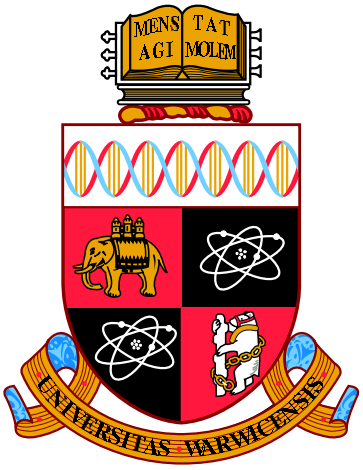

Emergent Patterns in Complex Networks

by

Janis Klaise

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in Mathematics and Complexity Science

Centre for Complexity Science

December 2017

WARWICK
THE UNIVERSITY OF WARWICK

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support and guidance from my supervisor Samuel Johnson. During the past three years I have enjoyed many intellectually stimulating discussions and I cannot thank Sam enough for the constant positivity and encouragement.

I would also like to thank the students and staff of the Centre for Complexity Science for providing a stimulating and friendly environment to work in. The frequent exchange of ideas as well as the general camaraderie and support have played an invaluable part in shaping the past few years of my life and undoubtedly the years to come. Special thanks goes to Alex Bishop and Elizabeth Buckingham-Jeffery who were there when the going was difficult.

Finally, I would like to thank my parents and family for understanding the importance of this endeavour and their unwavering support.

DECLARATIONS

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Parts of this thesis have been published by the author:

1. Chapter 3: J. Klaise and S. Johnson. The origin of motif families in food webs. *Scientific Reports*, 7:16197, 2017, doi:10.1038/s41598-017-15496-1.

2. Chapter 4: J. Klaise and S. Johnson. From neurons to epidemics: How trophic coherence affects spreading processes. *Chaos*, 26:065310, 2016, doi:10.1063/1.4953160.

3. Chapter 5: J. Klaise and S. Johnson. Relaxation dynamics of maximally clustered networks. *Physical Review E*, 97:012302, 2018, doi:10.1103/PhysRevE.97.012302.

ABSTRACT

Complex interacting systems permeate the modern world. Many diverse natural, social and human made systems—ranging from food webs to human contact networks, to the Internet—can be studied in the context of network science. This thesis is a compendium of research in applied network science, investigating structural and dynamical patterns behind the formation of networks and processes supported on them.

Trophic food webs—networks of who eats whom in an ecosystem—have fascinated network scientists since data from field observations of the gut content of species first became available. The empirical patterns in food webs reveal a rich hierarchy of feeding patterns. We study how global structure of food webs relates to species immediate diet over a range of 46 different ecosystems. Our finding suggest that food webs fall broadly into two different families based on the extent of species tendency towards omnivory.

Drawing inspiration from food webs, we investigate how trophic networks support spreading processes on them. We find that the interplay of dynamics and network structure determines the extent and duration of contagion. We uncover two distinct modes of operation—short-lived outbreaks with high incidence and endemic infections. These results could be important for understanding spreading phenomena such as epidemics, rumours, shocks to ecosystems and neuronal avalanches.

Finally, we study the emergence of structural order in random network models. Random networks serve as null models to empirical networks to help uncover significant non-random patterns but are also interesting to study in their own right. We study the effect of triadic ties in delaying the formation of extensive *giant components*—connected components taking over the majority of the network. Our results show that, depending on the network formation process, order in the form of a giant component can emerge even with a significant number of triadic ties.

CHAPTER 1

INTRODUCTION

We live in a connected world. Networks of nodes, representing agents, elements or entities, interacting via a set of connections or links are ubiquitous in nature, society and artificial systems. Ecosystems are often characterized by the trophic networks of food webs—feeding relationships between species [201]. Chemical reactions between molecules in cells lead to the study of interactomes—networks describing interaction relationships, typically between proteins [254]. A typical human brain consists of 10^{11} neurons connected together via 10^{14} synapses [80, 117] resulting in a large network over which neural activity propagates. People are connected socially via various kinds of relationships—social friendships in real life [255] and on social media such as *Facebook* [155] and *Twitter* [146], collaborations in science [180], physical contact based on the proximity of interactions which leads to the study of human contact networks in epidemiology of infectious disease [72, 141], and many others. Power grids consist of generators, transformers and substations all linked together with high-voltage transmission lines [5] while the Internet is made up of computers connected to each other [194]. The ubiquity of networked systems has lead to the rise of network science—the study of the formation, structure and function of these systems.

With the availability of data and increased computational power, network science has emerged over the past two decades as the *de facto* framework of studying complex, interacting systems. The main strength of the network science approach is its ability to abstract a variety of very diverse systems into the language of networks. The tools developed for studying generic networks are then readily applied to networks arising from the observation of empirical systems. Techniques ranging from

measurements of various network properties to building stochastic network models to explain the formation of real world systems have resulted in an unprecedented insight into the organization and function of complex networked systems [42].

In this thesis we study various structural and dynamical properties of networks—both with a view to studying empirical networks as well as random network models in their own right. The thesis is laid out as follows.

Chapter 2 introduces the minimum background information on network science to follow the arguments developed in later chapters. We focus on defining the most widely used network types and structural network properties in modelling real world systems. We also introduce the notion of random networks and their importance as null models of empirical networks. We motivate the study of random network models in their own right via highlighting some emergent properties (e.g. vanishing clustering coefficient, formation of a giant connected component) in two popular random network models—the Erdős-Rényi random graph and the configuration model. We close our discussion by mentioning a number of other widely used random network models that are beyond the scope of this thesis.

In Chapter 3 we study a particular example of a networked system—ecological food webs. Food webs, viewed as networks, are an abstract representation of who eats whom in an ecosystem. Theoretical ecologists have been fascinated by the intricate, non-random patterns observed in the structure of food webs as reconstructed by field ecologists by examining species gut content or by direct observation of animal diet. This interest has resulted in a number of theoretical models of food web formation from simple rules to explain the observed structured in nature, however the existing models cannot fully account for all of the patterns. In this chapter we study the structure of 46 empirical food webs. We focus on the characterization of food webs based on local preying patterns formalized in the concept of network motifs. We show that almost all food webs fall into one of two families distinguished by the overall extent of omnivory, the tendency of species to consume prey from multiple hierarchical levels. We formulate a new food web model—the generalized preferential preying model—that can successfully produce artificial food webs belonging to either family and can replicate the structure of empirical food webs remarkably well.

In Chapter 4 we move beyond structural analysis of networks by considering dynamical processes spreading through a network. Inspired by the model for food webs, we study how an initial pulse of activity spreads through these types of trophic networks. We study this under two paradigmatic processes—a complex contagion process inspired by epidemiology and a neural network process. Our results based on numerical simulations uncover that the topology of the network can have a significant

qualitative impact on the spreading properties of both processes. The underlying topology can determine whether the pulse of activity will percolate through the entire network or remain confined to a small subset of nodes, and also whether the activity will quickly die out or become endemic and endure indefinitely. These results could help our understanding of many spreading processes naturally supported on top of network topologies such as epidemics, rumours, shocks to ecosystems, neuronal avalanches and many others.

Finally, in Chapter 5 we combine both structure and dynamics of networks by studying how the structure of random networks changes as they undergo random link dynamics. A large number of empirical networks, particularly human social networks, exhibit a large number of triangles or triadic ties which confirms the common belief that a person's friends are likely to be friends with each other. This observation has spurred a large interest in studying random network models that can create a significant number of triangles. Despite the plethora of theoretical research, some questions about such highly clustered networks remain evasive due do the edge independence assumption being violated and techniques developed for networks without triangles or small loops in general do not apply. In this chapter we explore simple relaxation dynamics of highly clustered networks to an unclustered (uncorrelated) equilibrium state under two link rewiring schemes. We study the evolution of the clustering coefficient and the presence of a giant connected component. We find that under both dynamics a giant connected component emerges via a continuous phase transition at a different critical point than in equilibrium networks. Additionally, we derive time evolution equations for various general network properties. Our results could help us understand the properties of random network ensembles better, specifically when dealing with highly unusual samples.

CHAPTER 2

BACKGROUND

Network science has exploded in popularity in the last couple of decades. Networks are used to describe diverse complex, interacting systems, from power grids to human brain networks, and network science has become a paradigmatic framework of analysing complex datasets that admit a network interpretation.

Despite the recent peak of interest in network science, driven by the availability of data and increased computational power, the roots of modern network science date back to the early 18th century and Euler who studied the famous *Seven Bridges of Königsberg* problem [91] (see Biggs et al. [40] for a reprinted and translated version). This was the beginning of graph theory—the mathematical study of the properties of pairwise relationships of network structures.

While graph theory has traditionally studied fixed and immutable network structures, a probabilistic theory of networks was initiated by Erdős, Rényi and Gilbert in the 1960s [88, 89, 105]. This approach treats networks as random variables and any fixed network is assumed to originate from a probability distribution over networks or as a realization of some underlying random process. Network properties are then studied as averages over the whole probability distribution under consideration.

The probabilistic approach has been very successful at studying empirical networks because it recognizes the stochastic nature of network formation in the real world. Even though we only observe one realization of a real network structure at any given time (e.g. the structure of the Internet, while changing over time, has a fixed structure at any given point in time), the formation of it is assumed to be the result of some random process. The insight offered by network science is then to measure

the properties of the empirical network and compare it to expectation values of various random network models in a quest to find out the formation principles behind the empirical network. In many instances new random network models based on simple rules are formulated that can provide explanations for the empirical network structures we observe. This can lead to further insight about the function of the network.

While random network models are often motivated by the need to study empirical networks, they are objects of interest in their own right. Random networks often exhibit a rich structural phenomenology as parameters of the network formation process are changed. Furthermore, stochastic processes taking place on top of random networks can lead to new and interesting phenomena [77]. Tools from statistical physics are indispensable to studying random networks and random processes on random networks in their own right.

In this chapter we introduce the the terminology and the fundamental definitions of networks, starting from graph theory basics to introducing some of the most popular probabilistic network models. Often the networks studied in network science are called *complex networks*—this is to emphasize that most empirical networks as well as random network models have non-trivial structural features as opposed to *simple networks* such as regular lattices or trees. The majority of real-world examples such as the Internet, the World Wide Web, social networks—are all complex but possibly in different ways. Many random network models produce complex networks because of the emergence of non-trivial patterns in the network formation process. While there is no universal definition of what exactly makes a network complex, the term is used to emphasize the lack of regularity and a high degree of non-trivial topological features which in turn motivates the methodology of studying these networks.

2.1 Complex networks

A network is a collection of nodes, representing interacting agents such as people in a social network or species in an ecological food web, together with links, representing interactions between the agents, who is friends with whom in a social network or what species prey on each other in a food web. Networks are studied mathematically in the language of graph theory. Sometimes a distinction between networks and graphs is made in the literature, the former describing naturally occurring networked systems while the latter describing the abstract notion of a collection of nodes and links. In reality, both terms convey exactly the same meaning and information, so in the

context of this thesis we will use the terms interchangeably.

In graph theory terms, we represent a network as a tuple $G = (V, E)$ where V is the set of nodes (vertices) and E is the set of links (edges). It is customary to denote the number of nodes and the number of links as $N = |V|$ and $L = |E|$ respectively. We shall label the nodes by integer indices $V = \{1, \ldots, N\}$ and links by $e_{ij} \in E$ where $i, j \in \{1, \ldots, N\}$ are indices of the nodes at the ends of link e_{ij}.

A very convenient representation of a network is its *adjacency matrix* $\mathbf{A}$. This is an $N \times N$ matrix whose entries a_{ij} indicate a presence or absence of links between nodes i and j:

$$a_{ij} = \begin{cases} 1 \text{ if } e_{ij} \in E \\ 0 \text{ otherwise.} \end{cases} \tag{2.1}$$

A network is *undirected* if its adjacency matrix is symmetric, that is $\mathbf{A} = \mathbf{A}^\top$. Undirected networks are common models of systems in which interaction between agents are inherently devoid of directionality. Examples include the Internet, composed of computers (nodes) and network connection between them (links), co-authorship networks of scientific disciplines and sexual contact networks.

Links can also have directionality. For example, a survey asking people to name their friends would likely result in some pairs of friends i, j only one of whom named the other as a friend. The World Wide Web—a collection of websites with hyperlinks pointing from website to website—is a directed network since hyperlinks need not be reciprocated. Similarly, food webs—networks of whom-eats-whom in an ecological community—are directed because prey rarely feed on their predators. In directed networks the adjacency matrix is no longer symmetric because of the inherently asymmetric nature of directed links.

A network is *simple* if it does not allow self-edges (self-loops), i.e. $a_{ii} = 0$ for all $i \in V$, or multi-edges, i.e. E is a set rather than a multiset. In this thesis we will be mainly concerned with studying simple undirected networks (Chapter 5) and simple directed networks (Chapters 3 and 4).

More general types of networks are often studied. For example, *weighted networks* place different *weights* or *strengths* on links to emphasize the often heterogeneous importance of different links (e.g. the weights could be distances or throughput in a transportation network [151, 187], collaboration frequency in scientific collaboration networks [179] or strength of relationships in social networks [110] to name a few). *Temporal networks* assume that the connectivity of the network depends on time, i.e. the adjacency matrix is no longer *static* but is rather a function of an independent time variable $\mathbf{A}(t)$ [119]. This approach is useful in studying

evolving networks over time as opposed to static snapshots. Recently *multilayer* networks have become popular in modelling systems that consist of multiple distinct but interconnected networks [43, 142]. These can be useful for studying systems such as transportation networks in which interchanges are served by different types of transport links or social media networks in which users interact on several social media platforms. Many other types and generalizations of networks have been introduced, but we focus on simple undirected and directed networks in the following work.

We now turn to describing some basic network measures and models which provide the basis of quantitative research in network science. While there is a large body of literature relating to a multitude of network properties and network models, we have selected here some of the most fundamental concepts with particular attention devoted to those which will be the subject of study in later chapters. The material presented here can be found in a number of introductory books on network science [21, 54, 185]. A number of high quality reviews on research in network structure and dynamics that develop these concepts have also been published over the years [6, 42, 77].

2.2 Network properties

2.2.1 Network degrees

The node *degree* is the number of links attached to it. Using the adjacency matrix, the degree of node i is given by

$$k_i = \sum_{j \in V} a_{ij}. \tag{2.2}$$

The collection of all degrees $\{k_1, k_2, \ldots, k_n\}$ defines the *degree sequence* of the network.

In any undirected network we have the useful identity that the sum of all degrees equals twice the number of links (since any link contributes to the degree of two nodes):

$$\sum_{i \in V} k_i = 2L. \tag{2.3}$$

This is the so called *handshake lemma* first proven by Euler [91] as part of the resolution to the *Seven Bridges of Königsberg* problem (see Biggs et al. [40] for a reprinted and translated version).

In directed networks, the directionality of links naturally leads to the notions of *in-degree* and *out-degree*, counting the number of links pointing to and away from a node respectively:

$$k_i^{\text{in}} = \sum_{j \in V} a_{ji}, \tag{2.4}$$

$$k_i^{\text{out}} = \sum_{j \in V} a_{ij}. \tag{2.5}$$

The equivalent to the handshake lemma in directed networks does not have the factor of two since each link contributes one to total in-degree and out-degree count:

$$\sum_{i \in V} k_i^{\text{in}} = \sum_{i \in V} k_i^{\text{out}} = L. \tag{2.6}$$

The *mean degree* of an undirected network is defined as

$$\langle k \rangle = \frac{1}{N} \sum_{i \in V} k_i. \tag{2.7}$$

Combining the expression for the mean degree and the identity given by eq. (2.3) we get another identity, true for any undirected network:

$$\langle k \rangle = \frac{2L}{N}. \tag{2.8}$$

This simple identity will be useful in Chapter 5 for deriving analytical expressions for various properties in random network models.

The equivalent descriptions of mean in-degree and mean out-degree in directed networks turn out to be equal:

$$\langle k^{\text{in}} \rangle = \frac{1}{N} \sum_{i \in V} k_i^{\text{in}} = \frac{1}{N} \sum_{i \in V} k_i^{\text{out}} = \langle k^{\text{out}} \rangle =: \langle k \rangle, \tag{2.9}$$

leading to

$$\langle k \rangle = \frac{L}{N}. \tag{2.10}$$

The *density* or *connectance* of an undirected network measures how many links are present in the network out of all possible links. In undirected networks this is given by

$$\rho = \frac{2L}{N\,(N-1)} = \frac{\langle k \rangle}{N-1}. \tag{2.11}$$

For directed networks, the factor of two is absent from the definition. A network in which $\rho \to 0$ as $N \to \infty$ is called *sparse* while for *dense* networks ρ remains non-zero. Equivalently, in sparse networks the number of links scales at most linearly with the number of nodes, $L \sim N$, while in dense networks it must scale with the square of the number of nodes, $L \sim N^2$. Of course, the definitions of sparse and dense networks make sense only in the infinite network limit, so they are of limited use to empirical networks. Nevertheless, most real world networks are in fact sparse [21, Sec. 2.4] in the sense that $L \ll N^2$.

The mean degree $\langle k \rangle$ and the density ρ together with the number of nodes N and links L forms a very basic but useful set of summary statistics to gauge the large-scale properties of the network in question at a glance. However, we can gain a lot more insight by studying the full *degree distribution* defined as

$$p_k = \frac{N_k}{N}, \tag{2.12}$$

where N_k is the number of nodes with degree k. The degree distribution has the interpretation that p_k is the probability of selecting a random node with degree k. The degree distribution has become a crucial network metric because it was discovered that many network structures from diverse areas exhibit a common pattern—their degree distribution is a power law [20, 22, 194]:

$$p_k \propto k^{-\gamma}. \tag{2.13}$$

In directed networks, the degree distribution is bivariate: p_{kl} denotes the probability of a random node simultaneously having an in-degree k and an out-degree l. The marginals $p_k = \sum_l p_{kl}$ and $p_l = \sum_k p_{kl}$ give the in-degree and out-degree distributions respectively.

2.2.2 Components

Networks can be *connected*—meaning there exists a path between any two nodes i and j in the network—or *disconnected*, in which case the network consists of a union of connected components. Components are of interest because the size and number of these have a direct impact on any dynamics on the network. For example, a contact network that is fragmented into many small components is immune to endemic infections as any disease will not be able to spread beyond the small component it originated in. A disconnected component in a power grid could equally be immune from power outages originating somewhere else in the network or, to the contrary, suffer from a complete power outage due to having been disconnected from the main

grid.

There are two qualitatively different types of components—the so called *giant components* and small components. In an empirical network it is common to find that a large proportion of nodes form one, big component while the rest of the network is composed of much smaller components, disconnected from the largest component (e.g. see fig. 2.4). Sometimes the largest component in an empirical network is called the giant component—however this is a slight abuse of terminology as there is a specific definition of what it means to be a giant component in network science. A *giant component* of a network model is defined as a component whose size grows linearly with the number of nodes N while *small components* are those whose size remains constant.

The component structure of directed networks is much more complicated [78, 233] due to the fundamental fact that the reachability of node j from node i via a directed path does not guarantee the existence of a reciprocal path from j to i. Because of this, we distinguish several types of giant components in directed networks. A directed network can be decomposed into a *giant weakly connected component* (GWCC), corresponding to the giant component if we replaced all directed links with undirected ones, and small directed components. Then, the GWCC can be decomposed further into the *giant strongly connected component* (GSCC), in which there is a directed path between any pair of nodes, the *giant in-component* (GIN), a set of nodes from which it is possible to reach the GSCC, and finally the *giant out-component* (GOUT), formed by the nodes that can be reached from the GSCC. This is supplemented by a hierarchy of finite, directed structures called *tubes* and *tendrils* [233].

We will describe the formation of giant components in some random network models in Section 2.3 and the study of giant components in relation to network clustering will be the topic of Chapter 5.

2.2.3 Clustering

A network property that has attracted a lot of attention due to its prominence in social networks [246] is that of *clustering*[1] or *transitivity*. Clustering is concerned with quantifying the number of triangles—closed loops of length three—in networks. The importance attached to triangles in networks comes from the simple observation that in social networks it is highly likely that many friends of a given person will be friends amongst themselves—thus leading to the presence of triangles.

[1] Not to be confused with finding closely related clusters of nodes as in community detection.

The extent of clustering in an undirected network is generally measured by the *clustering coefficient* (sometimes called the *global clustering coefficient* or *transitivity*) [185] defined as

$$C = \frac{3N_{\triangle}}{N_{\wedge}} = \frac{\operatorname{tr}\left(\mathbf{A}^3\right)}{\sum_{i \neq j}\left(\mathbf{A}^2\right)_{ij}}, \tag{2.14}$$

where $N_{\triangle}$ is the number of triangles and $N_{\wedge}$ is the number of connected triples in a network. Note that each triangle contributes three to the number of connected triples so that the factor of three in this definition normalizes the clustering coefficient to be in $[0, 1]$. This definition also admits a probabilistic interpretation—it is the probability that a randomly chosen connected triple is closed.

An alternative to the clustering coefficient is to consider the local clustering of each node [248]. The *local clustering coefficient* of node i quantifies how likely any two neighbours of node i are neighbours between themselves:

$$c_i = \frac{e_i}{\binom{k_i}{2}} = \frac{\sum_{j,m} a_{ij} a_{jm} a_{mi}}{k_i\left(k_i - 1\right)}, \tag{2.15}$$

where e_i is the number of such existing connections between node i's neighbours and $\binom{k_i}{2}$ is the maximum possible number of such connections. Having computed this for all nodes, one can average to obtain the *average clustering coefficient* of the network:

$$\bar{C} = \frac{1}{N}\sum_{i \in V} c_i. \tag{2.16}$$

Local clustering is mainly interesting because it has been shown to correlate with degree in many empirical networks [67, 243] often as a power law $c_k \propto k^{-\alpha}$, where c_k is the average clustering coefficient across nodes with degree k. Note that despite their similarities, the global and the average clustering coefficients are not equivalent, see fig. 2.1. Furthermore, it is possible to design networks whose global and average clustering coefficients demonstrably converge to different limits [217, Apendix C]. We shall favour the global clustering coefficient because of its easier interpretation and calculation in Chapter 5.

Definitions of clustering have been extended to directed networks [93], but the directedness of links have resulted in a proliferation of alternative definitions [54]. Instead, it is more common to analyse directed networks in terms of its *motifs* which we discuss next.

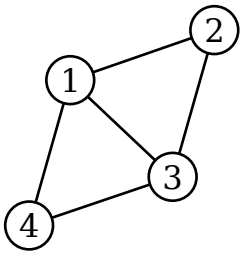

Figure 2.1: Example of a network in which $C = 0.75 \neq 0.8\dot{3} = \bar{C}$.

2.2.4 Network motifs

To reduce the complexities involved in studying large networks as a whole, a popular approach has been to focus on the occurrence patterns of much smaller subgraphs [171]. These smaller subgraphs called *motifs*[2] can offer insight into the evolution and organization of a networked system. This is because the recurrence of certain motifs can signal a particular functional role in the network [171, 222].

The study of network motifs is particularly popular in directed networks in which measures such as the clustering coefficient can be difficult to define. The smallest meaningful motifs (beyond single links) are the connected subgraphs on three nodes. In the undirected case there are only two such motifs—a triangle and a wedge which is not a triangle (also known as a connected triplet or a 2-star)—which naturally leads to the definition of the clustering coefficient—the proportion of wedges which are also triangles. On the other hand, in directed networks there are 13 possible motifs, see fig. 2.2. This number grows very quickly if we consider even larger structures—there are 6 motifs on four nodes in undirected networks [173] and 199 motifs in directed networks [206].

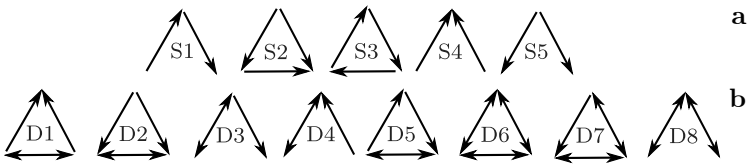

Figure 2.2: The 13 unique connected motifs on directed networks. These can be separated into two groups: (a) five triads, S1–S5, that only contain single links, (b) eight triads, D1–D8, that have double links.

Network motifs are interesting to study for the same reason as clustering—in many real networks some motifs are significantly more abundant than others,

[2]Sometimes the term is used to implicitly describe network subgraphs which occur at significantly higher numbers than expected according to some null model, here our use of the term is synonymous with *subgraph*.

especially when compared to some network null model. Practically, one measures the number of occurrences N_M of motif M in the real network, generates an ensemble of random networks to act as a null model and measure the occurrences of M in each of the random networks. We can then write down a statistical significance z-score for motif M:

$$z_M = \frac{N_M - \langle N_M \rangle_{\text{rand}}}{\sigma_{\text{rand}}}, \tag{2.17}$$

where $\langle N_M \rangle_{\text{rand}}$ and σ_{rand} are the randomized ensemble average and standard deviation of the appearance of M in the random ensemble, respectively. Thus, if z_M is significantly different from zero, it can be evidence of different mechanisms influencing the evolution of the real network.

The last question is, what null model to use? A popular choice is the configuration model which considers all networks with a fixed degree sequence. We discuss random network models next.

2.3 Random network models

Randomness is ubiquitous in nature. From a simple flipping of a coin to explaining the physical phenomena in thermodynamics, random numbers, random variables and the framework of statistical modelling is an indispensable tool. The same is true for network science in which random network models play a role analogous to random variables in statistics or ensembles in statistical physics.

The use of random network models in network science is twofold. First, they can serve as null models in hypothesis testing about real network formation. For example, comparing an empirical network such as a friendship network between schoolchildren to an ensemble of randomly generated networks with a fixed number of nodes and links (equal to the same quantities observed in the real network), we can test the hypothesis that social networks form completely randomly. If many of the network properties (degree distribution, clustering, etc.) in the artificial ensemble are not comparable to the same properties measured in the empirical network, we can reject the "full randomness" hypothesis and argue that there is evidence for some hidden mechanisms in social network formation or test the randomness hypothesis against some other random network model with more constraints [184].

Secondly, stochasticity is an inherent feature in empirical network formation. Indeed, at first sight real networks like the one in fig. 2.4 appear to be at least partially randomly connected and it is not unthinkable that in the real world there is a great degree of stochasticity when links of a network are formed. For example,

we only observe a single snapshot of the World Wide Web of hyperlinks pointing from website to website at any given moment. However, arguably the appearance of the observed network has been influenced by many small, random events and if we were to turn back time and repeat the evolution of the World Wide Web, we would ostensibly arrive at a different configuration. One of the major aims of network science is to formulate network models that reproduce properties of real networks thus shedding light on the evolution and function of networked systems but in doing so it is crucial to consider stochastic models of network formation. The term *complex networks* emphasizes the point that real networked systems often exhibit non-trivial, highly irregular properties, often driven by inherent stochasticity in model formation, as opposed to the structural regularity of lattice models (fig. 2.3) often studied in statistical physics [153].

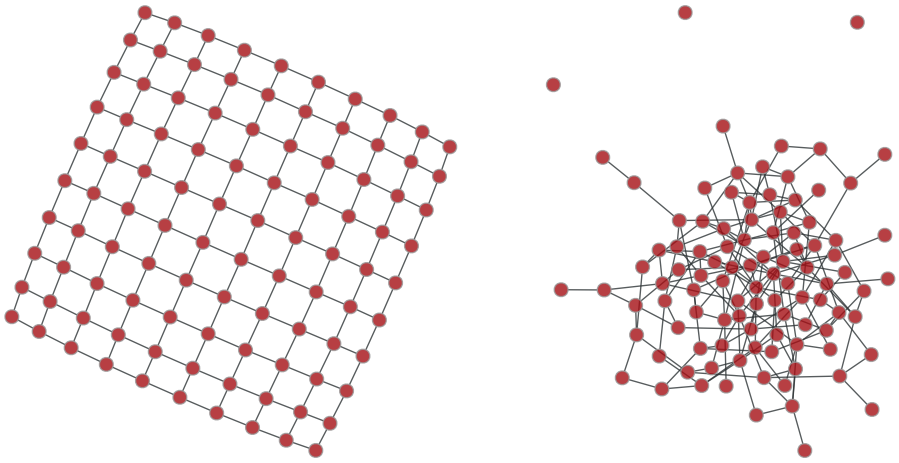

Figure 2.3: Left: regular 2D lattice. Right: Erdős-Rényi random graph. Both networks have $N = 100$ nodes and $L = 180$ links. Figure created using [198].

