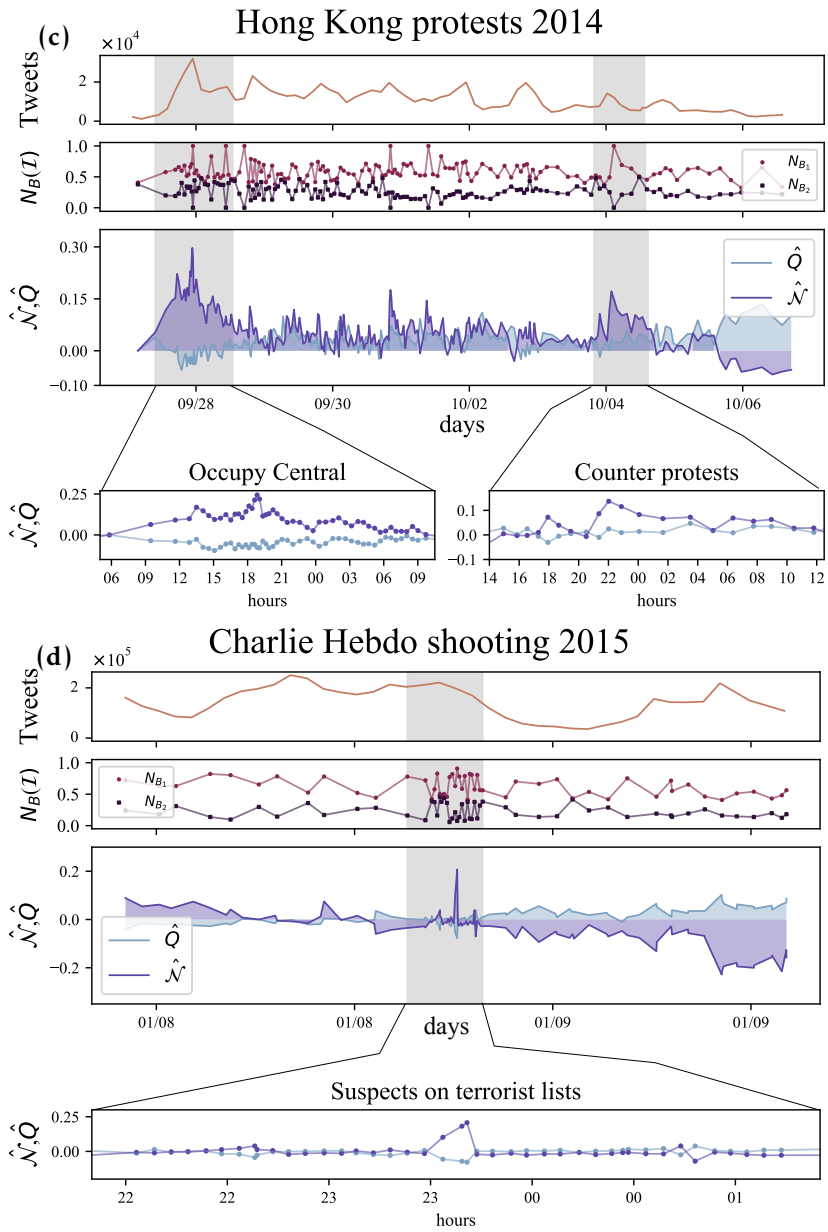


Figure B.1: **Structural measures over time for four different datasets.** Panel (a) corresponds to the 2014 Catalan self-determination referendum, panel (b) corresponds to the 2012 UEFA Football Championship, panel (c) corresponds to a series of streets protests that took place in Hong Kong in 2014. Finally, panel (d) correspond to the Charlie Hebdo shooting on 2015. In accordance with empirical results presented in Chapter 5, here we observe how a block organization dominates the system, reflecting the separate interests of users, until external events induce large-scale attention shifts, which rearrange completely the observed architecture towards a macroscale nested pattern. Once again, we highlight specific time windows in each dataset with some identifiable event happening in them.

B.3 Connectance of empirical networks

The connectance C of a network, is the percentage of existing interactions over all possible ones ($N_U \cdot N_H$). In our model, we wanted to build synthetic networks with approximately the same connectance of those emprirical ones –in order of magnitude– . Figure B.2, shows the values of network connectivity as a function of the number of nodes ($N_U + N_H$, in our case), for empirical matrices created applying the construction process described in Chapter 5, Section 5.2.2, at different N_U thresholds. All the results presented in Chapter 5 were obtained for connectance $C \propto 10^{-2}$, which provided a better match with the empirical connectance for network with $N_U = 100$, see blue triangles in Fig. B.2.

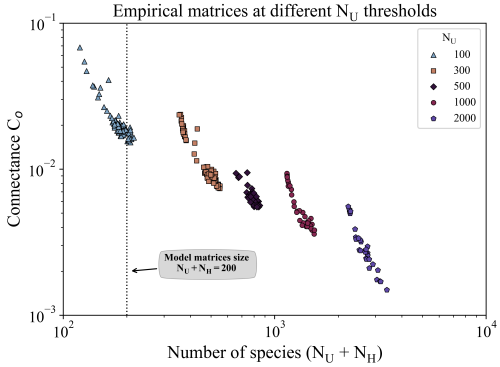


Figure B.2: Connectivity as a function of the number of species ($N_U + N_H$) for the empirical networks, at different N_U thresholds. Note the log-log scale.

B.4 Complementary figures:

B.4.1 Post- external event species survival

Here, we present the results for the survival rate of species after the introduction of an sudden external event.

As shown in Fig. B.3, for $\lambda = 0$ a considerable number of especies goes extinct by the end of the simulation time. This results is not surprising, since at the onset of the event the single topic configuration increases the competition among species. The λ parameter helps to balance the intense competition between the species, therefore, we observe a decrese on the amount of extinctions as λ goes higher.

B.4.2 Meso- and macroscale nested arrangements

At last, we explored the structural evolution of the system by means of the in-block nestedness function $\mathcal{I}$ [53]. This exploration confirms the general character of the fluctuating meso to macroscale nested organization described in Chapter 5, Section 5.5.3. Figure B.4 shows the relative size of the largest nested blocks $N_{B_1}(\mathcal{I})/N$, before (panel (a)) and after (panel (b)) the introduction of a sudden external event. Before the event,

we observe that in general, for all the parameter space, the size of the largest nested block constitutes a 25% of the whole network, approximately, i.e, the user are evenly aligned over the four predefined topics. After the event, we observed how a state of global consensus is emerging, as N_{B_1}/N increases over all the parameter space, representing more than 50% of the size of the network in most of the cases.

