Statistical Mechanics of Simplicial Complexes

Owen Courtney

A thesis submitted in partial fulfillment of the requirements of the Degree of

Doctor of Philosophy

2019

Declaration

I, Owen Courtney, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Date:

Details of collaboration and publications: parts of this work have been completed in collaboration with Ginestra Bianconi, and are published in the following papers:

- "Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes", *Physical Review E*, vol. 93, no. 6, 2016.
- "Weighted Growing Simplicial Complexes", *Physical Review E*, vol. 95, no. 6, 2017.

- "Dense Power-law Networks and Simplicial Complexes", *Physical Review E*, vol. 97, no. 5, 2018.

Abstract

Simplicial complexes are a generalization of networks that can encode *many-body* interactions between more than two nodes. Whereas networks represent the interactions between the parts of a complex system using nodes and links, the *simplices* in a simplicial complex represent interactions between any number of nodes.

In a number of applications these many-body interactions have been shown to carry important information about the complex system. Furthermore, the representation of systems as simplicial complexes has made it possible to characterise their structure in ways not possible with network representations, including using new tools inspired by algebraic topology or geometry.

However, the use of simplicial complexes as a network science tool is still new and there remains an urgent need for a theoretical framework that can allow us to interpret the highly complex, high-dimensional data associated with real simplicial complexes. How do the new measures of structure that are being introduced relate to each other? And what can they tell us about the evolution or function of the simplicial complex? How can we best model simplicial complexes based on incomplete information? And what constitutes 'interesting' or 'significant' structure?

In this thesis we tackle these problems through the development of stochastic models of simplicial complexes that can help us to disentangle the interactions, dependences and correlations between the structural

properties of simplicial complexes and their evolution.

First, we propose two maximum entropy models of a simplicial complex with given *generalised degrees of the nodes* (the number of simplices of a given dimension that a node participates in). These models have a clear use as null models for simplicial complexes as they are the most statistically appropriate models for simplicial complexes given knowledge of the generalized degrees. Importantly, they allow for a statistically rigorous understanding of the implications of particular choices of the generalized degrees for the structure of simplicial complexes and dynamics taking place upon them.

Second, we propose a model of a simplicial complex that is weighted and growing. This model follows in the tradition of growing network models that seek to characterize the relations between simple mechanisms of growth (of the network) and reinforcement (of the weights) and their structural properties. The model exhibits a very rich variety of topologies and weight distributions for different values of the model parameters and different dimensions of simplices. Remarkably each of these distributions and scalings could be exhibited simultaneously within a single simplicial complex for faces of different dimension. The model shows that simple, plausible mechanisms of growth and reinforcement in simplicial complexes can produce a broad range of topologies and distributions, and shows the important role that dimension plays in determining these properties.

Third, we propose a modelling framework for producing weighted networks and simplicial complexes which are both dense and scale-free. The growth mechanisms of the models contained within our framework are analogous to the Pitman-Yor process, a 'ball-in-the-box' process well-known among probability theorists for producing power-laws with exponent $\gamma \in (1, 2]$. Our framework demonstrates the difficulty of producing

a simple network which is both dense and scale-free. By relaxing the requirement for the network to be simple, either by a direction to the link or by reinterpreting the weight of a link as the number of multilinks between two nodes, we found that it was easy to create scale-free networks and simplicial complexes with tunable dense exponent $\gamma \in (1, 2]$.

Acknowledgements

First and foremost I would like to thank Ginestra for her support, guidance and encouragement during my time as a student. Her tireless enthusiasm for research has been a great inspiration to me, and her generosity towards me when I have needed help has been invaluable.

I'd also like to thank my parents, not just for the practical and emotional support they have given me throughout my PhD studies, but also for the ever present love and encouragement they have always provided throughout my life.

I want to thank my girlfriend Amelia for her patience and support, and for providing light in the harder times.

Finally, I also want to thank the friends I have made in my time at Queen Mary. You have made the PhD experience so much more fun. The challenge of undertaking a PhD would be so much harder without the community you provide.

Contents

Chapter 1

Introduction

The representation of complex systems as networks of interacting parts has provided insight into the workings of systems as diverse as on-line social networks, global trade networks, protein interaction networks, the brain and the financial system [1–6]. The nodes of a network represent the parts of a system and the links represent some kind of interaction between the parts. Networks thus encode the 'interaction structure' underlying real systems, and in network science the aim is to gain useful insight about these systems by studying their underlying network structures.

However the use of networks relies on the simplifying assumption that all of the information crucial to understanding how the parts of a system interact is encodable in a set of pairwise relations between them (i.e. as links between pairs of nodes). This assumption fails to recognise that for some systems, interactions can intrinsically involve more than two nodes. For example in the case of the interactions between proteins in a cell [5], the proteins interact in *complexes* composed of multiple proteins, which together have some function in the cell. It is possible that three proteins could interact together in a single complex involving all three of them, or they could also interact together in three separate complexes, each of which includes only two of them. The *network representation* of these two cases would be the same: three nodes connected into a triangle by three links. Networks thus map higher dimensional interactions that intrinsically involve more than two nodes to cliques, and so a large number of distinct higher dimensional structures can be mapped to

a single network. While for some applications this simplification may be appropriate, for others the ability to distinguish interactions of different dimension may be important.

Simplicial complexes [7–9] are a generalization of networks that are able to encode these many-body interactions between more than two nodes. Whereas networks are constructed from nodes and links, simplicial complexes are constructed from something called *simplices*. These simplices represent groups of nodes for which there is an interaction that intrinsically involves all of the nodes. Nodes and links are simplices of dimension 0 and 1 respectively, while triangles are simplices of dimension 2 and tetrahedra are dimension 3. In general d-dimensional simplices represent interactions involving $d+1$ nodes, and for $d > 3$ can be thought of as higher dimensional generalisations of tetrahedra.

Simplicial complexes thus extend the descriptive power of networks by distinguishing between interactions involving different numbers of nodes. In the protein interaction example, the three proteins interacting in a single complex would be represented by a single triangle simplex, while in the case of the three 'pairwise' interactions between each pair of nodes, the interactions would be represented by three links but no 2-dimensional simplex (this can be thought of as a triangular 'hole' in the network in contrast to the 'filled' triangle in the first case). Figure 1.1 illustrates this idea: panels A-F show a number of different simplicial complexes which all correspond to the same 'skeleton' network shown in panel A. Note that the network shown in panel A is itself a simplicial complex composed of simplices of dimension 0 and 1 only.

This extra descriptive power can be helpful in a variety of ways. On a simple level, modelling with simplicial complexes rather than networks could help us to understand the evolution of systems for which group membership may influence the dynamics, such as academic collaboration [10–12] amongst researchers or social networks [13–15].

On the other hand for some real world systems the *absence* of an interaction

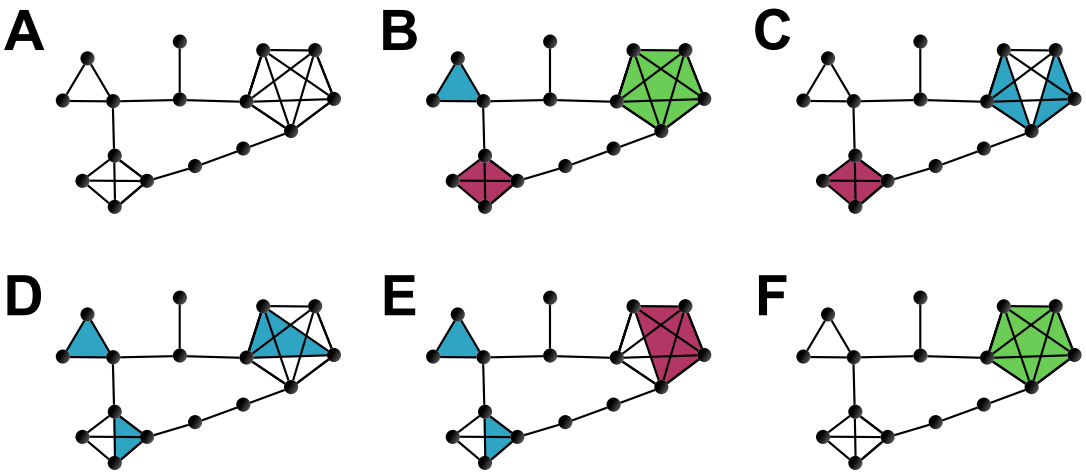

Figure 1.1: Figure showing a variety of simplicial complexes which all correspond to the same 'skeleton' network. Panel A is a 1-dimensional simplicial complex (a network) composed of 0-simplices (nodes) and 1-simplices (links). Panels B-F show higher dimensional simplicial complexes. The coloured shapes indicate simplices of different dimensions: blue indicates a 2-simplex (triangle), purple indicates a 3-simplex (tetrhadra) and green represents a 4-simplex. Many more simplicial complexes could be constructed corresponding to the network shown in panel A.

between a group of nodes may be just as important as the presence of one. Recently there has been much interest in the homological properties of complex systems [13, 16–18], which characterises the shape of data coming from the systems in terms of its cavities at different dimension. In [17] the authors examined simplicial complexes constructed from the fMRI data of patients who had either been given psilocybin (the pschoactive element of magic mushrooms) or a placebo. The two groups were distinguishable from each other by the homological structure of their simplicial complexes. Another study [18] found that the responses to different stimuli in a simulated neocortical circuit (a small region of a brain) could be distinguished from the number of simplices at different dimensions and the Betti numbers. Similar tools have also been applied to study social, infrastructure and biological networks in [13] resulting in the identification of new classifications for real networks not apparent using traditional network tools.

Simplicial complexes also have a use in the uncovering of hidden geometries. In [19] it was found that the shape of a maze being explored by a mouse could be

recovered by examining neuron coactivation data in the mouse's brain. A simplicial complex was constructed from the data on coactivation of a number of neurons in the mouse's brain that were known to be involved in the navigation of its surroundings. The topology of this simplicial complex was found to accurately reflect the topology of the maze. Meanwhile in the unrelated area of quantum gravity models of simplicial complexes have been developed with the aim of showing that discrete combinatorial objects can produce an 'emergent' geometry [20–23]. These examples demonstrate the relevance of simplicial complexes to another area of network science that is currently receiving a lot of attention, namely networks with an underlying geometry [24, 25].

As can be seen from the examples mentioned above, simplicial complexes are a useful tool for modelling many real systems. In comparison to networks they are able to capture an even richer variety of highly complex structures in real systems, and examining this structure has already enabled classification and distinction of systems and their phenomena that has previously not been possible using the traditional tools of network science.

However the use of simplicial complexes as tools for network science is still new, and there is a need for basic research to make such high-dimensional, high-complexity data interpretable. Firstly, there is a need for new measures of the structure of simplicial complexes that are relevant to the real systems being modelled. Then there is a need to understand how these new measures of structure relate together, and what they mean for the evolution or function of the systems they describe.

To do this we need simple models of simplicial complexes that can act as benchmarks or as null models to which real simplicial complexes can be compared or that function as toy models that help reveal the connections between simple rules for generating simplicial complexes and their resulting structure. By comparing real simplicial complexes to these models we can answer questions such as how might the simplicial complex have evolved? Does it have 'significant' higher level structure that does not have an explanation in terms of more simple aspects of their structure

(e.g. do higher level 'homological' properties follow naturally from simple 'connectivity' properties like the degree)? And how can we best model a simplicial complex based on incomplete data?

Recently models of simplicial complexes have been proposed that emulate their evolution [11, 26], that establish relations between their stochastic construction and their homology [27, 28] or other topological features [29–31], that have an emergent geometry [30], or that can act as simple null models [8, 9, 27, 32]. In this thesis we present a number of models of simplicial complexes that add to this combined literature and help to build a picture of how the structural properties of simplicial complexes relate to each other, how they emerge from simple construction rules, and how we can model simplicial complexes in a statistically rigorous way.

The structure of this thesis is detailed below.

In Chapter 2 we define simplicial complexes and a few simple measures of their structure. The purpose of this chapter is mainly to provide the basic concepts necessary for understanding the rest of the thesis, although we do present some simple results which have a relevance throughout the thesis.

Chapter 3 provides some background on models of networks and simplicial complexes. These models have a particular relevance to our own models of simplicial complexes which we present in Chapters 4, 5 and 6. Contained within this chapter is also a brief introduction to entropy maximisation, which is relevant in Chapter 4.

Chapter 4 contains research first published in our paper *Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes* [33]. In this chapter we present two maximum entropy models of simplicial complexes, which generalise the network configuration model and canonical ensemble (or soft configuration model). The primary purpose of these models is to act as null models for simplicial complexes with given generalised degrees of the nodes (definition in Chapter 2). We explore these models both analytically and numerically via algorithms we develop to sample simplicial complexes from the models.

We derive some important quantities associated with the models in terms of the generalised degrees, including the model entropies.

These *maximum entropy ensembles* have a clear use as null models as they are the most appropriate way to model simplicial complexes given knowledge of their generalized degrees. Furthermore they allow for a statistically rigorous understanding of the implications of particular choices of the generalized degrees on other aspects of the structure of simplicial complexes. We believe that the models constitute a first step in modelling simplicial complexes with equilibrium statistical mechanics tools and that this work will open up new perspectives for investigating a new generation of maximum entropy models of simplicial complexes.

Chapter 5 contains research first published in our paper *Weighted Growing Simplicial Complexes* [34]. In it we present a model of a simplicial complex that is weighted and growing. This model follows in the tradition of growing network models that seek to characterize the relations between simple growth mechanisms of networks and their structural properties. The model evolves stochastically via the addition of new simplices and the 'reinforcement' of existing simplices (increasing their weight), and our interest is in the kinds of topologies that can be produced and the distribution of weight over these topologies.

These 'growth' and 'reinforcement' dynamics differ in a significant way from those in network models. They act on simplices of dimension $d-1$ and d respectively, with probabilities that depend on properties belonging to those simplices. Thus the model dynamics is not limited to depending only on 'node properties' but instead considers the properties of groups of nodes represented by simplices. We study the model analytically, finding that relatively simple growth mechanisms for generating simplicial complexes can give rise to a rich variety of weight distributions and topologies. These mechanisms could provide a plausible explanation for the origins of such distributions in real simplicial complexes.

Chapter 6 contains research first published in our paper *Dense Power-law Networks and Simplicial Complexes* [35]. In it we present a modelling framework for

producing networks and simplicial complexes which are *dense* (have an average degree or generalised degree that grows with the system size) and have power-law distributions of the degrees or generalised degrees. The growth mechanisms of the models contained within our framework are analogous to the Pitman-Yor process, a stochastic process well-known among probability theorists for generating random partitions with power-law distributions of block sizes.

Our models address the question of to what extent it is possible to produce networks that are both dense and power-law. On the one hand there are numerous examples in the literature of real networks that appear to be dense and power-law. On the other hand in [36] it has been suggested that such degree distributions are *ungraphical*, i.e. no network can be produced with such a degree distribution without adding links that connect a node to itself or add more than one link between two nodes. Our models show that its is in fact possible to generate dense networks that for the most part appear to be power-law.

Finally, in Chapter 7 we give our conclusions.

Chapter 2

Basic definitions

In this chapter we introduce the basic concepts and definitions that will be used throughout the thesis. The simplicial complexes that we use in this thesis should more precisely be called *abstract simplicial complexes*. These are combinatorial structures indicating interactions (or the lack of them) between collections of nodes. An abstract simplicial complex defined on a set of nodes $\mathcal{V}$ is a collection of subsets of $\mathcal{V}$ such that if some $\alpha \subset \mathcal{V}$ is in the collection then every subset $\alpha' \subset \alpha$ is also in the collection. These sets are called simplices, and can be visualised as geometric objects: A node is a zero dimensional simplex, having the dimension of a point, while links, triangles and tetrahedra are simplices of dimension 1, 2 and 3 respectively. In general we call a d-dimensional simplex a 'd-simplex' or sometimes a 'd-face' when the simplex is itself a face of some higher dimension simplex.

The requirement that subsets of simplices must be simplices means that the faces of a simplex must always be present as simplices in a simplicial complex. For example if a 3-simplex on the nodes $\{1, 2, 3, 4\}$ exists in a simplicial complex then its lower dimensional faces $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{1, 3, 4\}$, $\{2, 3, 4\}$, $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, $\{3, 4\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$ must all also be included in the simplicial complex as well. This distinguishes simplicial complexes from the related concept of hypergraphs in which there is no such restriction.

A *d-dimensional simplicial complex* is one consisting of simplices of dimension

less than or equal to d with at least one simplex actually having dimension d. A *pure d-dimensional simplicial complex* is one constructed exclusively from d-simplices and their sub-simplices, i.e. every simplex of dimension less than d must be a face of a d-simplex, there are no 'independent' simplices of dimension less than d. Figure 2.1 shows examples of a pure 1-dimensional simplicial complex (a network) in panel A, and a pure 2-dimensional simplicial complex constructed from in panel B. In contrast, in figure 1.1 of the introduction the only pure simplicial complex is panel A. The structure of a pure d-dimensional simplicial complex is determined totally by its d-simplices, which determine whether or not simplices of lower dimensions are present in the simplicial complex. As such pure simplicial complexes are easier to work with both analytically and numerically in comparison to impure simplicial complexes. For this reason, the models of simplicial complexes that we explore in this thesis the simplicial complexes are always pure.

In network science it is typical to represent networks using *adjacency matrices* $\{a_{ij}\}_{i,j=1,\dots N}$ (where N is the total number of nodes in the network) with entries $a_{ij} = 1$ if nodes i and j are connected by a link and $a_{ij} = 0$ otherwise. We define an *adjacency tensor* of a simplicial complex similarly: Let $\mathcal{Q}_\delta(N)$ be the set of all possible simplices of dimension equal to δ in a d-dimensional simplicial complex with N nodes. The adjacency tensor $\{a_\alpha\}_{\alpha\in\mathcal{Q}_\delta(N)}$ has entries $a_\alpha = 1$ if the δ-simplex α is present or $a_\alpha = 0$ if it is not. We write the set of all δ-simplices actually present in the simplicial complex as $\mathcal{S}_\delta(N) = \{\alpha \in \mathcal{Q}_\delta(N) | a_\alpha = 1\}$. If the simplicial complex is also pure, then the adjacency tensors $\{a_\alpha\}_{\alpha\in\mathcal{Q}_\delta(N)}$ and sets $\mathcal{S}_\delta(N)$ for simplices of dimension $\delta < d$ are completely determined by $\{a_\alpha\}_{\alpha\in\mathcal{Q}_d(N)}$, as in this case a δ-simplex may only exist if it is a face of some d-simplex. A pure d-dimensional simplicial complex is thus entirely defined by its adjacency tensor $\{a_\alpha\}_{\alpha\in\mathcal{Q}_d(N)}$.

In the next sections we define some simple measures of the local structure of a simplicial complex: the generalised degrees and generalised strengths of the simplices.

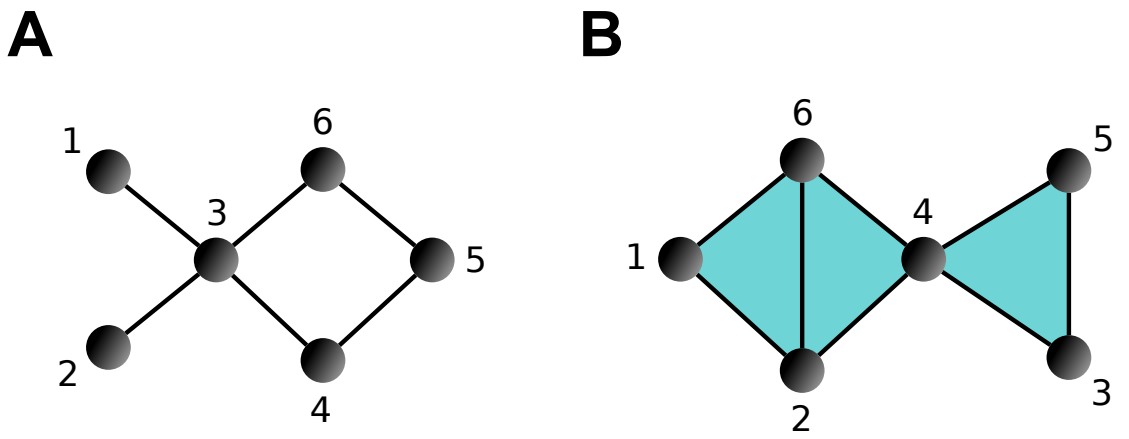

Figure 2.1: (Color online) Examples of simplicial complexes of dimension $d = 1$ (panel A) and $d = 2$ (panel B) are shown. Simplicial complexes of dimension $d = 1$ are simple networks. Simplicial complexes of dimension $d \geq 2$ characterize interactions occurring between more than two nodes, (specifically interactions occurring between $d + 1$ nodes).

2.1 Generalised degrees

The simplest measure of the local structure around a node in a network is its *degree*. The degree of a node is the total number of links it is a part of, and so is a simple measure of how connected it is to the rest of the network. In [31] the concept of degree was generalised to describe the connectivity of simplices in a simplicial complex. We use this definition of *generalised degree* extensively in the research presented in this thesis. Unlike the concept of degree in network science, it is not only nodes that have a generalised degree but simplices of higher dimension as well. The generalised degree $k_{d,\delta}(\alpha)$ of a δ-simplex α is the number of d-simplices that α is a face of, i.e.

$$k_{d,\delta}(\alpha) = \sum_{\alpha' \in \mathcal{Q}_d(N) | \alpha' \supseteq \alpha} a_{\alpha'}, \tag{2.1}$$

where $\mathcal{Q}_d(N)$ is the set of all possible d-simplices that could be constructed on N nodes, labelled 1 to N. The generalised degree $k_{d,\delta}(\alpha)$ therefore measures how connected the δ-simplex α is *at dimension* d. It should be noted that rather than saying

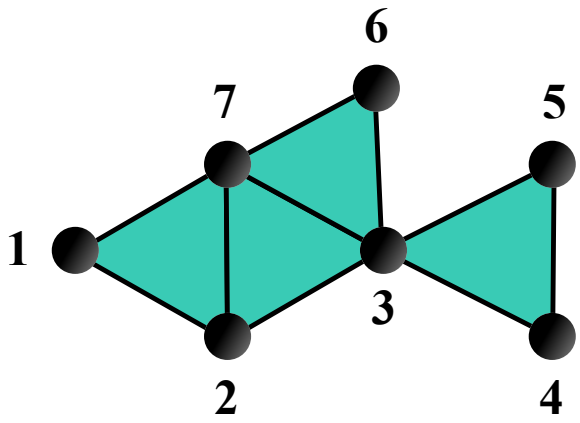

Figure 2.2: Figure showing a 2-dimensional simplicial complex composed of nodes, links and triangles. Node 7 has the generalised degrees $k_{2,0}(7) = 3$ and $k_{1,0}(7) = 4$ counting the number of triangles and links that 7 participates in respectively. In contrast node 3 has the same number of triangles as node 7, $k_{2,0}(3) = 3$ but a different number of links $k_{1,0}(7) = 5$. The generalised degree is also defined for the links in this simplicial complex. The link $(2, 7)$ participates in two triangles and so $k_{2,1}(2, 7) = 2$.

the generalised degree of α we should probably say *a* generalised degree of α as we can measure the connectivity of α at different dimensions by varying d (and in fact the generalised degree $k_{1,0}(i)$ of a 0-simplex i is simply the standard network degree of i). Figure 2.2 shows an example of the different generalised degrees that can be calculated for the simplices of a 2-dimensional simplicial complex.

The generalised degrees in a simplicial complex are defined for simplices of all dimensions and characterize the connectivity of each simplex at each higher dimension, and so therefore carry very detailed information about the connectivity structure of the simplicial complex in comparison to the information carried by the degrees of the nodes in a network. An obvious question is how do the generalised degrees of simplices of different dimension δ or measured at different dimension d relate to each other? When does fixing one fix the other, and can we find exact combinatorial expressions relating the quantities? If one generalised degree does not fix another, can we still make some kind of statistical argument about what values the other generalised degree could likely take, given our observation of the first? These questions provide a key part of the motivation for the research we present in Chapter 4, first

published in our paper *Generalized network structures: the configuration model and canonical ensemble of simplicial complexes* [33]. The bulk of our results from [33] are covered in Chapter 4, however we present here two simple but important results about the generalised degrees that have a broader importance in our research.

2.1.1 Combinatorial relation for the generalised degrees

For a δ-simplex α we found that its generalised degree $k_{d,\delta}(\alpha)$ may be related to the generalised degrees of the δ'-simplices of which α is a face with the following combinatorial expression:

$$k_{d,\delta}(\alpha) = \frac{1}{\begin{pmatrix} d-\delta \\ \delta'-\delta \end{pmatrix}} \sum_{\alpha' \in \mathcal{Q}_{\delta'}(N) | \alpha' \supseteq \alpha} k_{d,\delta'}(\alpha'). \tag{2.2}$$

The above equation is true for both pure and impure simplicial complexes. On the left hand side the generalised degree $k_{d,\delta}(\alpha)$ counts the number of d-simplices for which α is a face. On the right hand side in the sum, α' is a δ'-simplex of which α is a face and the generalised degree $k_{d,\delta'}(\alpha')$ counts the number of d-simplices incident to α'. Every d-simplex of which α' is a face, α must also be a face, as $\alpha \subset \alpha'$ and so it is the same d-simplices being counted on both sides of Eq. 2.2. The expression may be understood by considering how many times each of the d-simplices which are counted on the left side are counted on the right side. Each of these d-simplices contains $d+1$ nodes, and so therefore also contains $\binom{d+1}{\delta'+1}$ δ'-simplices, however not all of these δ'-simplices contain α. The total number of δ'-simplices in the d-simplex for which α is a face is instead $\binom{d-\delta}{\delta'-\delta}$, as we are asking how many distinct combinations of $\delta'-\delta$ nodes we can make from the $d-\delta$ nodes in the d-simplex that aren't also in α. Noting also that any δ'-simplex which is not a face of any d-simplex will contribute 0 to the sum on the right hand side of 2.2, we see that every d-simplex counted once on the left hand side will be counted exactly $\binom{d-\delta}{\delta'-\delta}$ times in the sum on the right hand side and so dividing the sum by this amount allows us to write our expression.