These considerations motivate the study of random network models, both as models of empirical network formation and in their own right. There is a plethora of of such models, but we focus our discussion on two of the most influential ones—the Erdős-Rényi random graph and the configuration model—which will also be the subject of Chapters 3 and 5.

2.3.1 Erdős-Rényi random graph

Erdős-Rényi random graph (henceforth ER model), sometimes just called the *random network model* or the *random graph*, is one of the simplest random network models and the subject of the first systematic studies of random network ensembles

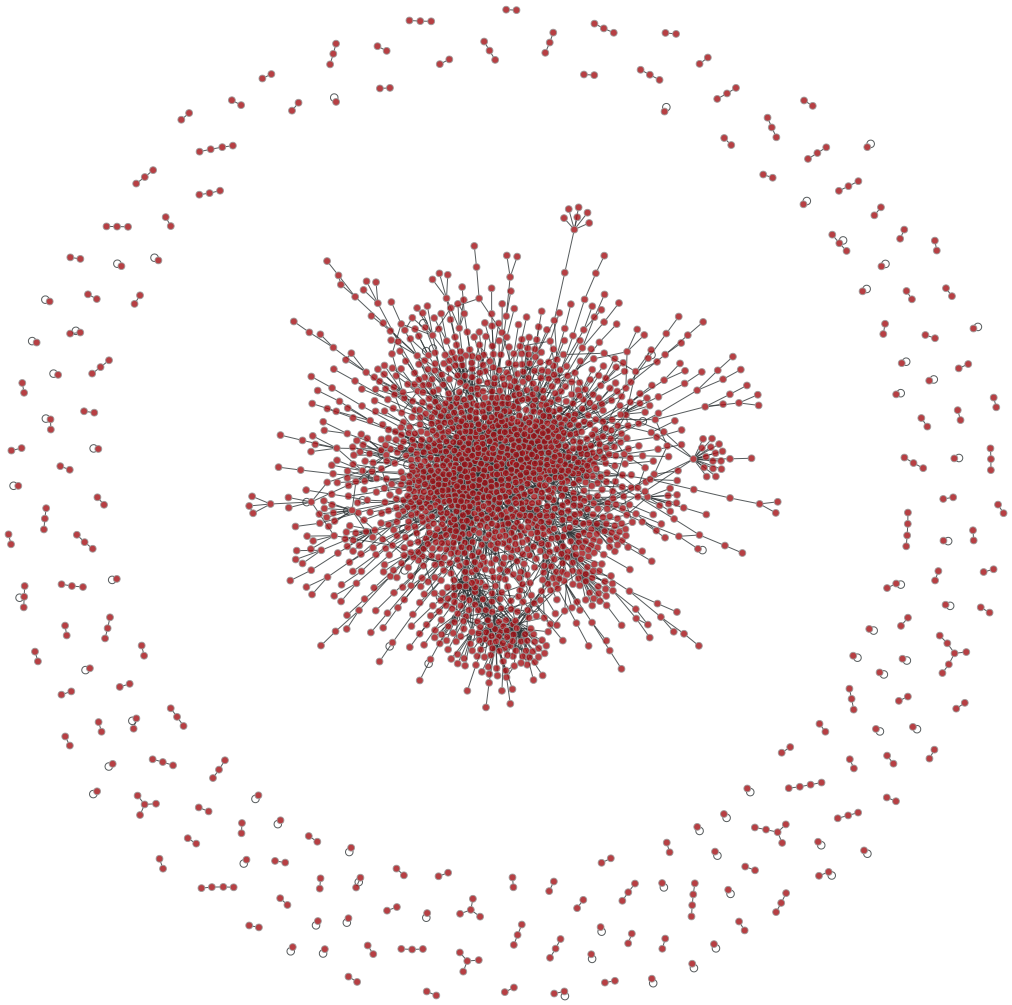

Figure 2.4: Protein interaction network of yeast *S. cerevisiae* with $N = 2018$ nodes representing proteins and $L = 2930$ links representing binding interactions [254]. Figure created using [198].

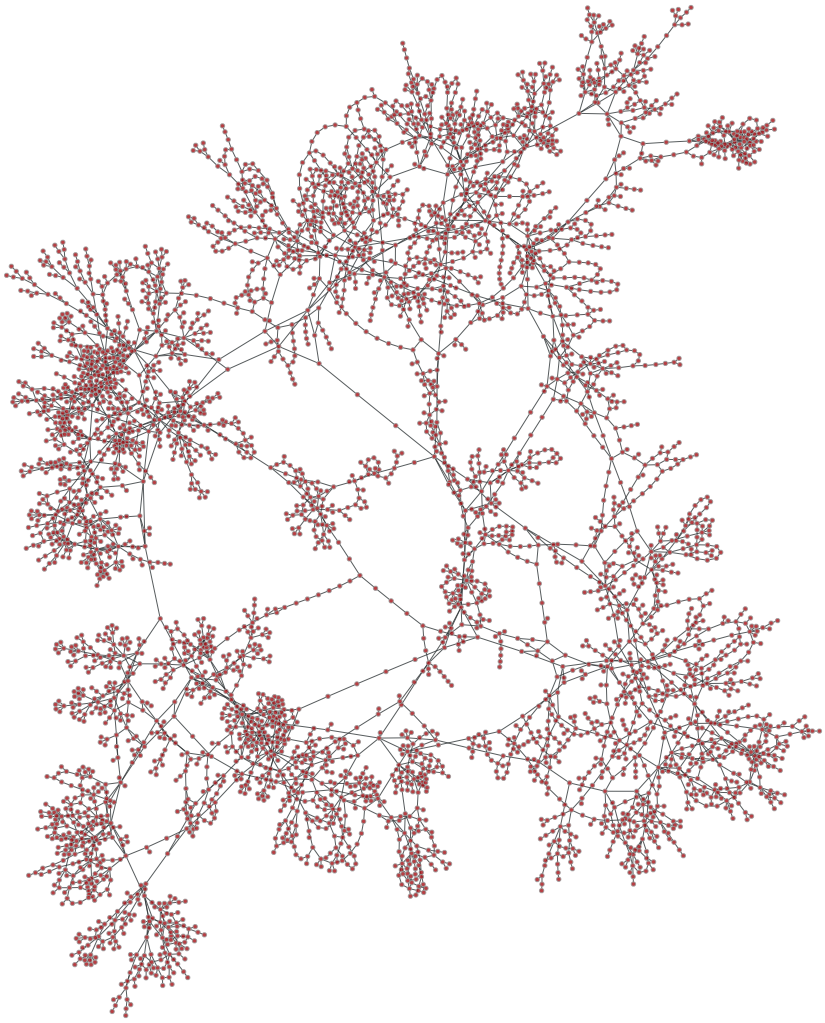

Figure 2.5: The network of the electrical power grid of the Western United States with $N = 4941$ nodes and $L = 6594$ links. The nodes represent generators, transformers and substations, and the links are high-voltage transmission lines between them [248]. Figure created using [198].

initiated by Erdős and Rényi [88, 89], and Gilbert [105]. The ER model, typically denoted by $G(N, L)$, as originally studied by Erdős and Rényi is defined by placing a fixed number of links L randomly between a fixed number of nodes N, prohibiting the placement of multiple links and self-loops so that the resulting network is a simple undirected network. Because there are $\binom{N}{2}$ choices for placing L links, each $G(N, L)$ network is only a specific realization amongst many potential candidates with different placements of links (see fig. 2.3 for an example realization). As such, a complete description of the properties of the $G(N, L)$ model would have to take into account the whole *statistical ensemble* of all possible realizations of networks. This is in fact a fundamental feature of all random network models—they are not defined in terms of a single randomly generated network but rather as an ensemble of networks by assigning a probability distribution over all possible network realizations. For example, in the case of the $G(N, L)$ model, the probability distribution $P(G)$ over all networks G is given by $P(G) = 1/\Omega_{N,L}$ for all simple networks G with exactly N nodes and L links, where $\Omega_{N,L}$ is the number of such networks, and zero otherwise [185].

A closely related model to the $G(N, L)$ model is the $G(N, p)$ model first studied by Gilbert [105]. In this model, instead of fixing the number of links L *a priori*, each of the potential $\binom{N}{2}$ links is realized with some fixed probability p. The $G(N, p)$ thus defines a different ensemble of random networks since, unlike in the $G(N, L)$ model, each network realization can potentially have a different number of links. The $G(N, p)$ model is usually preferred because it allows for easier calculations of network properties [21, 42, 185] and as such is the *de facto* version of the ER model. Finally, one can show that in the large network limit, both ensembles are equivalent [14, 225], i.e. when $N \to \infty$, most networks in the $G(N, p)$ ensemble have a number of links close to the expected number of links.

The ER model readily admits an extension to directed networks—the $G(N, L)$ model is exactly the same, but the links are directed, while the $G(N, p)$ model realizes $N(N-1)$ possible links with probability p as there are twice as many directed links possible compared to the undirected case.

We now describe some properties of the ER model that we will make use of in Chapter 5. Detailed derivations of these results can be found in any introductory book on network science such as Newman [185] or Barabási and Pósfai [21].

Probability of a network. As already discussed, the probability of generating a specific network in the $G(N, L)$ model is uniform in the total number of networks $\Omega_{N,L}$ with N nodes and L links, and zero otherwise. By contrast, in the $G(N, p)$

model any simple network can be generated with probability $p^L(1-p)^{\binom{N}{2}-L}$.

Mean degree. In the $G(N, L)$ model, the mean degree is just $\langle k \rangle = \frac{2L}{N}$ and is in fact the same for all network realizations. In the $G(N, p)$ model, the expected number of links is $\langle L \rangle = p\binom{N}{2}$ so that the mean degree is $\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$.

Degree distribution. In the $G(N, p)$ model the degree distribution is a binomial:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{2.18}$$

In many cases the $G(N, p)$ model is studied in the limit of large networks, i.e. as $N \to \infty$, since some of the analytical calculations become exact only in this limit. Typically the so called *sparse limit* is considered [185, p. 134] in which the mean degree $\langle k \rangle$ is fixed. In this limit the degree distribution is Poisson:

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \tag{2.19}$$

Because of this result, ER networks are sometimes called *Poisson random graphs*. This result is also the first to suggest that the ER model is inadequate for modelling real world networks because the degree distributions are rarely as homogeneous as a Poisson distribution and much more often feature heavy tails, often manifested as a power law degree distribution [20, 180].

Clustering coefficient. Because in the $G(N, p)$ model all links are independently realized with probability p, the clustering coefficient takes the same value $C = p = \frac{\langle k \rangle}{N-1}$. Thus, in the sparse limit the clustering coefficient vanishes and there are almost no triangles in a typical network. In fact, one can calculate the expected number of loops of any fixed size which turns out to be independent of network size [39]. This means that ER networks are locally *tree-like*—lacking in short loops as observed in many real networks. This constitutes a second major shortcoming of ER networks as a model for real networks.

Giant component. Consider the two limiting cases of the $G(N, p)$ model: for $p = 0$, there are no links in the network and it is completely disconnected while for $p = 1$, all of the possible $\binom{N}{2}$ links are present, the network is connected and is in fact the complete graph on N nodes. The largest connected component in each case is of size 1 and N respectively. However, if we increase N, in the first case for $p = 0$, the largest component is always of the same size 1, independent of network size,

while in the $p = 1$ case it is linear in N. Such a connected component that grows linearly with network size is called a *giant component*.

One of the most interesting properties of the ER model is that a giant component emerges spontaneously as we increase the connection probability p or equivalently the mean degree $\langle k \rangle$ in the sparse limit case. There exists a *critical value* of connection probability p_c (equivalently, a *critical mean degree* $\langle k \rangle_c$) at which the largest component stops being of fixed size and independent of N and becomes *extensive*—grows linearly with N. This phenomenon is called a *phase transition* and is ubiquitous in the natural sciences [226] and statistical mechanics [221], examples of which include the water-ice transition and the ferromagnetic-paramagnetic transition in magnetic materials, both manifesting under changing temperature. In the case of random networks, the connection probability p or the mean degree $\langle k \rangle$ are analogous to temperature.

It turns out that for the ER model, the transition to a giant component in the sparse limit occurs at $\langle k \rangle_c = 1$, that is, a giant component that grows linearly with N is present when the average degree of the network is at least one. Moreover, it is the only giant component as the second largest component only grows logarithmically with N [e.g. see 48, Ch. 6, or 83, Ch. 2], so most of the time we talk about *the giant component*. The emergence of the giant component is an example of phenomena studied in percolation theory [2].

It is important to note that the presence of a giant component is not the same as the whole network becoming connected. Indeed, a giant component can exist and only encompass a minority of nodes if the mean degree is only slightly larger than one. In fig. 2.6 we show the evolution of the size of the giant component in the $G(N, p)$ model in the sparse limit both analytically and via numerical simulation. We observe that there is no giant component present for any value of mean degree $\langle k \rangle < 1$, while for $\langle k \rangle > 1$ a giant component exists and occupies a fixed fraction of nodes in the network. Note, however, that for the network to become fully connected, the mean degree must be much larger than the critical degree $\langle k \rangle_c = 1$.

The emergence of a giant component in random network models with clustering will be our main focus in Chapter 5.

2.3.2 Configuration model

While the ER model was instrumental in launching the study of random networks, its application to modelling real network structures was quickly realized to be limited because many of its properties were not exhibited by real networks. Perhaps the most important roadblock to its success is that it produce networks with homo-

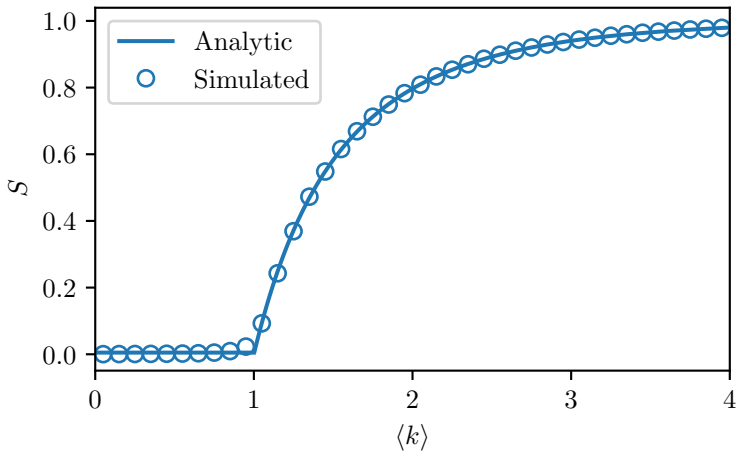

Figure 2.6: Emergence of the giant connected component in the $G(N, p)$ model. The solid line is the analytic solution for the fraction of nodes S in the giant component in the $N \to \infty$ limit while the circles are the mean values from a simulation of 100 networks with $N = 10^4$ nodes.

geneous, Poisson degree distributions, while real networks often exhibit heavy-tailed degree distributions such as the power-law $p_k \sim k^{-\gamma}$ [20, 22, 194]. To remedy this shortcoming, random network models with arbitrary degree distributions have been proposed, the most popular of which is the configuration model, sometimes also called the *generalized random graph model* as it is a generalization of the ER model.

The configuration model dates back to Bender and Canfield [32] and Bollobás [49] who define a random network ensemble with a fixed degree sequence. The idea is to fix the number of nodes N and the degree sequence $\mathbf{k} = \{k_1, k_2, \ldots, k_N\}$, and to construct realizations of networks with the given degree sequence by assigning to each node a number of *half-edges* or *stubs* equal to its degree and forming edges by randomly pairing the half-edges. This procedure generates a network with exactly the desired degree sequence. Importantly, this construction generates each network with degree sequence $\mathbf{k}$ with uniform probability [174] which is crucial for accurately calculating properties of the network ensemble in expected value.

Of course, not every integer sequence is a degree sequence—it has to satisfy some constraints to be able to generate a network, i.e. it has to be *graphic*. One trivial condition is the *handshake lemma* we already introduced in Section 2.2:

$$\sum_{i=1}^{N} k_i = 2L, \tag{2.20}$$

where L is the number of links. But this is not the only requirement for a degree

sequence to be graphic, a necessary and sufficient condition is given by the Erdős-Gallai theorem [101]. A popular algorithm for deciding whether a given degree sequence is graphic is the Havel-Hakimi algorithim [111, 114]. An improvement providing a linear time algorithm has recently been proposed [74].

The CM has some caveats, most importantly, because of the random matching of half-edges, it is prone to creating both self-edges and multiple connections leading to networks that are not simple. One might be tempted to reject any potential matchings of half-edges that would result in a self-edge or multi-edge, but doing so would violate the uniform sampling property and thus invalidate the analytical calculation of network properties [185]. Two other approaches to dealing with these shortcomings have been proposed—erasing any self-edges and multi-edges after the network construction or rejecting any completed networks that have these from the ensemble [51], but both methods generate slightly different ensembles with different network properties. In practice it is preferable to allow the creation of these unintended artefacts even if the real network systems that are being modelled do not have these features, the reason being that for a large class of degree sequences (i.e. those obeying the so called *structural cutoff* for which the maximum degree $k_{\max}$ is no larger than $(\langle k \rangle N)^{1/2}$ [45]) both the number of self-edges and multi-edges created during the CM process is fixed and independent of network size meaning that for large networks these constitute a negligible fraction of all links and can be safely ignored [185, p. 436].

These caveats have been resolved by a direct network construction method from a given degree sequence that has recently been proposed [41, 74]. This method generates biased samples of simple networks from the CM with appropriate weights, enabling the calculation of average properties in the network ensemble. This technique has also been extended to directed networks [140] and networks with prescribed degree correlations [27].

A slightly different definition of the CM was introduced by Newman et al. [183] who, instead of fixing the degree sequence $\mathbf{k}$ of the ensemble at the outset, prescribe drawing a valid degree sequence from a fixed degree distribution p_k for each new network realization. This means that the degree sequence will be different from network to network, but the resulting ensemble of networks will have the same degree distribution on average.

The two different ensembles are sometimes referred to as the *microcanonical ensemble* (with *hard constraints*) and the *canonical ensemble* (with *soft constraints*) in analogy to the two maximum entropy ensembles in statistical physics in which either all possible states with a fixed energy E are considered (microcanonical) or

the states can have different energies as long as the average energy $\langle E \rangle$ is fixed (canonical) [221]. The existence of two distinct CM ensembles also parallels the two distinct ER models—the fixed link model $G(N, L)$ and the average fixed link model $G(N, p)$, and the correspondence to the microcanonical and canonical ensembles of the ER models is the same. Interestingly, while the two ER models are equivalent in the thermodynamic $N \to \infty$ limit, the two CM ensembles are not equivalent even in the $N \to \infty$ limit due to the fact that the number of constraints in this case grows linearly with N unlike in the ER case [14, 225]. This motivates the need for a principled choice of exactly which model to use in applications as results obtained from either ensemble may differ significantly.

The CM model can also be generalized to directed networks. In this case, instead of specifying the degree sequence $\mathbf{k} = \{k_1, k_2, \ldots, k_N\}$, we specify the *bi-degree* sequence of pairs of in-degrees and out-degrees: $\mathbf{k} = \{(k_1^{\text{in}}, k_1^{\text{out}}), (k_2^{\text{in}}, k_2^{\text{out}}), \ldots, (k_N^{\text{in}}, k_N^{\text{out}})\}$ [90, 140]. As with the undirected version, one can consider instead the canonical ensemble by specifying the bivariate joint degree distribution p_{kl}.

Link probability Unlike in the ER model in which each link has the same probability p of being realized, the CM model has heterogeneous link probabilities which depend on the degree sequence. For a large class of degree distributions which are not heavy tailed [15, 36, 185], in the CM model the probability of creating a link between nodes i and j is given by

$$p_{ij} = \frac{k_i k_j}{\langle k \rangle N}. \tag{2.21}$$

This expression, however, is generally not valid for heavy-tailed degree distributions such as the power-law. In such cases, the expressions for p_{ij} take a more complicated form involving "hidden variables" [46, 60] related to Lagrange multipliers arising from the constrained maximisation problem when solving for p_{ij} [15, 224]. This slightly more complex situation corresponds to natural correlations arising in the CM for heavy-tailed degree distributions which are not present for more homogeneous degree distributions [36, 46, 129]. The link probabilities p_{ij} allows one to calculate the probability of generating any particular network structure as in the case of the ER model, although the precise expression is more involved [15].

Clustering coefficient. The clustering coefficient in the CM is given by

$$C = \frac{1}{N} \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)^2}{\langle k \rangle^3}. \tag{2.22}$$

Thus, clustering only depends on the first two moments of the degree distribution. Like in the ER model, the clustering coefficient scales as N^{-1} and vanishes in the large network limit, however, an important difference is that this more general expression involves the second moment of the degree distribution $\langle k^2 \rangle$ in the numerator which can be large in highly skewed degree distributions like the power law $p_k \sim k^{-\gamma}$ for $\gamma \in (2, 3)$ and thus make the clustering coefficient non-negligible in finite networks [66].

Giant component. The emergence of the giant component in the CM was first studied by Molloy and Reed [174, 175] who showed that the CM has a giant component if and only if

$$\langle k^2 \rangle - 2\langle k \rangle > 0. \tag{2.23}$$

This condition, now commonly known as the *Molloy-Reed condition*, thus generalizes the corresponding result for the ER model and, like the clustering coefficient, only depends on the first two moments of the degree distribution. Because of the appearance of the second moment of the degree distribution, as in the case of the clustering coefficient, for power law degree distributions with the exponent $\gamma < 3$ a giant component is always present. This fact has been used to show that power-law networks are exceptionally resilient to random failures of nodes [64] but at the same time susceptible to targeted attacks [65] as well as large scale invasions by infectious diseases [181, 192, 193].

2.3.3 Other important models

We briefly mention a few other influential random network models that will not be the subject of this thesis.

The *small-world* network model, also known as the *Watts-Strogatz* (WS) model is a model of networks that exhibits both the small-world property—the typical average distance between nodes is logarithmic in the network size—and high clustering [248]. It is based on a probabilistic rewiring procedure of edges and interpolates between a regular ring lattice and an Erdős-Rényi random graph. The main limitation of this model is the homogeneous degree distribution.

The *preferential attachment* model, also know as the *Barabási-Albert* (BA) model is a growing network model formulated to explain the formation of the World Wide Web [20]. The "growth" part refers to new nodes arriving in the system while "preferential attachment" ensures that they get preferentially linked with existing

high-degree nodes. The main appeal of this model is that it produces a true scale-free degree distribution $p_k \propto k^{-3}$, however its clustering coefficient still vanishes with system size as $C \propto N^{-0.75}$ albeit slower than in the ER random graph. The BA model has also been extended to be able to generate scale-free degree distributions with any exponent in the range $(-2, \infty)$ [144, 145]. It has also been modified to include intrinsic node fitness that influences the preferential attachment beyond just the node degree, this is known as the *Bianconi–Barabási* model [38] and can account for degree correlations in modelling the Internet. It also exhibits interesting condensation transitions [37].

The *stochastic block model* [118] is a network model for producing networks with community structure—subsets of nodes characterized by denser connections between them than with other "external" nodes [108]. This model is basically an extension of the ER model in that it partitions the nodes into sets of disjoint partitions (communities) and assigns links between nodes in communities and links between nodes from different communities with altered probabilities. Various extensions have been proposed to account for more realistic network structures: overlapping communities [4], degree heterogeneity [134], directed networks and multigraphs [197], and weighted networks [3] to name a few. Networks with community structure and community detection in empirical networks is a vast subject area in network science, comprehensive reviews of the subject include Fortunato [96], Fortunato and Hric [97].

Spatial networks or *geometric graphs* are models of networks in which nodes are explicitly embedded in some underlying coordinate space and links between nodes are made via some distance metric. The study of spatial networks goes back to Gilbert [106] and has attracted a lot of attention in the following decades [24, 199]. Spatial networks are attractive to model real systems in which the underlying space is relevant, examples include transportation networks [152], Internet [252], power grids (fig. 2.5) [5], neural networks in the brain [52] and many others.

Exponential random graphs are a class of generative network models that apply Jayne's principle of maximum entropy [126, 127] to generate maximally random networks with pre-defined constraints. This method borrows heavily from statistical mechanics in that it treats a realization of a specific network as a *microstate* of the statistical ensemble of all random networks with the desired macroscopic properties (enforced by constraints). Random networks are obtained by maximising the Gibbs entropy [70] of the probability distribution of graphs subject to constraints of interest (e.g. number of triangles/clustering [53, 191, 230], degree sequence [14, 15, 36], degree correlations [34], community structure [100]). This method is very flexible

because almost any network structure can be encoded as a constrained optimization problem. The drawback is that the models get analytically intractable very quickly when the constraint go beyond very simple [190] and even sampling the ensemble via Markov Chain Monte Carlo can yield unsatisfactory results [121, 191]. These types of random network models have been particularly popular in social sciences where they are also known as p^* models [213, 214].

CHAPTER 3

NOVEL PATTERNS IN FOOD WEB MODELLING

Food webs are abstract representations of which species consume which others in an ecosystem [62, 87, 189, 201]. In a network-based description, species are represented by nodes and their trophic interactions are represented by directed links, pointing from prey to predator [81, 82, 201]. In practice, food webs are inferred by field ecologists, typically by studying species gut content or by direct observation of animal diet (e.g. see [163] for a description of the resolution of Little Rock Lake food web). Figure 3.1 is an example of such a food web and illustrates why directed networks are the natural theoretical abstraction for the study of their properties.

Much work has been devoted to understanding the origin and meaning of the particular feeding patterns observed in various ecosystems [103, 167, 202]. Faced with the complexity of whole food webs, many researchers have focused on the interactions among subsets of species, through the analysis of small, connected subgraphs, or *motifs* [25, 50, 56, 196, 229]. However, the emergence of these motifs corresponding to specific preying behaviours is still not well understood. In this chapter we study the structure of 46 empirical food webs in the context of a new food web metric and accompanying model which allows us to classify food webs into two broad families according to their motif structure.