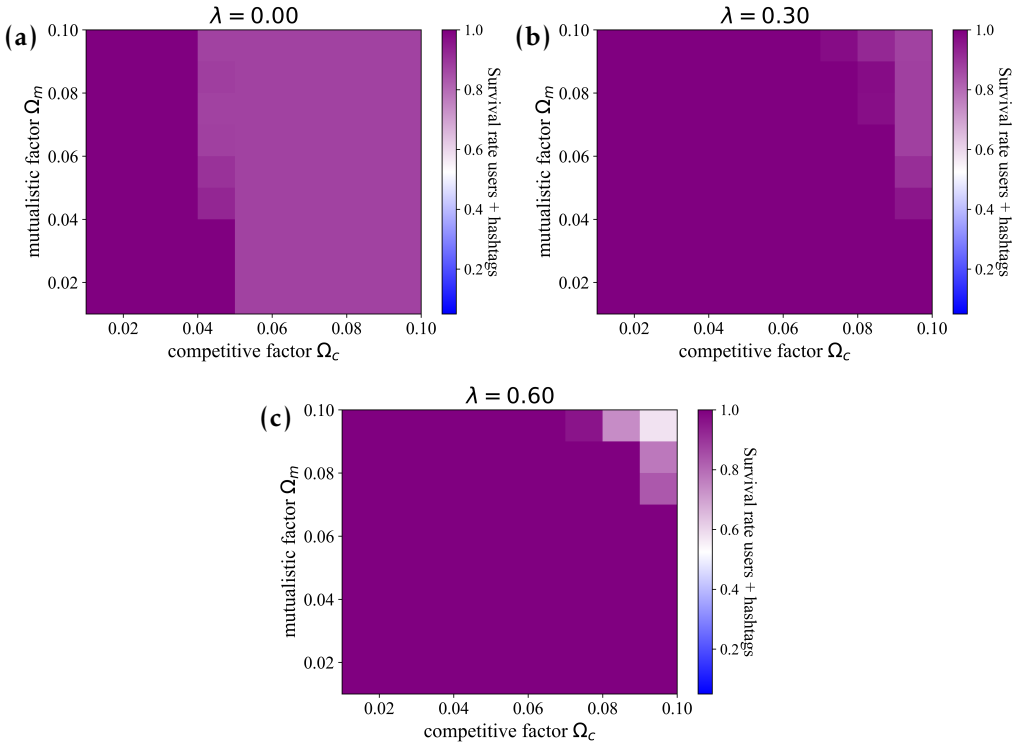


Figure B.3: **Survival rate at $t > t_E$:** two-dimensional plots in the Ω_m–Ω_c parameter space showing survival rate of the species at the end of the simulation, for different values of the inter-intra competition parameter λ.

B.5 Anti-correlated behaviour between Q and $\mathcal{N}$

In this section, we discuss in depth the observed anti-correlated behaviour between nestedness and modularity, for both the empirical data and the model's outcome.

We know from the results presented in Chapter 3, that there exists an upper bound for the co-existence of nested and modular structures, regardless of the size or the density of the network at stake. This bound implies that a highly modular structure can only "afford" a non-nested structure, and the other way around, which helps to explain the observed anti-correlated behaviour between Q and $\mathcal{N}$. Here, we statistically confirm such anti-correlated behaviour between Q and $\mathcal{N}$ by computing the Pearson correlation between both measures across time, see Table B.2. For the synthetic cases (last two rows), it is measured from $t = 1.5 \times 10^4$ onwards, to avoid the initial random fluctuations. For

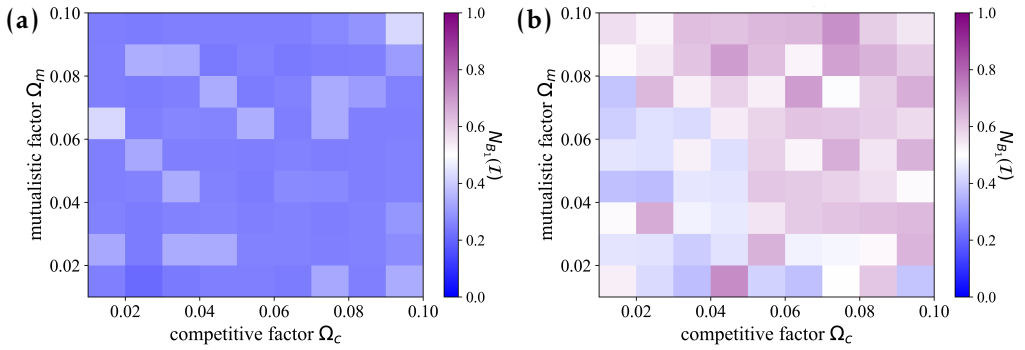


Figure B.4: **Relative size of the largest nested block** (N_{B_1}/N) **before and after the external event:** two-dimensional plots in the $\Omega_m - \Omega_c$ parameter space showing the relative size of the largest nested block before (panel (a)) and after (panel (b)) the external event with $\lambda = 0.6$.

smoother comparisons, correlation is measured over the whole period covered for the datasets, but also before/during/after the main events that we identify in Fig. 5.2, and Fig. B.1, respectively. Except for the Catalan dataset, the matching between empirical and synthetic results is remarkable, not only in the periods where the correlation is strong, but also during the pre-event stages, where, in most cases, both correlations are irrelevant, despite their opposed signs.

The mismatch in the Catalan dataset, can also be explained in terms of the upper bound described above, since such bound does not rule out other possible regimes. For example, it is quite common for both Q and $\mathcal{N}$ values in a network to be extremely low; or that both have intermediate values, which typically signals the presence of in-block nestedness. The Catalan political conflict has an associated extremely high polarization: the dataset contains both users in favour of and against a referendum, and a large fraction of Spanish users who think that a referendum should not even be discussed at all. As a consequence, the polling day, for example, contains not only tweets paying attention to the results of the (illegalised) referendum, but also many messages calling for political and legal action against the organizers, or simply appealing to political dialogue between the parts. In structural terms, all of this translates into an in-block nested structure, suggesting a sort of "partial consensus" among the different sides that participate in the conversation. Therefore, for this case, we can obtain intermediate values of both quantities (Q and $\mathcal{N}$), which explains the weak anti-correlation observed that day ($r = -0.3079$).

B.6 Statistical significance of Q and $\mathcal{N}$

To further strengthen the validity of our results, we now are focused on exploring a possible lack of statistical significance of the reported patterns Q and $\mathcal{N}$, which has been and is a controversial issue for these descriptors. Before explaining the randomization

Table B.2: Pearson coefficients for Q and $\mathcal{N}$ at different times, for both, the model and the data.

Data type	Dataset	Whole period	Pre event	Event	Post event
Empirical	2019 Spanish general elections	-0.8264	-0.2914	Debate: -0.9094 Polling: -0.7178	-0.8467
	2015 Nepal Earthquake	-0.7337	-0.26566	-0.9358	-0.74138
	2014 Catalan referendum	-0.1490	-0.1036	Diada: -0.21326 Polling: -0.3079	-0.7234
	2012 UEFA football championship	-0.7930	-0.42895	Semis: -0.8675 Finals : -0.7545	-0.8827
	2014 Hong Kong Protests	-0.5774	0.0034	Occupy central: -0.6194	-0.6180
	2015 Charlie Hebdo shooting	-0.9180	-0.4496	-0.8979	-0.8240
Numerical	"expected event"	-0.743	0.1625	-0.7611	-0.8641
	"sudden event"	-0.6651	-0.0558	-0.7390	-0.8533

procedure employed to assess the statistical significance of Q and $\mathcal{N}$, we want to stress here that, by definition, both descriptors incorporate a null model term. While this is not new for Q, whose quantification has always been in reference to a null term, it is so for the definition of $\mathcal{N}$ that we employ along this thesis, which differs from the classical ones (e.g. NODF [73]), that do not include a random expectation term; see eq. 1.9.