2.1.2 Structural cutoff for simplicial complexes

Another relatively simple result from our paper [33] is an expression for the structural cutoff for the generalised degrees of the nodes in a pure d-dimensional simplicial complex. In network science the structural cutoff [37] for the degrees of the nodes in a network is the maximum degree that a node can possess before the network must necessarily have degree correlations. We say a network is 'uncorrelated' when the degrees of the nodes at either end of a randomly chosen link are independent. This concept is elaborated on in detail in Chapter 4, but here we just say that this independence requirement on the degrees of nodes at opposite ends of a link imposes a restriction on the maximum degree that nodes in the network can have. Above this 'structural cutoff' it is impossible for the network to be uncorrelated.

The network structural cutoff K_1 is given by [37]

$$K_1 = (N\langle k\rangle)^{\frac{1}{2}} . \tag{2.3}$$

Networks where the maximum degree is less than K_1 can be correlated or uncorrelated while networks where the maximum degree is greater than K_1 have unavoidable *natural correlations*.

For *sparse* networks with a finite average degree that does not grow with number of nodes in the network the cutoff shown in Eq. (2.3) scales like $N^{\frac{1}{2}}$. Most real networks that have been studied are sparse, however *dense* networks with diverging average degree do exist [38–40] and are the subject of interest in Chapter 6 of this thesis. More specifically, in Chapter 6 we are interested in producing networks and simplicial complexes that are simultaneously dense and and *scale-free*, meaning that the degrees of the nodes follow a power-law distribution $P(k) \sim k^{-\gamma}$. As will be discussed in the next chapter and also in Chapter 6, in order for a network to be both dense and scale-free it must have a exponent in the range $\gamma \leq 2$. The structural cutoff in such a network has been studied in [39]. Unlike in the sparse case, for dense scale-free networks, imposing any kind of maximum degree k_{max} on the degree distribution will affect the average degree $\langle k\rangle$. In fact the average degree must scale like $\langle k\rangle \sim k_{max}^{2-\gamma}$ for $N \gg 1$. In [39] this fact was used in combination with

the equation for the structural cutoff that we have given here in Eq. (2.3) to obtain the scaling for the structural cutoff of a dense scale-free network which is $K_1 \sim N^{\frac{1}{\gamma}}$.

In [33] we derived a structural cutoff for pure d-dimensional simplicial complexes with given sequence of the generalised degrees of the nodes $\{k_{d,0}(i)\}_{i=1,\dots N}$. Each node i has a specified number of d-simplices that it takes part in, and we say that the simplicial complex is uncorrelated if the generalised degrees of the nodes of a randomly chosen d-simplex are independent of each other. Later in Chapter 4 we show that the maximum generalised degree K_d of the nodes in the simplicial complex for which this can be true is

$$K_d = \left[\frac{(\langle k \rangle N)^d}{d!} \right]^{1/(d+1)} . \tag{2.4}$$

The quantity K_d is the structural cutoff in a pure d-dimensional simplicial complex, above which natural correlations must exist in the simplicial complex.

Eq. (2.4) reduces to the network structural cutoff given in Eq. (2.3) for $d = 1$. Interestingly K_d scales like $(\langle k \rangle N)^{d/(d+1)}$, i.e. it is increasing with an exponent that is larger for larger dimensions d. For sparse networks the structural cutoff thus scales like $N^{d/(d+1)}$. For dense simplicial complexes with scale-free distribution of the generalised degrees the average generalised degree is the same as in the dense scale-free network, $\langle k \rangle \sim k_{max}^{2-\gamma}$. Setting $k_{max} = K_d$ yields the scaling $K_d \sim N^{\frac{d}{1-(1-\gamma)d}}$.

2.2 Generalised strengths

In some of the simplicial complexes considered in this thesis the simplices have an associated weight. These weights could represent intensities of interaction, or perhaps the number of instances of an interaction (e.g. the number of papers coauthored by a 'simplex' of academics).

Analogously to how we record the presence of simplices in a simplicial complex, we record the weights of the simplices in a weight tensor $\{w_\alpha\}_{\alpha \in \mathcal{Q}_d(N)}$. We can

then define the *generalised strength* $s_{d,\delta}(\alpha)$ of a δ-simplex in a way that mirrors our definition of the generalised degree:

$$s_{d,\delta}(\alpha) = \sum_{\alpha' \in \mathcal{Q}_d(N) | \alpha' \supseteq \alpha} a_{\alpha'} w_{\alpha'}. \tag{2.5}$$

The generalised strength above measures the total weight of the simplices of dimension d of which α is a face. Depending on the interpretation of the weights the inclusion of $a_{\alpha'}$ in the sum on the right hand side of Eq. 2.5 may not be necessary, for example if a non zero weight $w_{\alpha'} > 0$ implies $a_{\alpha'} = 1$ while $w_{\alpha'} = 0$ implies $a_{\alpha'} = 0$. Like the generalised degrees, the generalised strengths constitute a rich source of information about the structure of a simplicial complex. In order to interpret this information it would be useful to understand the basic relationships amongst the generalised strengths of different simplices at different dimension and also their relation with the generalised degrees.

In our paper *Weighted Growing Simplicial Complexes* [34] (presented in Chapter 5) we derive a combinatorial relation between the generalised strength $s_{d,\delta}(\alpha)$ of a δ-simplex α and the generalised strengths of the δ'-simplices of which α is a face. In particular we find

$$s_{d,\delta}(\alpha) = \frac{1}{\begin{pmatrix} d-\delta \\ \delta'-\delta \end{pmatrix}} \sum_{\alpha' \in Q_{\delta'} | \alpha' \supseteq \alpha} a_{\alpha'} s_{d,\delta'}(\alpha'). \tag{2.6}$$

The above expression is of course very similar to that given in Eq. 2.2 relating the generalised degrees, and is easily justified following the same logic as we used to justify Eq. 2.2.

The way that the generalised strengths relate to the generalised degrees conveys interesting information about the distribution of weight across the simplicial complex. In studies of real weighted networks [41–44] it has been found that many networks fall into one of two classes: class I networks in which the strengths of the

nodes scales linearly with the degrees:

$$s_i \propto k_i, \tag{2.7}$$

and class II networks in which the strength scales super-linearly with the degree:

$$s_i \propto k_i^{\theta} \quad \text{wth} \quad \theta > 1, \tag{2.8}$$

In the first of these classes, the weight must be distributed homogeneously across the links of the network, while in the second nodes with higher degree have links with higher average weight. We refer to these correlations as *weight-topology correlations*. Examples of class I networks include collaboration networks, while examples of networks in class II include the networks of flight routes between airports where the weights measure the number of passengers using each route.

In Chapter 3 we discuss as background a model first proposed in [43] that can produce weighted networks in either class I or class II depending on competing effects or *growth* and *reinforcement*. This model shows that the two distinct classes can be produced within a single framework based on simple growth mechansims. In Chapter 5 (reporting results of our paper [34]) we propose a model of a growing simplicial complex that generalises the network model of [43]. In our model the stochastic evolution depends on the generalised degrees and generalised strengths of the simplices, and we find that it is possible to produce linear and super-linear scalings of generalised strength with generalised degree as well as a third possible scaling: $s_{d,\delta}(\alpha) \propto e^{\beta k_{d,\delta}(\alpha)}$. Interestingly these different 'classes' can be observed within a single simplicial complex for faces of different dimension δ.

2.3 Skeleton networks

In the introduction to this thesis we highlighted the ability to distinguish many-body interactions from cavities as a key advantage of modelling real systems with simplicial complexes compared to networks. However, models of simplicial complexes can have a use even when we ignore this 'higher order' information and focus only on

the network structure that remains when we neglect all information about simplices with dimension greater than 1. We call these structures *skeleton networks* and they are formed from the 0-simplices (nodes) and 1-simplices (links) of a d-dimensional simplicial complex.

In this thesis we present a number of generative models of simplicial complexes which are designed to explore the relations amongst structural properties of the simplicial complexes such as the generalised degrees and generalised strengths, or to examine what effect growth processes depending on these properties have on the global structure of the simplicial complexes. On the one hand, studying the skeleton networks produced by these models could give insight into hidden simplicial complex structure underlying systems for which only network data is available, on the other hand the skeleton of a d-simplex is a clique of $d+1$ nodes and our models can be viewed as generative models based on the structural properties belonging to 'motifs' that in our case happen to be cliques.

In [33] (Chapter 4 of this thesis) and [34] (Chapter 5 of this thesis) we study the skeleton networks produced by the respective models presented in the papers.

In [33] the model we propose is a *maximum entropy* model based on constraints on the generalised degrees of the nodes. The theory behind this is discussed extensively in Chapters 3 and 4, but in short this means that the model is the 'least biased' assignment of probability to the set of all simplicial complexes such that the nodes have exactly the generalised degrees specified by the constraints. Other properties of this model can in a sense be considered to be a natural consequence of the generalised degrees of the nodes. We examine the effect that the generalised degrees have on the structure of the skeleton network, in particular on the clustering (which quantifies the tendency for the 'friends' of a node to be friends with each other). Our maximum entropy model provides us with a rigorous way to study the effect of the 'propensity to group membership' (characterised by the generalised degree of a node) on the clustering (or any other measurable property) of a network.

In [34], we propose a model of a growing and weighted simplicial complex. The

goal is to examine the effects that different growth mechanisms and simplex dimension have on the distributions of the generalised degrees and strengths. We also examine the effect these things have on the degree distribution of the skeleton networks. Pivotal early discoveries of network science were that power-law distributed degree distributions $P(k) \propto k^{-\gamma}$ were ubiquitous in real networks [1–3, 45] and that networks with this property can be generated through a combination of growth and preferential attachment of new nodes to existing nodes with high degree [45]. Since then a range of other models have been proposed which exploit a variety of mechanisms to produce power-law networks [24, 25, 46–55]. In [34] we show that a model in which the new nodes attach to simplices of dimension greater than 0 with probabilities dependent on the properties of these higher dimensional simplices can cause power-law distributions of the degrees in the skeleton network. This demonstrates that models exploiting properties belonging to simplices/cliques rather than solely node-centric properties can be useful for explaining the structure observed in real networks.

Chapter 3

Background: Models of networks

In the previous chapter we defined simplicial complexes and showed how simple measures of structure in networks could be extended for use in simplicial complexes. These measures of structure provide us with a language to describe simplicial complexes, and allow us to characterize the similarities or differences between them. We would like to go further though, and understand what the structural properties we have defined really mean for a simplicial complex.

For example, how do the rules governing the evolution of a simplicial complex affect its structure, and can we make an informed guess at these rules based on observing the structure of a real simplicial complex? How do different properties (e.g. the generalised degrees at different dimensions) relate to each other? What is the fairest way to model a simplicial complex based on incomplete data? And what effect do structural properties have on dynamic processes (e.g. diffusion, epidemics) occurring on a simplicial complex?

Models help us answer these kinds of questions. They allow us to understand the dependences and relations between the structure of simplicial complexes, their evolution and processes occurring upon them. They help us to form hypotheses about the origins of simplicial complexes, to make informed guesses at their structure in the absence of complete data, and to make generalisations and predictions about simplicial complexes sampled from the same source.

In order for models to be capable of achieving these goals they need to be well designed. In particular, models should incorporate the assumptions we make about simplicial complexes or their evolution in a transparent way, and shouldn't incorporate additional unintended assumptions or biases. Introducing an element of randomness into a model is a good way of reducing unintended biases. By controlling one aspect of a simplicial complex (such as a specific set of its structural properties, or the probability with which new simplices are added to it) but making everything else 'as random as possible', we can in a sense isolate the effects of the controlled aspect on the type of simplicial complex produced.

The models of simplicial complexes presented in this thesis are random. They fall into two categories: *explanatory models* and *null models*. Explanatory models are those in which our motivation is to develop a method for generating simplicial complexes with a given set of properties, and that could plausibly explain how such properties emerge in real simplicial complexes. Null models instead answer the question, what is the 'best' way to model a simplicial complex given it has a given set of properties?

Our models extend/generalise known models of networks. In this background chapter we discuss models of networks which have a relevance to the simplicial complex models we present in Chapters 4-6, and explain the key ideas behind them.

3.1 Explanatory network models

Network science seeks to build a common mathematical framework to understand real world complex systems that on a superficial level may not seem related. The belief that building such a framework might be possible has been motivated by the discovery around twenty years ago that networks arising in diverse areas often have a striking resemblance to each other [1–3].

The resemblances referred to here are the 'universal properties' present in almost all real networks that have been observed, namely power law distributions of the

degrees, short average path lengths, an abundance of short loops and non-trivial community structure [1–3].

The development of *explanatory* models of networks which generate one or more of these properties can provide insight into the underlying mechanisms that produce the same properties in real networks. The objective with these models is not necessarily to create the most 'realistic' networks but to isolate aspects of the generation process that result in the desired property(s). These 'aspects' can be refined and combined in future models in order to produce networks which better resemble real networks while at the same time remaining simple and interpretable.

Some famous early models which were shown to reproduce universal properties of complex networks include the Barabási-Albert model [45] which uses a 'rich-get-richer' mechanism to produce power-law degree distributions and the Watts-Strogatz model [56] which produces 'small-world' networks with low average path lengths and high clustering.

In the next sections we discuss a number of explanatory models of networks and simplicial complexes which have a particular relevance to our own explanatory models of simplicial complexes that we present in Chapters 5 and 6.

In the Section 3.1.1 we briefly discuss the Barabási-Albert model and other models that produce power-law distributions of the degrees of the nodes. These models and the preferential attachment mechanism that most of them rely on are of particular relevance to the simplicial complex models that we present in Chapters 5 and 6.

In Section 3.1.2 we discuss a model of a weighted network [43] closely related to the Barabási-Albert model, in which competing effects of *growth* and *reinforcement* produce interesting weight-topology correlations observed in real weighted networks. Our model of a weighted simplicial complex presented in Chapter 5 is a generalisation of this model, and the *mean-field* techniques used in [43] and Section 3.1.2 are of particular relevance to similar calculations in both Chapters 5 and 6.

In Section 3.1.3 we present a model of a growing d-dimensional simplicial complex [57]. This model explores the effects of dimension and attachment mechanisms on the topologies of the simplicial complexes, and is closely related to the model of a weighted simplicial complex presented in Chapter 5.

3.1.1 Producing Scale-free networks with growth and preferential attachment

In this section we briefly discuss the Barabási-Albert model and other models that produce power-law distributions of the degrees of the nodes. The degree distribution $P(k)$ of a network is power-law if it follows

$$P(k) \propto k^{-\gamma}, \tag{3.1}$$

for $k \gg 1$, where γ is the exponent of the distribution and for most real networks is in the range $(2, 3]$ (a much smaller class of networks have exponent $\gamma \in (1, 2]$, and are the subject of Chapter 6). Power-law networks with $\gamma \leq 3$ have a divergent second moment and for this reason are called *scale-free* as they have no 'typical' scale.

The Barabási-Albert (BA) model produces scale-free networks with exponent $\gamma = 3$ through a combination of *growth* of the network through the addition of new nodes and new links and the *preferential attachment* of those new links to existing nodes with high degree. The model progresses in discrete time steps, starting from some suitable initial network at time $t = 1$. At each subsequent time step a new node is added to the network with m links. Each of these links connects to an existing node with a probability proportional to its degree. For an existing node i with degree k_i at time t the probability $m\Pi_i$ of gaining a new link is equal to

$$m\Pi_i = m\frac{k_i}{\sum_j k_j} \approx \frac{k_i}{2t}, \tag{3.2}$$

where Π_i is the probability that a given link attaches to i. Eq. (3.2) implements the preferential attachment mechanism according to which nodes that already have

high degree k_i are more likely to further increase their degree by acquiring new links. The degree distribution $P(k)$ of the BA model can be evaluated exactly in the large network limit $N(t) \gg 1$ and is given by [48, 58]

$$P(k) = \frac{2\Gamma(m+2)}{\Gamma(m)} \frac{\Gamma(k)}{\Gamma(k+3)} \simeq 2m(m+1)k^{-3}, \tag{3.3}$$

where the latter approximated expression describes the tail of the distribution where $k \gg 1$.

The BA model is thus a simple model of network growth that produces scale-free networks and is also analytically amenable. At each time step of the model, the selection of existing nodes is random with a probability that is proportional to the degree of the node and has no dependence on any other aspect of the structure of the network or on any hidden variables or attributes belonging to the nodes. Importantly, the m existing nodes selected at each time step are *selected independently of each other*. In this sense the model can be thought of as isolating the effect of the linear preferential attachment mechanism from other considerations.

Another set of models that exploit the preferential attachment mechanism to produce scale-free networks include those that grow by 'node copying' [52–55] where new nodes copy a subset of the links of (uniformly) randomly selected nodes. This results in a form of preferential attachment, as the probability that an existing node gains a new link is equal to the probability that one of its neighbours is copied and so is proportional to its degree. Naturally, the nodes that a new node selects to be its neighbours are in these models *not independent* but are always a short distance from each other. Models of this type thus introduce a new 'bias' in comparison with the BA model which can have the effect of inducing shorter average path lengths and a higher clustering in the networks produced [59].

The BA model and these copying models form part of a family of models capable of producing scale-free networks, each with its own set of intentional biases that have been introduced with the aim of replicating some aspect of the structure of real networks. Other such models include those based on a hidden hyperbolic

geometry [24, 25], models that assume that nodes possess an intrinsic 'attractiveness' [46, 47, 58] that affects their ability to attract new links, models that use non-linear forms of preferential attachment [48] and models based on fractals or hierarchically arranged motifs [49–51]. By studying these models analytically or numerically a picture may be built of the effects of different construction methods on networks topologies that can lend insight into the possible processes underlying real networks and help guide the development of more sophisticated and realistic models.

3.1.2 Weight-topology correlations and the weighted BA model

In this section we discuss a model [43] of a network that is both weighted and growing. This model is an explanatory model designed to explore the effects of competing processes of *growth* and *reinforcement* on the emergence of the weight-topology correlations which were mentioned in Section 2.2 of the previous chapter.

In particular the model is an adaptation of the BA model, in which each link (i, j) in the network has a weight w_{ij} associated to it, and where these weights evolve in time. The model can produce class I networks with homogeneous distributions of the weights across the links or class II networks in which high degree nodes possess links with higher average weight than low degree nodes.

As in the BA model, this model progresses in discrete time steps, starting from some suitable initial network at time $t = 1$. At each subsequent time step two processes take place:

A) *Growth process*:
A new node arrives and m new links with initial weight w_0 are created between the new node and existing nodes. The probability Π_i that a given node i with degree k_i is selected by one of the new links is given by

$$\Pi_i = \frac{k_i}{\sum_i k_i}. \tag{3.4}$$

B) *Reinforcement process*:
At this step m' existing links are selected and their weights are increased by w_0. A link (i, j) with weight w_{ij} is selected for reinforcement with probability $\tilde{\Pi}_{ij}$ proportional to its weight, i.e.

$$\tilde{\Pi}_{ij} = \frac{w_{ij}}{\sum_{ij} w_{ij}}. \tag{3.5}$$

The growth process is independent of the weights of the links, and when ignoring these weights the model is exactly the BA model described in the previous section. Process B implements a preferential attachment mechanism for the weights according to which links with larger weights are more likely to gain additional weight. The parameters m and m' are the number of links added or reinforced at each time step respectively and the choice of values of these parameters relative to each other determines the distribution of the weights over the links [43]. In fact, the model produces class I networks ($s \propto k$) when $m > m'$ and class II networks ($s \propto k^\theta$, $\theta > 1$) when $m < m'$ [43]. In particular, for large t the strengths of the nodes scale with the their degrees like

$$s(k) \propto \begin{cases} k & \text{if} \quad m > m', \\ k \log k & \text{if} \quad m = m', \\ k^{2\lambda} & \text{if} \quad m < m', \end{cases} \tag{3.6}$$

where $\lambda = \frac{m'}{m+m'}$ is the proportion of the total weight added to the network via the reinforcement mechanism and $s(k)$ is the average strength of a node with degree k in the network. Note that $2\lambda > 1$ when $m < m'$ indicating that it this case the network belongs to class II.

In the rest of this section we show how these results were first derived in [43]. The calculations below are especially relevant to the model of a simplicial complex that we present in Chapter 5, which generalises the weighted BA model presented here to simplicial complexes of dimension d. The techniques employed are also relevant to calculations performed in Chapter 6.

The nodes of the weighted BA model are differentiated from each other by the time at which they arrived in the network, which we refer to as their birth time. In [43] the average strength $s(k)$ of nodes with degree k is approximated by making the *mean-field* assumption that the degree and strength of a node i with birth time t_i are well approximated by their expected values conditioned on the birth time, i.e.

$$k_i \approx \overline{k}(t_i) \quad , \quad s_i \approx \overline{s}(t_i). \tag{3.7}$$

As long as $\overline{k}(t_i)$ provides a one-to-one relation between birth times and (expected) degrees then $s(k)$ can be approximated by using $t_i(\overline{k})$ in $\overline{s}(t_i)$.

In [43] the mean-field degrees and strengths of nodes born at t_i are calculated by making the further approximation that the degrees, strengths and time t may be treated as continuous variables, resulting in differential equations that can be solved to obtain the degrees and strengths. For the degree of a node i born at t_i this differential equation is

$$\frac{\partial}{\partial t} k_i = m\Pi_i = m\frac{k_i}{\sum_i k_i} \approx \frac{k_i}{2t}, \tag{3.8}$$

where in the last equality we have used the fact that $\sum_i k_i$ is equal to 2 times the number of links in the network at time t and so $\sum_i k_i \approx 2mt$ for large t. Eq. (3.8) has initial condition $k_i = m$ at time $t = t_i$ and is solved to obtain

$$k_i = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}. \tag{3.9}$$

The above equation can also be used to derive the degree distribution of the model by observing that within the mean-field approximation the proportion of nodes with degree greater than some value k is equal to $\frac{t^*}{t}$ where t^* solves $k = m\left(\frac{t}{t^*}\right)^{\frac{1}{2}}$. It is simple to check that this results in the degree distribution

$$P(k) = \frac{d}{dk}P(k_i < k) = \frac{d}{dk}\left[1 - \frac{t^*}{t}\right] = 2m^2k^{-3}. \tag{3.10}$$

The closeness of this result to the exact result [48, 58] shown in Eq. (3.3) supports

the validity of the mean-field approach taken in [43] and which we present here.

To calculate the mean-field strength of a node born at t_i we note that $s_i = \sum_j w_{ij} a_{ij}$ becomes $s_i = \sum_j w_{ij} p_{ij}$ in the mean-field approximation, where p_{ij} is the probability of the link (i, j) and we make the mean-field approximation $w_{ij} = \overline{w}_{ij}$ (the expected value of w_{ij}). The probability that (i, j) increases at some time t is given by Eq. (3.5) and so similarly to the degrees a mean-field continuous approximation of the weight can be obtained from the differential equation

$$\frac{\partial}{\partial t} w_{ij} = w_0 \tilde{\Pi}_{ij} = w_0 m' \frac{w_{ij}}{\sum_{ij} w_{ij}} \approx \lambda \frac{w_{ij}}{t}, \tag{3.11}$$

where we have used $\sum_{ij} w_{ij} \approx (m' + m) w_0 t$ is the total weight that has been added to the network by time t and $\lambda = \frac{m'}{m+m'}$. All links start with weight w_0 and so the initial condition is $w_{ij} = w_0$ at time $t = \max(t_i, t_j)$. The solution is thus

$$w_{ij} = w_0 \left(\frac{t}{t_{ij}} \right)^{\lambda}, \tag{3.12}$$

where $t_{ij} = \max(t_i, t_j)$.

To calculate the probability p_{ij} that the link (i, j) actually exists in the network, assume that $t_i < t_j$ then the p_{ij} is the probability that at time $t = t_j$ the new node j attaches one of its m initial links to i. Using Eq. (3.4) and the mean-field degree given in Eq. (3.9) p_{ij} is found to be

$$p_{ij} = m \Pi_i(t = t_j) = \frac{k_i(t = t_j)}{2 t_j} = \frac{m}{2} (t_i t_j)^{-\frac{1}{2}}. \tag{3.13}$$

Notice that Eq. (3.13) is symmetric with respect to i and j so that it also holds when $t_i > t_j$.

Using Eq.s (3.12) and (3.13) the mean-field strength can be calculated:

$$\begin{aligned} s_i &= \sum_j w_{ij}p_{ij} \approx w_0\frac{m}{2}t^{\lambda}t_i^{-\frac{1}{2}}\int_0^t dt_j t_j^{-\frac{1}{2}}t_{ij}^{-\lambda} \\ &= w_0\frac{m}{2}t^{\lambda}t_i^{-\frac{1}{2}-\lambda}\int_0^{t_i} dt_j t_j^{-\frac{1}{2}} + w_0\frac{m}{2}t^{\lambda}t_i^{-\frac{1}{2}}\int_{t_i}^t dt_j t_j^{-\frac{1}{2}-\lambda} \end{aligned} \tag{3.14}$$

where in the first line the approximation being made is to integrate over the birth times of the nodes rather than summing over the nodes, and in the second line the two integrals correspond to the contributions when $t_{ij} = t_i$ and $t_{ij} = t_j$ respectively. Evaluating these integrals gives

$$s_i = \begin{cases} w_0 m\left(1-\frac{1}{1-2\lambda}\right)\left(\frac{t}{t_i}\right)^{\lambda} + w_0 m\frac{1}{1-2\lambda}\left(\frac{t}{t_i}\right)^{\frac{1}{2}} & \text{if} \quad \lambda \neq \frac{1}{2}, \\ w_0\frac{m}{2}\left(\frac{t}{t_i}\right)^{\frac{1}{2}}\left[1+\log\left(\frac{t}{t_i}\right)\right] & \text{if} \quad \lambda = \frac{1}{2}. \end{cases} \tag{3.15}$$

Keeping only the leading terms when $\frac{t}{t_i} \gg 1$, gives

$$s_i \propto \begin{cases} \left(\frac{t}{t_i}\right)^{\frac{1}{2}} & \text{if} \quad \lambda < \frac{1}{2}, \\ \left(\frac{t}{t_i}\right)^{\frac{1}{2}}\log\left(\frac{t}{t_i}\right) & \text{if} \quad \lambda = \frac{1}{2}, \\ \left(\frac{t}{t_i}\right)^{\lambda} & \text{if} \quad \lambda > \frac{1}{2}. \end{cases} \tag{3.16}$$

Finally, using the fact that the degree scales like $k \propto \left(\frac{t}{t_i}\right)^{\frac{1}{2}}$, Eq. (3.16) becomes Eq. (3.6).