3.1 Background

3.1.1 Trophic coherence

Consider a directed network of N nodes characterized by the $N \times N$ adjacency matrix $\mathbf{A}$ with elements $a_{ij} = 1$ if there is a directed link from i to j and $a_{ij} = 0$

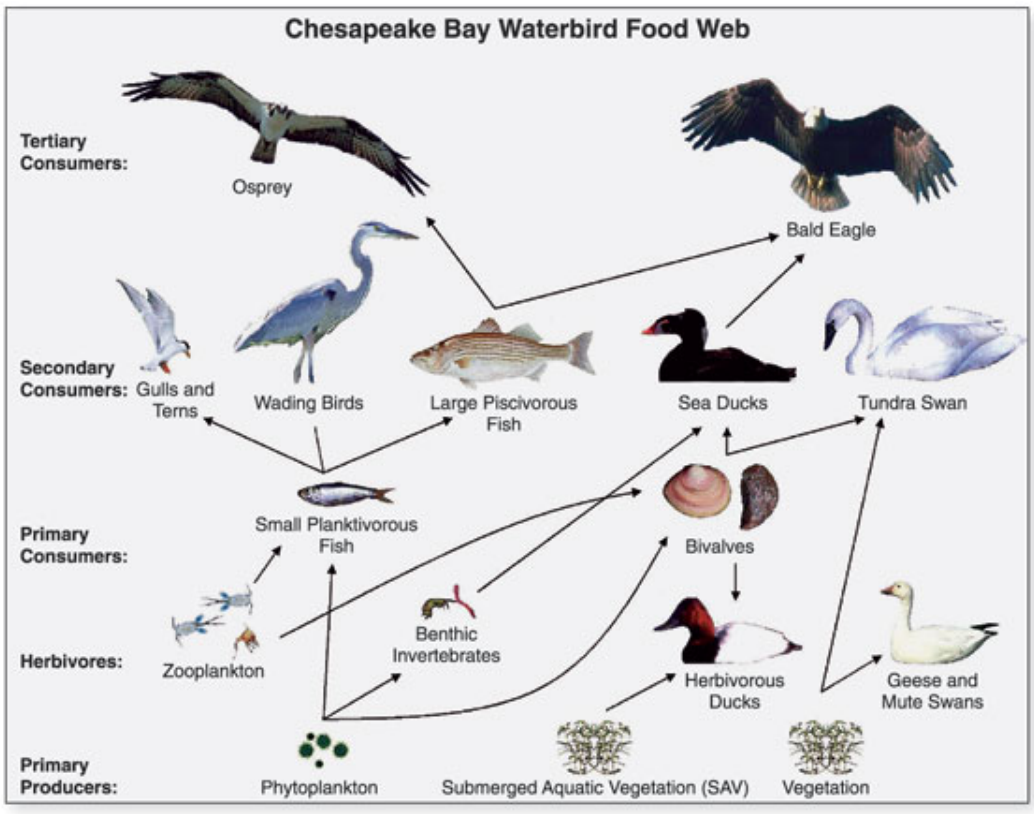

Figure 3.1: A subset of the Chesapeake Bay food web focusing on trophic interactions between water birds [200].

otherwise. The *in-degree* and the *out-degree* of node j are defined as $k_j^{\text{in}} = \sum_i a_{ij}$ and $k_j^{\text{out}} = \sum_i a_{ji}$ respectively. We shall assume that our network is weakly connected—there is only one weakly connected component—and that there is a positive number of nodes $B > 0$ with $k^{\text{in}} = 0$ which we will call *basal nodes*. It is standard in ecology [131, 154] to define the *trophic level* s_j of node j as

$$s_j = 1 + \frac{1}{k_j^{\text{in}}} \sum_i a_{ij} s_i \tag{3.1}$$

if $k_j^{\text{in}} > 0$ or $s_j = 1$ if $k_j^{\text{in}} = 0$. In other words, the trophic level of basal nodes (autotrophs in the ecological context) is set to be $s = 1$ by convention while the other nodes (consumer species) are assigned the average trophic level of their in-neighbours (prey, resources), plus one.

Equation (3.1) defines a linear system which can be solved for s_j to assign a trophic level to every node in the network. In order to do this, two conditions must hold. First, there must be at least one basal node, $B > 0$ which we have assumed is the case; and second, every node in the network must be reachable by a path from at least one basal node. Note that food webs in particular always satisfy both conditions so that a unique solution for trophic levels exists.

Next, following Johnson et al. [131], we associate a *trophic distance* to each link ij by $x_{ij} = s_j - s_i$.[1] Then consider the distribution of trophic distances $p(x)$ as measured on a network. By eq. (3.1), this always has mean $\langle x \rangle = 1$ and standard deviation $q = \sqrt{\langle x^2 \rangle - 1}$ which we will call the *trophic incoherence* parameter.

The trophic incoherence parameter is thus a measure of the homogeneity of the trophic distance distribution $p(x)$: the more similar the trophic distances of all links, the more *coherent* the network. For perfectly coherent networks we have $q = 0$ which translates to having only integer valued trophic levels. In the context of ecology, this would characterize a food web in which species feed on prey only one trophic level below their own and there are distinct groups of "herbivores" feeding only on basal species, "predators" feeding only on "herbivores" and so on. For less coherent networks, $q > 0$ indicates a less ordered trophic structure where the trophic levels can have fractional values and species tend to prey on a broader range of trophic levels. Thus trophic coherence can be seen as an average measure of trophic specialization of consumer species in a food web.

Figure 3.2 shows two examples of food webs with vastly different trophic structures: Crystal Lake (Delta) [239], a highly coherent network with $q = 0.17$ and Coachella Valley [204], a highly incoherent terrestrial food web with $q = 1.21$.

[1] This is not a distance in the mathematical sense since it can assume negative values.

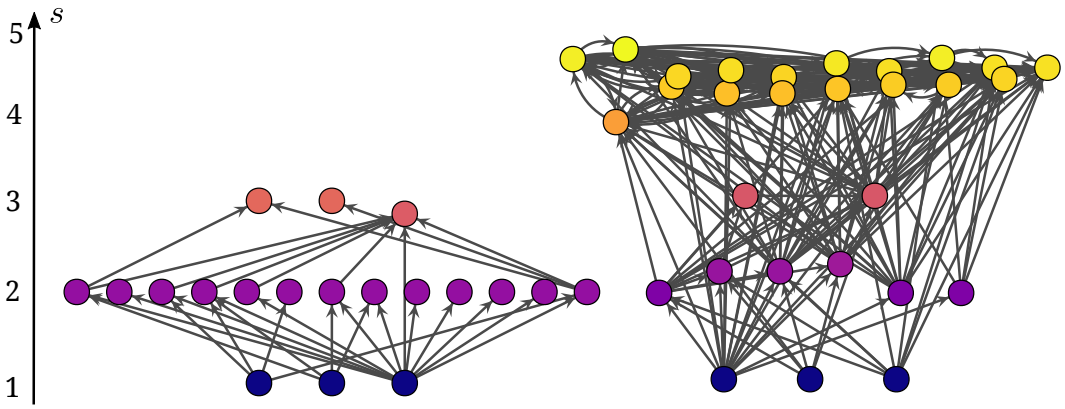

Figure 3.2: Examples of different degrees of trophic coherence in food webs. Left: Crystal Lake (Delta) [239]—a highly coherent network with $q = 0.17$, note that only one link prevents the network from being perfectly coherent ($q = 0$). Right: Coachella Valley [204]—an incoherent network with $q = 1.21$, note the high number of nodes falling between integer trophic levels due to the complex patterns of trophic links.

Trophic coherence as measured by q has been shown to be the best statistical predictor of linear stability as well as a number of structural properties of empirical food webs [76, 131]. It has also been related to the cycle structure and distribution of eigenvalues in directed networks [128].

3.1.2 Food web motifs

Faced with the complexities of studying large networked systems, scientists have sought to gain insight by looking at the role of smaller local modules or subgraphs embedded within the whole network structure [123], sometimes termed "building blocks" of networks [171]. There is a large scientific interest in studying the subgraph composition of both empirical networks [143, 206] and popular network models [18, 99, 123], studying correlations between subgraph counts and global network measures [243], and formulating network models that can prescribe different subgraph frequencies in a bid to uncover a new class of null models [95, 133, 156, 211, 212].

Because the number of possible non-isomorphic graphs on n vertices grows very quickly even for small n, it is customary to study the smallest possible subgraphs that would give rise to non-trivial behaviour. In particular, as we are studying directed networks, we will be focusing on the three-node connected subgraphs (henceforth *triads*) of which there are exactly 13 distinct ones (fig. 3.3).

Subgraphs such as these have received a lot of attention in food web modelling because many of these triads admit a straightforward ecological interpretation [229].

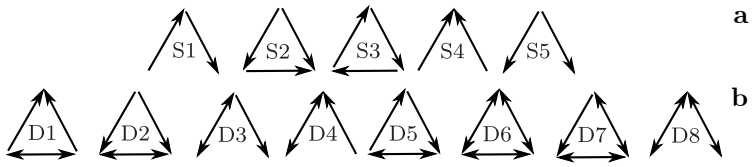

Figure 3.3: The 13 unique connected triads on directed networks. These can be separated into two groups: (a) five triads, S1–S5, that only contain single links, (b) eight triads, D1–D8, that have double links (corresponding to mutual predation).

The eight triads D1–D8 feature double links which correspond to mutual predation between two species. The five single link triads S1–S5 consist of some of the more basic building blocks of food webs. The triad S1 is a simple food chain [25, 62], S2 represents omnivory (a predator preying on two species at different trophic levels) [25, 204], triad S3 is a cycle (a relatively rare feature) [204, 229], and triads S4–S5 represent a predator preying on two species (apparent competition) and two predators sharing a prey species (direct competition), respectively [50].

One of the most successful methods for gaining insight into the origin and function of small substructures has been to study their *relative* prevalence in an empirically observed network against some null model of a synthetic network which incorporates some ground hypothesis about the formation process of the empirical network [171, 173]. Applying this technique to food webs, ecologists have developed several competing hypotheses for the relative prevalence of the three-node triads. The prevailing hypotheses are that these triads emerge as a result of physical constraints (e.g. species body size, abstracted by the "niche dimension") in the assembly of food webs [56, 229], that functional importance leads to the observed structural patterns [206], or that certain stability properties favour some subgraphs over others [50].

Attempts to explain subgraph patterns using the two most established food web models, the generalized cascade model [228] and the generalized niche model [250], have been unsatisfactory since either model produces food webs with rigid three-species subgraph patterns [56, 229] while real food-webs display a far richer array of local preying patterns [50, 229]. To remedy this disagreement between theory and observation, we study a new food-web model, the Generalized Preferential Preying Model (GPPM) [131] which takes into account the trophic coherence of networks and can accurately predict the three-species subgraph patterns across a wide array of distinct types of food-webs.

3.1.3 Food web models

We briefly describe the two most widely used structural food web models—the Generalized Cascade Model and the Generalized Niche Model—before introducing the starting point of modelling food webs with trophic structure via the Preferential Preying Model.

Cascade Model and Generalized Cascade Model The original Cascade Model [61] assigns a a random number n_i to each species i from the uniform distribution on $[0, 1]$. Then for all pairs (i, j), i is set to be the consumer of j with some constant probability p if $n_i > n_j$ and with probability zero otherwise. If the number of species is set to be N and the expected number of links to L, the constant probability takes the form

$$p = \frac{2L}{N\,(N-1)}. \tag{3.2}$$

The Cascade Model thus assumes a "pecking order" of species and a rule under which no species can consume prey higher than itself in the order. On the other hand, the probability of a species preying on a lower-ranked species is constant regardless of how far down the hierarchy the potential prey exists. This was the first attempt to show that networks similar to real food webs could be generated via simple rules and imposing a hierarchy of species.

The model was later modified so that the number of prey would be instead drawn from a Beta distribution as in the Niche Model [250] and the new model was called the Generalized Cascade Model [227]. This amendment improved the model's predictions with respect to the distribution of prey and predators (i.e. the empirically observed degree distributions).

Niche Model and Generalized Niche Model As in the Cascade Model, in the Niche Model [250] each species i is awarded a uniform random number in $[0, 1)$ called the *niche value*, but the model refines the Cascade Model by constraining a species i diet to a subset of species j with lower niche value. Specifically, species i is constrained to consume species j such that $c_i - r_i/2 \leq n_j < c_i + r_i/2$, i.e. all species within an interval of size r_i centred at c_i and no others. The range parameter is defined as $r_i = x_i n_i$ where x_i is drawn from the Beta distribution with parameters $(1, \beta)$. For a fixed number of species N and an expected number of links L, the parameter β takes the form

$$\beta = \frac{N\,(N-1)}{2L} - 1. \tag{3.3}$$

The centres of the intervals c_i are drawn from the uniform distribution on $[r_i/2, n_i]$.

The Niche Model is an improvement on the Cascade Model by allowing for cannibalism as well as looping—propensity to consume prey above a species trophic niche. The motivation behind the Niche Model is the so called *intervality hypothesis*—that there is an intrinsic ordering of species such that the prey of any given predator are contiguous [63]. Recent work has shown that food webs are biased towards intervality [228], although they are generally not perfectly interval, especially if species body size is taken as a proxy for niche value [57]. Nevertheless, the Niche Model has been very successful as it outperforms the Cascade Model in predicting many structural features of food webs [250].

The Niche Model was later modified to account for the fact that real food webs are not maximally interval by introducing an additional *contiguity* parameter c which determines the proportion of prey to be allocated according to the original Niche Model prescription and the rest according to the Generalized Cascade Model [228]. This modified model is known as the Generalized Niche Model and is implemented as the original Niche Model, but with reduced ranges $r_i = cx_in_i$. Then for each species a number of extra prey $k_i^{\text{cascade}} = (1-c)x_in_iN$ is drawn randomly from all the available species with niche values lower than n_i as in the Generalized Cascade Model. Thus, the Generalized Niche Model interpolates between the Niche Model ($c = 1$) and the Generalized Cascade Model ($c = 0$).

Preferential Preying Model (PPM) The Preferential Preying Model (PPM) was introduced by Johnson et al. [131] as the first food web model to explicitly model trophic coherence in food webs. It is a growing network model, inspired by the Barabási-Albert model of preferential attachment [20], but in the context of a growing ecosystem through immigration of new species. The model starts with B nodes (basal species) with trophic levels $s = 1$ and no links. New nodes are then added sequentially until the total number of nodes reaches N. Every new node i is first awarded a random prey node j from all the existing nodes in the network and assigned a temporary trophic level $\hat{s}_i = 1 + s_j$. Following this, another κ_i prey nodes l are chosen with a probability that decays with the tentative trophic distance $\hat{x}_{il} = s_l - \hat{s}_i$, specifically:

$$P_{il} \propto \exp\left(-\frac{|\hat{x}_{il} - 1|}{T}\right), \tag{3.4}$$

where T is a "temperature" parameter that tunes the trophic coherence: for $T = 0$, perfectly coherent networks are generated ($q = 0$), while q increases monotonically for $T > 0$.

The number of prey κ_i in the original PPM is obtained similarly to the Niche Model method [227] by setting $\kappa_i = x_i n_i$, where n_i is the number of nodes in the network when node i arrives and x_i is a Beta random variable with parameters $(1, \beta)$ such that

$$\beta = \frac{N^2 - B^2}{2L} - 1, \tag{3.5}$$

where L is the expected number of links in the resulting network. This way the Generalized Cascade Model is recovered in the $T \to \infty$ limit while the degree distributions are as in the Generalized Niche Model.

Even though the PPM captures many empirically observed patterns in food webs, it suffers from a major drawback, namely it can only generate acyclic networks. While this is a property that is shared by many food webs, it is far from universal [56] and, in particular, makes the PPM unsuitable for studying mutual predation or food web motifs. To remedy this, we now describe an extension of the PPM that can generate cycles.

3.2 Generalized Preferential Preying Model (GPPM)

In this section we extend the PPM model introduced by Johnson et al. [131] in a way that allows bidirectional links (corresponding to mutual predation) and cycles of higher order to form thus producing more realistic networks. Again, we denote by B, N and L the number of basal nodes (autotrophs), the number of total nodes and the number of links in the network respectively as input parameters.

The network construction begins with B basal nodes and no links. We assign trophic level $s = 1$ for all basal nodes. We then proceed to introduce $N - B$ new, non-basal nodes in the network sequentially. For each such new node j, we pick exactly one prey node or in-neighbour i at random from among all the existing node in the network thus creating a link from i to j. In doing so, we assign a *temporary* trophic level to node j as $\hat{s}_j = 1 + \hat{s}_i$. After this procedure finishes, we have a network of N nodes, $N - B$ links and each node i has a temporary trophic level $\hat{s}_i$ that is an integer (since every non-basal node has only one in-neighbour at this point).

Finally, we add the remaining links to the network to bring the total number of links up to L (in expected value). For this, we choose links to be placed among

all possible pairs of disconnected nodes (i, j) where j is not a basal node[2] with a probability P_{ij} that decays with the *tentative* trophic distance $\hat{x}_{ij} = \hat{s}_j - \hat{s}_i$ between them. Specifically, we set

$$P_{ij} \propto \exp\left(-\frac{(\hat{x}_{ij} - 1)^2}{2T^2}\right). \tag{3.6}$$

As in the original PPM, the "temperature" parameter T sets the amount of deviation from a perfectly coherent network and the probabilities are normalized so that the expected number of links in the final network is L.[3] At the end of the network creation process the trophic levels need to be recalculated according to eq. (3.1) as the addition of new links will have changed the network topology, and the trophic levels in the final network need not correspond to the temporary integer valued trophic levels.

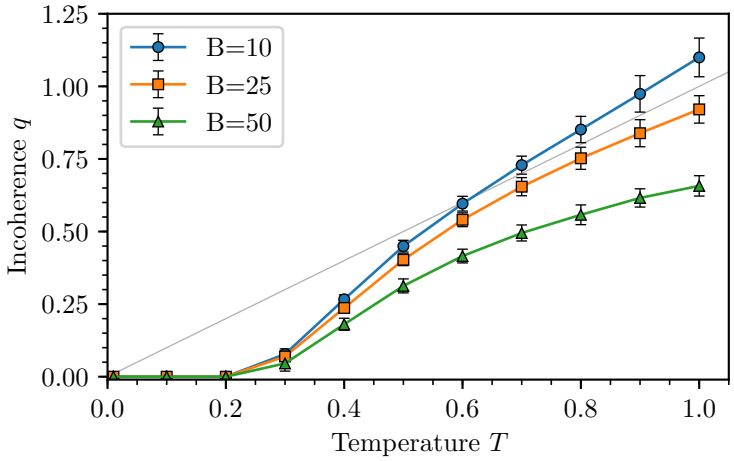

Figure 3.4: Dependence of the incoherence parameter q on the temperature parameter T. Simulated ensembles of networks have $N = 100$ nodes B of which are basal and average non-basal degree $\langle k \rangle = L/(N - B) = 10$. The averages are computed over at least 1000 networks and error bars are one standard deviation of the sample. The grey line is $f(x) = x$.

A sample dependence of the trophic incoherence parameter q on T is shown in fig. 3.4. The model exhibits a monotonic dependence of q on T which provides a basis for fitting the model to empirical food webs given the empirically observed value of q. We also note that the level of incoherence that is achieved at any given

[2]This ensures no incoming links to basal nodes which would make them non-basal.

[3]In principle, the link addition process can be amended so that the final number of links is exactly L not just on average. However, this can be detrimental because in some situations, e.g. for low T and high link density, the exact number of links L may not be attainable without distorting the probability distribution set by eq. (3.6).

temperature depends on B/N, the ratio of basal species to all species. Before fitting the model to empirical data, we explore some interesting properties of the GPPM in the following section.

3.3 Properties of the GPPM

3.3.1 Limits

The GPPM has two interesting limiting behaviours as we change the temperature parameter T. As $T \to 0$, the probability of forming links between trophic levels more than one apart tends to zero so the model generates perfectly layered networks with $q = 0$. In the other extreme, as $T \to \infty$, the probability of forming links between different trophic levels becomes homogeneous and the generated networks become increasingly more random. In this limit the model generates directed Erdős-Rényi random graphs with two restrictions (biases) that make it different from fully random directed ER graphs:

1. The basal nodes are still precluded from receiving any incoming connections.
2. The initial "skeleton" of $N - B$ links is still present, ensuring that the network is weakly connected.

This interesting deviation from the ER random graph has lead to the study of the *basal network ensemble* which is defined as a subset of the ER random graph ensemble in which a number of basal nodes B is fixed and cannot receive any incoming connections and for every non-basal node, the proportion of incoming links coming from basal nodes is the same. This is equivalent to fixing the mean degree $\langle k \rangle$ in the ER ensemble and leads to highly homogeneous networks in which the distribution of trophic levels is bimodal [128].

3.3.2 Distribution of trophic levels

We can calculate the distribution of trophic levels in the GPPM model when $T = 0$ exactly and the mean trophic level in the limit $T \to \infty$ which leads to an interesting crossover behaviour of the mean trophic level in coherent and incoherent networks.

Let $T = 0$ be fixed, then links can only be formed between integer valued trophic levels. Furthermore, it is only necessary to consider the first stage of introducing $N-B$ non-basal nodes and their links because the second stage of introducing the remaining $L-N+B$ links at $T = 0$ will not alter the values of the trophic levels.

As before, start with B basal nodes and consider time steps $t = 0, \ldots, N-B$, where at teach time step a new node is added with one incoming link chosen randomly from all the existing nodes already in the network. Let $n_s(t)$ be the expected number of nodes on level s at time t. Then the first level satisfies $n_1(t) = B$ for all time by the definition of basal nodes. For higher trophic levels we can write

$$\frac{d}{dt} n_{s+1}(t) = \frac{n_s(t)}{B+t}, \tag{3.7}$$

since the expected increase in the number of nodes at level $s+1$ is the fraction of nodes which are at level s. Rescaling the time variable as $x = (t+B)/B$ this reads

$$\frac{d}{dx} n_{s+1}(x) = \frac{n_s(x)}{x}. \tag{3.8}$$

This equation can be solved iteratively for any level s given the initial condition $n_s(x=1) = n_s(t=0) = 0$ for all $s \neq 1$. This gives the general solution

$$n_s(x) = \frac{B}{(s-1)!} \log^{s-1}(x). \tag{3.9}$$

Using this, the distribution of trophic levels $p_s(x)$ is obtained by dividing $n_s(x)$ by the total number of nodes at time t which is $B+t$. Solving for $t = N-B \Rightarrow x = N/B$, we obtain the distribution of trophic levels once the network has been assembled:

$$p_s = \frac{B}{N} \frac{1}{(s-1)!} \log^{s-1}\left(\frac{N}{B}\right). \tag{3.10}$$

Letting $\lambda = \log(N/B)$ we note that this is a shifted Poisson distribution with support $s \in 1, 2, \ldots$. This gives for the mean trophic level

$$\langle s \rangle = \log\left(\frac{N}{B}\right) + 1. \tag{3.11}$$

Thus, the mean trophic level for coherent networks is logarithmic in the ratio of total number of nodes to basal nodes. This suggests a slow growth in the mean trophic level and consequently the mean food chain length in coherent food webs as a function of the number of species.

What about when $T \neq 0$? We can't say much about the general case, but we can perform the calculation of the mean trophic level $\langle s^\infty \rangle$ when $T \to \infty$. Again consider B basal nodes and $N-B$ non-basal nodes connected via $N-B$ links at the end of the growth part of the model. The infinite temperature limit means that we can place the remaining $L-N+B$ links randomly between the nodes as long as we disallow incoming connections to basal nodes. For simplicity we consider the

effect of the $N - B$ initial links to have negligible impact on the final values of the trophic level, i.e. we assume that $L \gg N - B$.

Because all links can be placed randomly between the nodes, at the end of the network creation it must be the case that each non-basal node has, on average, the same incoming degree $\langle k_{\mathrm{NB}}^{\mathrm{in}} \rangle$ which is made up of the same proportion of links coming from basal nodes, L_{B}, and links coming from other non-basal nodes, L_{NB}:

$$\langle k_{\mathrm{NB}}^{\mathrm{in}} \rangle = \frac{L}{N-B} = \frac{L_{\mathrm{B}} + L_{\mathrm{NB}}}{N-B}. \tag{3.12}$$

This together with the definition of trophic level (eq. (3.1)) allows us to write down a self-consistent equation for the mean level of non-basal nodes:

$$\langle s_{\mathrm{NB}} \rangle = 1 + \frac{L_{\mathrm{NB}} \langle s_{\mathrm{NB}} \rangle}{L} + \frac{L_{\mathrm{B}}}{L}. \tag{3.13}$$

Rearranging we obtain

$$\langle s_{\mathrm{NB}} \rangle = \frac{L}{L_{\mathrm{B}}} + 1. \tag{3.14}$$

Thus, the mean trophic level, including basal nodes, is given by [128, SI Appendix]

$$\langle s^{\infty} \rangle = \frac{N-B}{N} \langle s_{\mathrm{NB}} \rangle + \frac{B}{N} = \frac{L}{L_{\mathrm{B}}} \left(1 - \frac{B}{N} \right) + 1. \tag{3.15}$$

If we make the assumption that the number of links L and the number of basal links L_{B} are proportional to N and B, respectively, (equivalently, assuming the average outgoing degree of basal nodes is equal to that of the mean degree of the network), this simplifies to

$$\langle s^{\infty} \rangle = \frac{N}{B}. \tag{3.16}$$

Thus, in contrast to the coherent $T = 0$ case, in the random network limit, the mean trophic level is linear in the ratio of total nodes to basal nodes suggesting a much more pronounced growth in the mean food chain length in incoherent food webs as a function of the number of species.

In fig. 3.5 we plot the mean trophic level $\langle s \rangle$ of 46 food webs (details to follow in Section 3.4.4) and compare it to the model prediction at $T = 0$. We can see that the model prediction is very accurate for more coherent food webs (low q) while it is not performing as well for more incoherent food webs. Measuring the absolute relative error of the prediction versus the actual value, 63% of predictions are within 10% and 78% of predictions are within 20% of the observed values. Furthermore,

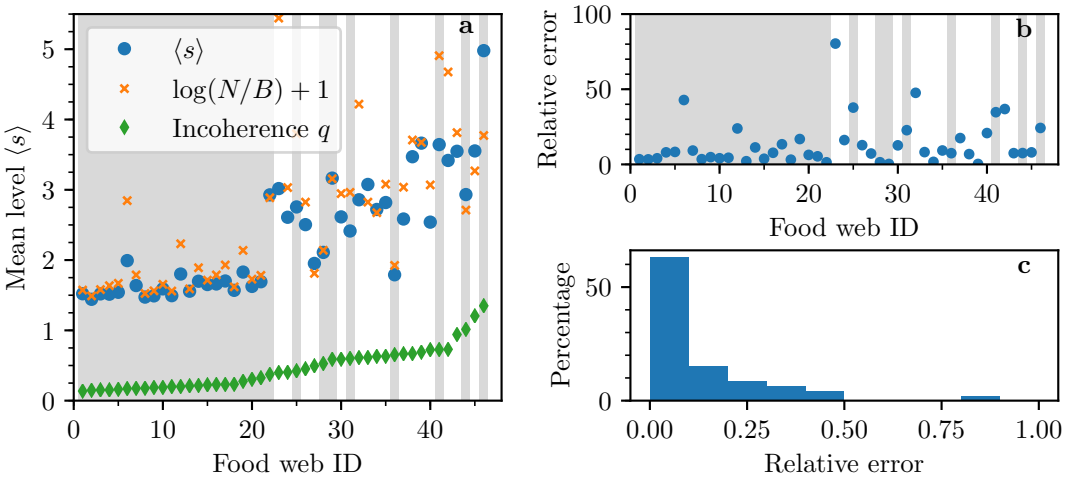

Figure 3.5: (a) Mean trophic level $\langle s \rangle$ of 46 food webs compared to the analytical prediction of the GPPM model at $T = 0$. Food webs are arranged by increasing trophic incoherence parameter q. Grey bars indicate food web membership to Family 1 as uncovered by a hierarchical clustering algorithm used to compare similarity of food web motif profiles in Section 3.4. Details of food webs can be found in Section 3.4.4. (b) Relative error of the analytical prediction. (c) Distribution of relative errors.

even in the case of more incoherent networks, the logarithmic dependence on the N/B ratio is a much better approximation than the linear dependence in the $T \to \infty$ limit. This suggests two things:

1. Even highly incoherent food webs (high q) are much less random than randomly assembled GPPM networks.

2. For coherent networks (low q) the agreement between the analytical prediction of the mean trophic level and the actual empirical value suggests that species "entering" an ecosystem and choosing prey randomly is a good mechanistic approximation on how some food webs evolve.

3.3.3 Degree distribution

To calculate the degree distribution of the GPPM at $T = 0$, we make use of the fact that the distribution of trophic levels, p_s, is a shifted Poisson distribution as derived in the previous section:

$$p_s = \frac{e^{-\lambda}\lambda^{s-1}}{(s-1)!} \text{ for } s = 1, 2, 3 \ldots, \tag{3.17}$$

where $\lambda = \log(N/B)$. Denote by N_s the expected number of nodes at level s. Then $B = N_0 = Np_0$.