For the sake of simplicity, and due to computational limitations, we only perform the statistical test for the Spanish dataset. For each one of the matrices of the Spanish dataset and the synthetic matrices under the expected event, we have generated 150 randomizations in which we preserve the link density. This form of randomization amounts to considering that the overall system's activity is kept, but users lack a preference for one or another meme for communication purposes. Figure B.5 shows the results of the z-scores of the two descriptors against the ensemble of randomized matrices. The black solid lines in the plots corresponds to a $z = 2$. The dotted lines show the actual Q and $\mathcal{N}$ values on the real matrices, that are indicated by the secondary y–axis in both panels.

From Fig. B.5 (a), we observe that during the debate the measured values for Q are no longer statistically significant at the selected confidence interval. For the second event (polling), we observe an abrupt decrease in the z-scores for the measured Q, although it is still significant under the considered threshold. In the case of $\mathcal{N}$, we observe that

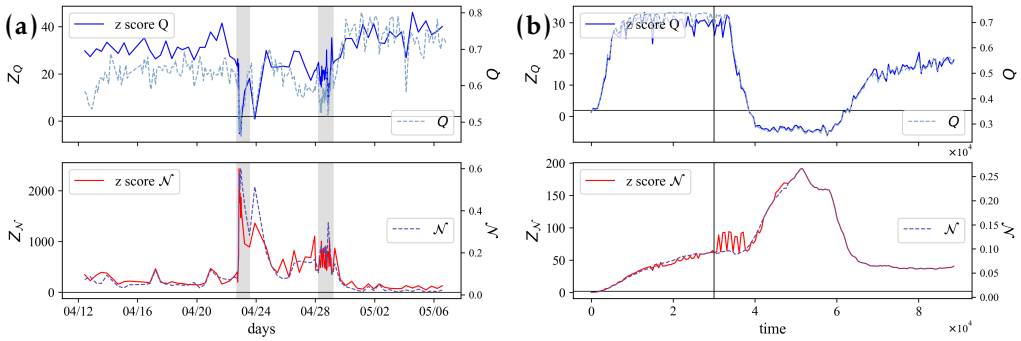


Figure B.5: **z-scores for modularity and nestedness, against an ensemble of 150 randomizations:** for the empirical case (panel (a)) and for the model's numerical simulation (panel (b)).

its values remain statistically significant for the whole dataset. Nonetheless, we find evident variances between their statistical significance during the crucial periods. In general, the z-scores for $\mathcal{N}$ are extremely high during the extreme event, despite low values (compared to the peak) outside the exogenous events. Pursuing stronger connections between findings in the data and the model, we have also analyzed, under the same scope, the outcome of our model. Results over the synthetic networks (Fig. B.5 (b)), show the same observed behaviour: during the artificially introduced event, the z-scores for Q fall below statistical significance; while $\mathcal{N}$ is significant overall the period, but with a marked surge after the exogenous introduction of an event at $t = 3 \times 10^4$, proving that, in general, the changes in Q and $\mathcal{N}$ are beyond reasonable expectation, and are thus statistically robust.

B.7 Disentangling the effects of a change in the activity

In this section, we explore in detail the effects of the activity changes on the system's properties. Our main interest in exploring this particular aspect is to avoid a possible confounding factor: the false impression that changes in activity can, on their own, explain the reported structural shifts from (to) modular to (from) nested arrangements. In other words, we intend to discard the idea that changes in the network's topology are simple by-products of changes in the activity.

Examining this aspect in more detail, we can observe that most of the increased activity during extraordinary events is due to new users entering the topic (Fig. B.6 top panel), and that these produce a very large amount of hashtags as well (Fig. B.6 bottom panel). Results from Fig. B.6 may be interpreted as a proof that the model cannot mimic the observed behaviour in the data, e.g. the number of user and hashtags is highly fluctuating in the data, while it remains constant in the model. In the following, however, we explore whether other system's quantities, such like the users' average activity or the amount of effective hashtags, are also affected by activity increases.

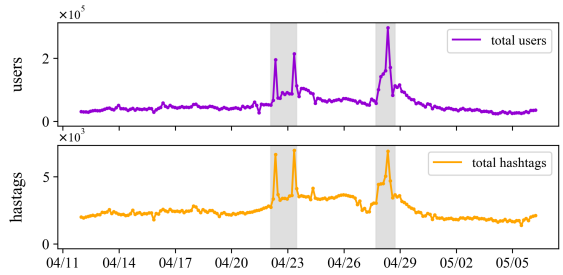


Figure B.6: Evolution of the number of users (top) and the the number of hashtags during the Spanish election cycle. Both quantities show remarkable increases during the identified exceptional events (debate, polling day).

B.7.1 Effects of activity increase on the users' and hashtags average quantities:

We start this section by exploring the effect of an activity increase from the users' perspective. Specifically, we tracked the users' average activity, i.e., the number of hashtags per user, $\langle h \rangle$, over time (Fig. B.7 (a)). We can observe that, even during an exceptional event, $\langle h \rangle$ remains relatively constant, this means that when the users' attention profiles are shifted, these do not significantly increase their activity (in terms of hashtag usage) on average, but rather start switching towards the topic on which that same activity is devoted. This finding provides a stronger link between actual data and the model for which $\langle h \rangle$ is constant by design, as can be seen from Fig. B.7 (b).

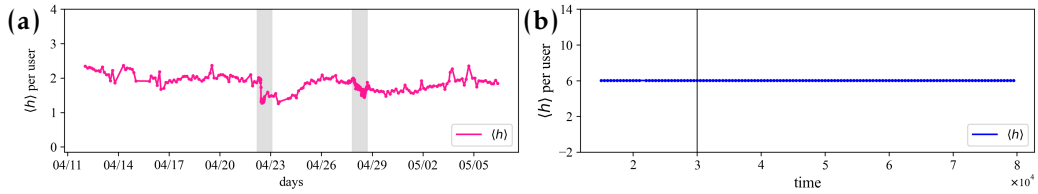


Figure B.7: **Average number of hashtags per user $\langle h \rangle$:** for the Spanish dataset (panel (a)), and for the model's numerical simulation (panel (b)).

Moving on, we have also explored the effects of an activity increase from the hashtags' side. Notably, from Fig. B.6 (bottom panel), it is clear that the absolute number of unique hashtags increases during the highlighted events, debate, and polling day, respectively. Nonetheless, when we try to quantify the minimum amount of hashtags needed to account for a large fraction of the users (99% of the users, in our case) in each time window, i.e. the diversity of the hashtags in terms of a cover set, we see that 99% of the users can be accounted for with no more than 400 hashtags, and with 100 hashtags or less during intense attention episodes. This behavior is qualitatively mimicked as well by the model, see panel (b) from Fig. B.8. We have computed this "hashtag cover-set" by applying the following iterative scheme: we count (and remove) all users who tweeted the most frequent hashtag; then we count (and remove) all users who used the

second most frequent hashtag; and so on, until we reach the desired threshold (99% of the users). Counted in this way, the hashtag coverset represents to what extent users are focused on only a few items, despite the presence of many more memes in the information system that may (or probably may not) get anyone's attention. Last but not least, these results show that, even though the model does not take into account the fluctuating behaviour of users and hashtags observed in the data, which cannot be incorporated without considering birth/invasion processes, it is still able to reproduce the evolution of *e*ffective hashtags in the system.