This model demonstrates that class I and class II networks can be produced within the same framework based on simple mechanisms of growth and reinforcement. These classes were originally observed in real weighted networks and so the model provides a possible hypothesis about how such networks might evolve. In [43], the mean-field results were compared to numerical simulations of the model and found to be in agreement, validating the mean-field continuous approach taken.

In Chapter 5 we apply a similar approach to an original model of a weighted

simplicial complex evolving via growth and reinforcement processes. Depending on the model parameters the simplicial complexes produced can produce a wide variety of topolgies and scalings of generalised strength with generalised degree. Our model shows the effects of dimensionality on weight-topology correlations and for particular choices of parameters in fact reduces to the weighted BA model discussed in this section.

3.1.3 Network Geometry with Flavor

In this section we discuss a model [57] of a growing simplicial complex which has particular relevance to the model we present in Chapter 5. This model is called the Network Geometry with Flavor (NGF), and it can produce pure d-dimensional simplicial complexes with a variety of topologies including chains, higher dimensional manifolds and scale-free networks with small-world properties and non-trivial community structure [57].

The model explores the effects of growth mechansims and dimensionality on the properties of simplicial complexes. In this section we introduce the NGF and discuss some of the results in [57] that have particular relevance to our original model presented in Chapter 5. All results shown in this section are the work of [57] unless stated otherwise.

As in the models discussed in the previous two sections, the NGF progresses in discrete time steps. At time $t = 1$ the model starts with a single d-simplex and its faces. At each time step a new node appears and forms a single d-simplex with a randomly selected $(d-1)$-face. The probability with which a face is selected depends on the generalised degree of the face and the model parameters.

In particular the probability that a $(d-1)$-face α is selected is equal to

$$\Pi_{d-1}(\alpha) = \frac{1}{\mathcal{Z}_t} e^{-\beta\epsilon_\alpha}(1 + sn_\alpha), \tag{3.17}$$

where s is a parameter called the *flavor*, $n_\alpha = k_{d,d-1}(\alpha)$ is called the *saturation* of

α, ϵ_α is an *energy* associated to the face α governing the propensity of α to attract new d-simplices, β is the *inverse temperature* which controls the relative effects of the energies of the faces in comparison to the effect of the saturation, and $\mathcal{Z}_t$ is the *partition function* which normalises Eq. (3.17) and is given by

$$\mathcal{Z}_t = \sum_{\alpha \in S_{d,d-1}(t)} e^{-\beta \epsilon_\alpha}(1 + s n_\alpha). \tag{3.18}$$

The energies associated to the faces allow for the introduction of a heterogeneity in the 'attractiveness' of simplices of the same dimension, with β regulating the strength of this effect. The authors explore the NGF for $\beta = 0$ and $\beta > 0$, however in this section we outline their results for $\beta = 0$ only, as these are the results relevant to our own model in Chapter 5. For certain parameter choices our model corresponds to the NGF with $\beta = 0$.

Setting $\beta = 0$ the Eq.s (3.17) and (3.18) become

$$\Pi_{d-1}(\alpha) = \tfrac{1}{\mathcal{Z}_t}(1 + s n_\alpha), \quad \mathcal{Z}_t = \textstyle\sum_{\alpha \in S_{d,d-1}(t)}(1 + s n_\alpha). \tag{3.19}$$

The flavor s has an important effect on the topological properties of the simplicial complexes produced. Selection of $s = -1$ imposes the constraint that the generalized degree $k_{d,d-1}(\alpha)$ of a $(d-1)$-face α can only take the values 1 and 2, or equivalently imposes that the saturation n_α can only take values 0 (unsaturated) and 1 (saturated), which leads to the simplicial complex produced being a d-dimensional manifold. Choosing $s = 0$ or $s = 1$ removes this constraint, and gives a selection probability $\Pi_{d-1}(\alpha)$ that is uniform on the set of all $(d-1)$-faces for $s = 0$ and a form of preferential attachment with $\Pi_{d-1}(\alpha) \propto k_{d,d-1}(\alpha)$ for $s = 1$.

In [57], asymptotically exact expressions for the generalised degree distributions are derived for faces of dimension $0 \leq \delta \leq d-1$ (i.e. for all faces). In [57] this is done using a master equation approach which we do not show here. In our model [34] that we present in Chapter 5 we instead take a mean-field approach similar to the one shown in the previous section on the weighted BA model, yielding the same results as found in [57]. In particular the distributions of the generalised degrees $k_{d,\delta}(\alpha)$ of

Table 3.1: Distribution of generalized degrees of faces of dimension δ in a d-dimensional NGF of flavor s at $\beta = 0$.

flavor	$s = -1$	$s = 0$	$s = 1$
$\delta = d - 1$	Bimodal	Exponential	Power-law
$\delta = d - 2$	Exponential	Power-law	Power-law
$\delta \leq d - 3$	Power-law	Power-law	Power-law

faces of dimension δ can be bimodal, exponential, or power-law. The dependence of the generalised degree distributions on s, d and δ is shown in table 3.1.

For $s = 1$ the generalised degree distributions of all faces is power-law, while for $s = 0$ the generalised degrees of the $(d-1)$-faces have an exponential distribution and all faces of lower dimension are power-law distributed. For $s = -1$, the simplicial complex produced is a d-manifold and so the generalised degrees of the $(d-1)$-faces is bimodal, while for faces of dimension $d-2$ the distribution is exponential and then for all lower dimensions the distributions are again power-law.

These distributions are understood in [57] in terms of the *effective* attachment mechanism felt by the faces of lower dimension. The probability that the generalised degree $k_{d,\delta}(\alpha)$ of some δ-face α is increased in a given time-step is dependent on the number of $(d-1)$-faces that α is a part of and on the generalised degrees of these $(d-1)$-faces. These can be related to $k_{d,\delta}(\alpha)$, which in turn allows for a calculation of the attachment probability for α in terms of its own generalised degree $k_{d,\delta}(\alpha)$. By effective attachment mechanism we mean the type of dependence this attachment probability has on the generalised degree of the face: a linear dependence on $k_{d,\delta}(\alpha)$ indicates preferential attachment leading to power-law distributions, while a constant dependence indicates a uniform attachment mechanism leading to an exponential distribution. A more complete discussion of this phenomena is given in Chapter 5.

These results show that there is a significant interplay between dimension, attachment mechanism, and generalised degree distributions. The model also presents a new way to produce scale-free networks. In the simplicial complexes produced by the model the degree K_i of a node i in the skeleton networks is related to its gener-

alised degree $k_{d,0}(i)$ by

$$K_i = d - 1 + k_{d,0}(i). \tag{3.20}$$

This fact can be seen from the fact that a newly created node has $k_{d,0}(i) = 1$ and $K_i = d$, and for each subsequent d-simplex it gains, both quantities are increased by exactly 1. The consequence of this fact is that for $d-s \geq 2$ the skeleton networks are always power-law (in fact for d and s in this range the distributions have a diverging second moment indicating that they are scale-free). The NGF is thus also part of the family of models discussed in Section 3.1.1 that can produce scale-free networks.

3.2 Null models of networks

In the previous section, the explanatory models of networks that we discussed were formulated as 'theories' about how a real network might be constructed. The aim was to isolate the effects of the construction methods on structural properties of the networks produced. In this section we instead seek to develop *null models* that can isolate the effects of network structural properties on each other.

These models are ensembles of simplicial complexes, i.e. a set of simplicial complexes $\{G\}$ with a probability distribution P defined on it. Many of the explanatory models discussed in the previous section exploited 'randomness' to reduce unintended bias in the construction method. Similarly, in the null models we discuss here, we aim to eliminate unintended bias by finding the 'maximally random' assignment of probability P such that the networks produced have a given set of *constrained properties*.

These null models allow us probe the effects of the constrained properties on other secondary properties by asking how probable it is to observe these secondary properties given the constrained properties. If networks with a given set of secondary properties occur with high probability in a given model then those properties can be considered to be a 'natural' consequence of the constrained properties, while if

they occur with low probability it suggests that the secondary properties cannot be considered to have arisen as a consequence of the constrained properties.

As we shall see later in this section, Erdös-Renyi random networks (in which the links are present independently of each other with equal probability) are a very simple example of one of these 'maximally random' network models in which the only constrained property is the total number of links. The probability in the Erdös-Renyi model of observing the scale-free degree distributions or high clustering found in many real networks is in fact very low, indicating that these properties do not have simple explanations in terms of the density of links in a network. Other null models based on different constrained properties that better describe real networks have been developed [60–63]. The hope is that by looking for the simplest set of constrained properties that generate the most realistic networks we can identify the most 'important' structure in real networks.

But what do we mean by a 'maximally random' assignment of probability? In the following section we introduce the Shannon entropy of a probability ensemble which can be thought of as a measure of the 'uncertainty' present in the ensemble and which formalises what we mean by 'maximally random'. An assignment of probability that maximises this entropy subject to constraints on a set of network properties is the rigorous best choice for modelling a network *based only on knowledge of the constrained properties.*

Maximum entropy models defined in this way have a number of uses, including the identification of non-trivial structure in real network data, the reconstruction of real networks based on partial information about their structure, the investigation of the dependence that dynamical processes taking place on a network have on a narrow set of structural properties and the identification of statistical relations between distinct sets of structural properties.

3.2.1 Shannon entropy

The Shannon entropy [64] of a probability distribution P defined on some discrete set X is defined as

$$S = -\sum_{x \in X} P(x) \ln(P(x)), \tag{3.21}$$

while for a continuous set with accompanying probability density function, the equivalent definition is

$$S = -\int_X dx p(x) \ln(p(x)). \tag{3.22}$$

the entropy is the average amount of 'information' acquired by observing an event x as a function of its probability of occurring. For a discrete probability distribution P the *information content* $I(x)$ of an event x is defined as [64]

$$I(x) = \ln\left(\frac{1}{P(x)}\right). \tag{3.23}$$

This definition fulfils a number of axiomatic assumptions made by Shannon about the properties that any definition of the 'amount' of information ought to have, including the assumption that the amount of information imparted by an event should be larger for less probable events, should be zero when the probability of an event is equal to 1 and that the amount of information imparted by observing two independent events should be equal to the sum of the amounts of information imparted when observing the events separately, i.e.

$$I(x, y) = -\ln\left(P(x, y)\right) = -\ln\left(P(x)\right) - \ln\left(P(y)\right) = I(x) + I(y). \tag{3.24}$$

The negative logarithm of the probability is the only function that satisfies these properties, with different choices of base for the logarithm amounting to a change of units.

An important result which provides further intuition about the usefulness of the definition of information content given in Eq. (3.23) concerns finding the 'shortest'

encoding of the data coming from a random variable taking values $x \in X$. By encoding, we mean labelling each event x with a number represented in binary, which can be thought of as the answers to a set of yes/no questions that specify x. A possible motivation for doing this would be so that the outcome of observing the random variable can be transmitted efficiently along a communication channel. The fewer yes/no questions we need to label each element of X distinctly, the less strain we put on the communication channel. There are $|X|$ elements in the set X so we need $|X|$ distinct numbers. The most intuitive labelling scheme would be to give each element its own number with length $\log_2(|X|)$. In this labelling scheme every element is represented by a number of the same length regardless of its probability of occurring. However, by allowing events with different probabilities to have different lengths it is actually possible to achieve shorter *expected* lengths of the labels than in the naive labelling. By 'shortest' encoding we mean a labelling of the elements $x \in X$ that minimises the expected length of the label. This labelling would need to take advantage of any heterogeneity in the probability distribution so that more probable outcomes have shorter labels while longer labels will be given to events with lower probabilities.

The entropy given in Eq. (3.21) in fact provides the expected length of the labels that could be achieved by an optimal encoding [64]. This fact allows us to interpret the information content of an event x defined in Eq. (3.23) as the minimum number of yes/no questions needed to specify x *given that we already know the probability distribution* P. The information content is therefore the remaining information we need in order to specify x given the information conveyed to us about x by P.

More homogeneous distributions of the probability give a higher entropy because they are less informative and so observation of the variable provides us with more information, while heterogeneous distributions that assign high probability to a small number of outcomes are more informative about what the likely outcome of observation would be and so the act of observing the variable imparts less information on average, corresponding to a low entropy.

In the next section we discuss how we can use entropy maximisation to find

the 'best' (least informative) probability distribution to model a variable based on partial knowledge about the variable.

3.2.2 Maximising the Shannon entropy

By varying P so that the entropy is maximised, we find the probability distribution that is least informative about the random variable it describes. In the absence of any constraint on P apart from normalisation, the maximum entropy distribution is always uniform on the set of possible outcomes. This clearly corresponds with our common sense idea of what the 'best' distribution would be for modelling a variable about which we know nothing apart from the values that it can take.

In the case that we do know *something* about the variable, we would like to incorporate this knowledge into our model, and we would like to do so in a way that doesn't accidentally include any unintended biases or assumptions. This knowledge could be a 'hard constraint' on the variable that confines it to some subset Y of the the larger set of outcomes X (by putting $P(x) = 0$ for all $x \in X \backslash Y$). In this case the problem is equivalent to maximising the entropy with no constraints but using the restricted set Y instead of X. Alternatively we may wish to incorporate knowledge in the form of a 'soft constraint' on the expected value of the variable or some statistic drawn from the variable.

The maximum entropy distribution in either of the above cases is the least informative given the constraints used. This guarantees that we haven't included any additional assumptions in the model as these assumptions would necessarily result in more heterogeneity in the probability distribution and hence the distribution would have a lower entropy than the maximum entropy solution. The entropy maximisation approach thus formalises our common sense intuition that in the absence of knowledge about a variable we should pick a uniform distribution to model it, and it allows us to extends this intuition further to apply to variables for which we *do* possess some prior knowledge.

Some examples include the 'fair die' distribution on the set $X = \{1, \ldots, 6\}$ or the uniform distribution on $X = [a, b]$ which are the maximum entropy distributions in the absence of constraints, given their respective domains X. Meanwhile, the Gaussian distribution is the maximum entropy distribution when $X = \mathcal{R}$ is the real line and there are soft constraints on the first and second moments of the distribution i.e. $\langle x \rangle$ and $\langle x^2 \rangle$ are constrained.

In the next section we discuss how this approach has been used to model networks based on constraints on their structural properties.

3.2.3 Maximum entropy models of networks

In network science there are a number of different models of networks which can be described as maximum entropy [60–63]. In these models the 'random variables' are networks G that occur with a probability that maximises the Shannon entropy subject to constraints on specific structural properties of the networks. These constraints incorporate information about network properties such as the total number of links [1], the degrees of the nodes [60], the prevalence of some motif in the network [61], the degree correlations [63] or a block structure [62].

Models of this type are the best model in the case where the only information possessed about the networks is contained in the constraints. This allows us to 'reconstruct' networks about which we only have partial information or to make well founded probabilistic statements about other properties these networks may have.

They also allow us to probe the effects of the constrained properties on other secondary properties of the networks. Maximum entropy models based on constraints on the degrees of the nodes allow us to isolate the effects of the degrees on properties such as the clustering, degree correlations or community structure. They can be used as a benchmark to which real networks may be compared: for example a real network can be said to have a 'trivial' level of clustering if its clustering is no higher than that found in the maximum entropy model based on the degrees of the

real network. A real network with clustering much higher than that expected in the maximum entropy model has instead significant clustering as it is unlikely that a model in which the connections between the nodes depend only on the degree would produce such a high number of triangles.

It is worth noting that when maximum entropy models are used for this purpose that 'realism' is not a major concern. These models in a sense isolate the effects of the constrained properties on the networks produced, and it is finding the similarities and differences between these models and real networks that helps us to identify the important structure in real networks.

The constraints used for these models may be either hard or soft. As discussed in the previous section, performing entropy maximisation using hard constraints is effectively the same as restricting the set upon which the maximisation is done without constraints. For a network with N nodes and a set of M structural observables $\{C_r(G)\}_{r\in\{1,\dots,M\}}$ with (hard) constraining values $\{c_r\}_{r\in\{1,\dots,M\}}$ the maximum entropy probability of a network G is

$$P(G) = \frac{\prod_{r=1}^{M} \delta(C_r(G), c_r)}{\mathcal{N}}, \tag{3.25}$$

where $\delta(x, y)$ is the Kronecker delta, and $\mathcal{N}$ is the total number of networks satisfying the constraints, or

$$\mathcal{N} = \sum_{G} \prod_{r=1}^{M} \delta(C_r(G), c_r). \tag{3.26}$$

Evaluation of (3.26) may be difficult for large networks, depending on the choice of properties that are being constrained. For this reason, working with these models often involves the construction of algorithms to sample individual networks from the ensemble [1, 65], or approximate analytical methods that exploit relations with *soft constraints* models with similar constraints [60].

In models using soft constraints it is straightforward to find the correct form of

the probability of a network G in terms of a set of Lagrangian multipliers associated with the constraints. For a set of M structural observables $\{C_r(G)\}_{r\in\{1,\dots,M\}}$ with (soft) constraining values $\{c_r\}_{r\in\{1,\dots,M\}}$ the Lagrangian is

$$\mathcal{L} = S - \mu\left(\sum_G P(G) - 1\right) - \sum_{r=1}^{M} \lambda_r \left(\sum_G P(G)C_r(G) - c_r\right), \tag{3.27}$$

where S is the entropy $-\sum_G P(G)\ln\left(P(G)\right)$, and μ, and $\lambda_1, ..., \lambda_M$ are Lagrangian multipliers. The equations to solve are:

$$\frac{\partial\mathcal{L}}{\partial P(G)} = 0, \quad \frac{\partial\mathcal{L}}{\partial\mu} = 0, \quad \frac{\partial\mathcal{L}}{\partial\lambda_r} = 0. \tag{3.28}$$

Solving the first two equations above allows us to write the probability of a network G in this ensemble as

$$P(G) = \frac{1}{Z} e^{-\sum_{r=1}^{M}\lambda_r C_r(G)}, \tag{3.29}$$

where Z normalises the distribution and is called the *partition function* and is given by

$$Z = \sum_G e^{-\sum_{r=1}^{M}\lambda_r C_r(G)}. \tag{3.30}$$

Models of this type are known as Exponential Random Graph Models (ERGMs) [60], and they have been used widely in the study of social networks [61]. Whether or not it is simple to obtain $P(G)$ or more specifically Z in terms of the constraining values $\{c_r\}_{r\in\{1,\dots,M\}}$ depends on what choice of observables $\{C_r(G)\}_{r\in\{1,\dots,M\}}$ we are constraining, but in practice it is very difficult for all but the simplest observables. For sets of constraints where calculation of Z is analytically intractable, Z may be approximated numerically through Markov Chain Monte Carlo methods [66].

In the next section we discuss two widely used maximum entropy network models based on hard or soft constraints on the degrees of the nodes. In Chapter 4 we generalise these models to maximum entropy models of simplicial complexes where the constraints are instead on the generalised degrees $k_{d,0}(i)$ of the nodes.

3.2.4 The configuration model and canonical ensemble of networks

The configuration model and canonical ensemble [60, 67] of networks are maximum entropy models of networks where the constraints are on the sequence of degrees of the nodes in the network. In the configuration model these constraints are hard so that for every network in the ensemble we must have:

$$\overline{k_i} = k_i = \sum_j a_{ij} \quad \text{for} \quad i = 1, 2, ..., N, \tag{3.31}$$

where k_i is the degree of node i in a particular instance of the network while $\overline{k_i}$ is its constrained value. For the canonical ensemble the constraints are instead enforced only on the expectations:

$$\overline{k_i} = \sum_G P(G) k_i = \sum_G P(G) \sum_j a_{ij} \quad \text{for} \quad i = 1, 2, ..., N, \tag{3.32}$$

where the sumations in the above equation are over all networks G on N nodes.

These models have been used for exploring the implications of particular degree distributions on other properties of networks such as the degree correlations, clustering and community structure [60, 68]. They have also been used for examining the effects of the degrees on processes that take place on networks such as epidemic spreading and diffusion [69].

The probability of a network G in the configuration model is uniform on the set of all networks which satisfy the constraints and zero everywhere else, i.e.

$$P(G) = \frac{\prod_{i=1}^{N} \delta(k_i, \overline{k_i})}{\mathcal{N}}, \tag{3.33}$$

where $\mathcal{N}$ is the total number of networks satisfying the constraints. There is no general formula for $\mathcal{N}$ in terms of the degrees, however, for large N and choices of the degrees below the structural cutoff discussed in Section 2.1.2 the following

formula is known [67, 70]:

$$\mathcal{N} = \frac{[(\langle \overline{k} \rangle N)!]^{(1/2)}}{\prod_{i=1}^{N} \overline{k_i}!} \exp\left[-\frac{1}{4} \left(\frac{\langle \overline{k}^2 \rangle}{\langle \overline{k} \rangle} \right)^2 + \mathcal{O}(\ln N) \right] \tag{3.34}$$

Furthermore, below the structural cutoff the configuration model lacks degree correlations and so the probability that two nodes i and j with degrees k_i and k_j are joined by a link must necessarily be proportional to their degrees, i.e.

$$p_{ij} = \frac{\overline{k_i k_j}}{\langle \overline{k} \rangle N}. \tag{3.35}$$

Aside from the above results, the configuration model is difficult to work with analytically and so in most applications algorithms are used that sample networks uniformly from the model. In one such algorithm [1], the nodes of the network are assigned a number of 'stubs' corresponding to the degree that we wish them to have. These stubs are then paired randomly until there are no remaining stubs and we have a network with the desired degree sequence. In Chapter 4 we present a generalisation of this algorithm for our configuration model of simplicial complexes.

On the other hand the canonical ensemble is easier to work with analytically. The probability of a network G in the canonical ensemble is

$$P(G) = \frac{1}{Z} e^{-\sum_i \lambda_i k_i}, \tag{3.36}$$

where the λ_i's are Lagrange multipliers associated with the constraints and Z is the *partition function* which normalises Eq. (3.36) and is given by

$$Z = \sum_G e^{-\sum_i \lambda_i k_i}. \tag{3.37}$$

The probability of a link between two nodes i and j in this ensemble is

$$p_{ij} = \frac{e^{-\lambda_i - \lambda_j}}{1 + e^{-\lambda_i - \lambda_j}}, \tag{3.38}$$

leading to the following relation between the constrained expected degree of a node i and the Lagrange multipliers

$$\overline{k_i} = \sum_j p_{ij} = \sum_j \frac{e^{-\lambda_i - \lambda_j}}{1 + e^{-\lambda_i - \lambda_j}}. \tag{3.39}$$

For choices of the expected degrees less than the structural cutoff, the probability of a link can be expressed directly in terms of the expected degrees of the nodes:

$$p_{ij} = \frac{\overline{k_i k_j}}{\langle \bar{k} \rangle N}, \tag{3.40}$$

which is the same probability as in the configuration model below the structural cutoff. Note however that in this ensemble the link probabilities are independent of each other so that rather than being fixed the degree of a node varies from sample to sample and has a Poisson probability distribution across the ensemble. The fact that the canonical ensemble is easier to work with analytically than the configuration model has made it attractive as a stand in for the configuration model in analytical calculations. Naturally, in order for canonical ensemble calculations to be valid for the configuration model, the effects of the correlations between links in the configuration model must have a minimal effect. For various reasons that we will now discuss it has in the past been widely assumed that for networks with a large number of nodes these correlations can be neglected and that the two ensembles are essentially equivalent. However, recent work [67,71,72] has shown that this is in fact not the case for ensembles such as the configuration model and canonical ensemble which impose an extensive (scaling like N) number of constraints.

The assumption of equivalence between the two models lies in their interpretation in statistical physics terminology as *conjugated ensembles*. Physical systems such as gasses composed of molecules will under certain circumstances converge to thermal equilibrium. When this happens, the system will be in one of many different *microstates* that describe the microscopic position and momentum of every molecule in the system. Once equilibrium is reached the probability distribution over these states is maximum entropy with respect to 'constraints' corresponding to

the *macrostates* of the system which could include the number of molecules, volume, temperature, pressure or total energy.

Two common ensembles in statistical physics are the microcanonical and canonical ensembles. The microcanonical ensemble is a gas with fixed number of molecules, volume and total energy. The canonical ensemble on the other hand has fixed number of molecules and volume, but instead of having a fixed total energy it has a fixed temperature. An analogy can be made between these ensembles and maximum entropy models of networks based on hard or soft constraints on a set of structural properties. The microcanonical ensemble enforces the total energy for every microstate and so every microstate with the correct energy has equal probability while all others have probability zero. The canonical ensemble instead enforces the temperature which is equivalent to enforcing the expected total energy.

When we pick the same total energy and expected total energy for these two ensembles we say they are *conjugated*. For the particular example of a gas under discussion here, the two ensembles are *asymptotically equivalent* as the number of molecules tends to infinity. Roughly speaking if conjugated ensembles are equivalent then their probability distributions over the microstates should converge and their macrostate properties should become the same. In this case the entropies per node of the two ensembles should converge in the limit $N \to \infty$. Most examples of conjugated ensembles encountered in statistical physics can be shown to be equivalent in this way, and in fact it is often assumed (wrongly) even in statistical physics textbooks that all such ensembles are equivalent.

In network science the simplest example of conjugated microcanonical and canoncial ensembles are the Erdös-Renyi random graphs in which we fix either the total number of links (the microcanonical ensemble) or the expected number of links (the canonical ensemble). In [71] it was shown that the difference in the entropy per node between the hard and soft Erdös-Renyi random graphs was vanishing as $N \to \infty$, i.e.

$$\frac{1}{N} S_{ER}^{can} - \frac{1}{N} S_{ER}^{micro} \to 0, \quad \text{as} \quad N \to \infty, \tag{3.41}$$

where S_{ER}^{can} is the entropy of the canonical Erdös-Renyi random graph with fixed expected link probability p and S_{ER}^{micro} is the entropy of the microcanonical Erdös-Renyi random graph with fixed number of links L. This indicates that the two ensembles are asymptotically equivalent in the thermodynamic limit. For the configuration model and its related canonical ensemble however, the differences between the entropies per node was shown to be finite as $N \to \infty$ [71], with

$$\frac{1}{N}S_{CE} - \frac{1}{N}S_{CM} \sim \mathcal{O}(1), \quad \text{as} \quad N \to \infty, \tag{3.42}$$

where S_{CE} is the canonical ensemble entropy and S_{CM} is the configuration model entropy. The failure of the entropies per node of these two ensembles to converge is due to to the extensive number of constraints placed on the ensembles [71], and indicates that the two ensembles are not equivalent [67, 71, 72].