In the $T = 0$ case, we need to distribute a fixed number of links L uniformly between any two adjacent trophic levels $s, s+1$. The total number of possible places to put links is

$$\sum_{s=1}^{\infty} N_s N_{s+1} = N^2 \sum_{s=1}^{\infty} p_s p_{s+1} =: N^2 C, \tag{3.18}$$

so that each of these places has uniform probability L/N^2C of being assigned a link. The expected number of links between any two adjacent levels $s, s+1$ is then given by

$$\bar{L}_{s,s+1} = \frac{N_s N_{s+1} L}{N^2 C} = p_s p_{s+1} \frac{L}{C}. \tag{3.19}$$

From here on we focus on the out-degree distribution (the calculation for the in-degree is analogous). The average out-degree at level s is given by

$$\bar{k}_s^{\text{out}} = \frac{\bar{L}_{s,s+1}}{N_s} = p_{s+1} \frac{L}{NC}. \tag{3.20}$$

The out-degree distribution of a random node at level s is then given by a binomial distribution:

$$\mathbb{P}\left[\deg\left(v_s = k\right)\right] = \binom{\bar{L}_{s,s+1}}{k} q_s^k \left(1 - q_s\right)^{\bar{L}_{s,s+1} - k}, \tag{3.21}$$

where q_s is the probability of attracting one of the $\bar{L}_{s,s+1}$ links emanating from the N_s nodes at level s:

$$q_s = \frac{1}{N_s} = \frac{1}{Np_s}. \tag{3.22}$$

Finally, the out-degree distribution for a random node in the whole network is given by a mixture of these binomial distributions where the weights are given by the shifted Poisson distribution of the node trophic levels:

$$\mathbb{P}\left[\deg\left(v = k\right)\right] = \sum_{s=1}^{\infty} p_s \mathbb{P}\left[\deg\left(v_s = k\right)\right]. \tag{3.23}$$

It is unclear whether there is a closed form expression for this distribution as the mixture of binomials differ not only in the probability parameter q_s, but also in the "number of trials" parameter $\bar{L}_{s,s+1}$.

3.4 Motif analysis of food webs

We now come to the main part of the chapter in which we study the effect of trophic coherence on local topological features in food webs. In particular, we show that the relative prevalence of three-species motifs, corresponding to local preying patters, can be explained by the level of trophic coherence in both empirical and model food webs. This result provides another viewpoint in the debate about the origin of motif prevalences in food webs and further evidence of the importance of global organization in food webs [131].

3.4.1 Quantifying triad significance

For any given network the exact number N_k of any of the $k = 1, \ldots, 13$ connected triads (fig. 3.3) is influenced by the network size and the degree distribution of the vertices. To test the statistical significance of any given triad k, the empirically observed number N_k is compared against appearances of the same triad in a randomized ensemble of networks serving as a null model [173]. This comparison gives a statistical significance or z-score

$$z_k = \frac{N_k - \langle N_k \rangle_{\text{rand}}}{\sigma_{\text{rand}}}, \tag{3.24}$$

where $\langle N_k \rangle_{\text{rand}}$ and σ_{rand} are the randomized ensemble average and standard deviation for triad k, respectively. The z-score of triad k thus measures the deviation of prevalence in the observed network with respect to the null model.

The z-scores of all 13 triads can be summarized in a triad significance profile (TSP) which is a vector $\mathbf{z} = \{z_k\}$ with components z_k for each triad k. Additionally, the normalized version of the TSP is often used to compare networks of different sizes and link densities [173]. This is given by

$$\hat{\mathbf{z}} = \left\{ \frac{z_k}{\sqrt{\sum_{k=1}^{13} z_k^2}} \right\}. \tag{3.25}$$

The randomization procedure used to obtain the randomized ensemble statistics is a matter of choice. A careful selection of null model is important to discern between real effects and artefacts present in the TSP [29]. In our analysis, we follow the configuration model (CM) prescription [183, 185], and preserve the number of incoming and outgoing links for each node (the degree sequence) while randomizing links via a Markov chain Monte Carlo switching algorithm [171, 173]. This preserves both the total number of nodes (species) and the links (trophic interactions) in the

network. The generation of randomized networks and counts of triads was carried out with *mfinder*, the algorithm used by Milo *et al.* in their seminal work on network motifs [136, 171].

It is important to emphasize that the TSP is a relative measure of which triads are over- and under-represented with respect to the null model provided by the randomized CM networks. The over-(under-)representation as indicated by a positive (negative) z-score indicates that these triads appear more (less) frequently than in the randomized networks but do not imply an absolute saturation (absence) of said triads. Nevertheless, the TSP is an adequate tool for comparing networks of different sizes and degree distributions.

3.4.2 Comparing networks based on triad significance

To quantitatively compare networks based on their TSP, we use Pearson's correlation coefficient r between the normalized z-score vectors $\hat{\mathbf{z}}^a$ and $\hat{\mathbf{z}}^b$ of networks a and b, respectively [173, 229]. This is defined as

$$r = \frac{\sum_{k=1}^{n} \left(\hat{z}_k^a - \bar{z}^a\right)\left(\hat{z}_k^b - \bar{z}^b\right)}{(n-1)\,\sigma_{\hat{\mathbf{z}}^a}\sigma_{\hat{\mathbf{z}}^b}}, \tag{3.26}$$

where

$$\bar{z}^a = \frac{\sum_{k=1}^{n} \hat{z}_k^a}{n} \tag{3.27}$$

and

$$\sigma_{\hat{\mathbf{z}}^a} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(\hat{z}_k^a - \bar{z}^a\right)^2} \tag{3.28}$$

are the mean and the standard deviation of the normalized z-score vectors, a and b specify the networks, k is an index over the triads and $n = 13$ is the total number of triads.

With this definition a value of r close to 1 indicates that the two networks have very similar TSPs and thus patterns of over- and under-represented triads, a value close to 0 indicates no similarity, and a value close to -1 indicates anti-similarity—i.e. triads over-represented in one network will typically be under-represented in the other (and vice versa).

Comparing the empirical networks is straightforward as we just calculate the r-coefficient pairwise for the z-score vectors of all 46 food webs in our database. On the other hand, for comparison with the model, for each empirical network we fit our food-web model to the data, generate 1000 instances of a model network and then compute the r-coefficient of the empirical z-score vector and the average z-score

vector of the model-generated ensemble.

3.4.3 Clustering food webs into families

To uncover clusters of food webs with similar TSPs, we use a hierarchical, agglomerative clustering algorithm [92] based on the Pearson's correlation coefficient r between TSPs. First, we need to convert this to a distance measure. We define

$$d = \sqrt{2(1-r)}. \tag{3.29}$$

This definition ensures that d is a Euclidean metric [241] and we can readily apply hierarchical clustering. We use the UPGMA (average linkage) algorithm [92] to uncover the full cluster hierarchy.

3.4.4 Empirical food web data

We study the triad significance profile (TSP) in 46 empirical food webs from a variety of environments: marine, freshwater (river and lake) and terrestrial. Table 3.1 gives the relevant summary statistics of each food web.

Table 3.1: An alphabetical list of the 46 food webs studied in the paper. From left to right, the columns are for: name, number of species N, number of basal species B, number of links L, ecosystem type, trophic incoherence parameter q, value of the temperature parameter T found to yield (on average) the empirical q with our model, references to original work, and the numerical ID.

Food web	N	B	L	q	T	Type	Reference	ID
Akatore Stream	84	43	227	0.16	0.40	River	[231, 232, 235]	5
Benguela Current	29	2	196	0.69	0.65	Marine	[251]	39
Berwick Stream	77	35	240	0.18	0.40	River	[231, 232, 235]	7
Blackrock Stream	86	49	375	0.19	0.42	River	[231, 232, 235]	9
Bridge Broom Lake	25	8	104	0.53	0.64	Lake	[115]	28
Broad Stream	94	53	564	0.14	0.37	River	[231, 232, 235]	1
Canton Creek	102	54	696	0.15	0.38	River	[235]	4
Caribbean (2005)	249	5	3302	0.73	0.69	Marine	[26]	41
Caribbean Reef	50	3	535	0.94	0.82	Marine	[186]	43
Carpinteria Salt Marsh Reserve	126	50	541	0.65	0.85	Marine	[150]	36
Caitlins Stream	48	14	110	0.20	0.41	River	[231, 232, 235]	12
Chesapeake Bay	31	5	67	0.45	0.62	Marine	[1, 240]	26
Coachella Valley	29	3	243	1.21	1.02	Terrestrial	[203]	45
Coweeta (1)	58	28	126	0.30	0.52	River	[231, 232, 235]	20
Crystal Lake (Delta)	19	3	30	0.17	0.43	Lake	[239]	6
Cypress (Wet Season)	64	12	439	0.63	0.66	Terrestrial	[236]	34
Dempsters Stream (Autumn)	83	46	414	0.21	0.43	River	[231, 232, 235]	13
El Verde Rainforest	155	28	1507	1.01	0.99	Terrestrial	[244]	44
Everglades Graminoid Marshes	64	4	681	1.35	1.10	Terrestrial	[238]	46
Florida Bay	121	14	1767	0.59	0.59	Marine	[237]	29

German Stream	84	48	352	0.20	0.43	River	[231, 232, 235]	11
Grassland (U.K.)	61	8	97	0.40	0.69	River	[164]	24
Healy Stream	96	47	634	0.22	0.42	River	[231, 232, 235]	15
Kyeburn Stream	98	58	629	0.18	0.41	River	[231, 232, 235]	8
LilKyeburn Stream	78	42	375	0.23	0.44	River	[231, 232, 235]	18
Little Rock Lake	92	12	984	0.67	0.65	Lake	[163]	37
Lough Hyne	349	49	5102	0.60	0.60	Lake	[86, 209]	31
Mangrove Estuary (Wet Season)	90	6	1151	0.67	0.63	Marine	[237]	38
Martins Stream	105	48	343	0.32	0.51	River	[231, 232, 235]	21
Maspalomas Pond	18	8	24	0.49	1.01	Lake	[11]	27
Michigan Lake	33	5	127	0.37	0.48	Lake	[165]	22
Narragansett Bay	31	5	111	0.61	0.68	Marine	[176]	33
Narrowdale Stream	71	28	154	0.23	0.44	River	[231, 232, 235]	17
N.E. Shelf	79	2	1378	0.73	0.66	Marine	[157]	42
North Col Stream	78	25	241	0.28	0.45	River	[231, 232, 235]	19
Powder Stream	78	32	268	0.22	0.42	River	[231, 232, 235]	14
Scotch Broom	85	1	219	0.40	0.54	Terrestrial	[168]	23
Skipwith Pond	25	1	189	0.61	0.54	Lake	[245]	32
St. Marks Estuary	48	6	218	0.63	0.67	Marine	[59]	35
St. Martin Island	42	6	205	0.59	0.63	Terrestrial	[109]	30
Stony Stream	109	61	827	0.15	0.38	River	[235]	3
Sutton Stream (Autumn)	80	49	335	0.15	0.40	River	[231, 232, 235]	2
Troy Stream	77	40	181	0.19	0.42	River	[231, 232, 235]	10
Venlaw Stream	66	30	187	0.23	0.44	River	[231, 232, 235]	16

Weddell Sea	483	61	15317	0.72	0.68	Marine	[124]	40
Ythan Estuary	82	5	391	0.42	0.50	Marine	[122]	25

3.5 Results

3.5.1 Motifs in empirical food webs

The main results are summarized in figs. 3.6 and 3.7.

Figure 3.6 shows the pairwise Pearson correlation coefficients of the triad significance profiles between all 46 food webs. The food webs are arranged by increasing incoherence parameter q so that more coherent food webs are assigned a lower ID. Red hue or warmer colours indicate a larger coefficient, while blue hue or colder colours indicate an anti-correlation in the TSPs.

We see that roughly two families of food webs emerge with similar TSPs. The first family (roughly ID 1-22) is characterized by relatively high coherence (low incoherence parameter q), for which the similarities in the TSPs are very high ($r \geq 0.8$).

There is a second family of food webs, characterized by a high incoherence parameter q, that also show high similarities in their TSPs. Membership to this second family is not as clear as there is a tighter core of food webs belonging to it, with a periphery that only shares some similarities.

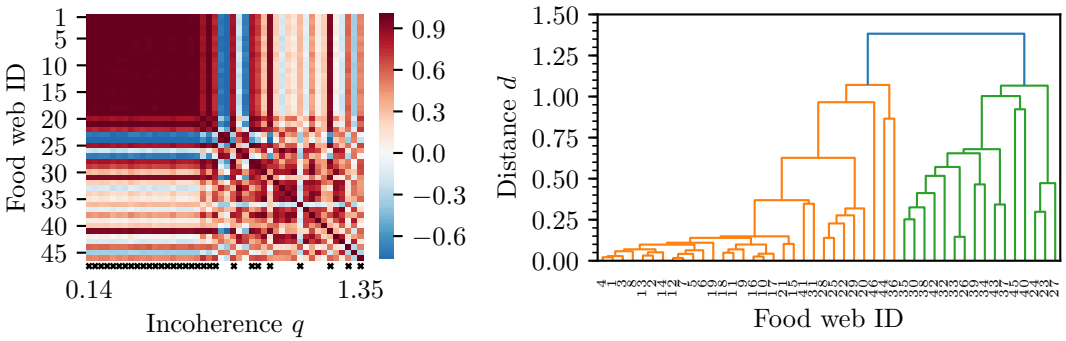

Figure 3.6: Pearson's correlation coefficient of the triad significance profiles (left) and clustering of food webs into two families (right). Left: The coefficient is measured pairwise between all pairs of empirical food webs. Warmer colours indicate greater similarity while colder colours indicate dissimilarity. The food webs are arranged according to increasing incoherence parameter (left to right and top to bottom). Black crosses just below the heatmap indicate membership to Family 1 according to a clustering algorithm. Right: Dendrogram of the hierarchical clustering algorithm applied to food webs based on the distance $d = \sqrt{2(1-r)}$. A threshold distance $d_c = 1.1$ uncovers two large families with smaller subclusters within.

To make these ideas more precise, we performed hierarchical clustering of food webs based on a distance metric derived from the pairwise Pearson correlation coefficients. The resulting clusters are shown as a dendrogram in fig. 3.6. By choosing

a threshold distance d_c, we can group food webs into a number of distinct families based on the similarities of their TSPs. Setting $d_c = 1.1$, we identify two families which include all webs. Family 1 consists of food webs with ID 1–22, 25, 28, 29, 31, 36, 41, 44, 46 whereas Family 2 contains webs with ID 23, 24, 26, 27, 30, 32-35, 37-40, 42, 43, 45. We also observe that these larger families contain within themselves smaller, even more closely related clusters (e.g ID 1–22 corresponding to very low q).

Setting a lower threshold distance could provide a more fine-grained classification of food webs in more than two distinct families but we now show that this coarse classification into two families allows us to qualitatively differentiate food webs based on species preying patterns, specifically the extent of omnivory. To this end, we look closer at the bulk behaviour of the TSPs for the two families. Figure 3.7 shows the normalized profiles of Family 1 (top) and Family 2 (bottom).

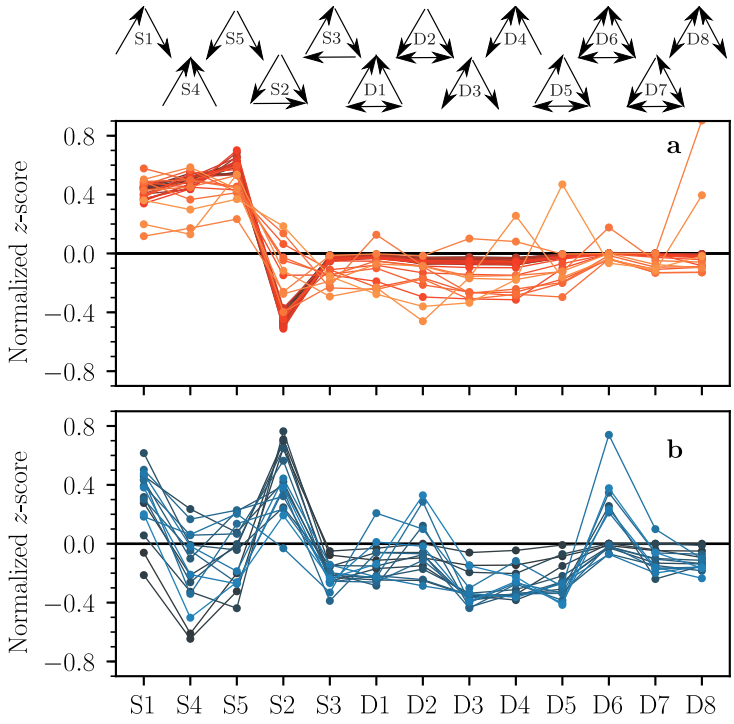

Figure 3.7: Triad significance profiles (TSP) as measured by the normalized z-score of the two groups of food webs. (a) Food webs in the first family (ID 1–22,25,28,29,31,36,41,44,46) with low incoherence parameter q characterized by an over-representation of triads S1, S4 and S5 and an under-representation of triad S2. (b) Food webs in the second family (ID 23,24,26,27,30,32-35,37-40,42,43,45) with high incoherence parameter characterized by an over-representation of triad S2.

We first consider Family 1. The bulk behaviour of food webs in this family

is characterized by an over-representation of triads S1, S4 and S5, as well as an under-representation of triad S2 (with the exception of ID 22 Michigan Lake, ID 29 Florida Bay and ID 46 Everglades Graminoid Marshes). We should find the pattern of under-representation of triad S2 (which represents omnivory) unsurprising, since the majority of food webs belonging to this family have a low incoherence parameter q, which limits the ability of species to feed on multiple different trophic levels. Equally, the over-representation of triads S1, S4 and S5 is to be expected as these are the only three triads out of 13 that can arise in a hypothetical food web with $q = 0$, which is a value close to the empirical values of q for food webs in this family. The double link triads D1-D8 are all under-represented or close to even, in agreement with our expectations.

We now turn to Family 2. Here the triads S1, S4 and S5 no longer follow a strong pattern of over-representation and the double link triads D1-D8 are not always under-represented. The most distinguishing feature, however, is the bulk over-representation of triad S2 (with the exception of ID 40 Weddell Sea), in stark contrast to Family 1. We will argue that this is the main feature that separates the two food web families.

This pattern of food webs based on the under- or over-representation of triad S2 was alluded to in previous work [229], however it is in disagreement with the predictions of the generalized cascade [228] and niche [250] models which can only produce food webs where S2 is over-represented [229]. Subsequently, we present results from our model which show that it is possible to change the pattern of under-representation to over-representation of triad S2 by increasing the incoherence parameter q, thus providing evidence that trophic coherence can naturally give rise to two food web families characterized by low or high prevalence of omnivory, respectively.

3.5.2 Comparison between empirical and model networks

We have also investigated the similarities of triad significance profiles between the empirical food webs and model generated food webs. To this end we study the similarity of the TSPs between each empirical food web and an ensemble of model food webs fitted to the data of the empirical one. The results are summarized in fig. 3.8. Averaging over an ensemble of 1000 model generated food webs fitted to each empirical food web, we measured the Pearson correlation coefficient between the TSP of the empirical food web and the TSP of the ensemble average. The results show that the model is able to reproduce empirically observed TSPs for the majority of food webs in both families with high accuracy. The model fails to produce accurate

TSPs for a number of food webs and sometimes even produces anti-correlated TSPs ($r < 0$). If we require that $r > 0.5$, eight food webs are not able to be reproduced accurately by our model, five in Family 1 (ID 31 Lough Hyne, ID 36 Carpinteria Salt Marsh Reserve, ID 41 Caribbean Reef, ID 44 El Verde Rainforest and ID 46 Everglades Graminoid Marshes) and three in Family 2 (ID 23 Bridge Broom Lake, ID 24 Grassland (U.K.) and ID 40 Weddell Sea). Recall that IDs are assigned in the order of increasing q so these particular food webs are unusual members of their respective families in that they tend to have extreme values of q with respect to the majority of networks in either family (higher than average in Family 1 and lower than average in Family 2). Because of the imperfect agreement between q and family membership, our model cannot replicate the structure of these sporadic webs. This suggests that for some food webs information about trophic coherence q may not be enough to reproduce realistic looking TSPs and there may be further mechanisms of prey selection at play [229].

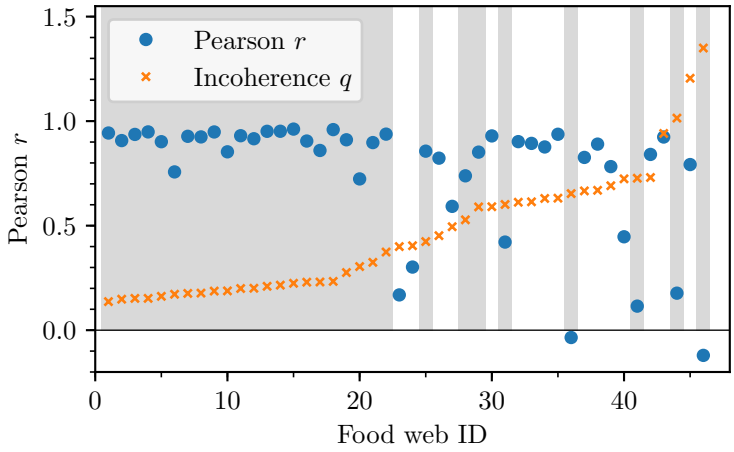

Figure 3.8: Pearson's correlation coefficient of the triad significance profiles (TSP). The coefficient is measured between the empirical TSP and the average TSP in the model ensemble over 1000 simulated networks. Food webs are arranged by increasing incoherence parameter q. The grey shading indicates membership to the first family.

3.5.3 The role of omnivory and basal species

We now focus on the claim that the main difference between the two families of food webs is the relative under- and over-representation of triad S2, or the degree of omnivory in a food web. A prevalence of triad S2 indicates that the species in a food web often feed on different trophic levels, contributing to an increased incoherence parameter q as discussed at the start of this section. A scarcity of triad

S2, on the other hand, indicates that species only tend to feed on prey with similar trophic levels, which in turn signals a low incoherence parameter. This suggests a relationship between the z-score of triad S2 and network incoherence as measured by q.

Furthermore, model results (fig. 3.4) suggest that a high proportion of basal species to all species, B/N, produces more coherent food webs (i.e. with a low incoherence parameter q). We take this as an additional predictive food web statistic for family membership.

Our findings are summarized in fig. 3.9. This is a scatter plot of all 46 food webs where we have plotted the fitted model temperature T and the measured incoherence parameter q against the ratio of basal species to all species B/N. We observe a clear anti-correlation between q and B/N (linear model $q = a\frac{B}{N} + b$: $a = -1.06, b = 0.77, R^2 = 0.53, p = 8.47 \cdot 10^{-9}$) that indicates a positive relationship between how coherent a network is (low q) and how many of its species are basal.

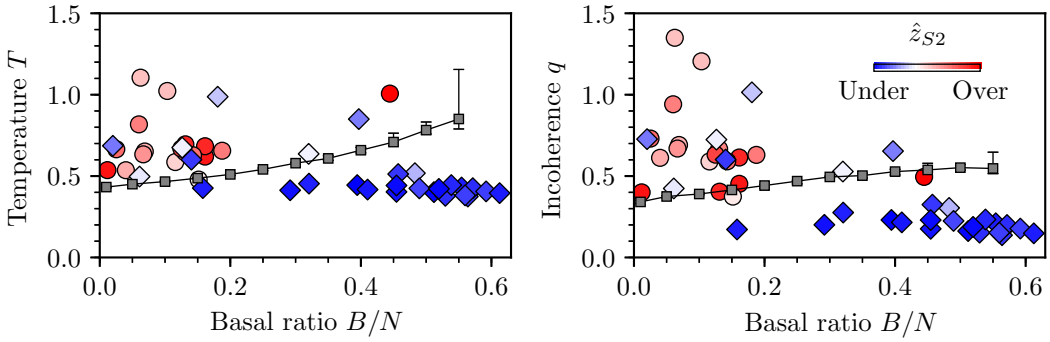

Figure 3.9: Scatter plots of the temperature T (left) and the incoherence parameter q (right) versus the basal species ratio B/N for all food webs. The gradient indicates the degree of over-representation (red circles) or under-representation (blue diamonds) of the feed-forward triad S2 as measured by the normalized z-score $\hat{z}_{\mathrm{S2}}$. The line shows the transition from over-representation (above) to under-representation (below) as observed in the model with $N = 100$, $\langle k \rangle = L/(N - B) = 10$ averaged over 100 runs. Error bars are approximate 95% confidence intervals.

We have also coloured the markers of each food web to indicate the level of over- or under-representation of triad S2 as measured by the normalized z-score $\hat{z}_{\mathrm{S2}}$. Red circles indicate an over-representation while blue diamonds indicate an under-representation of S2 in the respective food web. Remarkably, based on this measure, we uncover two clusters of food webs corresponding roughly to the two families based on TSP similarities. The first cluster is once again characterized by a high incoherence parameter q as well as a low ratio of basal species to all species

B/N. The second cluster is characterized by a low incoherence parameter and a high ratio of basal species to all species. The only exceptions are six food webs in the first family (ID 20 Coweeta (1), ID 21 Martins Stream, ID 31 Lough Hyne, ID 36 Carpinteria Salt Marsh Reserve, ID 41 Caribbean Reef and ID 44 El Verde Rainforest), four of which correspond to food webs poorly matched by our model (fig. 3.8). We conclude that, indeed, the main difference between the two families is the relative role of triad S2 as already observed in the bulk behaviour of the TSPs in fig. 3.7.

Finally, we study whether our model exhibits a similar transition from a relatively S2-poor to an S2-rich state which would explain the relatively good agreement between empirical and model generated TSPs for the two families (fig. 3.8). We find that for a given basal species ratio B/N there exists a critical temperature T_c, and thus a critical incoherence parameter q_c, which signifies such a transition. For T (and q) below these critical values, the model generates networks where S2 is under-represented, while for values above critical, the networks generated have either an even or an over-represented number of S2 triads. We include the transition line of the two regimes in fig. 3.9 for an ensemble of 100 model networks with $N = 100$ species and an average (non-basal) degree $\langle k \rangle = L/(N - B) = 10$. Networks with q below the line show an under-representation of S2 triads, while networks with q above the line show an over-representation as measured by $\hat{z}_{\mathrm{S2}}$.

Remarkably, the model results are in very good agreement with the empirical data despite the fact that both the network size N and the average degree $\langle k \rangle$ vary considerably between the empirical food webs. Almost all food webs with an under-represented number of S2 triads fall below the transition line of the model while those with an over-represented number reside above the line.

These findings suggest that the two families of food webs differ in the degree of omnivory present as measured by the prevalence of triad S2 which is itself intimately related to the incoherence parameter q. Interestingly, based on the strong anti-correlation between q and B/N, either parameter is a strong determinant of family membership. To our knowledge, the GPPM is the first food-web model able to reproduce triad significance profiles consistent with empirical observations. The ability to produce model networks belonging to either of the two families suggests that the parameters q and B/N are both important in the mathematical modelling of food webs and may, in fact, be fundamental for understanding local preying patterns in food webs.

3.6 Discussion

Our investigation of trophic interaction patterns in food webs has revealed significant correlations between the degree of omnivory, hierarchical organization of trophic species and the density of basal species.

The analysis of local trophic interactions via triad significance profiles in empirical food webs reveals two distinct families of food webs characterized by a relatively low or high incoherence parameter respectively. While certain differences across families of food webs based on their TSPs have been observed before [229], these are not predicted by any existing food web models, calling into question their use as null models given the academic significance attached to food web motifs [25, 50, 56, 196, 229]. Trophic coherence provides a network theoretic metric that enables us to classify and predict the relative prevalence of such motifs.

We have shown qualitatively that the the main difference between the two food web families is the extent of omnivory, as measured by the over- or under-representation of triad S2 (the "feed-forward loop"). This classification of food webs into two families according to the extent of omnivory is at odds with previous claims that omnivory occurs more often than one would expect to happen by chance across most food webs [229]. On the other hand, the existence of these families may be related to different ways omnivory emerges in food webs and influences their stability [10, 131, 177]. We have tested our prediction for the onset of omnivory using a new model for generating synthetic food webs with a given trophic coherence. We find that the model exhibits a transition from an under-representation of omnivory to an over-representation of omnivory as a function of trophic coherence. Our model results fit the food web data very well, providing evidence of the importance of trophic coherence as well as the basal species density in modelling realistic trophic interactions. We would like to emphasize that these findings are remarkably robust between food webs originating from vastly different habitats.