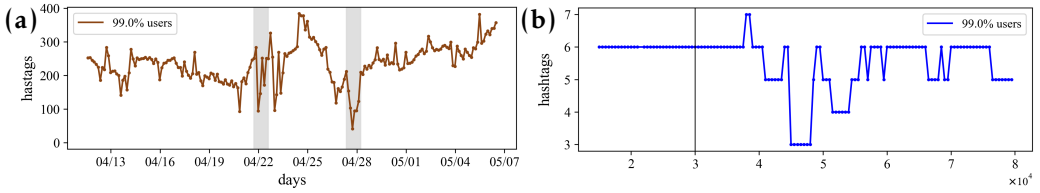


Figure B.8: **Hashtag coverset at the 99% threshold :** for the Spanish dataset (panel (a)), and for the model's numerical simulation (panel (b)).

B.7.2 Activity increase as a driver for nestedness and/or modularity:

We now want to investigate if an increase in activity can be a driving mechanism of the different structural transitions observed in the data. As explained in Chapter 3, given the mutual constraints that Q and $\mathcal{N}$ impose on each other, the growth of one implies the decline of the other. Nonetheless, the opposite is not necessarily true: a reduction in Q (due, for example, to larger density in the network) does not guarantee at all an automatic increase of nestedness. Conversely, a reduction in $\mathcal{N}$ will not imply, necessarily, a larger Q. Since we know *a* priori that an increase in activity may lead to a growth in network density, here, we want to fully address if such increment could be responsible for the increase in nestedness/decrease in modularity observed in the model and data.

We start our examination providing a simple example from the Spanish dataset: we take a snapshot of the system right before a large event occurs (debate), in which the system is clearly organized as a modular network and have a connectance (density) $C = 0.003$. Then, we start simulating an increase in activity (which translates to an increase in density), taking the system from the initial connectance to a final one of 0.005. We choose this network density, as this is precisely the connectance in the empirical data by the time we observe a maximum in nestedness (during the debate).

Figure B.9 summarises the results from the experiment. In the left panel, we see the network at its "starting point": Q takes a high value (0.677) while nestedness is negligible (0.058). Adding activity at random, up to $C = 0.005$, does not lead to a nested architecture. On the contrary, the process of adding links to the network has deteriorated both Q and $\mathcal{N}$. The second panel of Fig. B.9 shows the evolution of both quantities

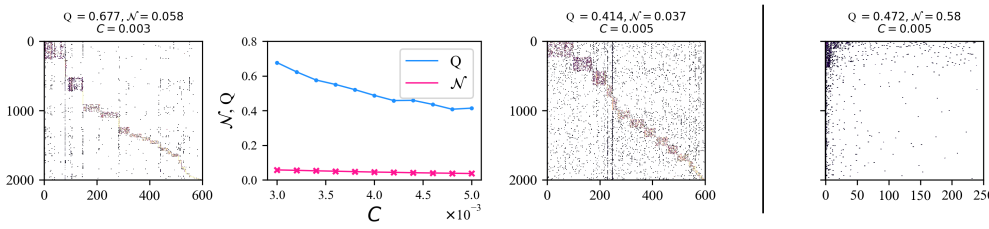


Figure B.9: **Expected effects of random increased activity in empirical data:** the system, organized in a modular pattern initially (left) transitions to a highly nested one (right); however, randomly increasing connectance (activity) does not imply per se increases in the values of nestedness (middle panels).

as links are added, and the third panel shows the resulting network. For the sake of comparison, the panel to the right shows the actual (empirical) network at the end of the process, in which nestedness peaks at 0.58 –an order of magnitude above what one would observe if activity is increased at random. We can observe that increased activity does not render, per se, any gains in $\mathcal{N}$ –rather, nestedness stays negligible to a value of in the order of 10^{-2}. To better grasp this relevant point, we further extend our experiment from Fig. B.9 to more general scenarios, employing a synthetic benchmark.

Starting from an initially modular, an initially nested, and an initially in-block nested network, each one of size $N_{col} = N_{row} = 150$ nodes, we randomly increase their densities, as a null model of activity growth, i.e., users are linking to more and more hashtags at random. Figure B.10 below shows the results from this study. Each point of the plot corresponds to an increase of 5% in the amount of links. From the initially modular network (top row), we observe that, as we move along the x-axis (added links), increasing activity at random decreases the modularity, and yet nestedness remains in the extremely low values that it showed initially. For the initially nested (middle row), and initially in-block nested (bottom row) networks, the results are very similar: no growth of the complementary pattern is observed at all.

Ultimately, it is clear that a sole increase in activity (in any situation) is not necessarily related to an implicit increment of any of the measures used in our study. In fact, nestedness for example, may emerge without a remarkable increase in activity: see panels (a), (c) and (d) of Fig. B.1. Shifts to nestedness in these cases are not directly related to increase in activity. Although we are aware that this result clashes with the idea that connectance underlies the emergence of nestedness [177], we want to stress that such increases in nestedness related to increases in connectance are usually quantified through the use of descriptors that does not discount the amount of overlap that two species may have due to random fluctuations, e.g., Almeida-Neto's NODF [73]. All in all, our results seem to indicate that, for all the three measures, it is required that some explicit driver on the network constituents and their interactions guides the changes at the macroscopic or mesoscopic levels.

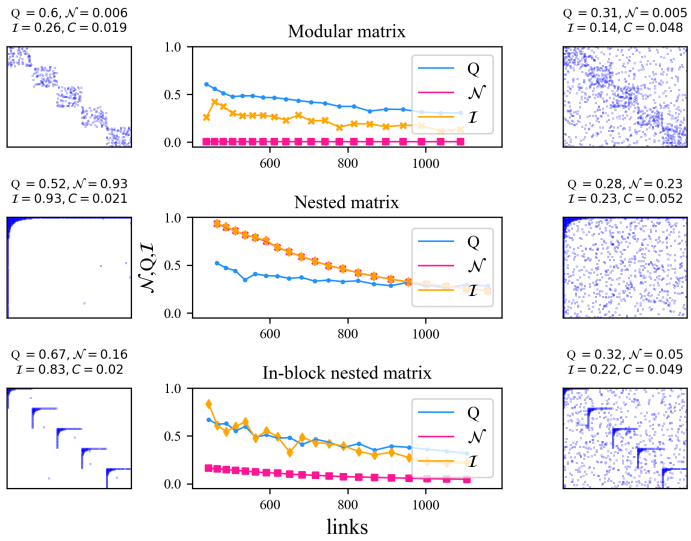


Figure B.10: **Effects of randomly increasing density on synthetic networks** with purely modular (top), nested (middle) and in-block nested (bottom) organizations. Random link addition harms the idealised initial structures, while it has no positive effects on the other descriptors.

Bibliography

[1] A. Arenas, A. Díaz-Guilera, et al., "Synchronization in complex networks", *Physics Reports* **469**, 93 (2008).