Eq. (3.42) shows that the entropy of the canonical ensemble is much larger than the entropy of the configuration model. This indicates that in a sense the number of networks with 'significant' probability in the canonical ensemble is much larger than in the configuration model [64]. In [67] the difference in entropies between conjugated network ensembles including the configuration model and canonical ensemble was characterised in terms of the large deviation properties of the canonical ensembles. In particular the following formula relating the entropies was derived:

$$S_{CE} - S_{CM} = \Omega, \tag{3.43}$$

where Ω is called the *entropy of large deviation*, and is given by

$$\Omega = -\log\left[\sum_{G} P_{CE}(G) \prod_{i=1}^{N} \delta(k_i, \overline{k_i})\right], \tag{3.44}$$

i.e. -1 times the logarithm of the probability of observing a network G in the canonical ensemble that has degrees exactly equal to their expectations (and therefore also fulfilling the hard constraints of the configuration model). This entropy was originally formulated in more general terms in [70] in the context of trying to 'fit' a canonical ensemble to a given network or network topology such as a degree

sequence. Small values of Ω indicate that the canonical ensemble typically produces networks obeying the hard constraints, while larger values indicate that significant probability is assigned by the canonical ensemble to networks not obeying the constraints. In [67] Ω is calculated via a cavity method approach for a number of conjugated network ensembles corresponding to different network constraints. For the case of the configuration model and canonical ensemble with degrees below the structural cut-off discussed in Chapter 2 they find that the entropy of large deviation is given by

$$\Omega = -\sum_{i=1}^{N} \ln\left[\pi_{\overline{k_i}}(\overline{k_i})\right] \tag{3.45}$$

where $\pi_{\overline{k_i}}(\overline{k_i})$ is the Poisson distribution with average $\overline{k_i}$ evaluated at $\overline{k_i}$, i.e.

$$\pi_{\overline{k_i}}(\overline{k_i}) = \frac{1}{\overline{k_i}!}\overline{k_i}^{\overline{k_i}} e^{-\overline{k_i}}. \tag{3.46}$$

In the canonical ensemble with expected degrees below the structural cut-off, the degree of a node i with expected value $\overline{k_i}$ has a Poisson distribution across the ensemble [71]. With this fact in mind Eq. (3.45) can be understood as the logarithm of the probability of the degrees equalling their expected values under the assumption that the degrees of the nodes are independent of each other. This quantity clearly scales with the number of nodes and so this result agrees with the claim given in [71] that conjugated ensembles with extensive number of constraints are not equivalent.

Similar results have been shown for conjugated ensembles defined on more general structures such as multiplex networks [73]. In Chapter 4 our configuration model and canonical ensemble of d-dimensional simplicial complexes are conjugated microcanonial and canonical ensembles of simplicial complexes. Following the methodology developed in [67,70,73] we derive an entropy relation analogous to the one shown in Eq. (3.45) and calculate the Shannon entropies of our ensembles and the entropy of large deviation. Our ensembles enforce an extensive number of constraints and our results support the results of [67, 71, 73] in that we find that our ensembles are not equivalent due to the extensive number of constraints.

Chapter 4

The configuration model and canonical ensemble of simplicial complexes

In this chapter we examine two maximum entropy models of simplicial complexes. The research presented in this chapter was published in our paper *Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes* [33].

As discussed in chapter 3, the entropy of a probability distribution is in a sense a measure of its 'unpredictability', with high entropy corresponding to hard to predict outcomes and low entropy corresponding to distributions which return small numbers of outcomes with higher probability. Optimising probability distributions to maximise the entropy subject to some kind of constraints is a way of incorporating the information contained in the constraints without including any other 'accidental' assumptions about the distribution.

Taking this kind of approach to modelling simplicial complexes has a number of different uses, namely the identification of non-trivial or 'interesting' structure in simplicial complex data, the uncovering of correlations between structural properties of the simplicial complexes, as a way of relating the behaviours of dynamical

processes taking place on the simplicial complexes to a narrow set of structural properties, and for the reconstruction of simplicial complexes from incomplete knowledge of their structure.

The models we proposed in [33] and that we devote this chapter to are ensembles of pure d-dimensional simplicial complexes. That is, simplicial complexes consisting only of d-simplices and their faces. These ensembles are maximum entropy models of simplicial complexes based on *hard and soft constraints on the generalized degrees of the nodes*.

As we saw in chapter 2, the generalized degree of a simplex is the number of other simplices of some higher dimension d that the first simplex is a part of, while specifically for a node r, the generalized degree $k_{d,0}(r)$ is the number of simplices of dimension d that are incident to r. This could for example represent the number of size $d+1$ research collaborations a scientist is involved in, or in a protein-protein interaction network (PPIN) it could represent the number of 'complexes' made from $d+1$ proteins that a single protein plays a role in. Our models allow us to explore the implications of any given sequence of the generalized degrees of the nodes for the structure and dynamics of simplicial complexes, and provide us with a framework through which data from real simplicial complexes may be analysed.

It should be noted that for dimension $d = 1$, the generalized degree $k_{d,0}(r)$ of a node r reduces to $k(r)$ the traditional concept of degree in networks, and so in this case our models in fact reduce to the configuration model and canonical ensemble of networks. The approach we take to investigating our configuration model and canonical ensemble of simplicial complexes, mirrors the approaches that have been taken previously to investigate the original network configuration model and canonical ensemble. All of the results that we present in this chapter are for general dimension d, and the fact that for $d = 1$ these results reduce exactly to the known network results is a validation of our approach. In the remainder of this section we define our models explicitly and give an overview of the structure of this chapter.

Our models are ensembles of pure d-dimensional simplicial complexes, i.e. the

set of all possible pure d-dimensional simplicial complexes on N nodes together with an assignment of probability $P(G)$ for each simplicial complex G. This assignment of probability is chosen in order to maximise the entropy:

$$S = -\sum_{G} P(G) \ln \left(P(G) \right) \tag{4.1}$$

subject to the condition that either the generalized degrees of the nodes (for the configuration model) or the *expected* generalized degrees of the nodes (for the canonical ensemble) are constrained to a predefined sequence $\{\hat{k_r}\}$. The sum in Eq. (4.1) is over all pure d-dimensional simplicial complexes G on N nodes, or equivalently, over all adjacency tensors $\mathbf{a}$. The two sets of constraints that we maximise S subject to can be expressed as

$$\hat{k_r} = k_{d,0}(r) = \textstyle\sum_{\alpha \in \mathcal{Q}_d(N) | r \subset \alpha} a_\alpha \quad \text{for} \quad r = 1, 2, ..., N \tag{4.2}$$

for the configuration model of simplicial complexes or

$$\hat{k_r} = \overline{k_{d,0}(r)} = \textstyle\sum_{G} P(G) \sum_{\alpha \in \mathcal{Q}_d(N) | r \subset \alpha} a_\alpha \quad \text{for} \quad r = 1, 2, ..., N \tag{4.3}$$

for the case of the canonical ensemble of simplicial complexes. These ensembles are thus the least biased ensembles of simplicial complexes obeying the above constraints and so for the reasons already discussed they are useful as null models for real simplical complexes for which the generalized degrees of the nodes is known.

There are two broad approaches that we can take when using these models. One is an analytical approach where we try to derive (exact or approximate) equations relating the constraints $\{\hat{k_r}\}$ to quantities such as S, $P(G)$ or p_α (the marginal probaility of a simplex α) or to the expected values of structural properties such as the generalized degrees of simplices of dimension $\delta > 0$ (the generalized degrees of links, triangles, tetrahedra etc). The other approach is to develop algorithms that sample simplicial complexes from the ensembles, i.e. algorithms that randomly return simplicial complexes with a probability equal to the probabilities implied by the models. Probability distributions or expected values can then be estimated from

a large number of simplicial complexes sampled from the algorithm.

As we shall see in the next sections, the canonical ensemble is far easier to work with analytically than the configuration model. Similarly to the canonical ensemble in networks, the probability of a simplicial complex in our canonical ensemble can be expressed in terms of a set of variables that have their origin as Lagrangian multipliers enforcing the constraints on the generalized degrees. In 4.2.1 we demonstrate that calculating S, $P(G)$, p_α, or expected generalized degrees of higher dimension faces as functions of these Lagrangian multipliers is relatively straightforward, while in 4.2.2 we will show that for choices of the constrained generalized degrees sufficiently below the structural cut-off introduced in chapter 2 that formulae for these quantities can be derived directly in terms of the expected generalized degrees of the nodes.

In the configuration model analytical results are harder to come by. As with the network configuration model the assignment of probability that maximises Eq. (4.1) subject to our hard constraints $\{\hat{k}_r\}$ is the one that assigns the same non-zero probability to every simplicial complex obeying the constraints and probability 0 to all other simplicial complexes. For this distribution to be normalised the non-zero probability must of course be $\frac{1}{\mathcal{N}}$, with $\mathcal{N}$ being the total number of simplicial complexes that have the exact sequence of the generalized degrees of the nodes specified by the constraints. The calculations of $\mathcal{N}$ or other ensemble quantities in terms of $\{\hat{k}_r\}$ are non-trivial combinatorial problems, and so for most applications use of our configuration model must be confined to the second 'stochastic' approach. With this in mind, in 4.3.2 we propose an algorithm for generating simplicial complexes stochastically from our configuration model.

An alternative approach to exploring the configuration model concerns the relation between the configuration model and the canonical ensemble when the same constraints $\{\hat{k}_r\}$ are chosen for both. In statistical physics terminology we say these ensembles are 'conjugated', and as discussed in chapter 3 it is a common claim even in standard statistical physics textbooks that conjugated ensembles are *thermodynamically equivalent*, although it has been shown that this is not the case for ensembles

with an extensive number of constraints [67, 71–73]. Thermodynamic equivalence in our case would imply that the probability of a simplicial complex in either ensemble should be the same in the limit $N \to \infty$, and that *macrostates* such as the values of the generalized degrees of simplices of any dimension, or global properties like average path length or the prevalence of some motif in the ensembles should also be the same for the two ensembles in the large system limit. In 4.3.4 we derive an equation relating the entropies of the configuration model and canonical ensemble when they are 'conjugated' and use this relation to prove that the ensembles are in fact *not* equivalent due to the fact that we are imposing an extensive number of constraints. The canonical ensemble can therefore *not* be used as a stand-in for the configuration model. However, that does not mean that we cannot exploit results from the canonical ensemble to understand the configuration model. In 4.3.5 we use the entropy relation derived in 4.3.4 along with canonical ensemble results described in 4.2.2 to calculate the asymptotic number of simplicial complexes $\mathcal{N}$ in the configuration model in terms of $\{\hat{k}_r\}$. The formula we propose is for simplicial complexes of general dimension d and is valid for choices of the generalized degrees much smaller than the structural cut-off for simplicial complexes. For $d = 1$ this result in fact reduces to the known Canfield-Bender [77] result for the number of networks with a given degree sequence below the network structural cut-off.

The research presented in this chapter gives a full account of the configuration model and canonical ensemble of simplicial complexes, which we characterize in statistical physics terms. Simplicial complexes are the ideal objects for representing many-body interactions between the parts of real complex systems and we argue that our two ensembles have an important role to play as null models for such systems. Below we outline the structure of the remainder of this chapter.

In 4.1 we show how the structural cutoff for simplicial complexes that we stated in Eq. (2.4) of Chapter 2 is derived. This cutoff plays an important role in both our canonical ensemble and configuration model; in 4.2 we explore the canonical ensemble of simplicial complexes enforcing a given sequence of expected generalized degrees of the nodes; in 4.3 we explore using statistical mechanics methods the configuration model of simplicial complexes with given sequence of generalized degrees

of the nodes; in 4.4 we discuss the natural correlations observed in our numerical realizations of the the configuration model of simplicial complexes; finally in 4.5 we give our chapter conclusions.

4.1 Structural cutoff for simplicial complexes

In this section we derive the structural cutoff K_d for a pure d-dimensional simplicial complex, which we already presented in Eq. (2.4) of Chapter 2. This cutoff plays an important role in the calculations for the canonical ensemble and configuration model shown in the rest of this chapter.

To calculate K_d, let us start by considering a pure d-dimensional simplicial complex with given sequence of the generalised degrees of the nodes $\{k_{d,0}(i)\}_{i=1,\dots N}$. Each node i has a specified number of d-simplices that it takes part in, and we say that the simplicial complex is uncorrelated if the generalised degrees of the nodes of a randomly chosen d-simplex are independent of each other. In such a simplicial complex the conditional probability that $d+1$ *randomly chosen* nodes $i_0, \dots, i_d$ are connected by a d-simplex $\alpha = \{i_0, \dots, i_d\}$, given their generalised degrees $k_{d,0}(i_0) = k_0, \dots k_{d,0}(i_d) = k_d$ is equal to

$$Prob(a_\alpha = 1 | k_{d,0}(i_0) = k_0, \dots, k_{d,0}(i_d) = k_d) = d! \frac{k_0 \dots k_d}{(\langle k \rangle N)^d}, \tag{4.4}$$

where in this case $\langle k \rangle = \frac{1}{N} \sum_i k_{d,0}(i)$ is the average generalised degree of the nodes in the network. It is probably wise to note here that the probability mass function $Prob(\cdot)$ shown in Eq. (4.4) does not refer to ensemble probabilities in either our canonical ensemble or our configuration model but instead refers to the probability of observing in a *single given simplicial complex* a simplex between a randomly selected set of real nodes given they have the specified generalised degrees. The 'randomness' here that we are quantifying with $Prob(\cdot)$ arises in the random selection of nodes in this fixed simplicial complex, rather than in probability distributions over an ensemble of different simplicial complexes.

To understand how Eq. (4.4) is derived, imagine we randomly select a d-simplex then randomly select one of the nodes of the d-simplex in a (correlated or uncorrelated) simplicial complex. The probability that the node has generalised degree k would be $\frac{kP_{d,0}(k)}{\langle k\rangle}$ where $P_{d,0}(k)$ is the proportion of nodes i with generalised degree $k_{d,0}(i) = k$. Thus if the simplicial complex is uncorrelated then the generalised degrees of each of the nodes in the randomly chosen d-simplex are independent and so their joint probability is

$$Prob(k_{d,0}(i_0) = k_0, \ldots, k_{d,0}(i_d) = k_d | a_\alpha = 1) = \frac{\prod_{r=0}^{d} k_r P_{d,0}(k_r)}{\langle k\rangle^{d+1}}. \tag{4.5}$$

Eq. (4.4) is obtained from Eq. (4.5) by applying Bayes theorem:

$$\begin{array}{l} Prob(a_\alpha = 1 | k_{d,0}(i_0) = k_0, \ldots, k_{d,0}(i_d) = k_d) \\ = \frac{Prob(k_{d,0}(i_0)=k_0,\ldots,k_{d,0}(i_d)=k_d|a_\alpha=1)Prob(a_\alpha=1)}{Prob(k_{d,0}(i_0)=k_0,\ldots,k_{d,0}(i_d)=k_d)}. \end{array} \tag{4.6}$$

In the above equation $Prob(a_\alpha = 1)$ is the probability that a randomly chosen *possible* simplex is *actually* a simplex and so is equal to the density of simplices in the simplicial complex, in particular $Prob(a_\alpha = 1) = \frac{\frac{1}{d+1}N\langle k\rangle}{\binom{N}{d+1}} \approx d!\frac{\langle k\rangle}{N^d}$, where the approximation is valid for large N. Meanwhile in the denominator $Prob(k_{d,0}(i_0) = k_0, \ldots, k_{d,0}(i_d) = k_d)$ is the probability that $d+1$ independently chosen nodes have generalised degrees $k_{d,0}(i_0) = k_0, \ldots, k_{d,0}(i_d) = k_d$. These are clearly independent so $Prob(k_{d,0}(i_0) = k_0, \ldots, k_{d,0}(i_d) = k_d) = \prod_{r=0}^{d} P(k_r)$. Combining everything, Eq. (4.6) becomes Eq. (4.4).

As already mentioned, the structural cutoff K_d is the upper limit on the generalised degrees above which the simplicial complex must have natural correlations in its generalised degree distribution. We derive it by assuming that the nodes in Eq. (4.4) all have K_d as their generalised degree, and by setting (4.4) equal to 1. For choices of the generalised degree larger than K_d, the probability (4.4) can take nonsensical values larger than 1, indicating that it is impossible to form an uncorrelated simplicial complex without adding more than one simplex between the same set of nodes. Solving for K_d gives us our expression for the structural cutoff which

we originally stated in Eq. (2.4):

$$K_d = \left[\frac{(\langle k \rangle N)^d}{d!}\right]^{1/(d+1)} . \tag{4.7}$$

4.2 Canonical ensemble of simplicial complexes

4.2.1 Canonical ensemble with given sequence of expected generalized degree of the nodes

In this section we discuss the canonical ensemble of simplicial complexes. This model is the maximum entropy model for pure d-dimensional simplicial complexes subject to the expected generalized degrees of the nodes being constrained to a given sequence $\{\hat{k_r}\}$. The assignment of probability is the one that maximises Eq. (4.1) subject to Eq. (4.3), and can be found using Lagrangian maximisation. Following the approach we took in Section 3.2, we write the Lagrangian as

$$\mathcal{L} = S - \mu\left(\sum_G P(G) - 1\right) - \sum_{r=1}^{N} \lambda_r \left(\sum_G P(G) k_{d,0}(r) - \hat{k_r}\right), \tag{4.8}$$

where S is the entropy as given in Eq. (4.1), μ is the Lagrangian multiplier enforcing normalisation, and $\lambda_1, ..., \lambda_N$ are Lagrangian multipliers enforcing the constraints on the generalised degrees of the nodes. To find $P(G)$ we must solve the equations:

$$\frac{\partial \mathcal{L}}{\partial P(G)} = 0, \quad \frac{\partial \mathcal{L}}{\partial \mu} = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda_r} = 0, \quad r = 1, ..., N. \tag{4.9}$$

Solving the first two equations gives us the following probability in terms of the remaining Lagrangian parameters,

$$P(G) = \frac{1}{Z} e^{-\sum_r \lambda_r k_{d,0}(r)} \tag{4.10}$$

where $k_{d,0}(r) = \sum_{\alpha | r \in \alpha} a_\alpha$ is the actual generalized degree of the node r in the simplicial complex G (as opposed to its expected value $\hat{k_r}$), and

$$Z = \sum_G e^{-\sum_r \lambda_r k_{d,0}(r)} \tag{4.11}$$

normalises the distribution and is called the partition function. The sum over simplicial complexes is equivalent to a sum over possible adjacency tensors $\mathbf{a}$ and can be written

$$\sum_G = \prod_{\alpha \in \mathcal{Q}_d(N)} \sum_{a_\alpha = 0,1}, \tag{4.12}$$

i.e. the sum over all possible d-simplices. We evaluate the sum in Eq. (4.11) by factorising it over the d-simplices:

$$\begin{aligned} Z &= \prod_{\alpha \in \mathcal{Q}_d(N)} \sum_{a_\alpha = 0,1} e^{-a_\alpha \sum_{m \subset \alpha} \lambda_m} \\ &= \prod_{\alpha \in \mathcal{Q}_d(N)} \left[1 + e^{-\sum_{m \subset \alpha} \lambda_m} \right], \end{aligned} \tag{4.13}$$

where in the first line we have used

$$\sum_r \lambda_r k_{d,0}(r) = \sum_r \lambda_r \sum_{\alpha \in \mathcal{Q}_d(N) | r \subset \alpha} a_\alpha = \sum_{\alpha \in \mathcal{Q}_d(N)} a_\alpha \sum_{m \subset \alpha} \lambda_m. \tag{4.14}$$

The probability of a simplicial complex G can then be written in the form

$$P(G) = \prod_{\alpha \in \mathcal{Q}_d(N)} \left[\frac{e^{-a_\alpha \sum_{m \subset \alpha} \lambda_m}}{1 + e^{-\sum_{m \subset \alpha} \lambda_m}} \right]. \tag{4.15}$$

With $P(G)$ in this form, it is easy to calculate the marginal probability p_α of a simplex α in the ensemble:

$$p_\alpha = \sum_G a_\alpha \prod_{\alpha' \in \mathcal{Q}_d(N)} \left[\frac{e^{-a_{\alpha'} \sum_{m \subset \alpha'} \lambda_m}}{1 + e^{-\sum_{m \subset \alpha'} \lambda_m}} \right]$$
$$= \frac{e^{-\sum_{m \subset \alpha} \lambda_m}}{1 + e^{-\sum_{m \subset \alpha} \lambda_m}}. \tag{4.16}$$

The d-simplices in the canonical ensemble are thus present independently of each other with probabilities dependent only on the Lagrangian multipliers associated with the nodes they contain. The 'event' of drawing a given simplicial complex G from the canonical ensemble is a composition of the independent events given by each entry a_α in the adjacency tensor $\mathbf{a}$. The probability of a simplicial complex G is then simply the product of the probabilities of observing each event $a_\alpha = 0$ or $a_\alpha = 1$ as specified by G:

$$P(G) = \prod_{\alpha \in \mathcal{Q}_d(N)} \Big[p_\alpha^{a_\alpha} (1 - p_\alpha)^{1 - a_\alpha} \Big]. \tag{4.17}$$

The above expression in combination with Eq. (4.16) gives $P(G)$ in terms of the Lagrangian multipliers $\{\lambda_r\}$ and in a form that factorises over the individual d-simplices. Inserting $P(G)$ into the formula for the entropy given in Eq. (4.1) gives

$$S = - \sum_{\alpha \in \mathcal{Q}_d(N)} \Big[p_\alpha \ln p_\alpha + (1 - p_\alpha) \ln(1 - p_\alpha) \Big], \tag{4.18}$$

i.e. the independence of the d-simplices leads to the total entropy of the ensemble being equal to the sum of the entropies of the probability distributions associated to each a_α.

Starting with $P(G)$ in the form given by Eq. (4.17) it is possible to calculate other ensemble properties such as the expected values of the generalized degrees of faces as a function of the Lagrangian multipliers. But how can we get these properties directly in terms of our constraints on the expected generalized degrees of the nodes? From any choice of the Lagrangian multipliers we get the maximum

entropy distribution subject to the constraints that the expected generalized degree of each node is equal to *some value determined by the multipliers*, but given a desired sequence of the expected generalized degrees of the nodes $\{\hat{k_r}\}$, can we obtain the correct set of multipliers $\{\lambda_r\}$ that gives $\overline{k_{d,0}(r)} = \hat{k_r}$ for every node r? To this end, let us calculate the expected generalized degree of a node r for *given* $\{\hat{k_r}\}$. This is given by

$$\overline{k_{d,0}(r)} = \sum_{\alpha \in \mathcal{Q}_d(N) | r \subset \alpha} \frac{e^{-\sum_{m \subset \alpha} \lambda_m}}{1 + e^{-\sum_{m \subset \alpha} \lambda_m}}. \tag{4.19}$$

With the above expression we have the expected generalized degree of any node r in terms of the multipliers $\{\lambda_r\}$. The correct set of multipliers to enforce a given sequence of the expected generalised degrees can be found numerically from this expression. A different approach that we take in 4.2.2 is to simplify Eq. (4.16) by restricting $e^{-\lambda_r} \ll 1$ for all nodes r. As we shall see in 4.2.2 this restriction allows us to obtain p_α in terms of $\{\hat{k_r}\}$ and is equivalent to requiring that the generalized degrees of the nodes are much smaller than the structural cut-off for simplicial complexes.

4.2.2 The canonical ensemble of simplicial complexes with structural cutoff

In this section we make the restriction that $e^{-\lambda_r} \ll 1$ for each node r. This restriction will allow us to eliminate the Lagrangian multipliers $\{\lambda_r\}$ and write the probability of a simplex p_α as the normalised product of the expected generalised degrees of its nodes. We also show that this restriction is equivalent to the maximum generalized degree being smaller than the structural cut-off for simplicial complexes given by Eq. (2.4).

With this assumption, the probability p_α given by Eq. (4.16) can be approxi-

mated by

$$p_\alpha \simeq \prod_{r\subset\alpha,\alpha\in\mathcal{Q}_d(N)} e^{-\lambda_r}, \tag{4.20}$$

while Eq. (4.19) simplifies to

$$\hat{k_r} = e^{-\lambda_r} \sum_{\alpha\in\mathcal{Q}_d(N)|r\subset\alpha} \prod_{m\subset\alpha|m\neq r} e^{-\lambda_m}. \tag{4.21}$$

For $N \gg 1$ (and $e^{-\lambda_r} \ll 1$) the following approximation can be made

$$\sum_{\alpha\in\mathcal{Q}_d(N)|r\subset\alpha} \prod_{m\subset\alpha|m\neq r} e^{-\lambda_m} = \sum_{m_1<m_2<\ldots<m_d} \prod_{j=1}^{d} e^{-\lambda_{m_j}} \simeq \frac{1}{d!}\left(\sum_m e^{-\lambda_m}\right)^d. \tag{4.22}$$

Rearranging Eq. (4.21) and using the above approximation we obtain an expression for $e^{-\lambda_r}$ in terms of $\hat{k_r}$ and the other Lagrangian multipliers:

$$e^{-\lambda_r} = \hat{k_r}\frac{d!}{(\sum_m e^{-\lambda_m})^d}. \tag{4.23}$$

Summing over all the nodes of the simplicial complex gives

$$\sum_r e^{-\lambda_r} = \left(\langle\hat{k}\rangle N d!\right)^{1/(d+1)}. \tag{4.24}$$

Finally, by combining Eq.(5.16) and Eq.(4.24) we get an expression for the Lagrangian multiplier λ_r directly in terms of the expected generalized degrees of the nodes:

$$e^{-\lambda_r} = \hat{k_r}\left[\frac{d!}{(\langle\hat{k}\rangle N)^d}\right]^{1/(d+1)}. \tag{4.25}$$

Using this result in Eq. (4.20) we see that in the current regime the probability p_α of a simplex $\alpha \in \mathcal{Q}_d(N)$ may be expressed as a normalized product of the expected

generalized degrees of its nodes:

$$p_\alpha = d! \frac{\prod_{r \subset \alpha} \hat{k_r}}{(\langle \hat{k} \rangle N)^d}. \tag{4.26}$$

This expression is valid for $N \gg 1$ and $e^{-\lambda_r} \ll 1$. It should be noted that in this regime, Eq. (4.25) implies

$$\hat{k_r} = e^{-\lambda_r} K_d \ll K_d, \tag{4.27}$$

where K_d is the structural cutoff for a d-dimensional simplicial complex given by Eq. (2.4), and which in Section 4.1 we showed was equal to

$$K_d = \left[\frac{\left(\langle \hat{k} \rangle N \right)^d}{d!} \right]^{1/(d+1)}. \tag{4.28}$$

This regime is the regime in which there are no correlations between the generalized degrees of the nodes. Moreover in this regime only a small number of simplices of dimension $\delta < d$ can be incident to more than one d-dimensional simplex. In fact, given the expression for p_α provided by Eq. (4.26), it is possible to evaluate in this ensemble the expected generalized degree of these simplices $\overline{k_{d,\delta}(\alpha')}$ for $\delta < d$. This is given by

$$\begin{aligned} \overline{k_{d,\delta}(\alpha')} = \sum_{\alpha | \alpha' \subset \alpha} p_\alpha &= d! \frac{\prod_{r \subset \alpha'} \hat{k_r}}{(\langle \hat{k} \rangle N)^d} \sum_{\alpha | \alpha' \subset \alpha} \prod_{m \subset \alpha | m \not\subset \alpha'} \hat{k_m} \\ &= \frac{d!}{(d-\delta)!} \frac{\prod_{r \subset \alpha'} \hat{k_r}}{(\langle \hat{k} \rangle N)^\delta}, \end{aligned} \tag{4.29}$$

where in the second line we have used

$$\sum_{\alpha | \alpha' \subset \alpha} \prod_{m \subset \alpha | m \not\subset \alpha'} \hat{k_m} = \sum_{m_1 < m_2 < \ldots < m_{d-\delta}} \prod_{j=1}^{d-\delta} \hat{k_{m_j}} = \frac{1}{(d-\delta)!} (\langle \hat{k} \rangle N)^{d-\delta}. \tag{4.30}$$

Therefore only δ-dimensional groups of nodes $(m_1, ..., m_{\delta+1})$ with generalized degrees

of the nodes $\hat{k_{m_1}}, ..., \hat{k_{m_{\delta+1}}}$ scaling like $N^{\frac{\delta}{\delta+1}}$ or faster are likely to share more than one d-dimensional simplex. This implies that at least for nodes r with generalized degrees $\hat{k_r} \ll N^{\frac{\delta}{\delta+1}}$, it is very unlikely for the distinct simplices incident to r to share any nodes apart from the node r itself. Without these overlaps, the number of δ-simplices incident to r for $\delta < d$ is determined by the number of r's d-simplices:

$$k_{\delta,0}(r) \approx \binom{d}{\delta} k_{d,0}(r). \tag{4.31}$$

All of the above calculations are for simplicial complexes of general dimension d. In order to have some concrete examples, in the following two sections we discuss simplicial complexes with $d = 1$ and $d = 2$ respectively, corresponding to the case of a network ($d = 1$) and a simplicial complex constructed exclusively from triangles and their sub-faces ($d = 2$).