This work has expanded on the importance of trophic coherence in predicting structural features in food webs [131], but the biological origin of trophic coherence remains elusive. Basal species density and its effect of suppressing highly incoherent structures in both empirical and model food webs may provide some clues. All other things being equal, a higher proportion of autotrophs in a food web will necessarily mean that a higher proportion of consumers will feed on these basal species. In turn, this would have a dampening effect on the formation of long food chains in the trophic hierarchy and hence fewer possibilities for a varied diet of species at the top. Figure 4.1 exemplifies how this hypothesis could lead to very different food

web structures. Established food web models do not treat basal species density as a predictor for emergent structure but rather as an emergent property itself. On the other hand, most food webs have been found to be significantly more trophically coherent than a random graph with the same density of basal species, so there must be other coherence-inducing mechanisms at play [128]. Further work is needed to elucidate the reasons behind this property of ecosystem structure.

CHAPTER 4

SPREADING PROCESSES ON TROPHIC NETWORKS

In the previous chapter we saw how analysing network structure can lead to new insights about complex food webs. It is perhaps surprising that just by looking at connection patterns between interacting agents, in this case feeding relations between species, we can draw conclusions about qualitative differences between the networks. However, network structure is only half the story. When we study dynamical processes on top of networked systems, we uncover a whole new world of interesting phenomena by incorporating a set of states that nodes in a network can take which are evolved according to some simple rules. Of course, the emergence and nature of these phenomena are also mediated by the underlying network structure. A major area of network science is to understand how different network structures affect dynamics.

In this chapter we study two simple dynamical processes on the trophic network model introduced in the previous chapter. Our aim is to investigate how the dynamics differ as we change the underlying structure of networks.

4.1 Background and related work

The spreading of activity through a network has been extensively studied in a wide variety of settings [23, 205]. Perhaps the most notable example is the study of infectious disease on human contact networks in epidemiology [71, 141, 195]. Mathematical modelling of infectious disease has a rich history because of its importance in informing public health interventions for the control of disease [16, 137, 139]. In the next section we briefly review basic models and concepts of epidemic models, their extensions on contact networks as well as some spreading processes which do

not originate from the study of epidemiology.

4.1.1 Compartmental epidemic models

Most of the classical models of epidemic diseases use *compartmentalization* of the population in which individuals are grouped into compartments according to their disease state (e.g. *susceptible, infected* or *recovered*) with transition rates for changing from state to state as the disease progresses [16, 137]. Without any underlying contact network structure such models are very simplistic and assume that all individuals in a given state are statistically indistinguishable from each other so it makes sense to only study the evolution of proportions of individuals in each compartment. This is the so called "well-mixed" population case because the absence of a contact network implicitly assumes that individuals constantly interact with each other in exactly the same way.[1]

The modern mathematical modelling of epidemics was initiated by Kermack and McKendrick [139] who define and study a compartmental SIR model (sometimes also called the Kermack-McKendrick model) with three disease states—susceptible, infected and recovered. They study the model in the deterministic setting using a set of coupled nonlinear ODEs:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{4.1}$$

where β is the infection rate and γ is the recovery rate. Infection is supported via the *law of mass action*—the assumption that rate of encounters between susceptible and infected individuals is proportional to the product of their numbers in society as indicated by the βSI terms.[2] This is just another way of saying that the population is well-mixed (homogeneous mixing assumption). The key quantity which determines whether the infection can spread is the so called epidemic threshold $R_0 = \beta S/\gamma$. If $R_0 < 1$, each infected person will in turn infect fewer than one person before

[1] In fact, the absence of an underlying contact network is equivalent to the case of individuals living on a complete network.

[2] It is important to note that this formulation of the SIR model has *density dependent* transmission—the number of infectious contacts is dependent on the population size. An alternative way to model SIR dynamics is via *frequency dependent* transmission where the number of infectious contacts is rescaled by the inverse population size N^{-1}. Both modelling assumptions can be justified depending on the disease dynamics under study, but frequency dependent transmission is usually more natural for infectious disease modelling even though it is different from the original Kermack-McKendrick model.

recovering and the disease will die out. On the other hand, if $R_0 > 1$, the opposite is true and the epidemic will spread. The epidemic threshold R_0 is perhaps the single most important quantity in epidemiology.

The SIR model can be made more realistic by making it stochastic using reaction diffusion equations [242]. In this case, the SIR model is governed by the reactions

$$\begin{aligned} S + I &\xrightarrow{\beta} 2I \\ I &\xrightarrow{\gamma} R, \end{aligned} \tag{4.2}$$

and the evolution of the disease is modelled as a continuous time Markov chain [216]. It is worth noting that although the Markov Chain formulation is exact, its use for acquiring analytic results is limited because of the sheer number of interdependent equations—an SIR model of 3 compartments and N individuals gives rise to a total of 3^N possible states of the epidemic which limits exact results to small systems and very simple dynamics. Because of this, a popular approach is to simulate the epidemic directly using Monte Carlo methods, typically using Gillespie's algorithm (also known as the stochastic simulation algorithm or kinetic Monte Carlo) [107].

One can show that the deterministic ODE model is an approximation of the full Markov Chain model in which all individuals in the same compartment are assumed to be statistically indistinguishable from each other and the interactions they experience can be described by the average interactions due to the full system [23]. This approach of reducing a stochastic model to its deterministic limit is sometimes called mean-field (MF) theory.

It is much more realistic to model epidemics on explicit contact network structures [19, 23, 71, 72, 195]. In this scenario, instead of implicitly assuming that individuals all interact with each other, they are identified as nodes of some predetermined network structure which prescribes the exact interactions between individuals via the link structure. As in the case of homogeneous mixing, exact results from a Markov Chain formulation are sparse, so practitioners rely on stochastic simulations and various mean-field approximation schemes (see [141, 195] for thorough reviews). Analogously to the homogeneous mixing case, there is a critical epidemiological threshold which separates diseases that die out quickly from diseases that spread to infect a considerable part of the network. In this case the threshold is dependent both on the effective infection rate and the network topology.

There are many different kinds of compartmental models studied in epidemiology and beyond [195]. The SI model only has two compartments and models diseases in which infected individuals remain infected indefinitely. The SIS model

(also known as the contact process in the statistical physics literature [112, 116]) allows for recovery, but does not grant immunity from the disease. The SEIR model incorporates a latent state (E for exposed) to deal with individuals exposed to the disease but not yet infectious while the SIRS model incorporates waning immunity. More complicated compartmental models may also include age structure, population demographics via birth and death rates and even attempt to move beyond the Markovian assumption to include memory effects and model more realistic infectious periods [44, 159, 160].

4.1.2 Complex contagion

Classical epidemic models are sometimes called "simple contagions" to reflect the fact that individuals have a constant probability of acquiring infection from infectious peers for every such exposure and interactions are generally assumed to be independent. Recently, however, there has been an interest in studying "complex contagions" in which various other assumptions are made to make transmission dynamics more realistic [188] and applicable to phenomena outside epidemiology such as in modelling rumour spreading or the diffusion of innovation. A subset of models that try to capture complex contagion are *threshold models*.

The most famous threshold model is the Watts Global Cascades Model or Watts Threshold Model (WTM) [247]. It is a binary state model with individuals adopting either of two states: S for susceptible and I for infected analogously to epidemic models. Each individual i also has a *threshold* R_i drawn from some distribution and fixed in time. The states of the nodes change in time according to an update rule in which a susceptible individual i becomes infected if at least a threshold fraction R_i of its neighbours are infected, else it remains susceptible. An infected individual remains infected indefinitely. The model exhibits a phase transition from local *cascades* in which only a finite number of individuals get infected to global cascades which involve infecting a finite fraction of the population.

Another example of a complex contagion is the so called Generalized Epidemic Process (GEP) introduced independently by Janssen et al. [125] and Dodds and Watts [75]. This model incorporates memory of past exposures to a contagious influence and can interpolate between the WTM and a simple SIR model. Recently, it has been shown to be, in fact, a special case of the WTM [170].

4.1.3 Other spreading processes

The literature on spreading processes on networks extends far beyond epidemic and contagion models typically considered in epidemiology. It is beyond the scope of this thesis to discuss the many different kinds of processes studied on networks, so we provide some examples together with selected reviews for the interested reader. Examples include random walks and diffusion on networks [8, 161, 166], spin models such as the Ising model, XY model or the Potts model [77], synchronization models [17], firing neural network models [13, 120], and models of social dynamics, including opinion and voter dynamics [58]. A recommended textbook that touches upon many of these types of processes has been published by Barrat et al. [23].

4.1.4 Generalized network structures

While most of the preceding discussion on spreading processes on networks has historically been focused on simple, undirected networks, the study of spreading processes on more general network structures such as directed or weighted networks as well as multi-layer networks [43, 142] and temporal networks [119] has become more prominent in recent years. There is good reason for considering more general network structures as they can confer an even higher degree of realism. For example, weighted networks are the natural framework for considering contagion which is not equally facilitated across all links [85, 102] [3] while directed networks become relevant in studying contagion with an intrinsic asymmetry in the propagation such as in some sexually transmitted diseases [169] or on some social media platforms such as *Twitter* [30].

In this chapter we continue the tradition of studying the effect of network structure on spreading dynamics by considering two different spreading processes on top of directed networks with a given trophic coherence.

4.1.5 A note on discrete vs continuous time dynamics

In the preceding discussion on spreading processes, we have implicitly assumed continuous time dynamics. Working in continuous time is equivalent to asynchronous updating of individual states as exemplified by Gillespie's algorithm [107]. Alternatively, one can choose to work in discrete time steps which naturally corresponds to parallel or synchronous updating of individual states. Synchronous updating of states has the advantage of allowing fast simulations, but care must be taken if

[3] This scenario is similar to studying contagion in which the rate of infection is individual dependent [138, 181].

discrete time simulations are used as approximations to genuinely continuous time dynamics [94]. For us, we have chosen the spreading processes defined in the following section to be naturally discrete so we do not face these approximation issues. For a more comprehensive discussion on these two approaches see [205, Sec. 3.5].

4.2 Two spreading processes

With a view of exploring the influence of trophic structure on how activity of some kind spreads through a directed network, we consider two different paradigmatic spreading processes: a model of complex contagion and a firing neural network model. In both models we take time to be discrete and update nodes in parallel (synchronously).

4.2.1 Complex contagion model

Our first model is an adaptation of the classical Susceptible-Infected-Susceptible (SIS) epidemic model [195] which bears resemblance to the global cascades model [247] in that spreading dynamics is governed by the proportion of a node's infected neighbours.

In this model each node i in a network at time t is characterized by a binary state variable $s_i(t)$ which can take two values: $s_i(t) = S$ if node i is susceptible or $s_i(t) = I$ if node i is infected. Next, we need to specify spreading dynamics. First, a node which was infected at time t automatically becomes susceptible again at time $t + 1$:

$$P[s_i(t+1) = S | s_i(t) = I] = 1. \tag{4.3}$$

In defining transmission dynamics, we incorporate the phenomenon of "complex contagion" whereby an "infection" event can be a non-linear function of the proportion of neighbours of the node in question who are already infected [12, 188]. Specifically, at each time step $t + 1$ a susceptible node i becomes infected with probability given by

$$P[s_i(t+1) = I | s_i(t) = S] = \left(\frac{n_i}{k_i}\right)^{\alpha}, \tag{4.4}$$

where n_i is the number of infected in-neighbours (i.e. nodes j such that $a_{ji} = 1$) at time t and k_i is the in-degree of node i. The parameter α introduces a non-linearity in the dynamics leading to different complex contagion scenarios. When α is between 0 and 1, the probability is a concave function of the proportion of infected

neighbours which means that the probability to be infected grows rapidly with the first few infected contacts of a susceptible node and slowly thereafter. On the other hand, if $\alpha > 1$ the probability is convex meaning that a large proportion of infected neighbours is required to trigger an infection event. Finally, if $\alpha = 1$, the probability of infection is linearly proportional to the fraction of infected neighbours. This is equivalent to the global cascades model [247] with uniformly distributed threshold values with the difference that nodes in our model recover immediately.

4.2.2 Neural network model

The second model we consider is a version of the Amari-Hopfield neural network [13, 120]. As in the contagion model, each node has a binary state variable $v_i(t) = \pm 1$ representing that at time t a neuron can either fire an action potential or not. The probability of node i updating its state at time $t + 1$ is given by a sigmoid

$$P[v_i(t+1) = \pm 1] = \frac{1}{2}\left(\pm \tanh[\beta h_i(t)] + 1\right), \tag{4.5}$$

where

$$h_i(t) = \sum_j a_{ji} v_j(t) \tag{4.6}$$

is a "field" experienced by node i whose strength is determined by the aggregate state of its in-neighbours. The parameter β allows for a degree of stochasticity, i.e. for low β the updates become increasingly random as nodes switch between states spontaneously and the field h_i plays no role while for high β the dynamics become more deterministic. The field h_i classically takes the adjacency matrix a_{ij} to be weighted to account for the effect of different synaptic weights, but for our purposes we consider all weights to be equal to unity.

4.3 Coherent networks

For completeness, we briefly recall the construction of the trophic network ensemble described in the previous chapter (see Section 3.2 for a full discussion) that serves as the network backbone that we study the spreading processes on. We start with B basal nodes (i.e. nodes with zero in-degree) and proceed to introduce $N - B$ non-basal nodes sequentially at each step choosing one random in-neighbour for each new node. At the end of this procedure the network has N nodes, each with a preliminary trophic level $\tilde{s}_i$, and $N - B$ links. Finally, to introduce the remaining $L - N + B$

links, we place a link between each pair of nodes (i, j) with probability proportional to

$$P[a_{ij} = 1] \propto \exp\left(-\frac{(\tilde{s_j} - \tilde{s_i} - 1)^2}{2T^2}\right), \tag{4.7}$$

where T is a temperature parameter modulating the degree of order or trophic coherence in the network. The completed networks are then characterized by the distribution of trophic levels s_i of nodes and the distribution of trophic distances $x_{ij} = s_j - s_i$ of links. We denote by q the standard deviation of the distribution of trophic distances so that $q = \sqrt{\langle x^2 \rangle - 1}$. q is called the *trophic incoherence parameter* and measures the disorder of the generated networks—a temperature parameter of $T = 0$ leads to $q = 0$ and perfectly coherent or "layered" networks while higher values of T lead to more disorder. Figure 4.1 shows two examples of networks generated using this model with identical parameters N, B and L but different temperatures T and different resulting trophic coherence q. Figure 4.2 shows the monotonic dependence of q on T as well as the effect of the number of basal species B. Our goal is to study the dynamics of spreading processes on networks that range from ordered (low T) to disordered (high T) networks.

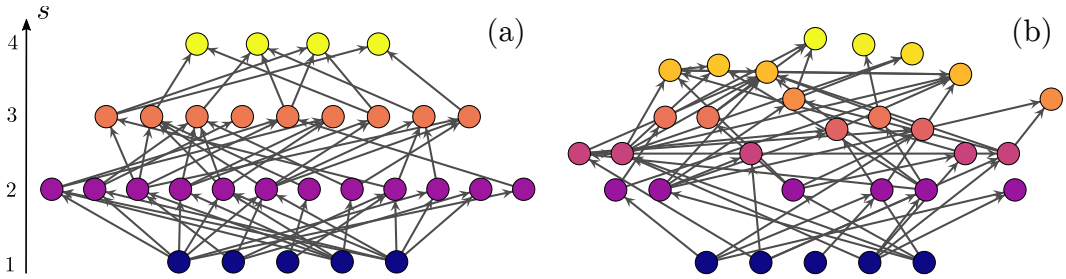

Figure 4.1: (a) An example of a maximally coherent network ($q = 0$). (b) A network with the same parameters N, B and L as the one on the left, but less trophically coherent ($q = 0.49$). In both cases, the hight of the nodes on the vertical axis represents their trophic level. The networks were generated with the preferential preying model as described in the main text, with $T = 0.001$ for the one on the left, and $T = 1$ for the one on the right.

4.4 Results

In order to numerically investigate the effects of trophic structure on spreading phenomena, we generate networks with a specified number of nodes N, basal nodes B and edges L, but varying temperature parameter T as described in Section 4.3 and perform Monte Carlo simulations of each of our dynamical systems on top of these

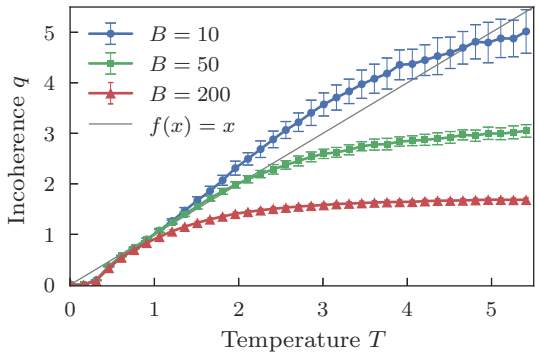

Figure 4.2: Trophic coherence, as given by q, against the temperature parameter T for networks generated with the preferential preying model described in the main text, for different numbers of basal nodes: $B = 10$, 50 and 200, as shown. In all cases, the number of nodes is $N = 1000$ and the mean degree is $\langle k \rangle = 5$. Averages are over 1000 runs.

networks. The initial condition for both the contagion and neural network model is to set all nodes to susceptible/inactive except for the basal nodes which are all infected/active, i.e. $s_i(t) = S$ or $v_i(t) = -1$ if $k_i^{\text{in}} > 0$ and $s_i(t) = I$ or $v_i(t) = 1$ if $k_i^{\text{in}} = 0$. For each simulation run we measure the *duration* of the infection, that is, the number of time steps until no nodes are infected, as well as the *incidence*, the proportion of nodes which have at any time been in the infected state.

Figure 4.3a shows the mean incidence against T for various value of α. On highly coherent networks ($T \simeq 0$), the infection spreads to the whole system for any value of α. On leas coherent topologies, however, whether contagion is sub- or super-linear has a strong influence on spreading: for $\alpha > 1$ the infection only reaches a fraction of the network, while for $\alpha < 1$ the effect of coherence on incidence is non-monotonic. In fig. 4.3b, where the mean incidence is plotted against α for various values of T, we can see how the effect of α on spreading is modulated by topology, becoming less severe the more coherent the networks. Hence, it is the interplay of both the trophic coherence of the underlying network and the form of the infection probability which determines whether the infection can reach a large proportion of nodes.

We also performed a similar investigation of the Amari-Hopfield neural model on networks generated in the same way. The duration is now the number of time steps taken before all nodes are in the inactive (not firing) state and the incidence is the proportion of nodes which at any moment during this period adopted the active (firing) state. As with the infection, whether this pulse will propagate throughout the whole network is determined by both the neural dynamics, as parametrised by

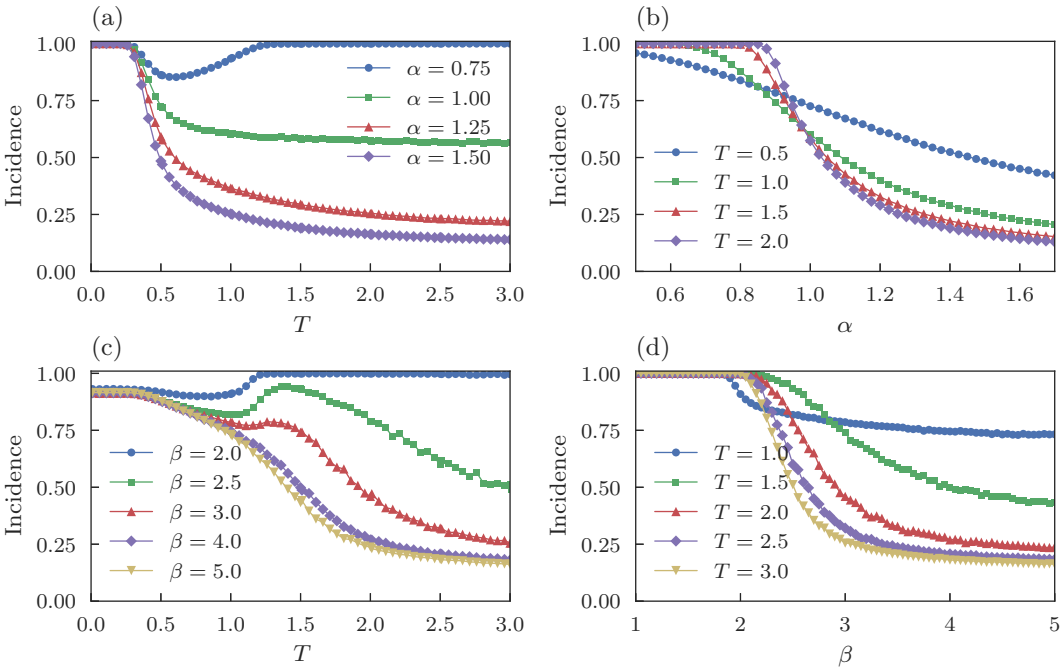

Figure 4.3: Average Incidence values from Monte Carlo simulations of the two spreading models on networks with varying trophic coherence, as described in the main text. (a) Incidence against T (smaller T means more coherent networks) in the complex contagion model for several values of the contagion parameter α, as shown. (b) Incidence against α in the complex contagion model for several values of T. (c) Incidence against T in the Amari-Hopfield neural network model for several values of the stochasticity parameter β. (d) Incidence against β in the Amari-Hopfield neural network model for several values of T. All networks have $N = 1000$, $B = 100$, and $\langle k \rangle = 5$. Averages are over 1000 runs.

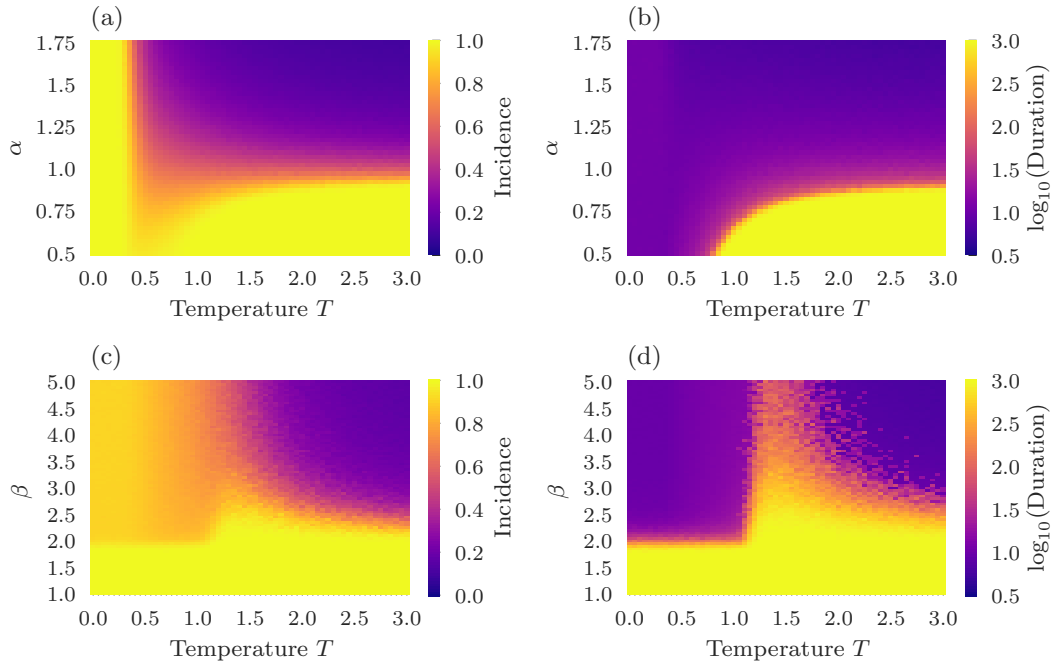

Figure 4.4: Heat-maps showing average values of incidence and of the common logarithm of duration on a colour scale; results are from Monte Carlo simulations of the two spreading models on networks with varying trophic coherence, as set by T. (a) and (b) Complex contagion model, where α is the contagion parameter. (c) and (d) Amari-Hopfield neural network model, where β is the stochasticity parameter. All networks have $N = 1000$, $B = 100$, and $\langle k \rangle = 5$. Averages are over 100 runs.

β, and the trophic coherence of the underlying network. Figure 4.3c shows the mean incidence against T for several values of β, while fig. 4.3d has β on the x-axis for various values of T. Despite the different dynamics, the curves bear a resemblance to the infection dynamics in panels (a) and (b). In both cases, a high trophic coherence ($T \simeq 0$) can ensure that the pulse of activity will reach most of the network irrespectively of other parameters, whereas if the network is incoherent ($T \gg 0$), propagation resulting in a large incidence requires low α (for the complex contagion model) or low β (in the neural network) respectively.

Figure 4.4 displays heatmaps of incidence and duration for both the complex contagion model (figs. 4.4a and 4.4b) and the neural network model (figs. 4.4c and 4.4d). Figures 4.4a and 4.4c show the mean incidence against T and the relevant model parameter (α for the complex contagion and β for the neural network), while figs. 4.4b and 4.4d show the logarithm of the duration against the same parameters. We performed simulations for a maximum of 10^3 Monte Carlo steps, so any duration above this can mean either a long but eventually finite (transient) period of

activation or an endemic state in which a degree of activity survives indefinitely.

By comparing incidence and duration, we can discern that both models exhibit three qualitatively different regimes of behaviour: at high T and high α or β, activity dies out quickly without reaching most of the system; at high T and low α or β, activity spreads to the whole system and remains indefinitely; finally, at low T, activity spreads to the whole system but dies out quickly. We can refer to these regimes as *inactive, endemic* and *pulsing* respectively. The main qualitative difference between the behaviour of the two models regards the endemic regime. In the complex contagion case, the endemic regime is confined to sufficiently incoherent networks and its range increases monotonically with T. On the other hand, in the neural network endemicity occurs for any T if $\beta \lesssim 2$, and the range is non-monotonic with T, peaking at intermediate values of T.

Why does trophic coherence affect the spreading processes as described and in such similar ways for both kinds of dynamics? Consider first the case of complex contagion on a perfectly coherent network (low T), like the one in fig. 4.1a. If the basal nodes are all initially infected, then we have from eq. (4.4) that in the next time step the probability of infection for nodes at level $s = 2$ is $P = 1$ for any value of α and thus the infections moves up a level. By the same process, one time step later the infections spreads to level $s = 3$ and continues to spread in this way until it has reached the whole system—at which point the infection dies out because the nodes at the top have no outgoing connections. On an incoherent network (high T), like the one in fig. 4.1b, as the infection moves up the trophic levels, the fraction of infected in-neighbours affecting a given node i, f_i becomes lower with increasing s_i. This is because with increasing T, the connections between nodes span a wider range of trophic levels and since infected nodes recover immediately, it is likely that at any given time a smaller proportion of a node's in-neighbours will bee infected than in the low T scenario. Hence, if the network is sufficiently incoherent, a pulse of activity will die out as it progresses up the levels and only reach a finite fraction of the nodes. This explains why the pulsing regime occurs at low T. According to eq. (4.5), the above considerations apply also to the neural network model at low T when β is sufficiently high that the probability of a node being activated when all its in-neighbours are active is $P \simeq 1$.