[2] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks", *Physical Review Letters* **86**, 3200 (2001).

[3] P. Holme and J. Saramäki, "Temporal networks", *Physics Reports* **519**, 97 (2012).

[4] S. Boccaletti, G. Bianconi, et al., "The structure and dynamics of multilayer networks", *Physics Reports* **544**, 1 (2014).

[5] M. Kivelä, A. Arenas, et al., "Multilayer networks", *Journal of Complex Networks* **2**, 203 (2014).

[6] P. Holme, "Modern temporal network theory: a colloquium", *The European Physical Journal B* **88**, 234 (2015).

[7] P. Erdos and A. Rényi, "On random graphs I", *Publicationes Mathematicae Debrecen* **6**, 290 (1959).

[8] P. Erdős and A. Rényi, "On the evolution of random graphs", *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17 (1960).

[9] P. Erdós and A. Rényi, "On the evolution of random graphs. II", *Bulletin of the International Statistical Institute* **38**, 343 (1961).

[10] B. Bollobás. *Random Graphs*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001. DOI: 10.1017/CBO9780511814068.

[11] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications", *Physical Review E* **64**, 26118 (2001).

[12] M. E. Newman et al., "Random graphs as models of networks", *Handbook of graphs and networks* **1**, 35 (2003).

[13] S. Milgram, "The small world problem", *Psychology today* **2**, 60 (1967).

[14] J. Travers and S. Milgram, "An exploratory study of the small world problem", *Sociometry* **32**, 425 (1969).

[15] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks", *Nature* **393**, 440 (1998).

[16] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks", *Science* **286**, 509 (1999).

[17] B. Karrer, E. Levina, and M. E. Newman, "Robustness of community structure in networks", *Physical Review E* **77**, 046119 (2008).

[18] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks", *Nature* **406**, 378 (2000).

[19] J. P. Gleeson, "Cascades on correlated and modular random networks", *Physical Review E* **77**, 46117 (2008).

[20] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks", *Scientific reports* **3**, 2522 (2013).

[21] T. M. Lewinsohn, P. Inácio Prado, et al., "Structure in plant–animal interaction assemblages", *Oikos* **113**, 174 (2006).

[22] W. Ulrich and N. J. Gotelli, "Pattern detection in null model analysis", *Oikos* **122**, 2 (2013).

[23] N. J. Gotelli, "Null model analysis of species co-occurrence patterns", *Ecology* **81**, 2606 (2000).

[24] M. Almeida-Neto, P. R Guimarães Jr, and T. M Lewinsohn, "On nestedness analyses: Rethinking matrix temperature and anti-nestedness", *Oikos* **116**, 716 (2007).

[25] G. Strona and J. A. Veech, "A new measure of ecological network structure based on node overlap and segregation", *Methods in Ecology and Evolution* **6**, 907 (2015).

[26] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures", *Social Networks* **21**, 375 (2000).

[27] P. Rombach, M. A. Porter, et al., "Core-periphery structure in networks (revisited)", *SIAM Review* **59**, 619 (2017).

[28] M. E. J. Newman, "Analysis of weighted networks", *Physical Review E* **70**, 56131 (2004).

[29] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization", *Physical Review E* **72**, 27104 (2005).

[30] M. J. Barber, "Modularity and community detection in bipartite networks", *Physical Review E* **76**, 66102 (2007).

[31] V. D. Blondel, J.-L. Guillaume, et al., "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

[32] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks", *Physical Review Letters* **100**, 118703 (2008).

[33] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", *Physical Review E* **69**, 26113 (2004).

[34] M. E. J. Newman, "Modularity and community structure in networks", *Proceedings of the national academy of sciences* **103**, 8577 (2006).

[35] F. Radicchi, C. Castellano, et al., "Defining and identifying communities in networks", *Proceedings of the National Academy of Sciences* **101**, 2658 (2004).

[36] W. W. Zachary, "An information flow model for conflict and fission in small groups", *Journal of Anthropological Research* **33**, 452 (1977).

[37] K. A. Eriksen, I. Simonsen, et al., "Modularity and extreme edges of the Internet", *Physical Review Letters* **90**, 148701 (2003).

[38] R. Guimerà and L. A. N. Amaral, "Functional cartography of complex metabolic networks", *Nature* **433**, 895 (2005).

[39] L. A. Adamic and N. Glance. "The political blogosphere and the 2004 US election: divided they blog". *Proceedings of the 3rd international workshop on Link discovery*. ACM. 2005, 36. DOI: 10.1145/1134271.1134277.

[40] S. Fortunato, "Community detection in graphs", *Physics Reports* **486**, 75 (2010).

[41] W. Atmar and B. D. Patterson, "The measure of order and disorder in the distribution of species in fragmented habitat", *Oecologia* **96**, 373 (1993).

[42] M. S. Mariani, Z.-M. Ren, et al., "Nestedness in complex networks: Observation, emergence, and implications", *Physics Reports* **813**, 1 (2019).

[43] B. D. Patterson and W. Atmar, "Nested subsets and the structure of insular mammalian faunas and archipelagos", *Biological Journal of the Linnean Society* **28**, 65 (1986).

[44] Y. L. Dupont, K. Trøjelsgaard, and J. M. Olesen, "Scaling down from species to individuals: a flower–visitation network between individual honeybees and thistle plants", *Oikos* **120**, 170 (2011).

[45] J. Bascompte, P. Jordano, et al., "The nested assembly of plant–animal mutualistic networks", *Proceedings of the National Academy of Sciences* **100**, 9383 (2003).

[46] S. Saavedra, D. B. Stouffer, et al., "Strong contributors to network persistence are the most vulnerable to extinction", *Nature* **478**, 233 (2011).

[47] S. Bustos, C. Gomez, et al., "The dynamics of nestedness predicts the evolution of industrial ecosystems", *PloS one* **7**, e49393 (2012).

[48] M. D. König, C. J. Tessone, and Y. Zenou, "Nestedness in networks: A theoretical model and some applications", *Theoretical Economics* **9**, 695 (2014).

[49] J. Borge-Holthoefer, R. A. Baños, et al., "Emergence of consensus as a modular-to-nested transition in communication dynamics", *Scientific Reports* **7** (2017).

[50] C. O. Flores, J. R. Meyer, et al., "Statistical structure of host–phage interactions", *Proceedings of the National Academy of Sciences* **108**, E288 (2011).

[51] C. O. Flores, S. Valverde, and J. S. Weitz, "Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages", *The ISME journal* **7**, 520 (2013).

[52] S. J. Beckett and H. T. P. Williams, "Coevolutionary diversification creates nested-modular structure in phage–bacteria interaction networks", *Interface Focus* **3**, 20130033 (2013).

[53] A. Solé-Ribalta, C. J. Tessone, et al., "Revealing in-block nestedness: Detection and benchmarking", *Physical Review E* **97**, 62302 (2018).

[54] S. Boccaletti, V. Latora, et al., "Complex networks: Structure and dynamics", *Physics Reports* **424**, 175 (2006).

[55] G. Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007. DOI: 10.1093/acprof:oso/9780199211517.001.0001.