4.2.3 Example: The canonical ensemble of simplicial complexes of dimension $d = 1$

For $d = 1$ our canonical ensemble of simplicial complexes reduces to the canonical ensemble of networks (exponential random graph) [74] with given expected degree sequence. The probability $P(G)$ of a given 1-dimensional simplicial complex (i.e. network) specified by the adjacency tensor $\{a_{rm}\}$ is given by Eq. (4.10) and in this case reduces to

$$P(G) = \frac{1}{Z} e^{-\sum_r \lambda_r k_{1,0}(r)} \tag{4.32}$$

where $k_{1,0}(r) = \sum_m a_{rm}$ is simply the degree of node r in the network G and the partition function Z is given by

$$Z = \prod_{r<m} \left(1 + e^{-\lambda_r - \lambda_m}\right). \tag{4.33}$$

The Lagrangian multipliers λ_r are fixed by the condition

$$\hat{k_r} = \overline{k_{1,0}(r)} = \sum_m p_{rm} \tag{4.34}$$

with p_{rm} indicating the probability that the link between the nodes r, m is present in the network. The probability p_{rm} are given by

$$p_{rm} = \frac{e^{-(\lambda_r+\lambda_m)})}{1+e^{-(\lambda_r+\lambda_m)}}. \tag{4.35}$$

The probability $P(G)$ of a simplicial complex G in this canonical ensemble can be expressed as a product of the marginal probabilities for the individual links:

$$P(G) = \prod_{r<m} \left[p_{rm}^{a_{rm}} (1-p_{rm})^{1-a_{rm}} \right]. \tag{4.36}$$

Therefore the entropy S of the ensemble is given by

$$S = -\sum_{r<m} \left[p_{rm} \ln p_{rm} + (1-p_{rm}) \ln(1-p_{rm}) \right]. \tag{4.37}$$

Finally in presence of the structural cutoff on the generalized degree of the nodes, i.e. if the maximal generalized degree of the nodes K_{max} satisfies

$$K_{max} \ll K_1 = \left(\langle \hat{k} \rangle N \right)^{1/2}, \tag{4.38}$$

the probabilities p_{rm} take a simple factorized expression given by

$$p_{rm} = \frac{\hat{k_r}\ \hat{k_m}}{\langle \hat{k} \rangle N}. \tag{4.39}$$

We note here that the structural cutoff of simplicial complexes of dimension $d = 1$ given by Eq. (4.28) reduces to the structural cutoff of simple networks [75] as expected.

4.2.4 Example: The canonical ensemble of simplicial complexes of dimension $d = 2$

In this section we summarize the results for the case of a canonical ensemble of two dimensional simplicial complexes where we constrain the expected generalized degree of the nodes to be $\hat{k_r} = \overline{k_{2,0}(r)}$. The simplicial complexes in this case are constructed

exclusively from triangles and their faces (nodes and links). The generalized degrees being constrained in this case are the number of triangles incident to each node. The probability $P(G)$ of a simplicial complex G, given by Eq. (4.10) in this case becomes

$$P(G) = \frac{1}{Z} e^{-\sum_r \lambda_r k_{2,0}(r)} \tag{4.40}$$

where $k_{2,0}(r) = \sum_{m<n} a_{rmn}$ and the partition function Z is given by

$$Z = \prod_{r<m<n} \left(1 + e^{-\lambda_r - \lambda_m - \lambda_n}\right). \tag{4.41}$$

The Lagrangian multipliers λ_r are fixed by the condition

$$\hat{k_r} = \overline{k_{2,0}(r)} = \sum_{m<n} p_{rmn} \tag{4.42}$$

where p_{rmn} is the probability that the triangle between the nodes r, m, n is present in the simplicial complex, and is given by

$$p_{rmn} = \frac{e^{-(\lambda_r + \lambda_m + \lambda_n)}}{1 + e^{-(\lambda_r + \lambda_m + \lambda_n)}}. \tag{4.43}$$

The probability $P(G)$ of a simplicial complex G in this canonical ensemble can be expressed as a product of the marginal probabilities for the individual triangles:

$$P(G) = \prod_{r<m<n} \left[p_{rmn}^{a_{rmn}} (1 - p_{rmn})^{1 - a_{rmn}} \right]. \tag{4.44}$$

And so the entropy S of the ensemble in this case may be written

$$S = -\sum_{r<m<n} \left[p_{rmn} \ln p_{rmn} + (1 - p_{rmn}) \ln(1 - p_{rmn}) \right]. \tag{4.45}$$

Finally in presence of the structural cutoff on the generalized degree of the nodes, i.e. if the maximal generalized degree of the nodes K_{max} satisfies

$$K_{max} \ll K_2 = \left(\frac{\langle \hat{k} \rangle N}{\sqrt{2}} \right)^{2/3} \tag{4.46}$$

the probabilities p_{rmn} take a simple factorized expression given by

$$p_{rmn} = 2 \frac{\hat{k_r}\ \hat{k_m}\ \hat{k_n}}{(\langle \hat{k} \rangle N)^2}. \tag{4.47}$$

Here the structural cutoff K_2 scales like $N^{2/3}$. It is therefore much larger than the structural cutoff for simple networks.

We note that this model is similar to the model of tagged social networks represented by hypergraphs presented in Ref. [32, 76]. Nevertheless it differs with respect to the cited work because in this work the three nodes linked in a given 2-dimensional simplex represent the same type of nodes. This difference is responsible for the factor two present in the right hand side of Eq. (4.47).

4.2.5 Generation of simplicial complexes by the canonical ensemble

As we have seen in Sections 4.2.1 and 4.2.2, the d-simplices in the canonical ensemble are present with independent probabilities that are determined by the expected generalized degrees of the nodes. Generating a simplicial complex from the ensemble is thus simply a matter of sampling every possible d-simplex with the appropriate probability. We propose the following algorithm for sampling from the canonical ensemble:

(a) Calculate the probabilities p_α of any d-dimensional simplex $\alpha \in \mathcal{Q}_d(N)$ given by Eq. (4.16) in absence of the structural cutoff K_d or by Eq. (4.26) in presence of the structural cutoff K_d.

(b) Draw every possible d-dimensional simplex $\alpha \in \mathcal{Q}_d(N)$ with probability p_α.

4.3 The configuration model of simplicial complexes

4.3.1 The configuration model of simplicial complexes with given generalized degree of the nodes

In this section we discuss the configuration model of simplicial complexes. This is the maximum entropy ensemble of simplicial complexes satisfying hard constraints on the generalized degrees of the nodes, i.e. the set of all pure d-dimensional simplicial complexes together with the assignment of probability that maximises Eq. (4.1) subject to the constraint that for any simplicial complex G not satisfying the constraints, $P(G)$ is equal to zero. This assignment of probability is constant on the set of simplicial complexes obeying the constraints and zero everywhere else. The probability of a simplicial complex G is thus

$$P(G) = \frac{1}{\mathcal{N}} \prod_r \delta\left(\hat{k_r}, k_{d,0}(r)\right), \tag{4.48}$$

where

$$\mathcal{N} = \sum_G \prod_r \delta\left(\hat{k_r}, k_{d,0}(r)\right) \tag{4.49}$$

is the total number of simplicial complexes obeying the constraints. Calculation of $\mathcal{N}$ is non-trivial. In the network configuration model (or equivalently our model with $d = 1$) an asymptotic expression for $\mathcal{N}$ is known only when the maximum degree is less than the structural cut-off [67,77]. In 4.3.5 we calculate an asymptotic expression for the total number of simplicial complexes in the configuration model $\mathcal{N}$, for sequences of the generalized degrees of the nodes below the structural cut-off for simplicial complexes K_d. This result is for general dimension d, and for $d = 1$ reduces to the result for the network configuration model found in [67].

Below the structural cut-off K_d the generalized degrees of the nodes are uncorrelated in the configuration model. This means that the probability of a set of $d+1$ nodes being connected by a d-simplex is proportional to the product of the generalized degrees of the nodes, i.e.

$$p_\alpha = d! \frac{\prod_{r \subset \alpha} \hat{k_r}}{(\langle \hat{k} \rangle N)^d}. \tag{4.50}$$

This marginal probability of a simplex is the same as the marginal probability of a simplex in the canonical ensemble below the structural cut-off. Unlike in the canonical ensemble these simplex probabilities are not independent as there must be exactly $\hat{k_r}$ simplices incident to a node r. Whether or not this dependence introduced by the hard constraints results in the two ensembles having significant differences between them for large N is the topic of the Section 4.3.4. In this section we derive an equation relating the entropies of the two ensembles and show that their difference scales like N below the structural cut-off indicating that even in the uncorrelated regime, the configuration model and canonical ensemble are not equivalent.

In order to explore the configuration model numerically we developed an algorithm to sample simplicial complexes from the ensemble. This algorithm generalizes the 'stub matching' algorithm commonly used to sample from the network configuration model. We present this algorithm in 4.3.2, as well as a modified version that is faster than the primary algorithm but that introduces small biases into the sampling process that we argue are likely to be negligible. In 4.4 we use this algorithm to explore the natural correlations arising in the configuration model above the structural cut-off.

4.3.2 Generation of the simplicial complexes by the configuration model

In this section we generalize the algorithm for the configuration model of networks with given degree sequence to the configuration model of d-dimensional simplicial

complexes with given sequence $\{\hat{k_r}\}_{r\leq N}$ of the generalized degrees of the nodes. It should be noted that for some sequences of the generalized degrees it is not possible to construct even a single simplicial complex. These sequences are called ungraphical. A graphical sequence is one for which at least one simplicial complex can be produced. For simple networks, i.e. for simplicial complexes of dimension $d = 1$, the conditions that a degree sequence must satisfy in order to be graphical have been fully identified [65, 78]. For simplicial complexes we know that the generalized degree of the nodes must satisfy

$$\sum_{r=1}^{N} k_{d,0}(r) = (d+1)M, \tag{4.51}$$

where M is the total number of d-simplices. In practice, it will often be useful to start from sequences of generalized degree of the nodes occurring in real datasets which are by definition graphical. In Figure 4.1 we show how from a given graphical sequence of generalized degree of the nodes it is possible in general to construct different simplicial complexes.

The algorithm we propose to sample from the configuration model of simplicial complexes generalizes the 'stub matching' process used to construct networks in the network configuration model. As discussed in Chapter 3, in the network process, each node is initially assigned a number of stubs equal to its degree. These stubs are then randomly matched together until a fully matched network is obtained. Of course, not all matchings produce networks with no loops, and no multi-edges, so for this reason these 'illegal' matchings are disallowed. In a naive construction of the algorithm, we might instruct the algorithm to randomly draw another pair of stubs when it encounters an illegal move, so that all of the pairs matched up to that point in the matching process are left unchanged. However, this kind of approach may lead to a bias in the algorithm so that not all networks are selected with equal probability. Roughly speaking, this is because legal configurations that are in some sense 'close' in structure to many different illegal configurations are more likely to be produced than those which are close in structure to fewer illegal configurations. In order to ensure that the construction algorithm samples with equal probability all networks with the given degree sequence, when an illegal matching is encountered

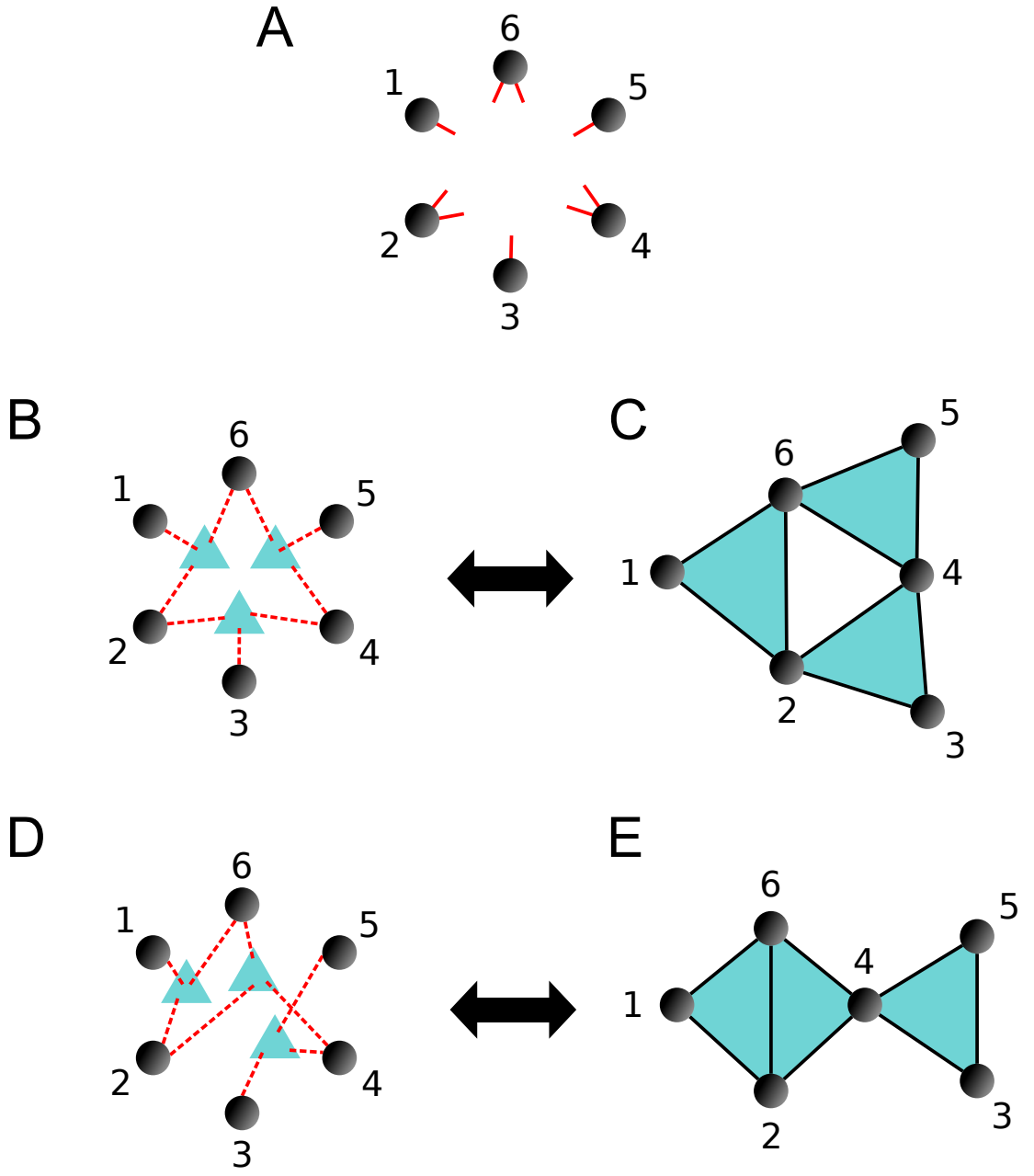

Figure 4.1: The figure shows the construction of two different $d = 2$ dimensional simplicial complexes belonging to the same configuration model of simplicial complexes. In panel A the $N = 6$ nodes are shown together with stubs indicating their generalized degree. In panel B triples of stubs are matched together to form 2-dimensional simplices. In panel C the corresponding simplicial complex is visualized. In panels D-E a different matching of the stubs is shown together with its corresponding simplicial complex. As is evident from the figure, a given generalized degree sequence of the nodes can give rise to different simplicial complexes. The logarithm of the total number $\mathcal{N}$ of simplicial complexes that can be constructed from a given generalized degree sequence of the nodes, is the Gibbs entropy Σ of the configuration model.

the algorithm must instead restart from the beginning with all stubs unmatched. Our algorithm for sampling from the configuration model of simplicial complexes generalizes the approach described above. Nodes are assigned stubs according to their generalized degree, and these stubs are randomly grouped into d-simplices formed from $d+1$ stubs. To conceptualize this we use a set of M auxiliary *factor nodes* $\mu = 1, 2 \ldots, M$ corresponding to the d-simplices where M is the total number of d-simplices and is found from the generalized degree sequence using

$$\sum_{r=1}^{N} \hat{k_r} = (d+1)M. \tag{4.52}$$

Every d-simplex is incident to exactly $d+1$ nodes, thus each factor node has $d+1$ stubs. The simplicial complex is then constructed by randomly pairing the stubs of nodes with the stubs of the factor nodes. The algorithm proceeds as follows:

(i) Initially, $\hat{k_r}$ stubs are assigned to each node $r = 1, 2, \ldots, N$, and $d+1$ stubs are assigned to each auxiliary factor node $\mu = 1, 2, \ldots M$. Initially these stubs are all unmatched.

(ii) A set of $d+1$ unmatched random stubs of the nodes is chosen with uniform probability. The nodes that these stubs belong to we label here as $(r_1, r_2, \ldots, r_{d+1})$.

(iii) If the nodes $(r_1, r_2, \ldots, r_{d+1})$ are all distinct and no factor node μ is already matched with the set of nodes $(r_1, r_2, \ldots, r_{d+1})$, we match the $d+1$ stubs of an unmatched random factor node to the nodes $(r_1, r_2, \ldots, r_{d+1})$. Otherwise we start again from Step (i).

(iv) If all the stubs are matched we construct the simplicial complex by placing a d-simplex between the nodes connected to each auxiliary factor node.

In Figure 4.2 we show an example of a possible matching of the stubs of nodes and factor nodes and the consequent construction of a simplicial complex.

The step (iii) rejects moves that are forbidden. These moves are described in Figure 4.3. As we have discussed already, this rejection procedure is necessary to

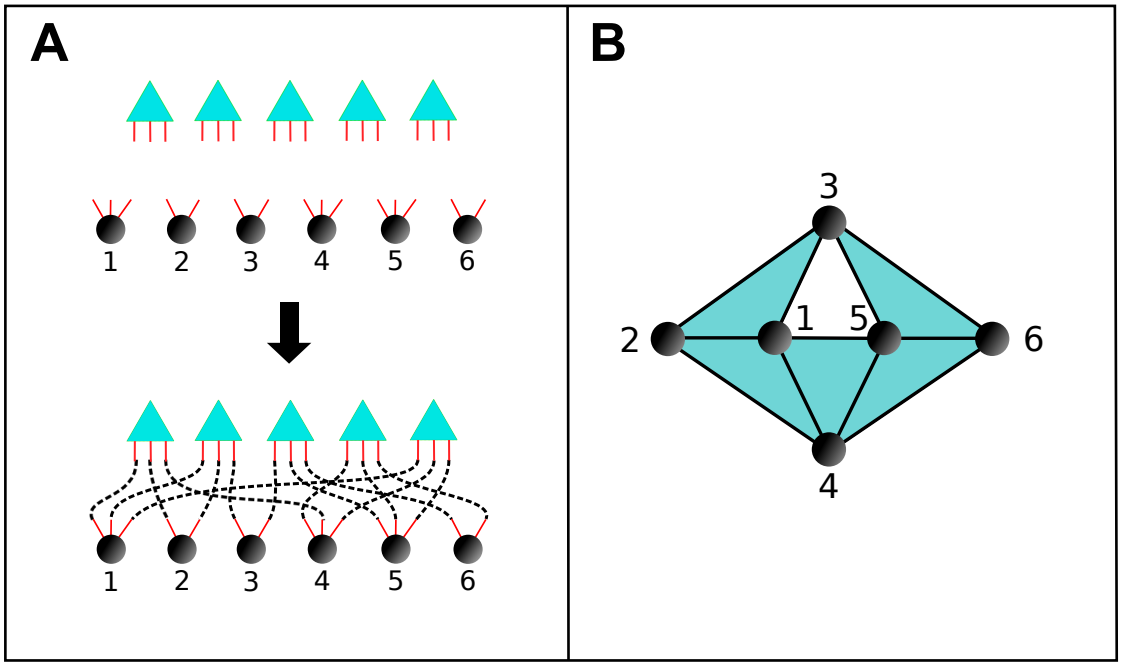

Figure 4.2: A scheme representing the algorithm for the construction of the configuration model is shown for the case $d = 2$. Panel A represents the Steps (i)-(ii)-(iii). To each node r with $r = 1, 2 \ldots, N = 6$ we assign $\hat{k_r}$ stubs. The nodes are represented with black circles. A set of M auxiliary factor nodes (cyan triangles) is considered. Each factor node has $d+1$ stubs. Subsequently an allowed matching of the stubs is found. Panel B shows how from the matching of the stubs we can construct a simplicial complex by adding a simplex between all of the nodes connected to a common factor node in panel A.

ensure that there are no spurious correlations in the structure of the simplicial complex, however for broad distribution of the generalized degrees of the nodes it might significantly slow down the algorithm.

In the context of the configuration model, more sophisticated algorithms have been proposed in Ref. [65, 79] and we believe that along these lines it could also be possible to optimize the code for the case of simplicial complexes in the future.

Here, when numerically implementing the algorithm, we have chosen to allow a rejection of a small number n_F of forbidden moves. Therefore we have modified the

above algorithm by substituting step (*iii*) with :

(iii)-a If the nodes $(r_1, r_2, \ldots, r_{d+1})$ are all distinct and no factor node μ is already matched with the set of nodes $(r_1, r_2, \ldots, r_{d+1})$, we match the $d+1$ stubs of an unmatched random factor node to the nodes $(r_1, r_2, \ldots, r_{d+1})$.

(iii)-b If the nodes $(r_1, r_2, \ldots, r_{d+1})$ are not all distinct or a factor node μ is already matched with the set of nodes $(r_1, r_2, \ldots, r_{d+1})$ we update a variable n_x that counts how many similar events have occurred so far. If $n_x \leq n_F$ we do not accept the move and we go back to Step (ii), if $n_x > n_F$ we go back to the initial Step (i).

This algorithm reduces to the one described before when $n_F = 1$, and when $n_F \ll N$ it speeds up the code significantly, without significantly altering the properties of the simplicial complexes.

4.3.3 Relation with bipartite network models

In this section we discuss how simplicial complexes may be related to bipartite networks by identifying the set of simplices with one of the two sets of nodes which we here call the 'factor nodes'.

In fact, the algorithm for sampling from the configuration model of simplicial complexes outlined in the previous Section 4.3.2 is equivalent to a bipartite network configuration model, with the added restriction that no two factor nodes can be connected to the exact same set of nodes.

Bipartite networks are formed by a set of nodes $r = 1, 2, \ldots, N$ and a set of factor nodes (or groups) $\mu = 1, 2, \ldots, P$ where a link may only exist between a node and a factor node and there are no links between two nodes or between two factor nodes. The adjacency matrix $\mathbf{A}$ of a bipartite network has elements $A_{r,\mu} = 1$ if the node r belongs to group μ, and $A_{r,\mu} = 0$ otherwise.