For the complex contagion to become endemic, given that nodes recover immediately after infection, there must be some degree of feedback, i.e. the network must contain cycles. As Johnson and Jones [128] have shown, the expected number of cycles in a network is a function of its trophic coherence q, and for q below a particular value (which depends on other topological properties), networks are al-

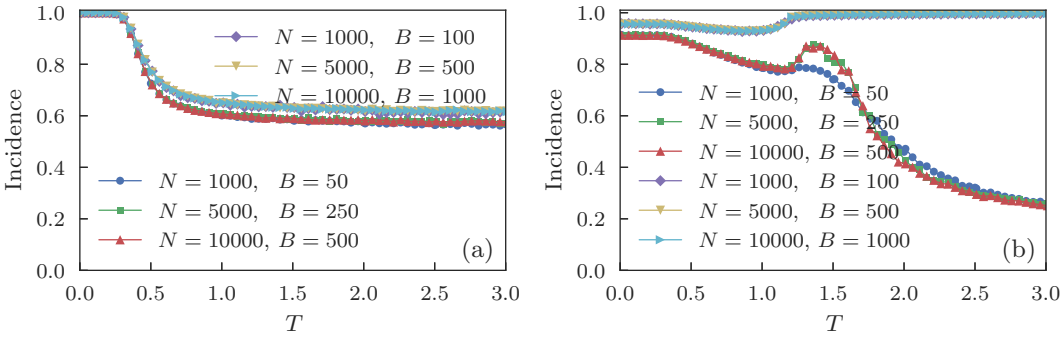

Figure 4.5: Average incidence values from Monte Carlo simulations of the two spreading models on networks with varying trophic coherence, as described in the main text. (a) Incidence against T (smaller T means more coherent networks) in the complex contagion model for $\alpha = 1$. (b) Incidence against T in the Amari-Hopfield neural-network model for $\beta = 3$. Symbols indicate different network sizes ($N = 1000, 5000$ and 10000) and proportions of basal nodes B ($N/B = 10$ and 20). In all cases, the mean degree is $\langle k \rangle = 5$. Averages are over 1000 runs.

most always acyclic. This explains the endemic phase which grows in range with T. However, the extent of node reinfection will depend on both the density of cycles and the infection probability as determined by eq. (4.4); reinfection is therefore more likely at lower α. Again, this argument extends to the neural-network model with a caveat. In the complex contagion model a node must have at least one infected in-neighbour to have non-zero probability of becoming infected, so for the endemic regime to arise, cycles in the network are a necessity. By contrast, the neural network model admits spontaneous node activation for any finite value of β. If β is low enough, the system enters the standard paramagnetic (or memoryless) phase of the model with continuous, random activation of nodes. This explains why the endemic regime for the neural model extends to the full range of T for low β.

Finally, the non-monotonic dependence of the endemic regime with T in the neural model seems to be cause by a balance between the two mechanisms we have described for activity propagation: a rapid pulse which can travel of coherent networks and the reverberation allowed for by the cycles present in incoherent ones. It is perhaps noteworthy that this effect of feedback in the neural model is similar to the mechanism of "cluster reverberation" put forward to explain short-term memory [130].

To conclude, we look into the effects of network size and the proportion of basal nodes on incidence. Figure 4.5a shows how incidence depends on T when we fix $\alpha = 1$ for the complex contagion model and fig. 4.5b shows the same for the neural network model with $\beta = 3$. Results are presented for three network sizes

($N = 1000, 5000$ and 10000), and two ratios of basal nodes, $N/B = 10$ and 20. In the complex contagion model, the lines for different N but fixed basal ratio collapse and there is only a small effect of the basal ratio on the incidence at high T. In the neural network model, however, there is a much more pronounced influence of the basal ratio on the incidence at high T. At high T, a ratio of $N/B = 10$ allows for the activity to reach the whole system while this is not possible if $N/B = 20$. This may be a consequence of the dependence of the mean trophic level on this ratio [128] which for random graphs has an expected value $\langle s \rangle = N/B$ (see eq. (3.16)). When $N/B = 20$, the non-monotonicity of incidence with T is also exacerbated slightly by N.

4.5 Discussion

We have shown that the trophic coherence of directed networks can have a significant impact on spreading processes taking place on top of the networks. In particular, our numerical investigation of two very different dynamics—one a model inspired by epidemics and the other a neural network model originally put forward to explain associative memory—indicates that this topological feature is relevant for any system in which some kind of signal is transmitted between agents in such a way that these signals interact. While there is not yet an analytical theory able to describe the precise spreading dynamics as a function of trophic coherence, it is clear that such a theory should take into account two effect: the transmission of pulses of synchronous activities which can occur on highly coherent topologies, and the maintenance of endemic states enabled by feedback loops on incoherent networks. Trophic coherence has already been shown to play an important role in determining various features of directed networks such as linear stability [131], feedback loops [128] and intervality [76]. We add here to such work by showing that spreading processes are also strongly influenced by this recently identified topological feature and submit that more research is required to determine its relationship to other network properties, to build a generalised understanding of its bearing on dynamical processes, and to discover by what mechanisms non-trivial trophic coherence comes about in nature.

CHAPTER 5

CLUSTERING AND ROBUSTNESS IN NETWORKS

Network robustness or resilience to random or targeted breakdown is a widely studied phenomenon in network science due to its crucial role for designing artificial networks such as power grids and the Internet as well as understanding connectivity patterns and stability of natural networks [7, 64, 65]. The tool of choice for studying robustness is percolation—the random (or targeted) removal of nodes or links on a network, which is among the most widely studied processes in statistical physics [2]. The robustness of the network is then measured as the fraction of of nodes remaining in the giant connected component.

The emergence of robustness in random networks is typically studied in equilibrium ensembles such as the ER and the CM ensembles. Much less is known about the emergence of robustness in systems far from equilibrium. While the equilibrium case only considers the emergence of robustness *on average*, in this chapter we explore how robustness arises starting from atypical, minimum entropy states of the ensemble and relaxing towards more typical equilibrium states.

5.1 Background and related work

Robustness of networks is generally studied in the context of percolation theory [2]. In its simplest form, percolation theory studies the robustness of interconnected systems such as networks by randomly removing a fraction of nodes or links (corresponding to *site percolation* and *bond percolation* respectively) and examining the structure of the remaining system. Conventionally, percolation is parametrized by the node or link occupation probability p so that each node or edge is deleted with probability $1 - p$. A robust or a resilient network is one which requires a large frac-

tion of node/link deletions to lose its giant connected component (GCC) while a non-robust network is one which does not require much damage to disintegrate into a union of many small components. This leads naturally to the concept of *percolation threshold* p_c which is simply the smallest value of the occupation probability for which a GCC exists with high probability in the $N \to \infty$ limit. It is worth noting that the study of percolation is intimately linked to the study of the SIR disease model on networks as discussed in Chapter 4 [138, 181].

Many other types of percolation processes beyond simple site and bond percolations have been proposed, examples include k-core percolation, in which nodes of degree less than k are progressively removed [79], bootstrap percolation, in which a collection of initially "activated" nodes activate their neighbours successively [28] and a generalized epidemic process which studies percolation via a contagion process [75, 125]. Recently these have all been shown to be special cases of the Watts Threshold Model discussed in Chapter 4 [170] .Other percolation processes include choice so as to maximise the damage mimicking targeted attacks [7, 65, 223] or to delay the percolation threshold such as in explosive percolation [84].

Network robustness via percolation can be studied in two ways—on single networks, usually of empirical origin, or on whole random network ensembles. In the first instance, there is only one underlying network structure, for example an empirical network such as a snapshot of the Internet [64] or a power grid [215]. Because of the stochastic nature of percolation, the percolation process is repeated many times for the same network structure and averages of the size of the GCC are computed for each value of p to determine the robustness and percolation threshold of the empirical network. By contrast, for studying percolation in random network ensembles the underlying network structure is also different from realization to realization, but the procedure for studying robustness is the same—generate many realization of random networks from some predefined ensemble (e.g. the configuration model with a fixed degree sequence), damage them via percolation and average the resulting sizes of the GCCs [55].

Studying percolation via simulations is time and resource consuming. In many cases analytical results can be derived to predict the percolation threshold, size of the GCC if it exists and even the size distribution of small components. In some cases these results are exact (in the $N \to \infty$ limit) but in others only approximate results have been obtained due to intrinsic correlations in the network structure (such as the existence of triangles). In the next two sections we briefly review approaches to studying network robustness via percolation for ensembles of random tree-like as well as triangle rich networks.

5.1.1 Robustness of tree-like networks

The study of network robustness started with the study of random network models itself by Erdős and Rényi [89] who studied the emergence of the giant connected component (GCC) in the Erdős and Rényi (ER) random graph model (see Section 2.3.1). In the $G(N, p)$ model the link occupation probability p is exactly the same as the bond occupation probability as studied in percolation theory [2]. The difference between traditional percolation theory and percolation on random graphs is the underlying structure—rather than considering rigid lattice models, more general structures such as the ER model are considered (see fig. 2.3). Erdős and Rényi [89] showed that there exists a critical link occupation probability (percolation threshold) $p_c = 1/N$ over which the $G(N, p)$ model almost surely has a unique GCC containing a positive fraction of nodes. Equivalently, in the large network limit ($N \to \infty$), there is a critical average degree $\langle k \rangle_c = 1$ over which a GCC emerges.

With the advance of generalized random networks such as the configuration model (CM), robustness results in terms of the existence of a GCC were extended [174] leading to the celebrated Molloy-Reed criterion which states that in infinite CM networks a GCC exists almost surely provided $\langle k^2 \rangle - 2\langle k \rangle > 0$. The size of the GCC when it exists was also derived by Molloy and Reed [175]. Cohen et al. [64] derive the same expression using a different approach and apply it to studying the resilience of the Internet. Schwartz et al. [218] generalizes this method for directed networks.

Slightly later in two seminal papers Callaway et al. [55] and Newman et al. [183] introduced the generating function (GF) formalism [249] which allows for analytical calculation of the critical point, the size of the GCC if it is present and numerical calculation of the size distribution of small components. The GF approach was later extended to directed networks which have a more complicated component structure [47, 78, 147]. Only very recently the GF formalism has yielded an analytical expression for the size distribution of small components in undirected networks [149] as well as directed and multiplex networks [148].

More recently, an alternative approach to assess network robustness has emerged—the so called *message passing algorithm*—first introduced in the context of general epidemic models [132] but later leading to a range of new results about the location of the phase transition for the emergence of the GCC as well as a new, fast algorithm for the numerical calculation of the component size distribution [135]. The message passing algorithm has also been extended to directed networks leading to new insights about the component structure [233]. It has also been extended to characterizing robustness of finite networks which highlights some limitations of the

approach when compared to direct numerical simulation [234].

The huge success of both the GF and the message passing approach has crucially hinged on the "locally tree-like assumption". Simply put, both approaches for calculating component sizes and the emergence of the GCC only yield exact results in the case of networks with no short loops such as triangles. As we saw in Section 2.3.2, large CM networks have exactly this property which have enabled the application of these two approaches to accurately describe the robustness properties of CM networks. However, real networks are rarely triangle-free which makes these methods inadequate for networks with high clustering. In the next section we briefly review approaches to studying robustness of clustered networks.

5.1.2 Robustness of clustered networks

Studying percolation in networks with a non-negligible number of short, closed loops is fundamentally non-trivial because of the structural correlations that do not exist in in tree-like networks. To illustrate this simple fact, consider that in tree-like networks one can safely assume that the second neighbours of a random node are always exactly two links away from the initial node. In networks with short loops such as triangles this is no longer true as the second neighbour of a node can also be a first neighbour via a transitive link, see fig. 5.1. Ignoring these correlations leads to overestimating the number of second neighbours and consequently the extent of the giant component [182]. Nevertheless, several attempts to extend the methods developed for tree-like networks to networks with non-negligible clustering coefficients have been made [33, 207, 220, 256], although exact results are limited to some special cases [73].

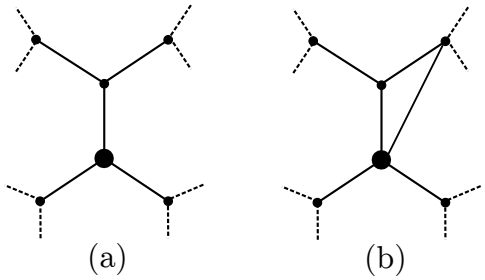

Figure 5.1: (a) A tree-like network in which no short loops are present. Both second neighbours of the central node are exactly two links away. (b) A network with a triangle. One of the second neighbours of the central node is also a first neighbour because of the transitive link.

Attempts to extend the GF method to networks with considerable clustering

have either yielded results for a limited set of networks, (e.g. networks in which triangles can have no overlapping edges [220]) or in contradictory results due to the violation of edge independence in clustered random networks [33]. Recently the message passing algorithm has also been extended to clustered networks [207], but it is inadequate in dealing with whole random network ensembles with clustering. Treating the full ensemble of random networks with a given degree sequence $\mathbf{k} = \{k_1, k_2, \dots\}$ and a fixed clustering coefficient $C > 0$ still proves to be intractable in the general case [256]. Instead, Del Genio and House [73] study the special case of $k-$regular networks, where every node has degree k, and a fixed clustering coefficient C. They show that regardless of the level of clustering, a GCC is always present except in the degenerate $C = 1$ case in which the network consists of a union of disconnected cliques of size $k + 1$.

Despite the large body of research surrounding the effect of clustering on the robustness of networks, its implications are still not fully understood which points to a need for more fundamental research of clustered networks. Here we take a step back and explore simple relaxation dynamics of highly clustered networks to an unclustered (uncorrelated) equilibrium state. Specifically, we study the evolution of the clustering coefficient under two edge rewiring schemes starting with fully clustered, degree-regular networks in which all nodes have the same number of neighbours and a maximal number of triangles. We find that under both dynamics whose equilibrium distributions correspond to the ER random graph and the CM respectively, a GCC emerges via a continuous phase transition. We provide an analytical prediction of the critical point for this transition as well as derive time evolution equations for various network properties.

5.2 Methods

5.2.1 Network metrics

We consider undirected graphs with N nodes and L edges described by a symmetric $N \times N$ adjacency matrix $\mathbf{A}$ with binary edge variables $a_{ij} \in \{0, 1\}$ for $i, j \in \{1, \dots, N\}$ with $a_{ij} = 1, i \neq j$ indicating an edge between nodes i and j so that $L = \sum_{i,j} A_{ij}$. The degree distribution of a network is defined as $p_k = N_k/N$, where N_k is the number of nodes with degree k. We denote the nth moment of the degree distribution by $\langle k^n \rangle$.

We define the *multiplicity* m_{ij} of an edge ij to be the number of triangles it participates in [219]. Similarly to the degree distribution, we define the edge multiplicity (or simply multiplicity) distribution as $q_m = L_m/L$, where L_m is the

number of edges with multiplicity m. We denote the nth moment of the multiplicity distribution by $\langle m^n \rangle$.

The clustering coefficient of a network is defined as three times the number of triangles divided by the number of connected triples, i.e. $C = 3N_{\triangle}/N_{\wedge}$ [185]. This measure of clustering is properly normalized so that $C \in [0, 1]$. It also admits a probabilistic interpretation—it is the probability that a randomly chosen connected triple of nodes is closed.

We can express the clustering coefficient in terms of the degree and multiplicity distributions. For any network we have

$$N_{\wedge} = \sum_k \binom{k}{2} N_k = N \sum_k \binom{k}{2} p_k = N \frac{\langle k^2 \rangle - \langle k \rangle}{2} \tag{5.1}$$

and

$$3N_{\triangle} = \sum_m L_m = L \sum_m m q_m = L \langle m \rangle. \tag{5.2}$$

Putting the above results together and noting that in any network $L = N\langle k \rangle/2$, we obtain the following general expression for the clustering coefficient:

$$C = \frac{\langle k \rangle \langle m \rangle}{\langle k^2 \rangle - \langle k \rangle}. \tag{5.3}$$

5.2.2 Random network ensembles

We study relaxation dynamics of k-regular networks under edge rewiring in two random network ensembles—the configuration model (CM) and the Erdős–Rényi random graph (ER).

The CM [48, 174, 185] is defined by drawing a valid degree sequence $\mathbf{k} = \{k_i\}_{i=1}^N$ from a degree distribution p_k and producing a network realization uniformly at random from all possible networks with that degree sequence [74, 185]. Provided the second moment of the degree distribution remains finite, it can be shown that the clustering coefficient scales as $C \sim 1/N$ so that in the thermodynamic limit ($N \to \infty$) the resulting networks are tree-like [185].

The ER random graph [88, 89, 105] is defined by placing L edges uniformly at random between N nodes[1]. If we require that the mean degree $\langle k \rangle = 2L/N$ be fixed, the degree distribution of the ER model in the thermodynamic limit is Poisson with mean $\langle k \rangle$ [185]. The ER model is thus a special case of the CM and has the same scaling behaviour of the clustering coefficient.

[1] Another common definition leading to a slightly different model is to place each of the possible $\binom{N}{2}$ edges with equal probability p, but this does not enforce a fixed number of edges.

Given that both the CM and ER random graphs are asymptotically triangle-free, it is natural to consider them as equilibrium ensembles for relaxation dynamics of highly clustered networks into an unclustered state. To this end we describe two edge rewiring mechanisms that have the CM and the ER random graphs as equilibrium distributions (see fig. 5.2 for a graphical demonstration).

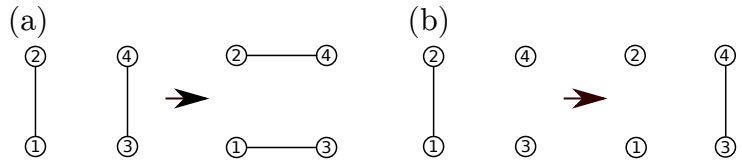

Figure 5.2: (a) Double edge swap or degree-preserving randomization. (b) Edge replacement or full randomization.

Double edge swap (CM). The double-edge swap [172, 208] is defined by choosing two existing edges in the network at random and rewiring their ends to produce two new edges while deleting the original two. This is also known as *degree-preserving* randomization and so naturally produces network realizations in the CM ensemble with a fixed degree sequence. The double-edge swap defines a Markov chain whose equilibrium distribution is the CM [208].

Edge replacement (ER). Alternatively, one can fully randomize a network by picking an edge at random and placing it anywhere in the network where there is no edge already [158, 178]. In this scheme the number of edges is preserved but the degrees of the nodes are not. Edge replacement defines a Markov chain whose equilibrium distribution is the ER ensemble.

A double-edge swap or an edge replacement constitutes an *elementary rewiring step.*

5.3 Results

To assess the evolution of network measures over time, we take into account the network size and the rewiring scheme (either CM or ER) to normalize the number of elementary rewiring steps per number of edges. If r_{CM} and r_{ER} are the number of elementary rewiring steps in the CM and ER ensembles respectively, we define the

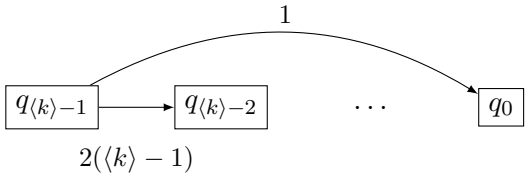

Figure 5.3: Transition rates in the multiplicity distribution for a single clique of size $\langle k\rangle + 1$.

corresponding time variables as

$$\begin{aligned} t_{\mathrm{ER}} &= \frac{r_{\mathrm{ER}}}{L} \\ t_{\mathrm{CM}} &= \frac{2r_{\mathrm{CM}}}{L}. \end{aligned} \tag{5.4}$$

These definitions have the useful interpretation that when $t_{\mathrm{scheme}} = 1$, the rewiring scheme has, on average, modified each edge in the network.

5.3.1 Multiplicity distribution

The multiplicity distribution evolves over time as edges are rewired and triangles are destroyed. The initial configuration of a k-regular network is a disjoint union of $N/(\langle k\rangle + 1)$ cliques of size $\langle k\rangle + 1$ which ensures maximal clustering $C = 1$. In other words, at time $t = 0$, the multiplicity distribution is

$$\begin{cases} q_{\langle k\rangle-1} & = 1 \\ q_m & = 0 \text{ if } m \neq \langle k\rangle - 1. \end{cases} \tag{5.5}$$

Consider the smallest informative time step $\Delta t_{\mathrm{CM}} = 2/L$ or $\Delta t_{\mathrm{ER}} = 1/L$ corresponding to exactly one elementary rewiring step. At $t = 0$ a clique of size $\langle k\rangle+1$ has exactly $\binom{\langle k\rangle+1}{2}$ edges all of which have maximal multiplicity $\langle k\rangle - 1$. Rewiring any single edge will destroy $\langle k\rangle - 1$ triangles leading to a decrease of $2(\langle k\rangle - 1) + 1$ edges with maximal multiplicity, one for the rewired edge and an additional two for each destroyed triangle. Assuming that no new triangles are created, the single rewired edge will have multiplicity zero. Figure 5.3 shows the transition rates in the multiplicity distribution of a single clique.

We now make the *ansatz* that this is the main way the multiplicity distribution changes over time—multiplicity is predominantly decreased by rewiring single edges from cliques and all such rewirings are independent. In this case, we can write down the full transition rate diagram between multiplicity classes as shown in

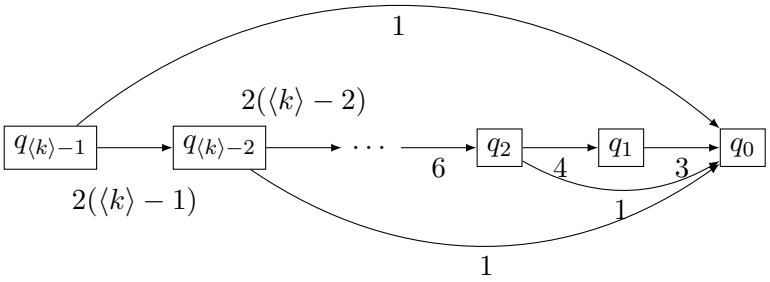

Figure 5.4: Transition rates in the full multiplicity distribution.

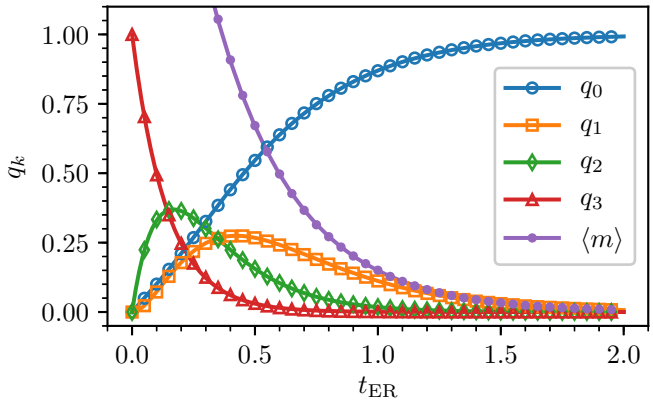

Figure 5.5: Evolution of the multiplicity distribution in an ER network with average degree $\langle k \rangle = 4$ and $N = 10^5$. The solid lines are numerical solutions of eq. (5.6) while the markers are simulation results. The purple line with filled circles indicates the average multiplicity $\langle m \rangle$.

fig. 5.4. This gives the following time evolution equations for q_m:

$$\begin{cases} \dfrac{dq_m}{dt} = -(2m+1)q_m + 2(m+1)q_{m+1}, & \text{for } m = \langle k \rangle - 1, \ldots, 1 \\ \dfrac{dq_0}{dt} = 3q_1 + \displaystyle\sum_{m=2}^{\langle k \rangle - 1} q_m. \end{cases} \tag{5.6}$$

Figure 5.5 shows the numerical solution of these ODEs which is in excellent agreement with simulation results. The calculations are valid both in the ER and the CM case.

Average multiplicity. Using the time evolution equations for the multiplicity distribution, we can derive exact expressions of its moments. Specifically we are interested in the average multiplicity $\langle m \rangle$ as it features in the expression for the

clustering coefficient. We have

$$\frac{d\langle m\rangle}{dt} = \sum_{m=1}^{\langle k\rangle -1} m\frac{dq_m}{dt}. \tag{5.7}$$

Inserting eq. (5.6) we obtain the simple expression

$$\frac{d\langle m\rangle}{dt} = -3\langle m\rangle. \tag{5.8}$$

Using the initial condition $\langle m\rangle(0) = \langle k\rangle - 1$, this has solution

$$\langle m\rangle = (\langle k\rangle - 1)e^{-3t}. \tag{5.9}$$

Figure 5.5 shows the analytic solution of the average multiplicity which is in perfect agreement with simulation results.

5.3.2 Degree distribution

In the case of the ER model, the degree distribution is also changing over time. Consider the degree distribution $p_k(t)$ as a function of time and a time step Δt_{ER}. We can calculate the rate at which $p_k(t)$ changes.

An edge replacement event in the ER model consists of two steps. First, a random edge is selected. Second, a random pair of nodes that are not linked by an edge (let us call this pair a *non-edge*) is selected and the edge selected in the first step is deleted while the non-edge becomes an edge.

When a random edge is selected, p_k can decrease if at least one end of the edge has degree k. Alternatively, p_k can increase if at least one end of the edge has degree $k+1$. The probability of reaching a node of degree k by following a randomly chosen edge is given by the so called excess degree distribution [185] which reads $s_k = kp_k/\langle k\rangle$. Given this and the fact that a randomly chosen edge can have 0,1 or 2 nodes of degree k, we can calculate the expected number of nodes of degree k at the ends of a random edge:

$$\mathbb{E}\left(k \to k-1\right) = 2s_k^2 + 2s_k\left(1 - s_k\right) = 2s_k = 2\frac{kp_k}{\langle k\rangle}. \tag{5.10}$$

This is the expected number of nodes whose degree would decrease from k to $k-1$ during a single edge selection step. Note that at the beginning of the process the degree distribution is regular so $\mathbb{E}\left(k \to k-1\right) = 2$ as expected.

Similarly, the expected number of nodes whose degree would decrease from

$k+1$ to k leading to an increase in p_k is:

$$\mathbb{E}\left(k+1 \to k\right) = 2s_{k+1} = 2\frac{(k+1)\,p_{k+1}}{\langle k\rangle}. \tag{5.11}$$

Now consider the second step in the edge replacement event, the selection of a non-edge. When a random non-edge is selected, p_k can also change in two ways. It can increase if at least one of the selected nodes has degree $k-1$ and it can decrease if at least one of the nodes has degree k. The calculation of the expected number of nodes changed as a result of this is similar to the previous case, but we must consider the distribution of *non-degrees* instead. To this end we study the graph complement of the original network defined as a network in which two nodes are linked if and only if they are not linked in the original network. From here on we denote by an overbar quantities in the graph complement.

It is easy to see that the degrees of nodes in the complement are given by $\bar{k} = N-1-k$ where k is the degree of a node in the original network and we have $p_{\bar{k}} = p_k$. Thus, the non-edges are selected proportionally to $\bar{k}$ not k as in the case of edge selection so we must work with the excess non-degree distribution given by $s_{\bar{k}} = \bar{k}p_k/\langle\bar{k}\rangle$. Note that the mean non-degree is given by

$$\langle\bar{k}\rangle = \sum_k \bar{k}p_k = N-1-\langle k\rangle. \tag{5.12}$$

As in the case of edge selection, the expected number of nodes whose degree would increase from k to $k+1$ thus reducing p_k during a single non-edge selection step is

$$\mathbb{E}\left(k \to k+1\right) = 2q_{\bar{k}} = 2\frac{\bar{k}p_k}{\langle\bar{k}\rangle} = \frac{2\left(N-1-k\right)}{N-1-\langle k\rangle}p_k. \tag{5.13}$$

When N is large we can approximate this by

$$\mathbb{E}\left(k \to k+1\right) \simeq 2p_k. \tag{5.14}$$

Similarly, p_k can increase if we select a non-edge with at least one node with degree $k-1$. The expected number of such nodes in a single non-edge selection is

$$\mathbb{E}\left(k-1 \to k\right) \simeq 2p_{k-1}. \tag{5.15}$$

Figure 5.6 describes pictorially the transition rates between degree classes as

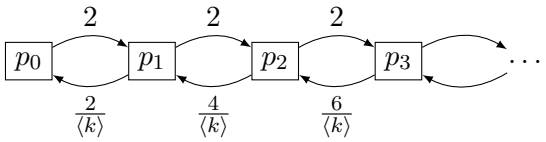

Figure 5.6: Transition rates in the degree distribution under the ER model.

derived here. This allows us to write down the time evolution equations for p_k:

$$\frac{dp_k}{dt} = 2p_{k-1} - 2\left(1 + \frac{k}{\langle k \rangle}\right) p_k + 2\frac{k+1}{\langle k \rangle} p_{k+1}, \tag{5.16}$$

for $k = 0, 1, \ldots$. This system of ODEs is not closed, so in order to solve it numerically, we must truncate the system at some p_{k^*} setting $p_k = 0$ for all $k > k^*$. The value of k^* should be set high enough so the probability mass unaccounted for is minimal for accurate predictions. We test our predictions by numerically solving the ODEs for a network with average degree $\langle k \rangle = 2$ and setting the cut-off $k^* = 8$. The results are shown in fig. 5.7. The numerical solution of the ODE system is in excellent agreement with simulation results. We also note that the cut-off is appropriate for this level of approximation as the total probability mass does not diverge from unity noticeably over the time period considered.