[56] M. Newman. *Networks: an introduction*. Oxford university press, 2018. DOI: 9780192527493.

[57] L. C. Freeman, "A set of measures of centrality based on betweenness", *Sociometry*, 35 (1977).

[58] M. E. J. Newman, "Assortative mixing in networks", *Physical Review Letters* **89**, 208701 (2002).

[59] D. A. Richard and M. Gwen, "Elite social circles", *Sociological Methods & Research* **7**, 167 (1978).

[60] D. Snyder and E. L. Kick, "Structural position in the world system and economic growth, 1955-1970: a multiple-network analysis of transnational interactions", *American Journal of Sociology* **84**, 1096 (1979).

[61] F. Luo, B. Li, et al. "Core and periphery structures in protein interaction networks". *IEEE Proccedings of the International Conference on Bioinformatics & Bioengineering*. Vol. 10. 2009.

[62] M. Cucuringu, P. Rombach, et al., "Detection of core–periphery structure in networks using spectral methods and geodesic paths", *European Journal of Applied Mathematics* **27**, 846 (2016).

[63] X. Zhang, T. Martin, and M. E. J. Newman, "Identification of core-periphery structure in networks", *Physical Review E* **91**, 32803 (2015).

[64] U. Bastolla, M. A. Fortuna, et al., "The architecture of mutualistic networks minimizes competition and increases biodiversity", *Nature* **458**, 1018 (2009).

[65] E. Thébault and C. Fontaine, "Stability of ecological communities and the architecture of mutualistic and trophic networks", *Science* **329**, 853 (2010).

[66] S. Suweis, F. Simini, et al., "Emergence of structural and dynamical properties of ecological mutualistic networks", *Nature* **500**, 449 (2013).

[67] S. Allesina and S. Tang, "Stability criteria for complex ecosystems", *Nature* **483**, 205 (2012).

[68] P. P. A. Staniczenko, J. C. Kopp, and S. Allesina, "The ghost of nestedness in ecological networks", *Nature communications* **4**, 1391 (2013).

[69] S. Valverde, J. Piñero, et al., "The architecture of mutualistic networks as an evolutionary spandrel", *Nature Ecology & Evolution* **2**, 94 (2018).

[70] D. S. Maynard, C. A. Serván, and S. Allesina, "Network spandrels reflect ecological assembly", *Ecology Letters* **21**, 324 (2018).

[71] S. Johnson, V. Domínguez-García, and M. A. Muñoz, "Factors determining nestedness in complex networks", *PloS one* **8**, e74025 (2013).

[72] C. Payrató-Borràs, L. Hernández, and Y. Moreno, "Breaking the spell of nestedness: The entropic origin of nestedness in mutualistic systems", *Physical Review X* **9**, 031024 (2019).

[73] M. Almeida-Neto, P. Guimarães, et al., "A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement", *Oikos* **117**, 1227 (2008).

[74] W. Ulrich, M. Almeida-Neto, and N. J. Gotelli, "A consumer's guide to nestedness analysis", *Oikos* **118**, 3 (2009).

[75] F. K. Bell, D. Cvetković, et al., "Graphs for which the least eigenvalue is minimal, I", *Linear Algebra and its Applications* **429**, 234 (2008).

[76] F. K. Bell, D. Cvetković, et al., "Graphs for which the least eigenvalue is minimal, II", *Linear Algebra and its Applications* **429**, 2168 (2008).

[77] A. Bhattacharya, S. Friedland, and U. N. Peled, "On the first eigenvalue of bipartite graphs", *The Electronic Journal of Combinatorics* **15**, 144 ().

[78] J. Galeano, J. M. Pastor, and J. M. Iriondo, "Weighted-interaction nestedness estimator (WINE): a new estimator to calculate over frequency matrices", *Environmental Modelling & Software* **24**, 1342 (2009).

[79] M. Almeida-Neto and W. Ulrich, "A straightforward computational approach for measuring nestedness using quantitative matrices", *Environmental Modelling & Software* **26**, 173 (2011).

[80] J. Podani, C. Ricotta, and D. Schmera, "A general framework for analyzing beta diversity, nestedness and related community-level phenomena based on abundance data", *Ecological Complexity* **15**, 52 (2013).

[81] W. Ulrich and N. J. Gotelli, "Null model analysis of species nestedness patterns", *Ecology* **88**, 1824 (2007).

[82] N. J. Gotelli and W. Ulrich, "Statistical challenges in null model analysis", *Oikos* **121**, 171 (2012).

[83] S. J. Beckett, C. A. Boulton, and H. T. P. Williams, "FALCON: a software package for analysis of nestedness in bipartite networks", *F1000Research* **3** (2014).

[84] T. Squartini and D. Garlaschelli, "Analytical maximum-likelihood method to detect patterns in real networks", *New Journal of Physics* **13**, 83001 (2011).

[85] J. Borge-Holthoefer and A. Arenas. "Navigating Word Association Norms to Extract Semantic Information". *roceedings of the Annual Conference of the Cognitive Science Society*. 2009. URL: http://csjarchive.cogsci.rpi.edu/Proceedings/2009/papers/621/paper621.pdf.

[86] D. B. Stouffer and J. Bascompte, "Compartmentalization increases food-web persistence", *Proceedings of the National Academy of Sciences* **108**, 3648 (2011).

[87] J. Borge-Holthoefer, A. Rivero, et al., "Structural and Dynamical Patterns on Online Social Networks: the Spanish May 15th Movement as a case study", *PloS One* **6**, e23883 (2011).

[88] S. Allesina, J. Grilli, et al., "Predicting the stability of large structured food webs", *Nature Communications* **6**, 7842 (2015).

[89] J. Grilli, T. Rogers, and S. Allesina, "Modularity and stability in ecological communities", *Nature Communications* **7**, 12031 (2016).

[90] X. Wu and Z. Liu, "How community structure influences epidemic spread in social networks", *Physica A: Statistical Mechanics and its Applications* **387**, 623 (2008).

[91] A. Nematzadeh, E. Ferrara, et al., "Optimal network modularity for information diffusion", *Physical Review Letters* **113**, 88701 (2014).

[92] S. Kojaku and N. Masuda, "Finding multiple core-periphery pairs in networks", *Physical Review E* **96**, 52313 (2017).

[93] S. Fortunato and D. Hric, "Community detection in networks: A user guide", *Physics Reports* **659**, 1 (2016).

[94] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001. DOI: 10.1007/978-0-387-84858-7.

[95] U. Von Luxburg, "A tutorial on spectral clustering", *Statistics and Computing volume* **17**, 395?416 (2007).

[96] J. B. McQueen. "Some methods for classification and analysis of multivariate data". *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. 1967, 281. URL: https://projecteuclid.org/euclid.bsmsp/1200512974.

[97] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks", *Proceedings of the National Academy of Sciences* **106**, 22073 (2009).

[98] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks", *Physical Review E* **83**, 16107 (2011).

[99] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks", *Physical Review X* **4**, 11047 (2014).

[100] M. E. Newman, "Equivalence between modularity optimization and maximum likelihood methods for community detection", *Physical Review E* **94**, 052315 (2016).