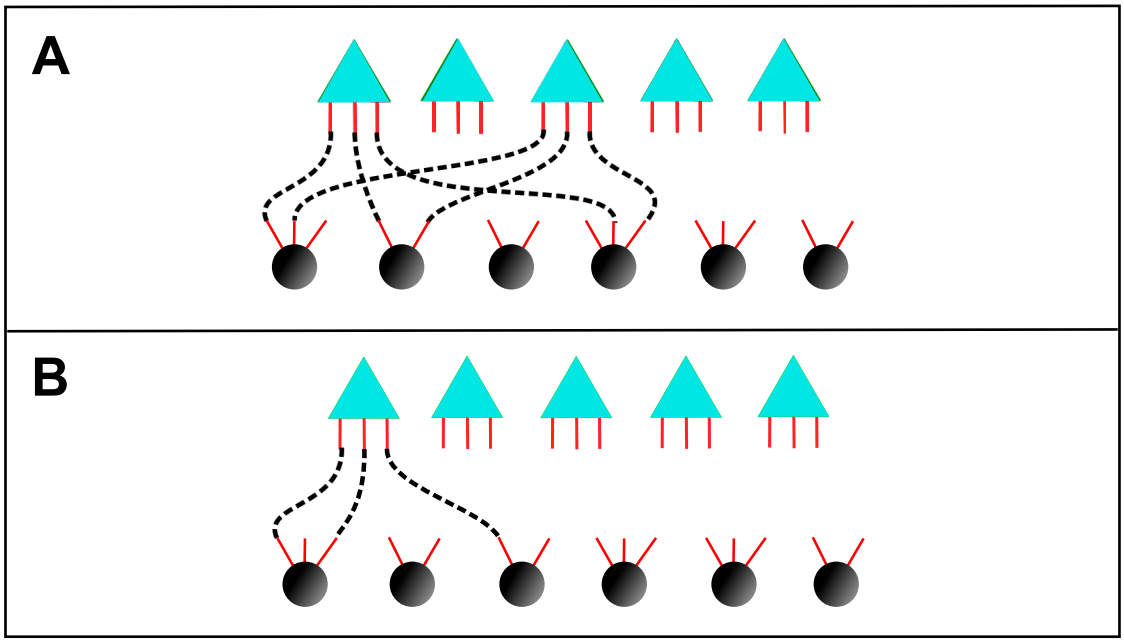

Figure 4.3: Two examples of forbidden moves are shown. In panel A the same set of nodes $(r_1, r_2, \ldots, r_{d+1})$ is selected more than once to form a simplex. In panel B the set of nodes $(r_1, r_2, \ldots, r_{d+1})$ selected to form a simplex is not formed by $d+1$ distinct nodes. Here the forbidden moves are shown for the configuration model of simplicial complexes of dimension $d = 2$.

As an example a bipartite network could be used to describe a network formed by scientists (the nodes) and by scientific papers (the groups) where links between scientists and papers indicate the author of a paper. Similar models have been proposed for social networks [15] and for immune networks [80].

We can map bipartite networks to simplicial complexes by mapping each factor node to a simplex between the nodes that the factor node is connected to. A factor node with degree $d+1$ then corresponds to a d dimensional simplex.

A bipartite configuration model where all of the factor nodes have degree $d+1$ and the degrees of the nodes are specified by some degree sequence $\{\hat{k_r}\}$, is similar to our configuration model of simplicial complexes where the generalized degrees are given by $\{\hat{k_r}\}$, but with the following differences:

(1) In the bipartite network more than one factor node can connect the same set of nodes.

(2) In the bipartite network the factor nodes are labelled.

An example illustrating these differences is that of a bipartite network between authors and papers co-authored by three authors and the corresponding simplicial complex describing the collaboration network between the authors. The difference between these two datasets is that bipartite networks distinguish between situations where three authors write only one or several papers together, and they also distinguish between papers with the same three authors (i.e. the papers are labelled). In contrast simplicial complexes indicate only whether a given set of three authors have co-authored at least one paper together, independently on the paper title and content.

4.3.4 Canonical ensemble conjugated to the configuration model of simplicial complexes

When we choose the same set of constraints $\{\hat{k_r}\}$ for both the configuration model and the canonical ensemble then we say that they are *conjugated ensembles*. This

is terminology that we have borrowed from statistical mechanics that is used to describe ensembles of dynamical systems either with a given energy (the micro-canonical ensemble corresponding to our configuration model) or with a given expected energy (the canonical ensemble).

As we discussed in Chapter 3, in many such systems the two ensembles are thermodynamically equivalent, i.e. their statistical properties are the same when one considers systems formed by a large number of particles, like for example a gas of molecules.

For network ensembles, the most fundamental example of conjugated micro-canonical and canonical ensembles are the Erdös-Renyi random graphs in which we fix either the total number of links (the micro-canonical ensemble) or the expected number of links (the canonical ensemble). Analogously to the example of a gas with fixed energy or fixed average energy, these network ensembles are equivalent in the thermodynamic limit, i.e. the limit as the number of nodes N tends to infinity. However, as we saw in Chapter 3 for examples of conjugated network ensembles with an extensive number of constraints (such as on the degree sequence) it has been shown that this equivalence does not hold [67, 71–73].

In this section we show that the configuration model of simplicial complexes and its conjugated canonical ensemble of simplicial complexes are not asymptotically equivalent. To do this we derive a relation between the entropies of the two ensembles and show that the difference between these entropies scales like the number of nodes. This implies that the two ensembles must have significantly different statistical properties in the thermodynamic limit and so they are not thermodynamically equivalent.

The calculation of the entropy relation is also useful because it allows us to use results obtained for the canonical ensemble to calculate the entropy of the configuration model explicitly in terms of its constraints $\{\hat{k_r}\}$, as long as we are in the presence of the structural cutoff.

The methodology we use to calculate the entropy relation mirrors that used in [67] to calculate similar relations for ensembles of simple networks. Our calculation of the entropy of large deviation is based on the calculations for networks given in [70] and later for multiplex networks in [73]. For $d = 1$ our models reduce to the network configuration model and canonical ensemble, and in this case our results reduce exactly to the results given in [67, 70].

In what follows we indicate the entropy of the configuration model with S_{CM} in order to distinguish it from the entropy of the canonical ensemble which we indicate with S_{CE}. Similarly we indicate the probabilities of a simplicial complex G in the configuration model and canonical ensemble with $P_{CM}(G)$ and $P_{CE}(G)$ respectively.

The expression that we derive relating these these two entropies is

$$S_{CM} = S_{CE} - \Omega. \tag{4.53}$$

In the above equation, Ω is the difference between the entropies of the two models, and below we shall show that it is in fact something known as the *entropy of large deviation* [70]. This quantity is the logarithm of the probability that in the canonical network model with expected generalized degree sequence $\{\hat{k_r}\}$, the generalized degrees of the nodes take exactly the values $k_{d,0}(r) = \hat{k_r}$. This is expressed as

$$\Omega = -\ln \left[\sum_{G} P_{CE}(G) \prod_{r} \delta \left(k_r, k_{d,0}(r) \right) \right]. \tag{4.54}$$

Large values of Ω indicate a high probability of simplicial complexes which have generalized degrees different to the constraints, while lower values indicate that the simplicial complexes are more likely to obey the constraints, and that the two ensembles are therefore more similar.

Following similar reasoning to [67] we now show that Eq. (4.53) holds with Ω given by Eq. (4.54). Firstly, we observe that in the canonical ensemble the

probability of a simplicial complex (given in Eq. (4.10) of Section 4.2.1) is

$$P(G) = \frac{1}{Z} e^{-\sum_r \lambda_r k_{d,0}(r)}, \tag{4.55}$$

and that this probability is constant on sets of simplicial complexes which have identical sequences of the generalized degrees of the nodes. The sum in (4.54) is actually the sum of the probabilities of all simplicial complexes for which the generalised degrees are exactly equal to $\{\hat{k_r}\}$ and so is simply equal to $\mathcal{N} P_{CE}(G^*)$, where G^* is any simplicial complex with generalised degrees equal to $\{\hat{k_r}\}$ and $\mathcal{N}$ is the total number of such simplicial complexes, defined earlier in Eq. (4.49). Eq. (4.54) thus becomes

$$\Omega = -\ln\left[\mathcal{N} P_{CE}(G^*)\right]. \tag{4.56}$$

To validate Eq. (4.53) we need to show that

$$S_{CE} - S_{CM} = -\ln\left[\mathcal{N} P_{CE}(G^*)\right]. \tag{4.57}$$

In particular we shall show that $S_{CE} = -\ln P_{CE}(G^*)$ and $S_{CM} = \ln \mathcal{N}$. For the configuration model, the entropy is

$$\begin{aligned} S_{CM} &= -\sum_G P_{CM}(G) \ln\left(P_{CM}(G)\right) && (4.58) \\ &= -\sum_G \frac{1}{\mathcal{N}} \prod_r \delta\left(\hat{k_r}, k_{d,0}(r)\right) \ln\left[\frac{1}{\mathcal{N}} \prod_r \delta\left(\hat{k_r}, k_{d,0}(r)\right)\right]. && (4.59) \end{aligned}$$

Note that in general, events with probability 0 are taken to contribute 0 to the Shannon entropy, i.e we use the convention $0 \ln 0 = 0$. This convention is justified by the fact that $\lim_{x\to 0} x \ln x = 0$, and because events which occur with probability 0 are impossible and so should contribute 0 to the entropy which as discussed in Chapter 3 is the 'average information' gained from sampling from the ensemble. Therefore, in the sum in Eq. (4.59) simplicial complexes with generalized degrees different to $\{\hat{k_r}\}$ contribute 0 to the entropy while simplicial complexes with generalized degrees equal to $\{\hat{k_r}\}$ each contribute $-\frac{1}{\mathcal{N}} \ln\left(\frac{1}{\mathcal{N}}\right)$. The total number of such simplicial complexes

is of course $\mathcal{N}$ so the entropy of the configuration model is simply equal to

$$S_{CM} = \ln \mathcal{N}. \tag{4.60}$$

The entropy of the canonical ensemble S_{CE} is given by

$$S_{CE} = -\sum_{G} P_{CE}(G) \ln \left(P_{CE}(G) \right). \tag{4.61}$$

Replacing the second instance of the probability $P_{CE}(G)$ with its form given in Eq. (4.55) we get

$$\begin{aligned} S_{CE} &= -\sum_{G} P_{CE}(G) \left[\ln \left(e^{-\sum_r \lambda_r k_{d,0}(r)} \right) - \ln \mathcal{Z} \right] && (4.62) \\ &= \sum_{r} \lambda_r \langle k_{d,0}(r) \rangle_{CE} + \ln \mathcal{Z}, && (4.63) \end{aligned}$$

where $\langle k_{d,0}(r) \rangle_{CE}$ indicates the expected value of the generalized degree of a node r in the canonical ensemble. This is of course just equal to $\hat{k_r}$, so that we have

$$\begin{aligned} S_{CE} &= \sum_{r} \lambda_r \hat{k_r} + \ln \mathcal{Z} && (4.64) \\ &= -\ln \left[\frac{e^{-\sum_r \lambda_r \hat{k_r}}}{\mathcal{Z}} \right]. && (4.65) \end{aligned}$$

The probability in the canonical ensemble of some simplicial complex G^* with sequence of the generalized degrees of the nodes equal to $\{\hat{k_r}\}$ is

$$P_{CE}(G^*) = \frac{e^{-\sum_r \lambda_r \hat{k_r}}}{\mathcal{Z}}, \tag{4.66}$$

so we have shown that $S_{CE} = -\ln \left[P_{CE}(G^*) \right]$ and thus our entropy relation given in Eq. (4.56) holds because

$$S_{CE} - S_{CM} = -\ln P_{CE}(G^*) - \ln \mathcal{N} = -\ln \left[\mathcal{N} P_{CE}(G^*) \right] = \Omega. \tag{4.67}$$

The entropy of large deviation is equal to the difference between the Shannon

entropies of our two conjugated ensembles. These entropies have an interpretation as the typical number of simplicial complexes in the ensembles, and so a large Ω would indicate that the canonical ensemble assigns significant probability to simplicial complexes not consistent with the configuration model, and so the canonical ensemble in a sense would have 'more' simplicial complexes than the configuration model. If Ω remains significant in the thermodynamic limit, i.e. as the number of nodes N tends to infinity, then our ensembles cannot be thermodynamically equivalent.

Naturally, the entropies of our ensembles will depend on our choice of constraints $\{\hat{k_r}\}$, and so we would like to obtain explicit expressions for them in terms of $\{\hat{k_r}\}$. For the canonical ensemble we have already achieved this in Section 4.2.2 for choices of $\{\hat{k_r}\}$ below the structural cutoff. For the entropy of large deviation Ω we follow the methodology developed in [70] and use the saddle point approximation to derive an expression in terms of $\{\hat{k_r}\}$ below the structural cutoff. The details of this calculation are given in Appendix A. We find that Ω may be written as

$$\Omega = -\sum_{r=1}^{N} \ln\left[\pi_{\hat{k_r}}(\hat{k_r})\right] \tag{4.68}$$

where $\pi_{\hat{k_r}}(\hat{k_r})$ is the Poisson distribution with average $\hat{k_r}$ evaluated at $\hat{k_r}$, i.e.

$$\pi_{\hat{k_r}}(\hat{k_r}) = \frac{1}{\hat{k_r}!}\hat{k_r}^{\hat{k_r}} e^{-\hat{k_r}}. \tag{4.69}$$

This expression is easily interpreted. In fact in the canonical ensemble the generalized degree of each node follows a Poisson distribution with average $\hat{k_r}$ (see Appendix B for details). The probability that the generalized degree of a node r takes exactly the value $k_{d,0}(r) = \hat{k_r}$ is given by $\pi_{\hat{k_r}}(\hat{k_r})$. From Eq. (4.68) we can infer that below the structural cutoff, the generalized degrees of the nodes are approximately independent from each other (in fact to show this properly requires only a trivial modification of the calculation in Appendix A so that we are calculating the probability of an arbitrary sequence of the generalized degrees rather than the sequence corresponding to expected generalized degrees). Our results in Eq.s (4.68) and (4.69) are validated by the fact that for $d = 1$ they reduce to the results for

networks given in [67, 70].

Finally, in order to assess whether or not the two ensembles are equivalent in the thermodynamic limit we look at how Ω scales with the number of nodes. Eq. (4.68) is valid for choices of the constraints $\{\hat{k_r}\}$ much smaller than the structural cutoff which scales like $(\langle\hat{k}\rangle N)^{\frac{d}{d+1}}$. For choices of the generalized degree in this range that are at least constant (i.e. they don't vanish in the limit $N \to \infty$) the entropy of large deviation Ω must be at least $\mathcal{O}(N)$. This difference between the entropies of the configuration model and canonical ensemble therefore grows at least as fast as the number of nodes and so the two ensembles are not equivalent. This result is thus in agreement with the findings of [67, 71–73], namely that extensive numbers of constraints cause conjugated ensembles not to be equivalent in the thermodynamic limit.

4.3.5 The asymptotic formula for the number of simplicial complexes in the configuration model with structural cutoff

The configuration model entropy S_{CM} can be evaluated below the structural cutoff using the entropy relation given by Eq. (4.53) together with Ω and the canonical ensemble entropy S_{CE} given by Eq.s (4.68) and (4.18) respectively, giving us

$$\begin{aligned} S_{CM} &= -\sum_{\alpha\in\mathcal{Q}_d(N)} [p_\alpha \ln p_\alpha + (1-p_\alpha)\ln(1-p_\alpha)] \\ &\quad + \sum_{r=1}^{N} \ln \frac{\hat{k_r}^{\hat{k_r}} e^{-\hat{k_r}}}{\hat{k_r}!}, \end{aligned} \tag{4.70}$$

where in presence of the structural cutoff the probabilities p_α are given by Eq. (4.26). Substituting the expression for p_α into Eq. (4.70) we get the asymptotic expression

for the configuration model entropy,

$$S_{CM} = \frac{d}{d+1}\ln(\langle\hat{k}\rangle N)! - \sum_{r=1}^{N}\ln\hat{k_r}! - \frac{\langle\hat{k}\rangle N}{d+1}\ln d! - \frac{d!}{2(d+1)(\langle\hat{k}\rangle N)^{d-1}}\left(\frac{\langle\hat{k}^2\rangle}{\langle\hat{k}\rangle}\right)^{d+1}. \quad (4.71)$$

As we saw in the previous section, the entropy of the configuration model is just the logarithm of $\mathcal{N}$, the total number of simplicial complexes in the model. Therefore, we get the asymptotic expression

$$\mathcal{N} = \frac{[(\langle\hat{k}\rangle N)!]^{d/(d+1)}}{\prod_{r=1}^{N}\hat{k_r}!}\frac{1}{(d!)^{\langle\hat{k}\rangle N/(d+1)}} \times \exp\left[-\frac{d!}{2(d+1)(\langle\hat{k}\rangle N)^{d-1}}\left(\frac{\langle\hat{k}^2\rangle}{\langle\hat{k}\rangle}\right)^{d+1} + \mathcal{O}(\ln N)\right]. \quad (4.72)$$

This expression in fact generalises the Canfield-Bender formula [77] for the ensemble of networks with given degree sequence, and for $d = 1$ it reduces exactly to the Canfield-Bender formula.

An interesting observation is that the asymptotic number $\mathcal{N}$ of simplicial complexes in the configuration model depends on the distribution of the generalized degrees of the nodes and that this dependency remains important even for generalized degree sequences with the same average $\langle\hat{k}\rangle$. This shows that the complexity of the ensemble depends strongly on the statistical properties of the generalized degree sequence. As observed in Ref. [81] in the context of simple networks, it can also be shown for simplicial complexes of dimension $d > 1$ that scale-free distributions of generalized degrees with the same average $\langle\hat{k}\rangle$ but with decreasing power-law exponent γ correspond to more complex ensembles of simplicial complexes. In fact they are characterized by a smaller entropy S_{CM} and a smaller asymptotic number $\mathcal{N}$ of simplicial complexes.

4.3.6 Combinatorial arguments for Eq. (4.72)

The asymptotic combinatorial expression (Eq. (4.72)) can be explained using combinatorial arguments, similar to the ones used to explain the Canfield-Bender formula in Ref. [82]. In fact the factor

$$\frac{[(\langle\hat{k}\rangle N)!]^{d/(d+1)}}{\prod_{r=1}^{N}\hat{k_r}!}\frac{1}{(d!)^{\langle\hat{k}\rangle N/(d+1)}} \tag{4.73}$$

counts all the possible combinations of the stubs of the nodes in groups of $d+1$ stubs when we disregard forbidden moves. In other words Eq. (4.73) counts all the possible structures that can be produced from the algorithm described in Section 4.3.2, if forbidden moves were allowed. Allowing forbidden moves means that not all of the structures produced are true simplicial complexes, as the stubs of a single node can appear more than once in a simplex, and multiple simplices can have the same set of stubs. The quantity in Eq. (4.73) is therefore larger than the total number of simplicial complexes that can be produced, and the exponential term in Eq. (4.72) can be interpreted as the term that corrects for the forbidden matchings.

We now justify the above claims by considering our algorithm when forbidden moves are allowed. In this version of the algorithm, we randomly draw stubs from the set of all of the stubs belonging to the nodes, and match them together into simplices of dimension d. As forbidden moves are allowed, this is equivalent to randomly drawing an ordered sequence of all of the stubs, and partitioning the sequence into 'blocks' of $d+1$ stubs, which become the d-simplices (see Fig. 4.4(A-C) for an illustration of this concept).

The total number of permutations of the stubs is

$$(\langle\hat{k}\rangle N)!, \tag{4.74}$$

and from each of these permutations we obtain a simplicial complex using the above approach. However, multiple permutations can produce the same simplicial complex, so to obtain the factor given in Eq. (4.73) we will need to calculate the multiplicity

of a single simplicial complex in the set of all stub permutations.

To explain how we do this, we start by making a few observations. First we note the quantity in Eq. (4.74) assumes that the stubs are 'distinuishable' from each other whereas in fact, the stubs belonging to a node should be indistinguishable from each other, so that permutations that leave nodes connected to the *same simplices* through *different stubs* should be counted as the same simplicial complex (an illustration of this concept is in Fig. 4.4(D)). Secondly the order in which stubs appear in a single block does not affect the simplex produced by the block (see Fig. 4.4(E)), so permutations with different orderings of the stubs but that nonetheless have the same set of stubs in each block produce the same simplicial complex (this is equivalent to saying that the stubs belonging to each factor node are indistinguishable). Lastly, we note that the simplices are unlabelled, and so are distinguished only by the nodes they connect to. This means that permuting the order of the blocks has no effect on the simplicial complex produced (see Fig. 4.4(F)).

From these observations it follows that the total number of simplicial complexes produced with given sequence of the generalized degrees and when forbidden moves are allowed is

$$\frac{(\langle\hat{k}\rangle N)!}{m_I m_{II} m_{III}}, \tag{4.75}$$

where

$$m_I = \prod_r \hat{k_r}! \tag{4.76}$$

accounts for the indistinguishability of the stubs of each node,

$$m_{II} = [(d+1)!]^{\frac{1}{d+1}\langle\hat{k}\rangle N} \tag{4.77}$$

accounts for the fact that stub order within blocks has no effect on the simplicial complex produced (recall that the total number of simplices or blocks is $\frac{1}{d+1}\langle\hat{k}\rangle N$),

and

$$m_{III} = \left(\frac{1}{d+1}\langle\hat{k}\rangle N\right)! \tag{4.78}$$

is the number of permutations of the blocks and accounts for the simplices being unlabelled. The quantity given in Eq. (4.78) can also be written in terms of a *multifactorial*:

$$\begin{aligned} m_{III} &= \frac{1}{(d+1)^{\frac{1}{d+1}\langle\hat{k}\rangle N}}\langle\hat{k}\rangle N\Big[\langle\hat{k}\rangle N-(d+1)\Big]\Big[\langle\hat{k}\rangle N-2(d+1)\Big]\ldots\Big[d+1\Big] \\ &= \frac{1}{(d+1)^{\frac{1}{d+1}\langle\hat{k}\rangle N}}\big(\langle\hat{k}\rangle N\big)!^{(d+1)}, \end{aligned}$$

where the notation $x!^{(n)}$ indicates the multifactorial

$$x!^{(n)} = \begin{cases} x & \text{if} \quad 0 < x \leq n \\ x\left[(x-n)!^{(n)}\right] & \text{if} \quad x > n \end{cases}.$$

We note that the multifactorial in Eq. (4.79) may be related to the single factorial by

$$\big(\langle\hat{k}\rangle N\big)! = \prod_{s=0}^{d}\big(\langle\hat{k}\rangle N - s\big)!^{(d+1)}, \tag{4.79}$$

and that for $N \gg d+1$ we can make the approximation $\big(\langle\hat{k}\rangle N\big)!^{(d+1)} \approx \big(\langle\hat{k}\rangle N - s\big)!^{(d+1)}$ for all $s = 1, 2, \ldots d$, so that we can write Eq. (4.79) as

$$m_{III} = \frac{1}{(d+1)^{\frac{1}{d+1}\langle\hat{k}\rangle N}}\Big[\big(\langle\hat{k}\rangle N\big)!\Big]^{1/(d+1)}. \tag{4.80}$$

Using this expression for m_{III} along with the expressions for m_I and m_{II} given by Eq.s (4.76) and (4.77) in (4.75) gives (4.73). We see that Eq. (4.73) is indeed the total number of simplicial complexes that can be produced by a given sequence of the generalized degrees when we disregard the possibility of forbidden matchings.

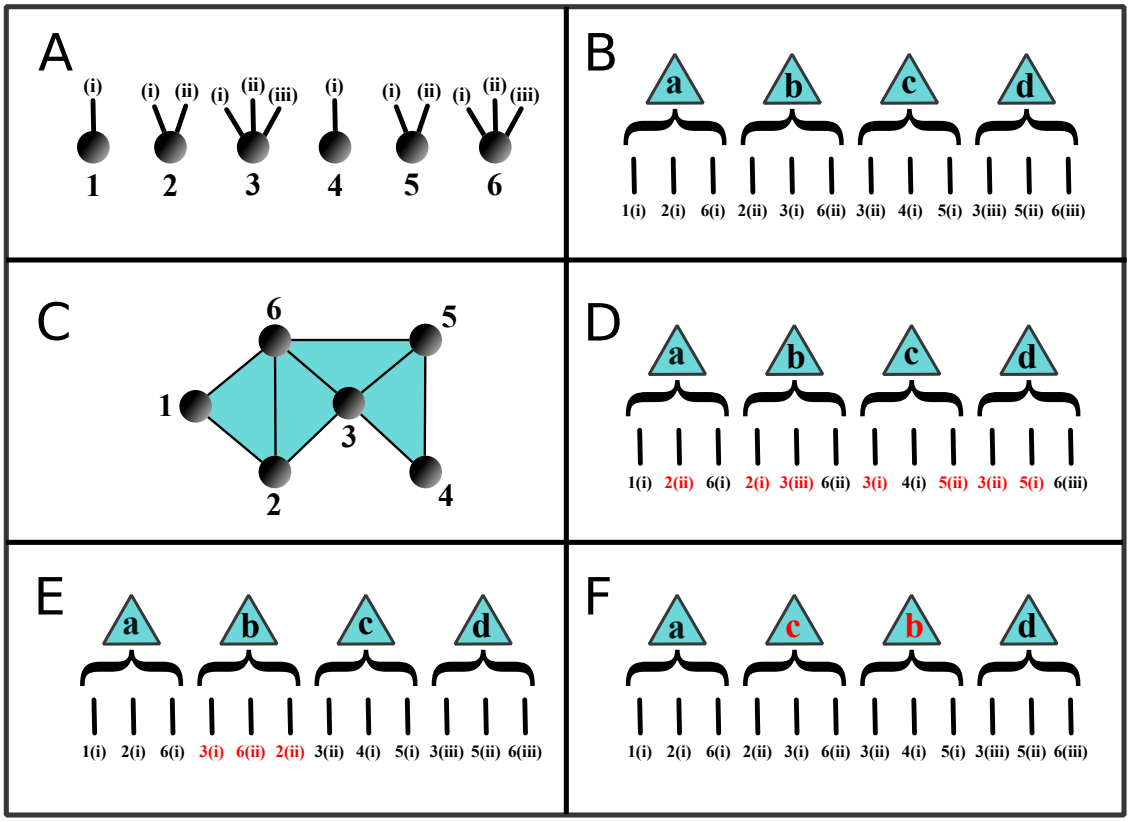

Figure 4.4: Panel A shows 6 nodes with labelled stubs that are initially unmatched. Panel B shows one possible random ordering of the stubs and their partitioning into blocks corresponding to 2-dimensional simplices. The triangles with letters a-d indicate the blocks while the labelled lines indicate the stubs (for example 2(i) is stub (i) of node 2). Panel C shows the simplicial complex implied by the partitioning in panel B. Panel D shows a different ordering of the stubs which leaves the same nodes connected to the same simplices, but through different stubs. This ordering produces the same simplicial complex as the original ordering shown in panel B. The stubs that have been permuted are shown in red. Panel E illustrates the fact that the order which stubs appear in a block does not affect the simplicial complex produced. In this case the stubs of block b have been permuted relative to the ordering in panel B. The simplicial complex produced by these orderings is the same. Panel F shows that permuting block labels does not affect the simplicial complex produced. In this case the blocks b and c have been permuted.

4.4 Natural correlations of the configuration model of simplicial complexes

In the preceding sections of this chapter we have introduced two maximum entropy models of simplicial complexes: the configuration model and canonical ensemble and have shown how we can use analytical and numerical techniques to explore these models.