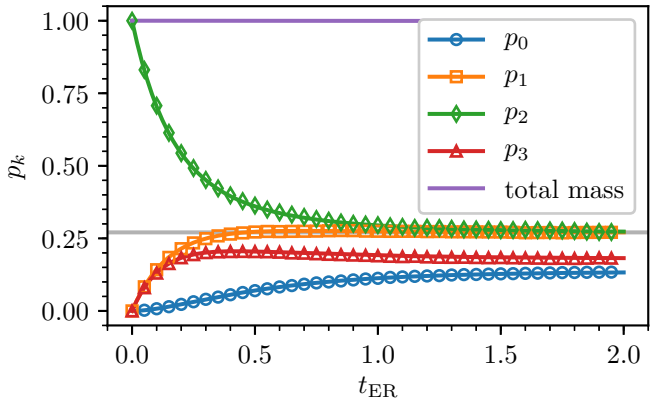

Figure 5.7: Evolution of the degree distribution in an ER network with average degree $\langle k \rangle = 2$ and $N = 10^5 - 1$. The solid lines are numerical solutions of eq. (5.16) while the markers are simulation results. The grey line indicates the equilibrium value of p_2 in an ER ensemble. The purple line indicates the total probability mass in the system accounted for by truncating the ODE system at $k^* = 8$.

Second moment of the degree distribution. Using the time evolution equations for the degree distribution, we can derive exact expressions of its moments.

Specifically, we are interested in the second moment $\langle k^2 \rangle$. We have

$$\frac{d\langle k^2 \rangle}{dt} = \sum_k k^2 \frac{dp_k}{dt}. \tag{5.17}$$

Inserting eq. (5.16) we obtain the simple expression

$$\frac{d\langle k^2 \rangle}{dt} = -4\frac{\langle k^2 \rangle}{\langle k \rangle} + 4\langle k \rangle + 4. \tag{5.18}$$

Using the initial condition $\langle k^2 \rangle(0) = \langle k \rangle^2$ and recalling that the average degree $\langle k \rangle$ is constant, this has solution

$$\langle k^2 \rangle = \langle k \rangle \left(\langle k \rangle + 1 - e^{-\frac{4t}{\langle k \rangle}} \right). \tag{5.19}$$

5.3.3 Clustering coefficient

Putting together the results for the multiplicity and degree distributions, and using eq. (5.3), we obtain exact expressions for the clustering coefficient as a function of time in both the CM and ER ensembles:

$$\begin{aligned} C_{\mathrm{CM}} &= e^{-3t} \\ C_{\mathrm{ER}} &= \frac{(\langle k \rangle - 1)\, e^{-3t}}{\langle k \rangle - e^{\frac{-4t}{\langle k \rangle}}}. \end{aligned} \tag{5.20}$$

We note that in the CM ensemble, the clustering coefficient has no dependence on the average degree while this is not the case for the ER ensemble. This is because the number of connected triples $N_\wedge$ in the CM ensemble is constant by virtue of having a fixed degree sequence while it is dependent on the evolving degree sequence in the ER ensemble.

5.3.4 Giant connected component

We find that under both rewiring schemes there is an emergence of global connectivity via the appearance of a giant connected component (GCC) at some critical time t^c (equivalently, critical clustering coefficient C^c). We confirm from simulation results that a GCC emerges in a continuous phase transition (figs. 5.8 and 5.9 for the CM and figs. 5.10 and 5.11 for the ER ensembles). Note that the large fluctuations in the 2-regular case is due to the fact that 2-regular networks are exactly at the point of criticality in the unclustered CM case ($C = 0$). This phenomenon has been studied in the context of reversible polymerization of rings [31].

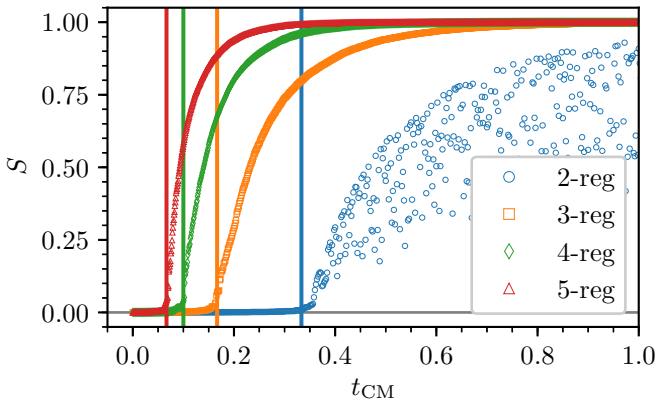

Figure 5.8: Proportion of nodes S in the giant connected component as a function of time t_{CM} for a few select k-regular networks. We observe a continuous phase transition at a critical point $t^{\mathrm{c}}_{\mathrm{CM}}$ which depends on the average degree of the network as explained in the main text. Vertical lines correspond to the analytically calculated critical points.

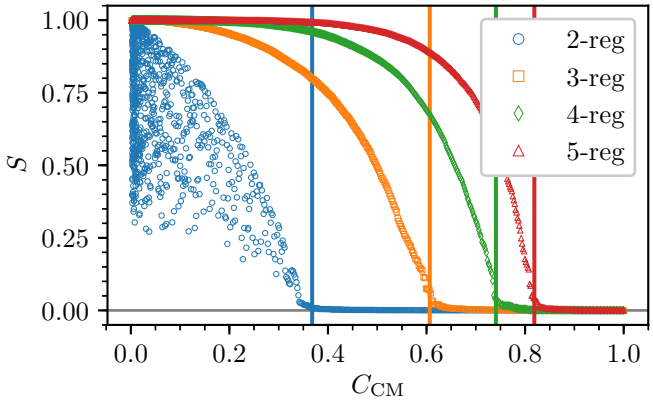

Figure 5.9: Proportion of nodes S in the giant connected component as a function of clustering C_{CM} for a few select k-regular networks under the CM rewiring scheme. We observe a continuous phase transition at a critical point $C^{\mathrm{c}}_{\mathrm{CM}}$ which depends on the average degree of the network as explained in the main text. Vertical lines correspond to the analytically calculated critical points.

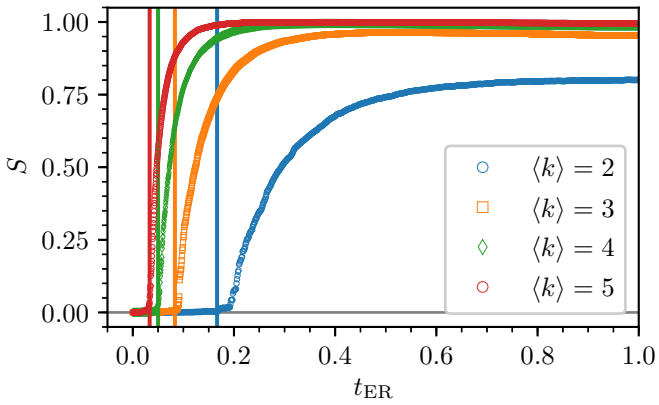

Figure 5.10: Proportion of nodes S in the giant connected component as a function of time t_{ER} for a few select mean degree $\langle k \rangle$ networks. We observe a continuous phase transition at a critical point $t_{\mathrm{ER}}^{\mathrm{c}}$ which depends on the average degree of the network as explained in the main text. Vertical lines correspond to the analytically calculated critical points.

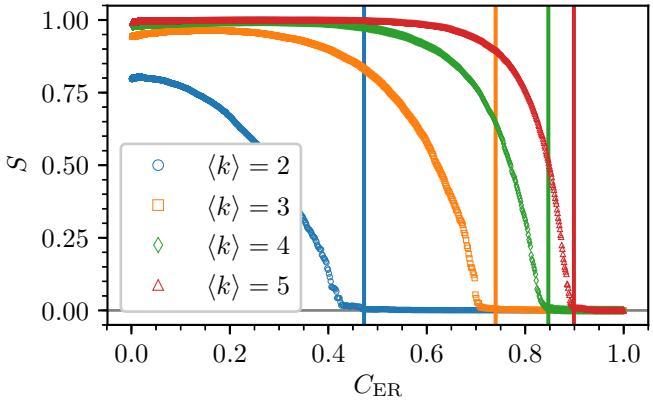

Figure 5.11: Proportion of nodes S in the giant connected component as a function of clustering C_{ER} for a few select mean degree $\langle k \rangle$ networks under the ER rewiring scheme. We observe a continuous phase transition at a critical point $C_{\mathrm{ER}}^{\mathrm{c}}$ which depends on the average degree of the network as explained in the main text. Vertical lines correspond to the analytically calculated critical points.

We can calculate the critical point analytically by using the known result that a GCC in an ER random graph emerges when $\langle k \rangle = 1$ [185]. We conjecture that a GCC induced by edge rewiring emerges when the average number of external edges between the original $N/(\langle k \rangle + 1)$ cliques of size $\langle k \rangle + 1$ exceeds one. If this is the case, the critical number of elementary rewiring steps is

$$r^{\mathrm{c}} = \frac{N}{2\left(\langle k \rangle + 1\right)}. \tag{5.21}$$

Expressing this in terms of the time variable, we obtain the critical time for both the CM and the ER rewiring schemes:

$$\begin{aligned} t^{\mathrm{c}}_{\mathrm{CM}} &= \frac{2}{\langle k \rangle \left(\langle k \rangle + 1\right)} \\ t^{\mathrm{c}}_{\mathrm{ER}} &= \frac{1}{\langle k \rangle \left(\langle k \rangle + 1\right)}. \end{aligned} \tag{5.22}$$

Note that these differ by a factor of two. This is because in the CM rewiring scheme, even though every elementary rewiring step involves two edges, the two rewirings are not independent—during one rewiring step it is possible to connect at most two disconnected components.

Expressed in terms of the clustering coefficient, the critical thresholds read:

$$\begin{aligned} C^{\mathrm{c}}_{\mathrm{CM}} &= e^{-6/\langle k \rangle (\langle k \rangle + 1)} \\ C^{\mathrm{c}}_{\mathrm{ER}} &= \frac{(\langle k \rangle - 1) e^{-3/\langle k \rangle (\langle k \rangle + 1)}}{\langle k \rangle - e^{-4/\langle k \rangle^2 (\langle k \rangle + 1)}}. \end{aligned} \tag{5.23}$$

Figures 5.8 and 5.9 confirm that these are in excellent agreement with simulations in the CM case and figs. 5.10 and 5.11 confirm a good agreement in the ER case which improves as the mean degree increases.

What is the cause of the discrepancy of the analytical result for the critical point and the numerical simulations, particularly for low mean degree ER networks? We conjecture that this is due to some edges being rewired multiple times while others are not rewired at all. This would have the effect of increasing the critical time because we have to wait slightly longer until the average number of rewired edges *discounting edges rewired multiple times* reaches the point where long range connectedness emerges. Figure 5.10 seems to confirm this to be the case. Let us calculate this revised critical time in the ER case.

During an edge replacement step, the probability of any edge being chosen for rewiring is $1/L$. So after r rewiring events the probability that a specific edge

has not been rewired is

$$\mathbb{P}\left(\text{not rewired}\right) = \left(1 - \frac{1}{L}\right)^{r}. \tag{5.24}$$

Substituting $r = Lt$ since we are in the ER case and taking the limit as $L \to \infty$, we get

$$\mathbb{P}\left(\text{not rewired}\right) = e^{-t}. \tag{5.25}$$

The new revised time for the emergence of the GC, call it t^{r}, is then the time at which point this probability drops below a certain threshold. What is this threshold? It should be when the proportion of edges that have been rewired gives rise to a GCC which is precisely given by t^{c}. We can then write

$$e^{-t^{\mathrm{r}}} = 1 - t^{\mathrm{c}}. \tag{5.26}$$

Note that by Taylor expansion we have $t^{\mathrm{r}} \simeq t^{\mathrm{c}}$ if this time is small as in the case when the average degree $\langle k \rangle \to \infty$. This explains why the t^{c} value becomes a better predictor for the critical threshold as the mean degree increases as seen in fig. 5.10.

The revised critical point in the ER case is thus

$$t^{\mathrm{r}} = -\log(1 - t^{\mathrm{c}}) = \log\left(\frac{\langle k \rangle(\langle k \rangle + 1)}{\langle k \rangle(\langle k \rangle + 1) - 1}\right). \tag{5.27}$$

Figure 5.12 confirms that t^{r} is a better predictor of the location of the phase transition. The difference between t^{c} and t^{r} becomes negligible as the mean degree increases.

Another aspect that could influence the position of the critical point is the possibility of having rewired more edges that needed to connect previously disconnected components. We show in Section 5.A that this should have no bearing on the critical point in large networks.

5.4 Discussion

In this chapter we studied the evolution of highly clustered networks under random edge rewiring dynamics. Our main result is showing the existence of a phase transition in which a giant connected component emerges. Del Genio and House [73] showed that equilibrium ensembles of degree-regular networks with prescribed clustering always admit a giant connected component. As a consequence, spreading processes such as infectious diseases in contact networks could always become en-

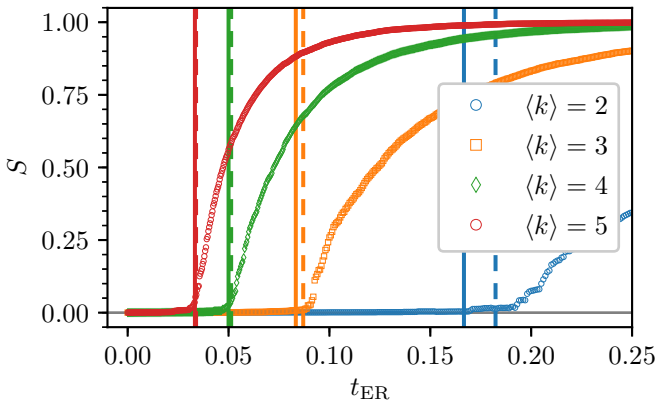

Figure 5.12: Proportion of nodes in the giant connected component S as a function of time t_{ER} for a few select mean degree $\langle k \rangle$ networks. Solid vertical lines correspond to the critical time $t^{\mathrm{c}}_{\mathrm{ER}}$ while dashed vertical lines correspond to the revised critical time $t^{\mathrm{r}}_{\mathrm{ER}}$.

demic regardless of the level of clustering. By contrast, our work implies that this need not be the case in non-equilibrium systems. Depending on the precise mechanisms of time evolution of real networks and the level of clustering maintained, a giant connected component facilitating spreading processes may or may not exist. We have studied a model in which highly clustered populations undergo fully random connectivity changes and even in this simple scenario we observe two different modes of global connectivity.

Another interesting aspect of our work is from the perspective of statistical mechanics. A maximally clustered network is essentially the lowest entropy microstate in the context of the random network ensembles studied here. This is because such a network, under relabelling of nodes, is unique and least likely to be produced by chance at equilibrium. By using this configuration as a starting state for network dynamics, we have shown that the emergence of global connectivity is effectively delayed. This raises several other research questions, for example, is random rewiring the most or least effective method of delaying the onset of global connectivity? It is probable that more sophisticated rewiring methods involving choice, such as those studied in explosive percolation [84], would lead to different critical thresholds. We have also limited ourselves to studying rewiring that consistently destroys triangles, but what about rewiring with a view to increase the number of triangles? A number of greedy as well as equilibrium algorithms exist and are widely applied to model highly clustered networks [98, 210], but it is unclear how they cover the space of all networks and can lead to interesting behaviour such as hysteresis loops [98]. Indeed, clustering in networks still leaves much to be explored.

5.A Extraneous edges

Another mechanism that could change the location of the critical point t^{c} is the number of extraneous edges between already connected components. A GCC is formed when there are enough external edges between the initial cliques. Only one external edge is needed to connect two cliques, but there are multiple ways to do it and sometimes multiple edges end up linking together the same cliques. For example, we need only one edge two join two disconnected triangles, but there are a total of 9 ways to do it, moreover there is no guarantee that we will not end up with multiple edges between these triangles.

More generally, let the average degree $\langle k \rangle$ be fixed, then at $t = 0$ there are $n = N/(\langle k \rangle + 1)$ cliques of size $\langle k \rangle + 1$. Any two cliques can therefore be connected in $(\langle k \rangle + 1)^2$ ways.

Suppose we never want to make more than one external edge to connect disconnected components. Then at $t = 0$ the number of choices for placing an external edge is given by

$$(\langle k \rangle + 1) \binom{n}{2} = \frac{N\left(N - \langle k \rangle - 1\right)}{2}. \tag{5.28}$$

After each rewiring event, the number of choices decreases by $(\langle k \rangle + 1)^2$, so after $r - 1$ rewires, the probability of placing an extraneous edge on the next rewire, r, is

$$\mathbb{P}\left(\text{extra edge on step } r\right) = \frac{2\left(\langle k \rangle + 1\right)^2 r}{N\left(N - \langle k \rangle - 1\right)}. \tag{5.29}$$

Thus, the expected number of extraneous edges after r rewiring events is

$$\mathbb{E}\left(\text{extra edges by step } r\right) = \sum_{r'=0}^{r} \frac{2\left(\langle k \rangle + 1\right)^2 r'}{N\left(N - \langle k \rangle - 1\right)}. \tag{5.30}$$

In particular, setting $r = r^{\mathrm{c}} = N/2(\langle k \rangle + 1)$ we get

$$\begin{aligned} &\mathbb{E}\left(\text{extra edges by step } r^{\mathrm{c}}\right) \\ &\qquad = \frac{2\left(\langle k \rangle + 1\right)^2}{N\left(N - \langle k \rangle - 1\right)} \cdot \frac{N}{4\left(\langle k \rangle + 1\right)} \left(\frac{N}{2\left(\langle k \rangle + 1\right)} + 1 \right) \\ &\qquad = \frac{N + 2\langle k \rangle + 2}{4\left(N - \langle k \rangle - 1\right)}. \end{aligned} \tag{5.31}$$

Taking the limit $N \to \infty$, we get

$$\mathbb{E}\left(\text{extra edges by step } r_c\right) \simeq \frac{1}{4}, \tag{5.32}$$

which is fixed and independent of network size. Therefore, the formation of extraneous edges does not affect the location of the critical point in the large network limit.

CHAPTER 6

CONCLUSIONS AND OUTLOOK

The studies considered in this thesis provide us with insights about the structure, organization and dynamics of complex networked systems. Some research questions in network science are motivated by an ever increasing amount of data while others consider fundamental properties of network models as their starting point of investigation. Here we have presented a synthesis of both data driven research (Chapter 3) and theoretical inquiry (Chapters 4 and 5) into network science.

Studying the make-up of ecological interactions may prove to be the cornerstone in understanding the relationship between human activity and nature, and is especially relevant in the age of ever more prevalent markers of climate change. Food webs have fascinated ecologists for decades due to their structural features resulting, for example, in counter-intuitive stability properties in the face of ever increasing complexity [131, 167]. In Chapter 3 we investigated how local preying patterns, conceptualized as network motifs, relate to large-scale organization described by trophic coherence. We found that the motif corresponding to omnivory—the tendency to prey on species from several trophic levels—is strongly correlated with trophic coherence and provides a basis of clustering food webs into two families—ones where omnivory is widespread and the others where it is sparse. Our network model which incorporates tunable trophic coherence can successfully generate artificial food webs in either family, suggesting the importance of trophic coherence as well as basal species density in modelling realistic food webs. However, the biological origin of trophic coherence and the fragmentation of food webs into distinct families depending on the extent of ommnivory is still not understood. This is an important research question both for network scientists and ecologists in order to understand how food

webs assemble in nature.

Network structures are not only interesting because of the complex structural patterns often found in empirically observed networks, but also because of their propensity to support dynamical processes on top of their topologies. In Chapter 4 we studied two paradigmatic spreading processes—a complex contagion inspired by epidemics and a neural network process—on directed networks with tunable trophic coherence. Our results uncovered a rich interplay between network structure and the fate of outbreaks ranging from short-lived but global infections to endemic infections confined to a small subset of nodes. These results could help our understanding of a variety of network processes taking place on directed networks—from epidemics and rumour spreading to shocks to ecosystems.

The study of the relationships between local and global network properties and how change in one influences another is a widely studied topic in network science. However, many attempts in quantifying these effects end up with inconclusive results due to the inherent difficulties in studying random networks that exhibit strong link dependence. In Chapter 5 we laid the groundwork for studying how clustering influences the formation of a giant connected component in ensembles of random networks. Contrary to previous approaches that study the problem at equilibrium, we use a non-equilibrium approach by selecting unlikely network states (maximal clustering) and studying their relaxation to common network states (low clustering) within the ensemble of interest. We show that by selecting highly unlikely configurations with a maximal number of triangles, the onset of the giant connected component can be delayed, unlike in the case of equilibrium networks. Our research readily raises several other research questions about the organization of complex systems. For example, is random rewiring the most effective method of delaying the formation of a GCC or are there methods, possibly including choice in the rewiring akin to explosive percolation [84], that would effect the position of the phase transition? When considering directed networks, an analogous research question would be to study the emergence of a strongly connected component when starting from highly coherent network configurations as considered in Chapters 3 and 4.

Our focus in this thesis has been on simple directed and undirected networks, however, in recent years several generalizations of network structures have been proposed to better capture the interactions in real world systems. Multilayer networks [43, 142] is a framework suited for studying complex systems which are made up of of a number of interacting, but fundamentally distinct networks. Examples include transportation multilayer networks [9] in which nodes represent stops or interchanges and links in each layer correspond to different transportation types.

Another example is social media networks—people are represented by nodes but connections between them will vary between different social media platforms and real life connections [155, 162]. Simplicial complexes, originally studied in the context of algebraic topology [113], have become an interesting tool of modelling complex systems in which more than two nodes can have interactions between them thus generalizing the inherent dyadic interactions represented by links in networks [68, 69, 253, 257]. The quintessential example is scientific collaboration networks. A transitive relationship between three authors in a classical network description cannot distinguish between the case that all three authors have collaborated on the same paper and the case that they have collaborated only pairwise on distinct papers. A simplicial complex description, however, allows for both kinds of interactions to be present and distinguishable from one another. Hypergraphs take the notion of simplicial complexes and extends it even further allowing very general interactions between any set of nodes [35, 104]. Generalized network structures is a flourishing field with many recent results obtained as generalizations of simpler results on networks. This is sure to become an exciting new paradigm for studying complex, interacting systems.

BIBLIOGRAPHY

[1] Luis G Abarca-Arenas and Robert E Ulanowicz. The effects of taxonomic aggregation on network analysis. *Ecological Modelling*, 149(3):285–296, 2002.

[2] Amnon Aharony and Dietrich Stauffer. *Introduction to percolation theory*. Taylor & Francis, 2003.

[3] Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2): 221–248, 2015.

[4] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[5] Réka Albert, István Albert, and Gary L Nakarado. Structural vulnerability of the north american power grid. *Physical Review E*, 69:025103, 2004.

[6] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[7] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[8] David Aldous and Jim Fill. Reversible markov chains and random walks on graphs, 2002.

[9] Alberto Aleta, Sandro Meloni, and Yamir Moreno. A multilayer perspective for the analysis of urban transportation systems. *Scientific Reports*, 7:44359, 2017.

[10] Stefano Allesina, Antonio Bodini, and Mercedes Pascual. Functional links and robustness in food webs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1524):1701–1709, 2009.

[11] J Almunia, G Basterretxea, J Arístegui, and R E Ulanowicz. Benthic-pelagic switching in a coastal subtropical lagoon. *Estuarine, Coastal and Shelf Science*, 49(3):363–384, 1999.

[12] Aamena Alshamsi, Flavio L Pinheiro, and Cesar A Hidalgo. When to target hubs? Strategic Diffusion in Complex Networks. *arXiv:1705.00232 [physics]*, 2017.

[13] S I Amari. Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements. *IEEE Transactions on Computers*, C-21(11):1197–1206, 1972.

[14] Kartik Anand and Ginestra Bianconi. Entropy measures for networks: Toward an information theory of complex topologies. *Physical Review E*, 80(4):045102, 2009.

[15] Kartik Anand, Dmitri Krioukov, and Ginestra Bianconi. Entropy distribution and condensation in random networks with a given degree distribution. *Physical Review E*, 89(6):062807, 2014.

[16] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*, volume 28. Oxford University Press, Oxford, 1992.

[17] Alex Arenas, Albert Díaz-Guilera, Jurgen Kurths, Yamir Moreno, and Changsong Zhou. Synchronization in complex networks. *Physics Reports*, 469(3): 93–153, 2008.

[18] V A Avetisov, S K Nechaev, and A B Shkarin. On the motif distribution in random block-hierarchical networks. *Physica A: Statistical Mechanics and its Applications*, 389(24):5895–5902, 2010.

[19] Shweta Bansal, Bryan T Grenfell, and Lauren Ancel Meyers. When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of The Royal Society Interface*, 4(16):879–891, 2007.

[20] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[21] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, Cambridge, 2016.

[22] Albert-László Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, and T Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614, 2002.

[23] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, Cambridge, 2008.

[24] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[25] Jordi Bascompte and Carlos J Melián. Simple trophic modules for complex food webs. *Ecology*, 86(11):2868–2873, 2005.

[26] Jordi Bascompte, Carlos J Melián, and Enric Sala. Interaction strength combinations and the overfishing of a marine food web. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5443–5447, 2005.

[27] Kevin E Bassler, Charo I Del Genio, Péter L Erdős, István Miklós, and Zoltán Toroczkai. Exact sampling of graphs with prescribed degree correlations. *New Journal of Physics*, 17(8):083052, 2015.

[28] G J Baxter, S N Dorogovtsev, A V Goltsev, and J F F Mendes. Bootstrap percolation on complex networks. *Physical Review E*, 82:011103, 2010.

[29] Moritz Emanuel Beber, Christoph Fretter, Shubham Jain, Nikolaus Sonnenschein, Matthias Müller-Hannemann, and Marc-Thorsten Hütt. Artefacts in statistical analyses of network motifs: general framework and application to metabolic networks. *Journal of The Royal Society Interface*, 9(77):3426–3435, 2012.

[30] Mariano Beguerisse-Díaz, Guillermo Garduño-Hernández, Borislav Vangelov, Sophia N. Yaliraki, and Mauricio Barahona. Interest communities and flow roles in directed networks: the Twitter network of the UK riots. *Journal of The Royal Society Interface*, 11(101), 2014.

[31] E Ben-Naim and P L Krapivsky. Kinetics of ring formation. *Physical Review E*, 83(6):061102, 2011.

[32] Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296 – 307, 1978.

[33] Yakir Berchenko, Yael Artzy-Randrup, Mina Teicher, and Lewi Stone. Emergence and Size of the Giant Component in Clustered Random Graphs with a Given Degree Distribution. *Physical Review Letters*, 102(13):138701, 2009.

[34] Johannes Berg and Michael Lässig. Correlated Random Networks. *Physical Review Letters*, 89(22):228701, 2002.

[35] Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

[36] Ginestra Bianconi. The entropy of randomized network ensembles. *EPL (Europhysics Letters)*, 81(2):28005, 2008.

[37] Ginestra Bianconi and Albert-László Barabási. Bose-Einstein Condensation in Complex Networks. *Physical Review Letters*, 86(24):5632–5635, 2001.

[38] Ginestra Bianconi and Albert-László Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

[39] Ginestra Bianconi and Matteo Marsili. Loops of any size and Hamilton cycles in random scale-free networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(06):P06005, 2005.

[40] Norman Biggs, E Keith Lloyd, and Robin J Wilson. *Graph Theory, 1736-1936*. Oxford University Press, 1976.

[41] Joseph Blitzstein and Persi Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics*, 6(4):489–522, 2011.

[42] S Boccaletti, V Latora, Y Moreno, M Chavez, and D Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[43] S Boccaletti, G Bianconi, R Criado, C I del Genio, J Gómez-Gardeñes, M Romance, I Sendiña-Nadal, Z Wang, and M Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[44] Marian Boguñá, Luis F Lafuerza, Raúl Toral, and M Ángeles Serrano. Simulating non-markovian stochastic processes. *Physical Review E*, 90:042108, 2014.