[101] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure", *Proceedings of the National Academy of Sciences* **105**, 1118 (2008).

[102] S. Fortunato and M. Barthelemy, "Resolution limit in community detection", *Proceedings of the National Academy of Sciences* **104**, 36 (2007).

[103] S. Gómez, P. Jensen, and A. Arenas, "Analysis of community structure in networks of correlated data", *Physical Review E* **80** (2009).

[104] P. J. Mucha, T. Richardson, et al., "Community structure in time-dependent, multiscale, and multiplex networks", *Science* **328**, 876 (2010).

[105] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, "Modularity from fluctuations in random graphs and complex networks", *Physical Review E* **70**, 25101 (2004).

[106] M. E. J. Newman, "Fast algorithm for detecting community structure in networks", *Physical Review E* **69** (2004).

[107] S. Boettcher and A. G. Percus, "Optimization with extremal dynamics", *complexity* **8**, 57 (2002).

[108] M. A. R. Mello, G. M. Felix, et al., "Insights into the assembly rules of a continent-wide multilayer network", *Nature ecology & evolution*, 1 (2019).

[109] S. Kojaku and N. Masuda, "Core-periphery structure requires something else in the network", *New Journal of Physics* **20**, 43012 (2018).

[110] S. Kojaku, M. Xu, et al., "Multiscale core-periphery structure in a global liner shipping network", *Scientific Reports* **9**, 404 (2019).

[111] B.-B. Xiang, Z.-K. Bao, et al., "A unified method of detecting core-periphery structure and community structure in networks", *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28**, 013122 (2018).

[112] B. Yan and J. Luo, "Multicores-periphery structure in networks", *Network Science* **7**, 70 (2019).

[113] J. M. Olesen, J. Bascompte, et al., "The modularity of pollination networks", *Proceedings of the National Academy of Sciences* **104**, 19891 (2007).

[114] M. A. Fortuna, D. B. Stouffer, et al., "Nestedness versus modularity in ecological networks: two sides of the same coin?", *Journal of Animal Ecology* **79**, 811 (2010).

[115] S. H. Lee et al., "Network nestedness as generalized core-periphery structures", *Physical Review E* **93**, 22306 (2016).

[116] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance", *Journal of Machine Learning Research* **11**, 2837 (2010).

[117] *Web of Life: ecological networks database*. http://www.web-of-life.es/. 2012.

[118] M. J. Palazzi, J. Cabot, et al., "Online division of labour: emergent structures in Open Source Software", *Scientific Reports* **9**, 13890 (2019).

[119] G. G. de Bunt, M. A. J. Van Duijn, and T. A. B. Snijders, "Friendship networks through time: An actor-oriented dynamic statistical network model", *Computational & Mathematical Organization Theory* **5**, 167 (1999).

[120] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research", *Proccedings of The European Conference on Machine Learning (ECML)*, 217 (2004).

[121] T. A. B. Snijders, C. E. G. Steglich, and G. G. van de Bunt, "Introduction to actor-based models for network dynamics", *Social Networks* (2008).

[122] V. A. Traag, P. Van Dooren, and Y. Nesterov, "Narrow scope for resolution-limit-free community detection", *Physical Review E* **84**, 16114 (2011).

[123] *https://github.com*.

[124] T. Okuyama and J. N. Holland, "Network structural properties mediate the stability of mutualistic communities", *Ecology Letters* **11**, 208 (2008).

[125] W. Cai, J. Snyder, et al., "Mutualistic Networks Emerging from Adaptive Niche-Based Interactions", *Nature Communications* **11**, 5470 (2020).

[126] *Open Source initiative*. https://opensource.org/.

[127] R. Schuwer, M. van Genuchten, and L. Hatton, "On the Impact of Being Open", *IEEE Software* **32**, 81 (2015).

[128] L. Dabbish, C. Stuart, et al. "Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository". *Proccedings of the Conference on Computer-Supported Cooperative Work and Social Computing ACM*. 2012, 1277. DOI: 10.1145/2145204.2145396.

[129] R. Padhye, S. Mani, and V. S. Sinha. "A Study of External Community Contribution to Open-source Projects on GitHub". *Working Conf. on Mining Software Repositories*. 2014, 332.

[130] L. Dabbish, C. Stuart, et al., "Leveraging Transparency", *IEEE Software* **30**, 37 (2013).

[131] A. Lima, L. Rossi, and M. Musolesi. "Coding Together at Scale: GitHub as a Collaborative Social Network". *International Conference on Web and Social Media*. 2014, 10.

[132] S. Fitz-Gerald, "Book Review of:'Internet success: a study of Open-Source Software commons' by CM Schweik and RC English", *International Journal of Information Management* **32**, 596 (2012).

[133] V. Cosentino, J. L. C. Izquierdo, and J. Cabot. "Assessing the bus factor of Git repositories". *International Conference on Software Analysis, Evolution, and Reengineering*. 2015, 499. DOI: 10.1109/SANER.2015.7081864.

[134] K. Yamashita, S. McIntosh, et al. "Revisiting the Applicability of the Pareto Principle to Core Development Teams in Open Source Software Projects". *International Workshop on Principles of Software Evolution*. 2015, 46. DOI: 10.1145/2804360.2804366.

[135] G. Avelino, L. Passos, et al. "A Novel Approach for Estimating Truck Factors". *International Conference on Program Comprehension*. 2016, 1. DOI: 10.1109/ICPC.2016.7503718.

[136] R. Pham, L. Singer, et al. "Creating a Shared Understanding of Testing Culture on a Social Coding Site". *International Conference on Software Engineering*. 2013, 112.

[137] K. Yamashita, Y. Kamei, et al., "Magnet or Sticky? Measuring Project Characteristics from the Perspective of Developer Attraction and Retention", *Journal of Information Processing* **24**, 339 (2016).

[138] H. Hata, T. Todo, et al. "Characteristics of Sustainable OSS Projects: a Theoretical and Empirical Study". *International Workshop on Cooperative and Human Aspects of Software Engineering*. 2015, 15.

[139] A. P. O. Bertholdo and M. A. Gerosa. "Promoting Engagement in Open Collaboration Communities by Means of Gamification". *International Conference on Human-Computer Interaction*. 2016, 15. DOI: 10.1007/978-3-319-40542-1_3.

[140] I. Steinmacher, T. U. Conte, et al. "Overcoming Open Source Project Entry Barriers with a Portal for Newcomers". *International Conerence. on Software Engineering*. 2016, 273.

[141] I. Steinmacher, M. A. G. Silva, et al., "A Systematic Literature Review on the Barriers Faced by Newcomers to Open Source Software Projects", *Information & Software Technology* **59**, 67 (2015).

[142] V. Cosentino, J. L. C. Izquierdo, and J. Cabot, "A systematic mapping study of software development with GitHub", *IEEE Access* **5**, 7173 (2017).

[143] S. Valverde and R. V. Solé, "Self-organization versus hierarchy in open-source social networks", *Physical Review E* **76**, 46118 (2007).

[144] C. Bird, D. Pattison, et al. "Latent social structure in open source projects". *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*. 2008, 24. DOI: 10.1145/1453101.1453107.