In the chapter introduction we outlined the possible uses for our models, namely the reconstruction of simplicial complexes based on knowledge of their generalized degrees, the identification of non-trivial or 'interesting' structure in simplicial complex data, the uncovering of correlations between structural properties of the simplicial complexes and as a way of relating the behaviours of dynamical processes taking place on the simplicial complexes to a narrow set of structural properties.

In this section we focus on the third reason listed above: the uncovering of correlations between structural properties of the simplicial complexes. In particular we show how using our configuration model, it is possible to explore the correlations that arise between the degrees of the nodes as a consequence of the sequence of the generalized degrees of the nodes.

Note that here we are distinguishing between the *degree* of a node r (the number of links r is a part of) and its generalized degree $k_{d,0}(r)$ (the number of d-simplices it is a part of).

Our configuration model is the maximum entropy model for a simplicial complex based on the sequence of the generalized degrees of the nodes, and so any correlations of the degrees that we observe in this model can be seen as the natural correlations arising from the generalized degrees. For simplicial complexes with $d = 1$ the degree and generalized degree of a node are the same and our model reduces to the network configuration model. In this case the question that we are asking is "how does the propensity of each individual node to form links affect the local structure of the network around the node"? By extending our model to $d > 1$ we are extending this

question to "how does the propensity of a node to participate in groups of size $d+1$ (d-simplices) affect the local (network) structure around the node".

We have chosen to use simplicial complexes constructed by the configuration model with scale-free distribution $P_{d,0}(k)$ of the generalized degree of the nodes $k_{d,0} = k$. The distribution $P_{d,0}(k)$ of the generalized degree of the nodes is given by

$$P_{d,0}(k) = Ck^{-\gamma}, \tag{4.81}$$

with minimal generalized degree $m = 1$.

From these simplicial complexes we extract their 1-dimensional 'skeleton networks' which consist of just their nodes and links. We indicate with $\hat{\mathbf{a}}$ the adjacency matrix of such a network and note that it can be obtained from the adjacency tensor $\mathbf{a}$ by putting $\hat{a}_{rl} = 1$ if there is at least one d-simplex α with $a_\alpha = 1$ and $\hat{a}_{rl} = 0$ otherwise. We indicate the degree of a node r in the network with κ_r.

These networks can be analysed by means of well established tools used in network theory. In particular we characterise the correlations existing in this network by means of the average degree $knn(\kappa)$ of the neighbours of the nodes of degree κ, and the average clustering coefficient $C(\kappa)$ of the nodes of degree κ. The average degree of nodes of degree κ is given by

$$knn(\kappa) = \frac{\sum_r \delta(\kappa_r, \kappa) \sum_l \hat{a}_{rl}\kappa_l}{\kappa N_\kappa}, \tag{4.82}$$

where $N_\kappa = \sum_r \delta(\kappa_r, \kappa)$ is the number of nodes with degree κ. The clustering coefficient of a single node r is the proportion of potential links between the neighbours of that node that are actually present in the network i.e.

$$C_r = \frac{\sum_{lj} \hat{a}_{rl}\hat{a}_{rj}\hat{a}_{lj}}{\kappa_r (\kappa_r - 1)}, \tag{4.83}$$

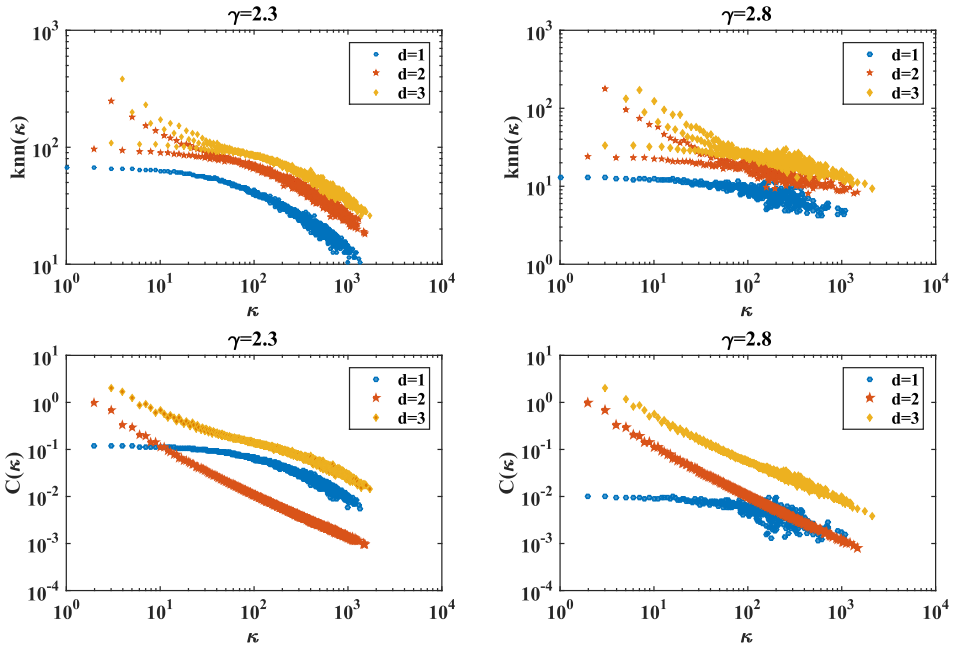

Figure 4.5: The average degree $knn(\kappa)$ of the neighbours of the nodes of degree κ and the average clustering coefficient $C(\kappa)$ of the nodes of degree κ, for simplicial complexes of dimension $d = 1, 2, 3$ constructed according to the configuration model with distribution of the generalized degrees of the nodes $P_{d,0}(k)$ given by Eq. (4.81) and $\gamma = 2.3, 2.8$. The simplicial complexes have $N = 10^4$ nodes, and $n_F = 70$. The data points have been averaged over 100 realizations.

and so the average clustering coefficient $C(\kappa)$ of the nodes of degree κ is

$$C(\kappa) = \frac{\sum_r C_r \delta(\kappa_r, \kappa)}{n_\kappa}. \tag{4.84}$$

These functions are plotted in Figure 4.5 for simplicial complexes with generalized degree distribution $P_{d,0}(k)$ given by Eq. (4.81) and $\gamma = 2.3, 2.8$. These results show that natural degree correlations occur in these models. The average clustering coefficient $C(\kappa)$ increases with the increasing dimensionality d of the simplicial complex, and the shape of the function $C(\kappa)$ is also strongly dependent on the dimensionality d. On the contrary, $knn(\kappa)$ does not appear to change so dramatically with the dimensionality d of the simplicial complex.

4.5 Conclusions

The research presented in this chapter, originally published in [33], constitutes a significant first step into the modelling of simplicial complexes. Seen as generalisations of networks, simplicial complexes encode many-body interactions between the parts of a complex system. They allow for a characterisation of the structure of such systems that is not limited to node properties, but also provides a language to describe the higher order properties of simplices with dimension $d > 1$. Additionally they are also ideal mathematical objects to which tools from Topological Data Analysis may be applied, or as discretisations of geometry from which the underlying geometries of real-world data could be inferred.

Our models are maximum entropy models of d-dimensional simplicial complexes, based on the generalised degrees of the nodes. As such, our models have a clear use as null models for simplicial complexes. They are the most appropriate models for simplicial complexes given knowledge of these generalized degrees, and allow for a statistically rigorous understanding of the implications of particular choices of the generalized degrees for the structure of simplicial complexes and dynamics taking place upon them.

In this chapter we have explored these models within a statistical physics framework. We have discovered the existence of a structural cutoff for simplicial complexes above which correlations occur between the generalized degrees of the nodes. This structural cutoff scales like $N^{\frac{d}{d+1}}$ and reduces to the network structural cutoff for $d = 1$.

The structural cutoff plays an important role in our understanding of our models. In the canonical ensemble, for choices of the generalized degrees of the nodes below the structural cutoff we derived simplex probabilities and expected generalized degrees of higher dimensional faces in terms of the generalized degrees of their nodes. Making an analogy with conjugated ensembles in statistical physics we derived a relation between the entropies of our two models, and showed that in statistical physics terms these models are not equivalent in the large system limit. Below the

structural cutoff we also calculated the entropies of our two models explicitly in terms of the generalized degrees of the nodes. These entropies have a number of uses in network analysis and network inference [83, 84].

The entropy of the configuration model is simply the logarithm of the total number of simplicial complexes that can be constructed with the given sequence of generalized degrees of the nodes. From the entropy we therefore obtained the total number of simplicial complexes with given sequence of the generalized degrees of the node, provided these generalized degrees were below the structural cutoff. Crucially, for $d = 1$ this result reduces to the known result for networks given in [77].

In addition to the analytical approach to investigating our models described above, we also developed algorithms to sample from our ensembles stochastically. For the configuration model we used this algorithm to investigate the statistical implications of the generalized degrees of the nodes on the structure of the skeleton network implied by the simplicial complex. We found that for generalized degree distributions without the structural cutoff can cause significant correlations to emerge in the skeleton network.

In conclusion we believe that the research presented in this chapter provides a full account of two of the most fundamental equilibrium models of simplicial complexes which can be used as null models for investigating the structure of simplicial complexes, or for studying dynamical processes. We believe that these models constitute only the first step in modelling simplicial complexes with equilibrium statistical mechanics tools and that our work will open up new perspectives for investigating a new generation of maximum entropy models of simplicial complexes.

Chapter 5

Weighted Growing Simplicial Complex (WGSC)

In this chapter we present a model of a simplicial complex that is weighted and growing. The research presented in this chapter was published in our paper *Weighted Growing Simplicial Complexes* [34].

A number of important examples of 'real' simplicial complexes are in fact weighted, in the sense that each of their simplices has a weight associated to it. One example is simplicial complexes constructed from data about academic collaboration between scientists where the number of papers shared by a set of scientists is the weight of the simplex between them [10–12]. Another example is simplicial complexes obtained from fMRI data which measures correlations in activity between regions of the brain [16, 17]. In this example the regions of the brain are the nodes of the simplicial complex and correlations between them are assumed to imply some kind of functional relation between the regions. A weighted simplicial complex can be constructed by placing simplices between functionally related regions with weights determined by the strength of the correlations.

In the above two examples as well as in other applications, weighted simplicial complexes are rather natural representations of data. Understanding how the weight is distributed across the simplices of a simplicial complex could reveal important in-

formation about the system being represented, especially when such distributions are inhomogeneous. In particular, characterizing the relation between the topology of a simplicial complex and its weights could be important for understanding the evolution of the simplcial complex or the functions of its components.

In the research presented in this chapter we characterize the structure of weighted simplicial complexes in terms of their generalised degrees and generalised strengths, and present a growing model of a weighted simplicial complex.

This model follows in the tradition of non-equilibrium network models that seek to characterize the relations between growth mechanisms of networks and their structural properties. As such this model falls into the category of models called explanatory models. As discussed in Chapter 3, explanatory models of networks seek to identify mechanisms for constructing networks that can explain a given set of structural properties observed in a real network.

In contrast to the null models presented in the previous chapter, an explanatory model does not necessarily attempt to give a 'realistic' representation of a network or simplicial complex based on the observed properties but instead tries to provide a plausible hypothesis about how a network or simplicial complex with such properties might have grown or been constructed.

For certain choices of the parameters our model reduces to the weighted BA model of [43] or the NGF of [57] both of which were discussed in Chapter 3. The model is designed to explore how simple growth mechanisms can affect the distribution of weight across a simplicial complex. In particular, we investigate the scaling of the generalised strength of the simplices scales with their generalised degree. We find that depending on a small number of parameters, our model can generate linear, super-linear or exponential scalings of generalised strength with generalised degree. In the next section we will explain how these different scalings correspond to various homogeneous and inhomogeneous distributions of weight across the simplices. Our model reveals that relatively simple growth mechanisms for generating simplicial complexes can give rise to a rich variety of weight distributions, and provides

a plausible explanation for the origins of such distributions in real simplicial complexes.

Our approach extends similar approaches to explaining the emergence of linear and super-linear scalings of strength and degree in simple networks [43]. In simple networks [41–44] it has been shown through the analysis of a vast set of real datasets that weights are not always distributed uniformly over the links of the network.

Specifically in some networks, high degree 'hub' nodes can have connections with on average stronger weights than the typical connections of low degree nodes. The way to characterize these weight-topology correlations is by studying the scaling of the average strength of nodes as a function of their degree.

In particular, if the strength grows linearly with the degree the weights are uniformly distributed among the nodes of the network. If, instead the observed scaling is superlinear then hubs typically have links with stronger weights than low degree nodes.

Both linear and superlinear scaling have been observed in real-world networks [41]. While early models of growing weighted networks exclusively captured the linear scaling [42] it was later shown that the emergence of the weight-topology correlations can be described within the framework of growing network models, including growth of the network through the addition of new links while at the same time increasing the weights of the links driven by a reinforcement dynamics [43].

Our research extends this idea for growing simplicial complexes, showing that weight-topology correlations emerge not just at the network level, but also for δ-faces of higher dimension. In order to identify these correlations, we develop a mean-field approach for calculating the generalised degrees and generalised strengths of simplices as a function of their birth time and use this to approximate the generalised strength-degree scalings. We also compare the results obtained in this mean-field approximation with extensive numerical simulations.

This chapter is organized as follows. In Section 5.1 we recall the basic definitions for weighted simplicial complexes and their main structural properties, and introduce notation specific to this chapter. In Section 5.2 we define our model of a growing weighted simplicial complex. In Section 5.3 we present our the mean-field solution of the model. In Section 5.4 we compare our theoretical prediction with the results of the numerical simulations. Finally in Section 5.5 we give our conclusions.

The code we used for numerical simulations in this chapter have been published online at the URL address [85].

5.1 Definitions and notation

As in the previous chapter, in this chapter the simplicial complexes we consider are pure d-dimensional, i.e. simplicial complexes constructed exclusively from d-dimensional simplices and their sub-faces.

The structure of these d-dimensional simplicial complexes of N nodes is determined by the adjacency tensor $\mathbf{a}$ with elements $a_\alpha = 1, 0$ indicating whether the simplex $\alpha \in Q_d(N)$ is present ($a_\alpha = 1$) or absent ($a_\alpha = 0$) from the simplicial complex. Occasionally it will be useful to refer to the set $S_\delta(N)$ which is the set of all δ-simplices actually present in a given simplicial complex, i.e. $S_\delta(N)$ is the set of all δ-simplices $\alpha \in Q_d(N)$ for which $a_\alpha = 1$.

Unlike in the previous chapter, the simplicial complexes considered here are weighted. This means that each d-dimensional simplex α has an a weight w_α associated with it. Similar to the adjacency tensor, we define the weight tensor $\mathbf{w}$ as having elements w_α for each d-simplex α.

In a simplicial complex representing co-authorship, for example, a simplex could represent a set of co-authors that have collaborated on at least one paper together, while the weight of that simplex corresponds to the total number of papers that have been co-authored by the team.

We characterize the properties of the simplicial complex using the *generalized degrees* and *generalized strengths* of the δ-faces, which we defined in Chapter 2. In contrast to the last chapter, in this chapter the generalized degrees evolve in time. For this reason we choose slightly different notation for the generalised degrees than in the previous chapter. The change of notation is shown below:

$$k_{d,\delta}(\alpha) \to k^{\alpha}_{d,\delta}(t) = \sum_{\alpha' \in Q_d(N,t) | \alpha' \supseteq \alpha} a_{\alpha'}. \tag{5.1}$$

The generalized strength $s_{d,\delta}(\alpha)(t)$ of a δ-face $\alpha \in S_\delta$ is the sum of the weights of the d-dimensional simplices incident to it, and we use similar notation to the generalised degree

$$s^{\alpha}_{d,\delta}(t) = \sum_{\alpha' \in Q_d(N,t) | \alpha' \supseteq \alpha} a_{\alpha'} w_{\alpha'}. \tag{5.2}$$

In weighted networks, it has been shown that it is possible to characterize the interplay between the network topology and the weights of the links by classifying networks depending on the scaling of the strength as a function of the degree of the nodes.

Specifically it has been shown that for some networks the weights of the links are distributed rather uniformly, resulting in a linear dependence of the strength of the nodes with its degree,

$$S_i \propto K_i \tag{5.3}$$

while in other networks hub nodes tend to have links with higher weights than low degree nodes. This latter scenario results in a superlinear scaling of the strength versus the degree, i.e.

$$S_i \propto (K_i)^{\theta}, \tag{5.4}$$

with $\theta > 1$.

An example of networks with linear dependence of the strength versus degree are collaboration networks, while an example of non-linear dependence of strength on degree are for instance airport networks where the weights measure the number of passengers for each flight connection.

In Ref. [43] it has been shown that a simple growing network model with reinforcement of the links is actually able to generate networks with linear and superlinear scaling of the strength versus degree depending on the rate at which new links are added with respect to the rate at which links are reinforced.

In this chapter we propose a model for growing simplicial complexes which shows a very rich phenomenology, and we show evidence that in simplicial complexes it is possible to characterize the correlations between weights and topology by exploring the dependence of the generalized strength $s^{\alpha}_{d,\delta}$ versus the generalized degree $k^{\alpha}_{d,\delta}$.

Specifically we are able to predict three alternative possible scalings: linear, superlinear and exponential, i.e.

$$s^{\alpha}_{d,\delta} \propto \begin{cases} k^{\alpha}_{d,\delta}, \\ \left(k^{\alpha}_{d,\delta}\right)^{\theta}, \\ \exp\left[\beta k^{\alpha}_{d,\delta}\right], \end{cases} \tag{5.5}$$

with $\theta > 1$ and β indicating a constant greater than zero. In this case the superlinear scaling indicates weight-topology correlations, and these correlations are even more pronounced for the exponential scaling.

5.2 The Model

In this section we present our model of a growing weighted simplicial complex. The model combines a growth process by which a new node and a set of new simplices

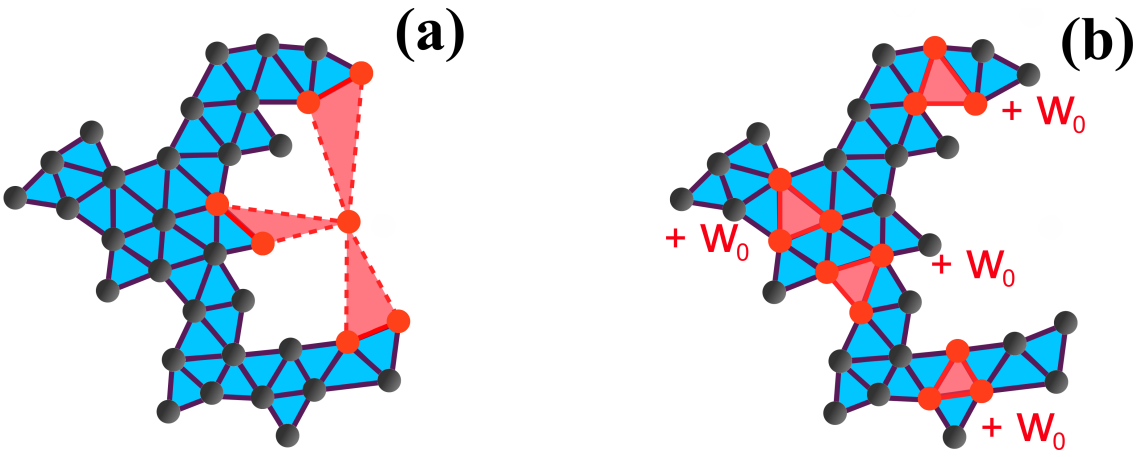

Figure 5.1: Graphical representation of process A (panel (a)) and process B (panel (b)) for a 2-dimensional simplicial complex with $m = 3$ and $m' = 4$ starting from a given initial condition.

are introduced into the simplicial complex at every time step of the model, and a reinforcement process where the weights of some simplices are reinforced (increased) at each time step. Both processes are stochastic, and the probabilities by which they operate are governed by a small number of parameters. Depending on the choice of these parameters the simplicial complex is capable of displaying a variety of interesting weight-topology correlations.

The evolution of the topology of these simplicial complexes, governed by the growth process, is based on the previously proposed framework of Network Geometry with Flavor (NGF) [57]. The NGF is a model of growing simplicial complex able to generate networks with different complex topologies, including hyperbolic manifolds, scale-free networks, and networks with relevant modularity. However, the model presented in this chapter includes two important new elements with respect to the NGF model: i) the simplicial complexes generated by this model are weighted, ii) the simplicial complexes generated by this model can have non-trivial homology.

The weighted simplicial complexes are generated as follows. We start at time $t = 1$ from an initial finite simplicial complex that comprises $m_0 > m$ d-dimensional

simplices of total weight ω_0. At each time-step $t > 1$ two processes take place:

A) *Add m simplices (growth process)*:
A new node arrives and m new d-simplices with initial weight w_0 are created between the node and pre-existing $(d-1)$-faces. The probability $\Pi_{d-1}(\alpha)$ that a given $(d-1)$-face α is selected by one of the new d-simplices is given by

$$\Pi_{d-1}(\alpha) = \frac{1}{\mathcal{Z}_t}(1 + sn_\alpha), \tag{5.6}$$

where $n_\alpha = k^\alpha_{d,d-1} - 1$ is called the *saturation* and where s is a parameter called *flavor* which takes the values $s = -1, 0, 1$ and controls the simplicial complex topology. Note that in Eq. (5.6), $\mathcal{Z}_t$ is a normalization constant given by $\mathcal{Z}_t = \sum_{\alpha \in S_{d-1}(t)}(1 + sn_\alpha)$.

B) *Reinforce m' simplices (reinforcement process)*:
At this step m' existing d-simplices are selected and their weights are increased by w_0. A d-simplex α with weight w_α is selected for reinforcement with probability $\tilde{\Pi}_d(\alpha)$ proportional to its weight, i.e.

$$\tilde{\Pi}_d(\alpha) = \frac{w_\alpha}{\tilde{\mathcal{Z}}_t}, \tag{5.7}$$

where $\tilde{\mathcal{Z}}_t = \sum_{\alpha \in S_d(t)} w_\alpha$.

Figure 5.1 shows an illustration of the two processes (process A and process B) for a 2-dimensional simplicial complex starting from a given initial simplicial complex. The flavor s has an important effect on the topological properties of the simplicial complexes produced. Selection of $s = -1$ imposes the constraint that the generalized degree $k_{d,d-1}(\alpha)$ of a $(d-1)$-face α can only take the values 1 and 2, or equivalently imposes that the saturation n_α can only take values 0 (unsaturated) and 1 (saturated), which leads to the simplicial complex produced being a d-dimensional manifold. Choosing $s = 0$ or $s = 1$ removes this constraint, and gives a selection probability $\Pi_{d-1}(\alpha)$ that is uniform on the set of all $(d-1)$-faces for $s = 0$ and a form of preferential attachment with $\Pi_{d-1}(\alpha) \propto k^\alpha_{d,d-1}$ for $s = 1$.

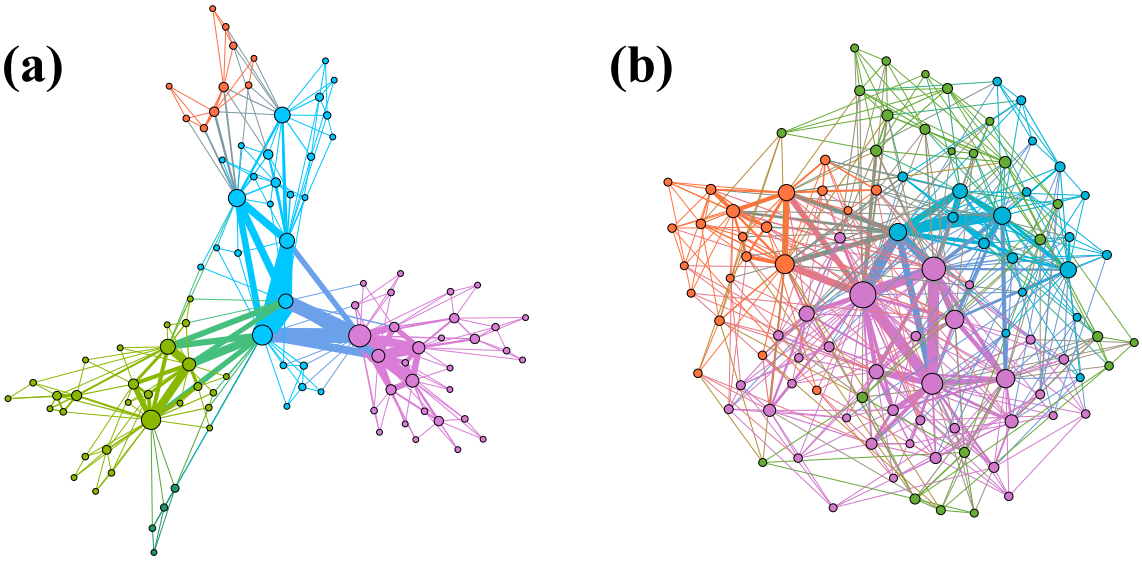

Figure 5.2: Skeleton networks of simplicial complexes generated by the model for $d = 3$, $N = 100$ and $s = -1$. Node sizes indicate their degrees while link widths indicate their generalized strength. Node and edge colorings indicate community structure calculated according to the Louvain algorithm [86]. Panel (a) shows the skeleton of a simplicial complex with $m = 1$ and $m' = 2$ while panel (b) shows the skeleton of a simplicial complex with $m = 2$ and $m' = 1$.

In Figure 5.2 we plot the weighted skeleton networks of two simplicial complexes generated by the model in the case $d = 3$ and $s = -1$ for $(m, m') = (1, 2)$ and $(m, m') = (2, 1)$. The weights of the links in these networks indicate the generalized strengths of the links in their corresponding simplicial complexes. While in the case $(m, m') = (1, 2)$ nodes with high degree have typically links with stronger weights than the weights of low-degree nodes, the weights are more homogeneously distributed in the case $(m, m') = (2, 1)$.

5.3 Mean-field solution of the model

In this section we derive mean-field approximations for the generalised degrees and generalised strengths of δ-faces as a function of their birth times and use these approximations to calculate the generalised degree distributions of the faces as well

as the scaling of the generalised strengths of the faces with their generalised degrees.

To do this we make use of a very well established framework for simple networks [1–3], where the generalised degrees, generalised strengths, simplex weights and the time t are approximated as continuous variables. Mean-field approximations for these quantities are then obtained by solving differential equations.