[45] M Boguñá, R Pastor-Satorras, and A Vespignani. Cut-offs and finite size effects in scale-free networks. *The European Physical Journal B*, 38(2):205–209, 2004.

[46] Marián Boguñá and Romualdo Pastor-Satorras. Class of correlated random networks with hidden variables. *Physical Review E*, 68(3):036112, 2003.

[47] Marián Boguñá and M Ángeles Serrano. Generalized percolation in random directed networks. *Physical Review E*, 72(1):016106, 2005.

[48] Béla Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001.

[49] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311 – 316, 1980.

[50] Jonathan J Borrelli. Selection against instability: stable subgraphs are most frequent in empirical food webs. *Oikos*, 124(12):1583–1588, 2015.

[51] Tom Britton, Maria Deijfen, and Anders Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6):1377–1397, 2006.

[52] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186, 2009.

[53] Z Burda, J Jurkiewicz, and A Krzywicki. Network transitivity and matrix models. *Physical Review E*, 69(2):026106, 2004.

[54] Guido Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.

[55] Duncan S Callaway, Mark E J Newman, Steven H Strogatz, and Duncan J Watts. Network Robustness and Fragility: Percolation on Random Graphs. *Physical Review Letters*, 85(25):5468–5471, 2000.

[56] J Camacho, D B Stouffer, and L A N Amaral. Quantitative analysis of the local structure of food webs. *Journal of theoretical biology*, 246(2):260–268, 2007.

[57] José A Capitán, Alex Arenas, and Roger Guimerà. Degree of intervality of food webs: From body-size data to models. *Journal of Theoretical Biology*, 334:35 – 44, 2013.

[58] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81:591–646, 2009.

[59] Robert R Christian and Joseph J Luczkovich. Organizing and understanding a winter's seagrass foodweb network through effective trophic levels. *Ecological Modelling*, 117(1):99 – 124, 1999.

[60] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

[61] J E Cohen and C M Newman. A stochastic theory of community food webs: I. models and aggregated data. *Proceedings of the Royal Society of London B: Biological Sciences*, 224(1237):421–448, 1985.

[62] Joel Cohen, Frédéric Briand, and Charles Newman. *Community food webs: data and theory*, volume 20 of *Biomathematics*. Springer-Verlag, Berlin, Germany, 1990.

[63] Joel E Cohen. *Food webs and niche space*. Number 11. Princeton University Press, 1978.

[64] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85: 4626–4628, 2000.

[65] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Breakdown of the internet under intentional attack. *Physical Review Letters*, 86: 3682–3685, 2001.

[66] Pol Colomer-de Simón and Marián Boguñá. Clustering of random scale-free networks. *Physical Review E*, 86(2):026120, 2012.

[67] Pol Colomer-de Simón, M Ángeles Serrano, Mariano G Beiró, J Ignacio Alvarez-Hamelin, and Marián Boguñá. Deciphering the global organization of clustering in real complex networks. *Scientific Reports*, 3:2517, 2013.

[68] Owen T Courtney and Ginestra Bianconi. Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. *Physical Review E*, 93(6):062311, 2016.

[69] Owen T Courtney and Ginestra Bianconi. Weighted growing simplicial complexes. *Physical Review E*, 95(6):062301, 2017.

[70] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[71] Leon Danon, Ashley P Ford, Thomas House, Chris P Jewell, Matt J Keeling, Gareth O Roberts, Joshua V Ross, and Matthew C Vernon. Networks and the Epidemiology of Infectious Disease. *Interdisciplinary Perspectives on Infectious Diseases*, 2011:e284909, 2011.

[72] Leon Danon, Thomas A House, Jonathan M Read, and Matt J Keeling. Social encounter networks: collective properties and disease transmission. *Journal of The Royal Society Interface*, 9(76):2826–2833, 2012.

[73] Charo I Del Genio and Thomas House. Endemic infections are always possible on regular networks. *Physical Review E*, 88(4):040801, 2013.

[74] Charo I Del Genio, Hyunju Kim, Zoltán Toroczkai, and Kevin E Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLOS ONE*, 5(4):e10012, 2010.

[75] Peter Sheridan Dodds and Duncan J Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92:218701, 2004.

[76] Virginia Domínguez-García, Samuel Johnson, and Miguel A Muñoz. Intervality and coherence in complex networks. *Chaos*, 26(6):065308, 2016.

[77] S Dorogovtsev, A Goltsev, and J Mendes. Critical phenomena in complex networks. *Reviews of Modern Physics*, 80(4):1275–1335, 2008.

[78] S N Dorogovtsev, J F F Mendes, and A N Samukhin. Giant strongly connected component of directed networks. *Physical Review E*, 64(2):025101, 2001.

[79] S N Dorogovtsev, A V Goltsev, and J F F Mendes. k. *Physical Review Letters*, 96:040601, 2006.

[80] David A Drachman. Do we have brain to spare? *Neurology*, 64(12):2004–2005, 2005.

[81] Barbara Drossel and Alan J McKane. Modelling food webs. In *Handbook of Graphs and Networks: From the Genome to the Internet*, pages 218–247. Wiley-VCH GmbH & Co. KGaA, 2005.

[82] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Network structure and robustness of marine food webs. *Marine Ecology Progress Series*, 273:291–302, 2004.

[83] Rick Durrett. *Random Graph Dynamics*. Cambridge University Press, Cambridge, 2007.

[84] Raissa M D'Souza and Jan Nagler. Anomalous critical and supercritical phenomena in explosive percolation. *Nature Physics*, 11(7):531–538, 2015.

[85] Ken T D Eames, Jonathan M Read, and W John Edmunds. Epidemic prediction and control in weighted networks. *Epidemics*, 1(1):70–76, 2009.

[86] A Eklöf, U Jacob, J Kopp, J Bosch, R Castro-Urgal, B Dalsgaard, N Chacoff, C deSassi, M Galetti, P Guimaraes, S Lomáscolo, A Martín González, M A Pizo, R Rader, A Rodrigo, J Tylianakis, D Vazquez, and S Allesina. The dimensionality of ecological networks. *Ecology Letters*, 16:577–583, 2013.

[87] Charles S Elton. *Animal ecology*. University of Chicago Press, 1927.

[88] Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

[89] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

[90] Péter L Erdős, István Miklós, and Zoltán Toroczkai. A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs. *The Electronic Journal of Combinatorics*, 17(1):R66, 2010.

[91] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

[92] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Hierarchical Clustering*, pages 71–110. John Wiley & Sons, 2011.

[93] Giorgio Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007.

[94] Peter G Fennell, Sergey Melnik, and James P Gleeson. Limitations of discrete-time approaches to continuous-time contagion dynamics. *Physical Review E*, 94:052125, 2016.

[95] Rico Fischer, Jorge C Leitão, Tiago P Peixoto, and Eduardo G Altmann. Sampling Motif-Constrained Ensembles of Networks. *Physical Review Letters*, 115(18):188701, 2015.

[96] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3): 75–174, 2010.

[97] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

[98] David Foster, Jacob Foster, Maya Paczuski, and Peter Grassberger. Communities, clustering phase transitions, and hysteresis: Pitfalls in constructing network ensembles. *Physical Review E*, 81(4):046115, 2010.

[99] Christoph Fretter, Matthias Müller-Hannemann, and Marc-Thorsten Hütt. Subgraph fluctuations in random graphs. *Physical Review E*, 85(5), 2012.

[100] Piotr Fronczak, Agata Fronczak, and Maksymilian Bujok. Exponential random graph models for networks with community structure. *Physical Review E*, 88 (3):032810, 2013.

[101] T Gallai and P Erdős. Graphs with prescribed degree of vertices (Hungarian). *Matematikai Lapok*, 11:264–274, 1960.

[102] Yan Gang, Zhou Tao, Wang Jie, Fu Zhong-Qian, and Wang Bing-Hong. Epidemic spread in weighted scale-free networks. *Chinese Physics Letters*, 22(2): 510, 2005.

[103] Diego Garlaschelli, Guido Caldarelli, and Luciano Pietronero. Universal scaling relations in food webs. *Nature*, 423(6936):165–168, 2003.

[104] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and M E J Newman. Random hypergraphs and their applications. *Physical Review E*, 79:066118, 2009.

[105] E N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144, 1959.

[106] E N Gilbert. Random plane networks. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):533–543, 1961.

[107] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[108] M Girvan and M E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[109] Lloyd Goldwasser and Jonathan Roughgarden. Construction and analysis of a large caribbean food web. *Ecology*, 74:1216–1233, 1993.

[110] Mark S Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[111] S Louis Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, 1962.

[112] T E Harris. Contact interactions on a lattice. *The Annals of Probability*, 2: 969–988, 1974.

[113] Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2002.

[114] Václav Havel. A remark on the existence of finite graphs (Czech). *Časopis pro pěstování matematiky*, 80:477–480, 1955.

[115] Karl Havens. Scale and structure in natural food webs. *Science*, 257(5073): 1107–1109, 1992.

[116] Malte Henkel, Haye Hinrichsen, Sven Lübeck, and Michel Pleimling. *Non-equilibrium phase transitions*, volume 1. Springer, 2008.

[117] Suzana Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3:31, 2009.

[118] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[119] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519 (3):97–125, 2012. Temporal Networks.

[120] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79 (8):2554–2558, 1982.

[121] Szabolcs Horvát, Éva Czabarka, and Zoltán Toroczkai. Reducing Degeneracy in Maximum Entropy Models of Networks. *Physical Review Letters*, 114(15): 158701, 2015.

[122] M Huxham, S Beaney, and D Raffaelli. Do parasites reduce the chances of triangulation in a real food web? *Oikos*, 76:284–300, 1996.

[123] S Itzkovitz, R Milo, N Kashtan, G Ziv, and U Alon. Subgraphs in random networks. *Physical Review E*, 68(2):026127, 2003.

[124] U Jacob, A Thierry, U Brose, W E Arntz, S Berg, T Brey, I Fetzer, T Jonsson, K Mintenbeck, C Möllmann, O L Petchey, J O Riede, and J A Dunne. The role of body size in complex food webs. *Advances in Ecological Research*, 45: 181–223, 2011.

[125] Hans-Karl Janssen, Martin Müller, and Olaf Stenull. Generalized epidemic process and tricritical dynamic percolation. *Physical Review E*, 70:026114, 2004.

[126] E T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[127] E T Jaynes. Information theory and statistical mechanics. II. *Physical Review*, 108:171–190, 1957.

[128] Samuel Johnson and Nick S Jones. Looplessness in networks is linked to trophic coherence. *Proceedings of the National Academy of Sciences*, 114(22):5618–5623, 2017.

[129] Samuel Johnson, Joaquín J Torres, J Marro, and Miguel A Muñoz. Entropic Origin of Disassortativity in Complex Networks. *Physical Review Letters*, 104 (10):108702, 2010.

[130] Samuel Johnson, J Marro, and Joaquín J Torres. Robust Short-Term Memory without Synaptic Learning. *PLOS ONE*, 8(1):e50276, 2013.

[131] Samuel Johnson, Virginia Domínguez-García, Luca Donetti, and Miguel A Muñoz. Trophic coherence determines food-web stability. *Proceedings of the National Academy of Sciences*, 111(50):17923–17928, 2014.

[132] Brian Karrer and M E J Newman. Message passing approach for general epidemic models. *Physical Review E*, 82(1):016101, 2010.

[133] Brian Karrer and M E J Newman. Random graphs containing arbitrary distributions of subgraphs. *Physical Review E*, 82(6):066118, 2010.

[134] Brian Karrer and M E J Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[135] Brian Karrer, M E J Newman, and Lenka Zdeborová. Percolation on Sparse Networks. *Physical Review Letters*, 113(20):208702, 2014.

[136] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Mfinder tool guide. *Department of Molecular Cell Biology and Computer Science and Applied Math., Weizmann Inst. of Science, Rehovot Israel, technical report*, 2002.

[137] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[138] Eben Kenah and James M Robins. Second look at the spread of epidemics on networks. *Physical Review E*, 76:036113, 2007.

[139] W O Kermack and A G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.

[140] Hyunju Kim, Charo I Del Genio, Kevin E Bassler, and Zoltán Toroczkai. Constructing and sampling directed graphs with given degree sequences. *New Journal of Physics*, 14(2):023012, 2012.

[141] István Z Kiss, Joel C Miller, and Péter L Simon. *Mathematics of Epidemics on Networks*. Springer, 2017.

[142] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[143] Arun S Konagurthu and Arthur M Lesk. On the origin of distribution patterns of motifs in biological networks. *BMC Systems Biology*, 2:73, 2008.

[144] P L Krapivsky and S Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

[145] P L Krapivsky, S Redner, and F Leyvraz. Connectivity of Growing Random Networks. *Physical Review Letters*, 85(21):4629–4632, 2000.

[146] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *Proceedings of the First Workshop on Online Social Networks*, pages 19–24. ACM, 2008.

[147] Ivan Kryven. Emergence of the giant weak component in directed random graphs with arbitrary degree distributions. *Physical Review E*, 94(1):012315, 2016.

[148] Ivan Kryven. Finite connected components in infinite directed and multiplex networks with arbitrary degree distributions. *Physical Review E*, 96:052304, 2017.

[149] Ivan Kryven. General expression for the component size distribution in infinite configuration networks. *Physical Review E*, 95:052303, 2017.

[150] K D Lafferty, R F Hechinger, J C Shaw, K L Whitney, and A M Kuris. Food webs and parasites in a salt marsh ecosystem. In Sharon K. Collinge and Chris Ray, editors, *Disease ecology: Community structure and pathogen dynamics*, pages 119–134. 2006.

[151] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87:198701, 2001.

[152] Vito Latora and Massimo Marchiori. Is the Boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1):109–113, 2002. Horizons in Complex Systems.

[153] David Anthony Lavis. *Equilibrium statistical mechanics of lattice models*. Springer, 2015.

[154] Stephen Levine. Several measures of trophic structure applicable to complex food webs. *Journal of Theoretical Biology*, 83(2):195–207, 1980.

[155] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4):330–342, 2008.

[156] Shuguang Li, Jianping Yuan, Yong Shi, and Juan Cristóbal Zagal. Growing scale-free networks with tunable distributions of triad motifs. *Physica A: Statistical Mechanics and its Applications*, 428:103–110, 2015.

[157] J Link. Does food web theory work for marine ecosystems? *Marine Ecology Progress Series*, 230:1–9, 2002.

[158] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.

[159] A L Lloyd. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1470):985–993, 2001.

[160] A L Lloyd. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theoretical Population Biology*, 60(1):59–71, 2001.

[161] L Lovász. Random walks on graphs: A survey. In D Miklós, V T Sós, and T Szőnyi, editors, *Combinatorics, Paul Erdős is Eighty*, volume 2, pages 353–398. János Bolyai Mathematical Society, Budapest, 1993.

[162] Matteo Magnani and Luca Rossi. *Formation of Multiple Networks*, pages 257–264. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[163] N D Martinez. Artifacts or attributes? Effects of resolution on the Little Rock Lake food web. *Ecological Monographs*, 61:367–392, 1991.

[164] Neo D Martinez, Bradford A Hawkins, Hassan Ali Dawah, and Brian P Feifarek. Effects of sampling effort on characterization of food-web structure. *Ecology*, 80:1044–1055, 1999.

[165] D M Mason. Quantifying the impact of exotic invertebrate invaders on food web structure and function in the great lakes: A network analysis approach. *Interim Progress Report to the Great Lakes Fisheries Commission- yr 1*, 2003.

[166] Naoki Masuda, Mason A Porter, and Renaud Lambiotte. Random walks and diffusion on networks. *Physics Reports*, 716-717(Supplement C):1–58, 2017. Random walks and diffusion on networks.

[167] Robert McCredie May. *Stability and complexity in model ecosystems*, volume 6. Princeton University Press, 1973.

[168] J Memmott, N D Martinez, and J E Cohen. Predators, parasitoids and pathogens: species richness, trophic generality and body sizes in a natural food web. *Journal of Animal Ecology*, 69:1–15, 2000.

[169] Lauren Ancel Meyers, M E J. Newman, and Babak Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240 (3):400–418, 2006.

[170] Joel C. Miller. Equivalence of several generalized percolation models on networks. *Physical Review E*, 94:032313, 2016.

[171] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594): 824–827, 2002.

[172] R Milo, N Kashtan, S Itzkovitz, M E J Newman, and U Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv:0312028 [cond-mat]*, 2003.

[173] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[174] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[175] Michael Molloy and Bruce Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7(3):295–305, 1998.

[176] Mark E Monaco and Robert E Ulanowicz. Comparative ecosystem trophic structure of three U.S mid-atlantic estuaries. *Marine Ecology Progress Series*, 161:239–254, 1997.

[177] Angelo B Monteiro and Lucas Del Bianco Faria. The interplay between population stability and food-web topology predicts the occurrence of motifs in complex food-webs. *Journal of Theoretical Biology*, 409:165–171, 2016.

[178] Tamas Nepusz and Tamas Vicsek. Controlling edge dynamics in complex networks. *Nature Physics*, 8(7):568–573, 2012.

[179] M E J Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64:016132, 2001.

[180] M E J Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[181] M E J Newman. Spread of epidemic disease on networks. *Physical Review E*, 66:016128, 2002.

[182] M E J Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25(1):83–95, 2003.

[183] M E J Newman, S H Strogatz, and D J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[184] M E J Newman, D J Watts, and S H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1): 2566–2572, 2002.

[185] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[186] S Opitz. Trophic interactions in Caribbean coral reefs. *ICLARM Tech. Rep.*, 43:341, 1996.

[187] Tore Opsahl, Vittoria Colizza, Pietro Panzarasa, and José J Ramasco. Prominence and control: The weighted rich-club effect. *Physical Review Letters*, 101: 168702, 2008.

[188] David J P O'Sullivan, Gary James O'Keeffe, Peter G Fennell, and James P Gleeson. Mathematical modeling of complex contagion on clustered networks. *Frontiers in Physics*, 3:71, 2015.

[189] Robert T Paine. Food web complexity and species diversity. *The American Naturalist*, 100(910):65–75, 1966.

[190] Juyong Park and M E J Newman. Statistical mechanics of networks. *Physical Review E*, 70(6):066117, 2004.

[191] Juyong Park and M E J Newman. Solution for the properties of a clustered network. *Physical Review E*, 72(2):026136, 2005.

[192] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, 2001.

[193] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63:066117, 2001.

[194] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2004.

[195] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, 2015.

[196] Pavel V Paulau, Christoph Feenders, and Bernd Blasius. Motif analysis in directed ordered networks and applications to food webs. *Scientific Reports*, 5:11926, 2015.

[197] Tiago P Peixoto. Entropy of stochastic blockmodel ensembles. *Physical Review E*, 85(5):056122, 2012.

[198] Tiago P Peixoto. The graph-tool python library. *figshare*, 2014.

[199] Mathew Penrose. *Random geometric graphs*. Number 5 in Oxford Studies in Probability. Oxford University Press, 2003.

[200] Scott W Phillips. *Synthesis of US Geological Survey science for the Chesapeake Bay ecosystem and implications for environmental management*, volume 1316. Geological Survey (USGS), 2007.

[201] Stuart L Pimm. *Food Webs*. Springer Netherlands, 1982.

[202] Stuart L Pimm, John H Lawton, and Joel E Cohen. Food web patterns and their consequences. *Nature*, 350(6320):669–674, 1991.

[203] G Polis. Complex trophic interactions in deserts: an empirical critique of food-web theory. *Am. Nat.*, 138:123–125, 1991.

[204] Gary A Polis. Complex trophic interactions in deserts: An empirical critique of food-web theory. *The American Naturalist*, 138(1):123–155, 1991.

[205] Mason A Porter and James P Gleeson. *Dynamical systems on networks: A tutorial*, volume 4. Springer, 2016.

[206] Robert J Prill, Pablo A Iglesias, and Andre Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLOS Biology*, 3(11):e343, 2005.

[207] Filippo Radicchi and Claudio Castellano. Beyond the locally treelike approximation for percolation on real networks. *Physical Review E*, 93(3):030302, 2016.

[208] A Ramachandra Rao, Rabindranath Jana, and Suraj Bandyopadhyay. A markov chain monte carlo method for generating random (0, 1)-matrices with given marginals. *Sankhyā: The Indian Journal of Statistics, Series A*, 58(2): 225–242, 1996.

[209] J O Riede, U Brose, B Ebenman, U Jacob, R Thompson, C Townsend, and T Jonsson. Stepping in Elton's footprints: a general scaling model for body masses and trophic levels across ecosystems. *Ecology Letters*, 14:169–178, 2011.

[210] Martin Ritchie, Luc Berthouze, Thomas House, and Istvan Z Kiss. Higher-order structure and epidemic dynamics in clustered networks. *Journal of Theoretical Biology*, 348:21–32, 2014.

[211] Martin Ritchie, Luc Berthouze, and Istvan Z Kiss. Beyond clustering: mean-field dynamics on networks with arbitrary subgraph composition. *Journal of Mathematical Biology*, 72(1):255–281, 2016.

[212] Martin Ritchie, Luc Berthouze, and Istvan Z Kiss. Generation and analysis of networks with a prescribed degree sequence and subgraph family: higher-order structure matters. *Journal of Complex Networks*, 5(1):1–31, 2017.

[213] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173– 91, 2007.

[214] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2):192–215, 2007.

[215] Martí Rosas-Casals, Sergi Valverde, and Ricard V Solé. Topological vulnerability of the european power grid under errors and attacks. *International Journal of Bifurcation and Chaos*, 17(07):2465–2475, 2007.

[216] Sheldon M Ross. *Introduction to Probability Models*. Academic Press, 2014.

[217] Thomas Schank and Dorothea Wagner. Approximating clustering coefficient and transitivity. *Journal of Graph Algorithms and Applications*, 9(2):265–275, 2005.

[218] N Schwartz, R Cohen, D ben Avraham, A-L Barabási, and S Havlin. Percolation in directed scale-free networks. *Physical Review E*, 66(1):015104, 2002.

[219] M Ángeles Serrano and Marián Boguñá. Clustering in complex networks. I. General formalism. *Physical Review E*, 74(5):056114, 2006.

[220] M Ángeles Serrano and Marián Boguñá. Percolation and Epidemic Thresholds in Clustered Networks. *Physical Review Letters*, 97(8):088701, 2006.

[221] James Sethna. *Statistical Mechanics: Entropy, Order Parameters, and Complexity*, volume 14 of *Oxford Master Series in Statistical, Computational, and Theoretical Physics*. Oxford University Press, 2006.

[222] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68, 2002.

[223] Ricard V Solé, Martí Rosas-Casals, Bernat Corominas-Murtra, and Sergi Valverde. Robustness of the european power grids under intentional attack. *Physical Review E*, 77:026102, 2008.

[224] Tiziano Squartini and Diego Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13(8): 083001, 2011.

[225] Tiziano Squartini, Joey de Mol, Frank den Hollander, and Diego Garlaschelli. Breaking of Ensemble Equivalence in Networks. *Physical Review Letters*, 115 (26):268701, 2015.

[226] H E Stanley. *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press, 1987.

[227] D B Stouffer, J Camacho, R Guimerà, C A Ng, and L A Nunes Amaral. Quantitative patterns in the structure of model and empirical food webs. *Ecology*, 86(5):1301–1311, 2005.

[228] Daniel B Stouffer, Juan Camacho, and Luís A Nunes Amaral. A robust measure of food web intervality. *Proceedings of the National Academy of Sciences*, 103 (50):19015–19020, 2006.

[229] Daniel B Stouffer, Juan Camacho, Wenxin Jiang, and Luís A Nunes Amaral. Evidence for the existence of a robust pattern of prey selection in food webs. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1621): 1931–1940, 2007.

[230] David Strauss. On a general class of models for interaction. *SIAM review*, 28 (4):513–527, 1986.

[231] R M Thompson and C R Townsend. Impacts on stream food webs of native and exotic forest: An intercontinental comparison. *Ecology*, 84:145–161, 2003.

[232] R M Thompson and C R Townsend. Energy availability, spatial heterogeneity and ecosystem size predict food-web structure in stream. *Oikos*, 108:137–148, 2005.

[233] G Timár, A V Goltsev, S N Dorogovtsev, and J F F Mendes. Mapping the Structure of Directed Networks: Beyond the Bow-Tie Diagram. *Physical Review Letters*, 118(7):078301, 2017.

[234] G Timár, R A da Costa, S N Dorogovtsev, and J F F Mendes. Nonbacktracking expansion of finite graphs. *Physical Review E*, 95(4):042322, 2017.

[235] Townsend, Thompson, McIntosh, Kilroy, Edwards, and Scarsbrook. Disturbance, resource supply, and food-web architecture in streams. *Ecology Letters*, 1(3):200–209, 1998.

[236] Robert Ulanowicz, C Bondavalli, and M S Egnotovich. Network Analysis of Trophic Dynamics in South Florida Ecosystems, FY 96: The Cypress Wetland Ecosystem. 1997.

[237] Robert Ulanowicz, C Bondavalli, and M S Egnotovich. Network Analysis of Trophic Dynamics in South Florida Ecosystems, FY 97: The Florida Bay Ecosystem. 1998.

[238] Robert Ulanowicz, Johanna Heymans, and M S Egnotovich. Network Analysis of Trophic Dynamics in South Florida Ecosystems, FY 99: The Graminoid Ecosystem. 2000.

[239] Robert E Ulanowicz. Identifying the structure of cycling in ecosystems. *Mathematical Biosciences*, 65(2):219–237, 1983.

[240] Robert E Ulanowicz and Daniel Baird. Nutrient controls on ecosystem dynamics: the Chesapeake mesohaline community. *Journal of Marine Systems*, 19(1–3):159–172, 1999.

[241] S van Dongen and A J Enright. Metric distances derived from cosine similarity and Pearson and Spearman correlations. *arXiv:1208.3145 [stat]*, 2012.

[242] N G Van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, 1981.

[243] A Vázquez, R Dobrin, D Sergi, J-P Eckmann, Z N Oltvai, and A-L Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proceedings of the National Academy of Sciences*, 101(52):17940–17945, 2004.

[244] R B Waide and D P Reagan. *The Food Web of a Tropical Rainforest*. University of Chicago Press, Chicago, 1996.

[245] P H Warren. Spatial and temporal variation in the structure of a freshwater food web. *Oikos*, 55:299–311, 1989.

[246] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, 1994.

[247] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[248] Duncan J Watts and Steven H Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440, 1998.

[249] Herbert S Wilf. *Generatingfunctionology*. A K Peters, Ltd., 2006.

[250] Richard J Williams and Neo D Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–183, 2000.

[251] Peter Yodzis. Local trophodynamics and the interaction of marine mammals and fisheries in the Benguela ecosystem. *Journal of Animal Ecology*, 67(4): 635–658, 1998.

[252] Soon-Hyung Yook, Hawoong Jeong, and Albert-László Barabási. Modeling the Internet's large-scale topology. *Proceedings of the National Academy of Sciences*, 99(21):13382–13386, 2002.

[253] Jean-Gabriel Young, Giovanni Petri, Francesco Vaccarino, and Alice Patania. Construction of and efficient sampling from the simplicial configuration model. *Physical Review E*, 96:032312, 2017.

[254] Haiyuan Yu, Pascal Braun, Muhammed A Yıldırım, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-François Rual, Amélie Dricot, Alexei Vazquez, Ryan R. Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E Hudson, Juyong Park, Xiaofeng Xin, Michael E Cusick, Troy Moore, Charlie Boone, Michael Snyder, Frederick P Roth, Albert-László Barabási, Jan Tavernier, David E Hill, and Marc Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[255] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

[256] V Zlatić, D Garlaschelli, and G Caldarelli. Networks with arbitrary edge multiplicities. *EPL (Europhysics Letters)*, 97(2):28005, 2012.

[257] Konstantin Zuev, Or Eisenberg, and Dmitri Krioukov. Exponential random simplicial complexes. *Journal of Physics A: Mathematical and Theoretical*, 48 (46):465002, 2015.