[145] Q. Hong, S. Kim, et al. "Understanding a developer social network and its evolution". *IEEE 27th International Conference on Software Maintenance, ICSM*. 2011, 323.

[146] R. Dunbar, "Neocortex size as a constraint on group size in primates", *Journal of Human Evolution* **22**, 469 (1992).

[147] B. Gonçalves, N. Perra, and A. Vespignani, "Modeling users' activity on twitter networks: Validation of dunbar's number", *PloS one* **6**, e22656 (2011).

[148] *Using the request:* https://api.github.com/search/repositories?q=stars:>1&sort=stars&order=desc&per_page=100.

[149] V. Cosentino, J. L. Cánovas Izquierdo, and J. Cabot. "Gitana: A SQL-Based Git Repository Inspector". *International Conference on Conceptual Modeling*. 2015, 329. DOI: 10.1007/978-3-319-25264-3_24.

[150] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root", *Journal of the American Statistical Association* **74**, 427 (1979).

[151] M. J. Palazzi, J. Borge-Holthoefer, et al., "Macro-and mesoscale pattern interdependencies in complex networks", *Journal of the Royal Society Interface* **16**, 20190553 (2019).

[152] R. Dunbar. *How many friends does one person need?: Dunbar's number and other evolutionary quirks*. Faber & Faber, 2010.

[153] M. Derex and R. Boyd, "Partial connectivity increases cultural accumulation within groups", *Proceedings of the National Academy of Sciences* **113**, 2982 (2016).

[154] M. Derex, C. Perreault, and R. Boyd, "Divide and conquer: intermediate levels of population fragmentation maximize cultural accumulation", *Philosophical Transactions of the Royal Society B* **373**, 20170062 (2018).

[155] U. Olsson, "Confidence intervals for the mean of a log-normal distribution", *Journal of Statistics Education* **13** (2005).

[156] H. A. Simon, "Theories of bounded rationality", *Decision and organization* **1**, 161 (1972).

[157] D. Kahneman. *Attention and effort*. Englewood Cliffs, N.J: Prentice-Hall., 1973.

[158] M. I. Posner, "Cumulative development of attentional theory.", *American Psychologist* **37**, 168 (1982).

[159] A. Bruns and T. Highfield. "Is Habermas on Twitter?: Social media and the public sphere". *The Routledge companion to social media and politics*. Routledge, 2015, 56. DOI: 10.4324/9781315716299-5.

[160] L. Weng, A. Flammini, et al., "Competition among memes in a world with limited attention", *Scientific Reports* **2** (2012).

[161] J. P. Gleeson, J. A. Ward, et al., "Competition-induced criticality in a model of meme popularity", *Physical Review Letters* **112**, 48701 (2014).

[162] J. P. Gleeson, K. P. O'Sullivan, et al., "Effects of network structure, competition and memory time on social spreading phenomena", *Physical Review X* **6**, 021019 (2016).

[163] P. Lorenz-Spreen, B. M. Mønsted, et al., "Accelerating dynamics of collective attention", *Nature communications* **10**, 1759 (2019).

[164] T. Gross, C. J. D. D'Lima, and B. Blasius, "Epidemic dynamics on an adaptive network", *Physical Review Letters* **96**, 208701 (2006).

[165] P. Holme and M. E. Newman, "Nonequilibrium phase transition in the coevolution of networks and opinions", *Physical Review E* **74**, 056108 (2006).

[166] F. Vazquez, V. M. Eguıluz, and M. San Miguel, "Generic absorbing transition in coevolution dynamics", *Physical Review Letters* **100**, 108702 (2008).

[167] S. Sheykhali, J. Fernández-Gracia, et al., "Robustness to extinction and plasticity derived from mutualistic bipartite ecological networks", *Scientific reports* **10**, 1 (2020).

[168] A. Bruns and J. E. Burgess. "The use of Twitter hashtags in the formation of *a*d hoc publics". *Proceedings of the 6th European Consortium for Political Research General Conference*. 2011.

[169] A. Chadwick, "The political information cycle in a hybrid news system: The British prime minister and the "Bullygate" affair", *The International Journal of Press/Politics* **16**, 3 (2011).

[170] P. R. Guimarães Jr, M. M. Pires, et al., "Indirect effects drive coevolution in mutualistic networks", *Nature* **550**, 511 (2017).

[171] S. Pilosof, M. A. Porter, et al., "The multilayer nature of ecological networks", *Nature Ecology & Evolution* **1**, 101 (2017).

[172] S. González-Bailón, N. Wang, et al., "Assessing the bias in samples of large online networks", *Social Networks* **38**, 16 (2014).

[173] A. Zubiaga, "A longitudinal assessment of the persistence of twitter datasets", *Journal of the Association for Information Science and Technology* **69**, 974 (2018).

[174] R. Alarcón, N. M. Waser, and J. Ollerton, "Year-to-year variation in the topology of a plant–pollinator interaction network", *Oikos* **117**, 1796 (2008).

[175] T. Petanidou, A. S. Kallimanis, et al., "Long-term observation of a pollination network: fluctuation in species and interactions, relative invariance of network structure and implications for estimates of specialization", *Ecology Letters* **11**, 564 (2008).

[176] C. Díaz-Castelazo, P. R. Guimarães, et al., "Changes of a mutualistic network over time: reanalysis over a 10-year period", *Ecology* **91**, 793 (2010).

[177] A. James, J. W. Pitchford, and M. J. Plank, "Disentangling nestedness from models of ecological complexity", *Nature* **487**, 227 (2012).

[178] C. S. Holling, "Engineering resilience versus ecological resilience", *Engineering within ecological constraints* **31**, 32 (1996).

[179] A. R. Ives and S. R. Carpenter, "Stability and diversity of ecosystems", *Science* **317**, 58 (2007).

[180] C. Folke, S. Carpenter, et al., "Resilience thinking: integrating resilience, adaptability and transformability", *Ecology and society* **15** (2010).

[181] S. Suweis, J. Grilli, et al., "Effect of localization on the stability of mutualistic ecological networks", *Nature communications* **6**, 10179 (2015).

[182] J.-F. Arnoldi, M. Loreau, and B. Haegeman, "Resilience, reactivity and variability: A mathematical comparison of ecological stability measures", *Journal of theoretical biology* **389**, 47 (2016).

[183] R. P. Rohr, S. Saavedra, and J. Bascompte, "On the structural stability of mutualistic systems", *Science* **345**, 1253497 (2014).

[184] J. Grilli, M. Adorisio, et al., "Feasibility and coexistence of large ecological communities", *Nature communications* **8**, 14389 (2017).

[185] J. Lehmann, B. Gonçalves, et al. "Dynamical classes of collective attention in twitter". *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, 251. DOI: 10.1145/2187836.2187871.

[186] R. J. Williams and N. D. Martinez, "Simple rules yield complex food webs", *Nature* **404**, 180 (2000).

[187] J. Bascompte and P. Jordano, "Plant-animal mutualistic networks: the architecture of biodiversity", *Annual Review of Ecology, Evolution, and Systematics* **38**, 567 (2007).

[188] S. Saavedra, R. P. Rohr, et al., "Nested species interactions promote feasibility over stability during the assembly of a pollinator community", *Ecology and evolution* **6**, 997 (2016).