In Section 5.3.1 we use this approach to calculate the generalised degree of a face, and subsequently derive the distributions of the generalised degrees for faces of all dimensions $0 \leq \delta \leq d-1$. The distributions can be power-law, exponential, or bimodal (for $\delta = d-1$ in the d-dimensional manifold case). Remarkably, even within a single simplicial complex we find that for certain choices of parameters it is possible to have all three of the distributions within the simplicial complex for faces of different dimension. We show how this phenomenon can be understood in terms of the 'effective' attachment mechanisms experienced by the faces of lower dimension, where one type of attachment mechanism at dimension $\delta = d-1$ is felt as a different type of attachment mechanism at lower dimensions (for example, we show that attaching simplices with uniform attachment to $(d-1)$-faces is felt as a kind of preferential attachment for faces of dimension $d-2$). A further implication of this effect is that for all three attachment mechanisms (manifold, uniform, preferential), if the dimension of the simplicial complex is $d \geq 3$, then the skeleton network is always scale-free. This fact demonstrates that scale-free networks can be created through attachment mechanisms that focus not on node properties but properties belonging to higher order groups of nodes (the simplices and faces).

The calculation of the mean-field approximations for the generalised strengths is more convoluted than for the generalised degrees. The generalised strength of a face is the sum of the independently evolving weights of the simplices to which it is incident. In order to calculate the generalised strength we first derive in Section 5.3.2 the probability that a given simplex exists in the simplicial complex at time t, and then in Section 5.3.3 derive the mean-field weight of the simplex at t. In Section 5.3.4 we combine these quantities to calculate the generalised strength of a face at time t, and subsequently find the scaling relation between the generalised strength

and generalised degree.

5.3.1 Mean-field solutions for the generalized degrees

In this section we use a mean-field approximation to derive the time evolution of the generalized degrees of the faces of any dimension δ. In fact, we will approximate the generalized degree $k^{\alpha}_{d,\delta}$ of a δ-face α with its expected value, and also make the further approximation that this expected generalized degree evolves continuously rather than in discrete time steps. As we shall see, this allows us to obtain the time evolution of the generalized degrees by solving differential equations of the form

$$\frac{\partial}{\partial t} k^{\alpha}_{d,\delta} = f(m, s, t, k^{\alpha}_{d,\delta}), \tag{5.8}$$

where $f(m, s, k^{\alpha}_{d,\delta})$ is some function of the model parameters, the time t and $k^{\alpha}_{d,\delta}$. The appropriate choice of function f is the expected increase in $k^{\alpha}_{d,\delta}$ in the interval $[t, t+1]$ conditioned on its expected generalised degree at time t.

In order to obtain this expected increase as a function of $k^{\alpha}_{d,\delta}$, we first need to calculate the probability that at time t, one of the new d-simplices attaches itself to a $(d-1)$-face incident to α. If $\delta = d-1$ then α is itself a $(d-1)$-face, and the probability is simply $\Pi_{d-1}(\alpha)$ as given in Eq. (5.6). For $\delta < d-1$, we write the probability as $\Pi_{\delta}(\alpha)$, and is simply the sum of the probabilities that any $(d-1)$ face $\alpha' \supseteq \alpha$ is chosen for attaching the new simplex, i.e.

$$\begin{aligned} \Pi_{\delta}(\alpha) &= \sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} \Pi_{d-1}(\alpha') \\ &= \frac{1}{\mathcal{Z}_t} \sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} \left(1 - s + s k^{\alpha'}_{d,d-1}\right). \end{aligned} \tag{5.9}$$

Note that $\Pi_{\delta}(\alpha)$ is the probability that a particular d-simplex out of the m d-simplices created at time t attaches to α (via one of the $(d-1)$-faces that α is incident to). The expected increase in $k^{\alpha}_{d,\delta}$ will instead be equal to the sum of these

probabilities for each of the m simplices, leading to

$$\frac{\partial}{\partial t} k_{d,\delta}^{\alpha} = m\Pi_{\delta}(\alpha). \tag{5.10}$$

We thus need an expression for $\Pi_{\delta}(\alpha)$ as a function only of $k_{d,\delta}^{\alpha}$, the time t and the model parameters. Let us start by turning our attention first to $\mathcal{Z}_t$, which normalises Eq. (5.6). We have

$$\begin{aligned} \mathcal{Z}_t &= \sum_{\alpha' \in S_{d-1}} \left(1 - s + s k_{d,d-1}^{\alpha'}\right) && (5.11) \\ &= (1-s) \left(\sum_{\alpha' \in S_{d-1}} 1 \right) + s \left(\sum_{\alpha' \in S_{d-1}} k_{d,d-1}^{\alpha'} \right). && (5.12) \end{aligned}$$

The first sum in the above equation is simply the total number of $(d-1)$-faces in the simplicial complex at time t. At each time step we add m new d-dimensional simplices each one contributing d new $(d-1)$-faces (neglecting the unlikely event that there is any overlap between the new simplices). For $t \gg 1$, the number of $(d-1)$-faces is thus well approximated by $\sum_{\alpha' \in S_{d-1}} 1 \simeq mdt$. Additionally we have that $\sum_{\alpha' \in S_{d-1}} k_{d,d-1}^{\alpha'} \simeq m(d+1)t$ for $t \gg 1$ because any new simplex increases by one the generalized degree of each of its $d+1$ faces of dimension $d-1$. We can thus make the approximation

$$\mathcal{Z}_t \simeq m(d+s)t, \tag{5.13}$$

i.e. an expression only dependent on the time and the model parameters. We now consider the sum in the second line of Eq. (5.9):

$$\begin{aligned} \textstyle\sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} \left(1 - s + s k_{d,d-1}^{\alpha'}\right) &= (1-s) \left(\textstyle\sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} 1 \right) \\ &\quad + s \left(\textstyle\sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} k_{d,d-1}^{\alpha'} \right). \end{aligned} \tag{5.14}$$

Using the combinatorial expression first derived in [33] and given in Eq. (2.2) we

can write the second sum as

$$\sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} k^{\alpha'}_{d,d-1} = \big(d - \delta\big) k^{\alpha}_{d,\delta}. \tag{5.15}$$

The first sum in Eq. (5.14) is the number of $(d-1)$-faces incident to α, i.e. the generalized degree $k^{\alpha}_{d-1,\delta}$ where the $d-1$ in the subscript indicates that the simplices being counted have dimension $d-1$. There is no combinatorial expression for $k^{\alpha}_{d-1,\delta}$ in terms of $k^{\alpha}_{d,\delta}$ that holds for *all* simplicial complexes, however for the simplicial complexes produced by the particular dynamics of our model it is possible to show that

$$\sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} 1 = k^{\alpha}_{d-1,\delta} = \begin{cases} 1 + \big(d - \delta - 1\big) k^{\alpha}_{d,\delta} & \text{for} \quad \delta > 0, \\ m + \big(d - 1\big) k^{\alpha}_{d,\delta} & \text{for} \quad \delta = 0. \end{cases} \tag{5.16}$$

In order to understand Eq. (5.16), first consider the generalised degrees $k^{\alpha}_{d-1,\delta}$ and $k^{\alpha}_{d,\delta}$ of α when it is initially created at time t_α. If α is a node (i.e. if $\delta = 0$), then when it is initially created it has exactly m d-simplices so that $k^{\alpha}_{d,\delta} = m$. For $t_\alpha \gg 1$, the probability that any of these new d-simplices have any overlapping $(d-1)$-faces (due to the simplices attaching to adjacent faces) is vanishingly small so that $k^{\alpha}_{d-1,\delta} \simeq md$. If, as the model progresses, the node α subsequently gains any additional simplices, then for each additional d-simplex obtained α also gains $d-1$ new $(d-1)$-faces. These faces are the $(d-1)$-faces formed by α, the new node created at that time step and each combination of exactly $d-2$ of the other $d-1$ nodes in the $(d-1)$-face of which α is a part of, and to which the new simplex is attaching. For the case $\delta = 0$ we thus have $k^{\alpha}_{d-1,0} = md + (d-1)\left(k^{\alpha}_{d,\delta} - m\right) = m + \big(d-1\big) k^{\alpha}_{d,\delta}$. A similar argument may be made for $\delta > 0$. In this case every δ-face is initially created incident to a single d-simplex and so must initially be incident to d $(d-1)$-faces. As with the case of a node, each subsequent d-simplex it gains contributes $d-1$ new $(d-1)$-faces, leading to $k^{\alpha}_{d-1,\delta} = d + (d-1)\left(k^{\alpha}_{d,\delta} - 1\right) = 1 + \big(d - \delta - 1\big) k^{\alpha}_{d,\delta}$.

Using Eq.s (5.15) and (5.16) in (5.14) gives

$$\sum_{\alpha' \in S_{d-1} | \alpha' \supseteq \alpha} \left(1 - s + s k^{\alpha'}_{d,d-1}\right) = (1-s) c_\delta + \big(d + s - \delta - 1\big) k^{\alpha}_{d,\delta}, \tag{5.17}$$

where

$$c_\delta = \begin{cases} 1 & \text{for} \quad \delta > 0, \\ m & \text{for} \quad \delta = 0. \end{cases} \tag{5.18}$$

With Eq. (5.17) and $\mathcal{Z}_t$ as given in Eq. (5.13) we can now write the probability $\Pi_\delta(\alpha)$ in terms of $k^\alpha_{d,\delta}$ and thus we obtain the following differential equation for the evolution of $k^\alpha_{d,\delta}$ in time:

$$\frac{\partial k_{d,\delta}(t,t_\alpha)}{dt} = m\Pi_\delta(\alpha) = \frac{(1-s)c_\delta + \big(d+s-\delta-1\big)k_{d,\delta}(t,t_\alpha)}{\big(d+s\big)t}, \tag{5.19}$$

where we have applied the change of notation $k^\alpha_{d,\delta} \to k_{d,\delta}(t,t_\alpha)$ to emphasise the fact that in our mean-field approximation the generalised degree of α depends only on t and α's birth time t_α. The differential equation in (5.19) has initial condition

$$k_{d,\delta}(t_\alpha,t_\alpha) = c_\delta = \begin{cases} 1 & \text{for} \quad \delta > 0, \\ m & \text{for} \quad \delta = 0. \end{cases} \tag{5.20}$$

The solution of this equation is

$$k_{d,\delta}(t,t_\alpha) = \begin{cases} c_\delta \frac{d-\delta}{d+s-\delta-1}\left(\frac{t}{t_\alpha}\right)^{\lambda_\delta} + c_\delta \frac{s-1}{d+s-\delta-1} & \text{for} \quad \delta - s \neq d-1, \\ c_\delta \frac{1-s}{d+s}\log\left(\frac{t}{t_\alpha}\right) + c_\delta & \text{for} \quad \delta - s = d-1, \end{cases} \tag{5.21}$$

where

$$\lambda_\delta = \frac{d+s-\delta-1}{d+s}. \tag{5.22}$$

For the case $\delta = d-1$ with $s = -1$ (i.e. the $(d-1)$-faces in the random d-manifold version of the model) the generalized degree distribution $P_{d,\delta}(k)$ is bimodal, because only the generalized degrees $k^\alpha_{d,d-1} = 1, 2$ are allowed for faces of dimension $d-1$. For all other cases it is possible to derive the generalized degree distribution using the mean-field solution given by Eq. (5.21). We show that the generalized degree distribution is exponential for $d-1+s-\delta = 0$ and power-law for $d-1+s-\delta > 0$.

In order to derive these results we first note that for fixed time t the generalised degree as given in Eq. (5.21) is monotone decreasing as a function of the birth time of the δ-face t_α. Within our mean-field approximation, all δ-faces born before some time τ have a generalised degree greater than $k_{d,\delta}(t, t_\alpha = \tau)$, therefore the proportion of δ-faces with a generalised degree greater than some value k is equal to the proportion of δ-faces born before $\tau(k)$, where $\tau(k)$ is such that $k_{d,\delta}(t, \tau(k)) = k$, thus

$$P(k_{d,\delta}(t, t_\alpha) \geq k) = \hat{P}_\delta(t_\alpha < \tau(k)). \tag{5.23}$$

The number of δ-faces added at each time step neglecting overlaps is constant, so for $t \gg 1$ the proportion of δ-faces born before τ is just equal to $\hat{P}_\delta(t_\alpha < \tau) = \frac{\tau}{t}$.

Using this result and Eq. (5.21), it is not hard to show that the probability that the generalized degree $k_{d,\delta}(t, t_\alpha)$ is greater than k is given by

$$P(k_{d,\delta}(t, t_\alpha) \geq k) = \begin{cases} \left(\frac{c_\delta(d-\delta)}{k(d+s-\delta-1)}\right)^{\frac{1}{\lambda_\delta}} & \text{for} \quad \delta - s < d-1, \\ \exp\left[-\frac{d+s}{(1-s)c_\delta}k\right] & \text{for} \quad \delta - s = d-1. \end{cases} \tag{5.24}$$

The generalized degree distribution $P_{d,\delta}(k)$ is found by differentiating Eq. (5.24) with respect to k leading to

$$\begin{aligned} P_{d,\delta}(k) &= -\frac{dP(k_{d,\delta}(t,t_\alpha)\geq k)}{dk} \\ &= \begin{cases} \frac{d+s}{d+s-\delta-1}\left(c_\delta \frac{d-\delta}{d+s-\delta-1}\right)^{\frac{1}{\lambda_\delta}} k^{-\frac{1}{\lambda_\delta}-1} & \text{for} \quad \delta - s < d-1, \\ \frac{d+s}{(1-s)c_\delta}\exp\left[-\frac{d+s}{(1-s)c_\delta}k\right] & \text{for} \quad \delta - s = d-1, \end{cases} \end{aligned} \tag{5.25}$$

valid as long as $\delta - s < d$. Therefore the generalized degree distribution of δ-dimensional simplices in growing simplicial networks with flavor s follows Table 5.1.

In particular, Table 5.1 shows that in the simplicial complexes produced when $s = -1$, the generalized degree distribution for the $(d-1)$-faces is bimodal, while for the $(d-2)$-faces and $(d-3)$-faces (assuming $d \geq 3$ so that such faces exist) the

distributions are exponential and power-law respectively. For $s = 0$ on the other hand the distribution for the $(d-1)$-faces is exponential and all other distribution for the lower dimensional faces is power-law, while for $s = 1$ all the distributions are power-law. For $m = 1$ (and ignoring weights) our model reduces to the NGF of [57], and indeed we find that in this case our mean-field results coincide with the exact results for the generalised degree distribution given in [57].

In general, for the right set of parameters it is possible to observe all three distributions within a single simplicial complex. The reason this occurs can be understood in terms of the effective attachment mechanisms felt by the δ-faces of dimension $\delta < d-1$. To be precise, in the manifold case $s = -1$, the probability that a δ-face gains a new d-simplex is proportional to the number of *unsaturated* $(d-1)$-faces it has. Every time a new d-simplex is formed incident to this δ-face, one of these unsaturated $(d-1)$-faces becomes saturated but at the same time it gains $d-\delta-1$ newly created unsaturated $(d-1)$-faces. For $\delta = d-2$ the number of unsaturated $(d-1)$-faces a δ-face has is therefore constant in the manifold case, and so these faces experience an effective uniform attachment equivalent to that experienced by $(d-1)$-faces in the $s = 0$ case. For $\delta < d-2$ in the manifold case, a δ-face gains $d-\delta-2$ new unsaturated faces each time its generalized degree goes up by 1. For these faces the effective attachment mechanism is therefore a form of linear preferential attachment.

A similar argument can be made for $s = 0$. In this case it is the total number of $(d-1)$-faces incident to a δ-face that determines the probability that the δ-face will attract a new simplex. Each time the δ-face gains a new simplex, the number of $(d-1)$-faces it is incident to goes up by $d-\delta-1$ and so for all $\delta < d-1$ the effective attachment mechanism is linear preferential.

For the case $s = 1$, the argument is slightly different as the effective attachment felt by a δ-face is proportional to the sum of the generalised degrees of the $(d-1)$-faces it is incident to. Fortunately, in Eq. (5.15) we saw that this quantity is proportional to the generalised degree of the δ-face, so for this reason all δ-faces experience linear preferential attachment when $s = 1$.

These distinct attachment mechanisms can be observed in the attachment probability $\Pi_\delta(\alpha)$ for a δ-face α given in Eq. (5.19). Across the entire model the δ-faces experiencing the manifold attachment mechanism have a bimodal distribution of the generalised degrees, those experiencing uniform attachment have an exponential distribution while those experiencing preferential attachment have a power-law distribution.

In this last case, the generalized degree distribution $P_{d,\delta}(k)$ given by Eq. (5.25) decays as a power-law $P_{d,\delta}(k) \propto k^{-\gamma_{d,\delta}}$ with the power-law exponent

$$\gamma_{d,\delta} = 1 + \frac{1}{\lambda_\delta} = 1 + \frac{d+s}{d+s-\delta-1} \tag{5.26}$$

as long as $\delta - s < d - 1$. These distributions are scale-free if $\gamma_{d,\delta} \leq 3$, or equivalently they are scale-free if

$$d \geq d_c^{[\delta,s]} = 2(\delta+1) - s. \tag{5.27}$$

In the above discussion, we have focused entirely on the generalized degrees of the faces. In particular the distributions discussed above all consider only the number of d-simplices incident to a face, and do not consider the network structure that the model creates. We now turn our attention to the skeleton network consisting of the nodes and links of the simplicial complex. The degree of any node K_i that does not belong to the small initial condition is given by

$$K_i = k_{d,0}^i + (d-1)m. \tag{5.28}$$

In fact each node initially has degree dm, and subsequently the degree increases by one for each d-simplex glued to one of the $(d-1)$-faces of the node. It follows then that if the generalized degree distribution of the nodes is scale-free then the degree distribution of the skeleton network is also scale-free. As a result, growing simplicial complexes of flavor $s = 1$ are scale-free for any $d \geq 1$, the ones of flavor $s = 0$ are scale-free for $d \geq 2$ and the ones of flavor $s = -1$ are scale-free for $d \geq 3$.

Table 5.1: Distribution of the generalized degrees of faces of dimension δ in a d-dimensional simplicial produced by the model with flavor s. For $d \geq d_c^{[\delta,s]} = 2(\delta+1) - s$ the power-law distributions are scale-free, i.e. the second moment of the distribution diverges.

flavor	$s=-1$	$s=0$	$s=1$
$\delta = d-1$	Bimodal	Exponential	Power-law
$\delta = d-2$	Exponential	Power-law	Power-law
$\delta \leq d-3$	Power-law	Power-law	Power-law

We have thus seen how a variety of attachment mechanisms at the level of the d-simplex can induce different attachment mechanisms at the level of the δ-face and in particular the node. Our model produces bimodal, exponential and power-law distributions of both the generalised degree and network degree. Our model also illustrates more generally that scale-free networks may be produced by models of growth that focus not only on node properties but also on properties belonging to higher-order structures such as (in our case) simplices or perhaps other network motifs.

In the following section we consider the probability that a given simplex exists in the simplicial complex, which we use later in Section 5.3.4 to help us calculate the strength of a simplex.

5.3.2 Probability of a simplex

Unlike for nodes, the existence of specific simplices or faces of dimension greater than 0 in a simplicial complex is not guaranteed. For these faces we are therefore interested in calculating their probabilities.

Towards this end, let us represent each *possible* δ-face α_δ by the sequence of its nodes $\alpha_\delta = [j_0, j_1, ..., j_\delta]$ where the nodes are ordered according to the time of their arrival in the simplicial complex, i.e. $t_{j_0} < t_{j_1} < ... < t_{j_\delta}$.

We want to calculate the probability that at time t, α_δ exists in the simplicial complex, given the arrival times of its nodes $t_{j_0} < t_{j_1} < ... < t_{j_\delta}$, i.e. we want to

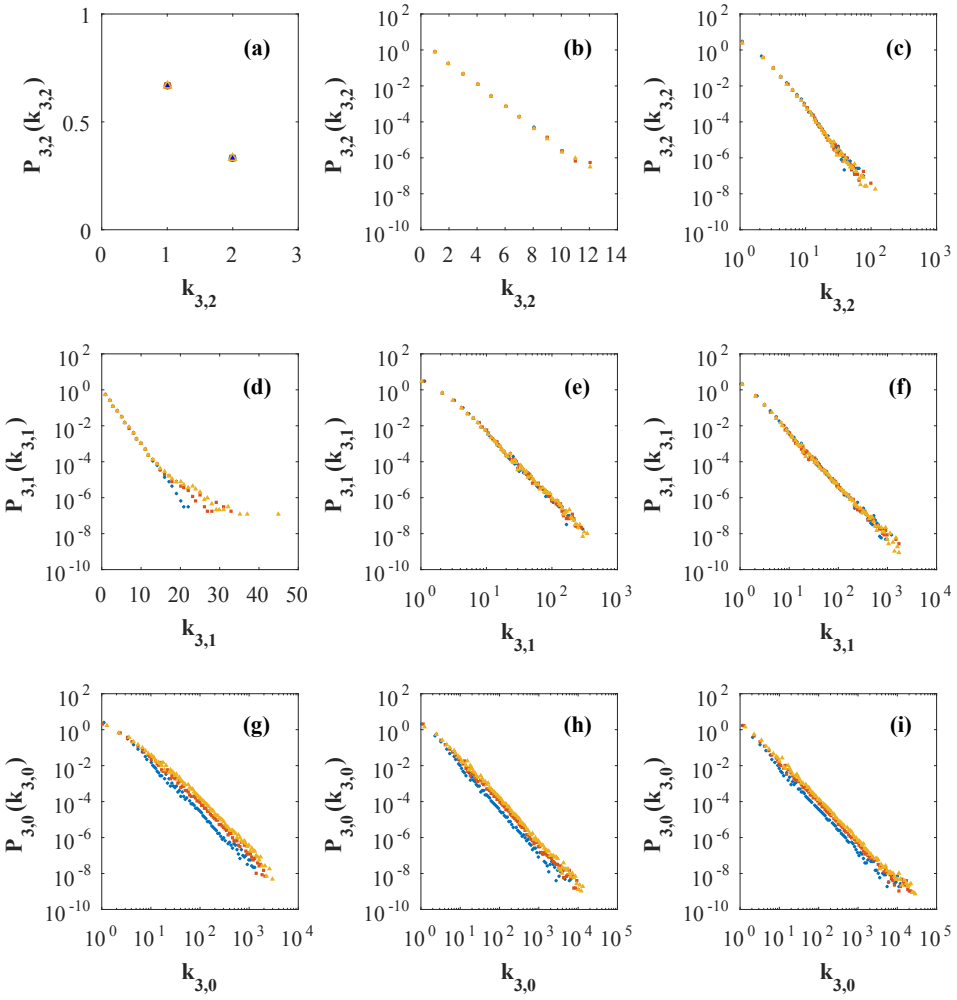

Figure 5.3: Generalized degree distributions $P_{d,\delta}(k_{d,\delta})$ are shown for simplicial complexes of dimension $d = 3$, and flavor $s = -1$ (panels a, d, g), $s = 0$ (panels b, e, h) and $s = 1$ (panels c, f, i), and for faces of dimension $\delta = 2$ (panels a, b, c), $\delta = 1$ (panels d, e, f) and $\delta = 0$ (panels g, h, i). The results of simulations are shown for $m = 1, 2$ and 3 (blue circles, red squares and yellow triangles respectively). The simulated simplicial complexes have $N = 10^5$ nodes and the results are averaged over 10 simplicial complex realizations.

calculate $P\left(\alpha_\delta \in S_\delta(t)|t_{j_0}, t_{j_1}, ..., t_{j_\delta}\right)$. As we described in Section 5.2, in our model each new d-simplex is created from a single new node and an existing $(d-1)$-face. Similarly, for $\delta \leq d-1$, each new δ-face is created from a new node and one of the $(\delta-1)$-dimensional sub-faces of a $(d-1)$-face to which the new node has attached. Furthermore once a simplex or face has arrived in the simplicial complex, it never 'dies' or 'leaves' the simplicial complex.

With these facts in mind, we make the following observations about the process by which new faces arrive in the simplicial complex:

(i) Let $\alpha_\delta = [j_0, j_1, ..., j_\delta]$ be a possible δ-face with node arrival times $t_{j_0} < t_{j_1} < ... < t_{j_\delta}$. Then α_δ is only a face if the $(\delta-1)$-face $\alpha_{\delta-1} = [j_0, j_1, ..., j_{\delta-1}]$ already exists at time t_{j_δ} and at this time the 'new' node j_δ forms a d-simplex incident to some $(d-1)$-face α with $\alpha_{\delta-1} \subset \alpha$.

(ii) Simplices and faces only arrive in the simplicial complex at the arrival time of their youngest node t_{j_δ}, and cannot 'die', so for any time $t > t_{j_\delta}$, the probability that α_δ is a face at time t is equal to the probability that it was a face at time $t_{j_\delta} + 1$, i.e.
$P\left(\alpha_\delta \in S_\delta(t)|t_{j_0}, t_{j_1}, ..., t_{j_\delta}\right) = P\left(\alpha_\delta \in S_\delta(t_{j_\delta}+1)|t_{j_0}, t_{j_1}, ..., t_{j_\delta}\right)$. In the rest of this chapter we write this probability with the short hand p_{α_δ}.

The probability p_{α_δ} can thus be calculated using the recurrence relation

$$\begin{aligned} p_{\alpha_\delta} &= p_{\alpha_{\delta-1}} \pi_{\delta-1}(t_{j_\delta}, t_{j_{\delta-1}}) \\ &= \prod_{n=0}^{\delta-1} \pi_n(t_{j_{n+1}}, t_{j_n}) \end{aligned} \tag{5.29}$$

where $\pi_0(t_{j_1}, t_{j_0}) = p_{\alpha_1}$ is the probability of the link $\alpha_1 = [j_0, j_1]$, while for $\delta > 1$, $\pi_{\delta-1}(t_{j_\delta}, t_{j_{\delta-1}}) = P\left(\alpha_\delta \in S_\delta(t_{j_\delta})|\alpha_{\delta-1} \in S_{\delta-1}(t_{j_{\delta-1}})\right)$ is the conditional probability that the node j_δ forms a δ-face with the $(\delta-1)$-face $\alpha_{\delta-1}$ *given* $\alpha_{\delta-1}$ exists. Within

our mean-field approximation we can write these probabilities as

$$\begin{aligned}\pi_\delta(t_{j_{\delta+1}}, t_{j_\delta}) &= m\Pi_\delta^{t_{j_{\delta+1}}}(\alpha_\delta) = \frac{(1-s)c_\delta + \big(d+s-\delta-1\big)k_{d,\delta}(t_{j_{\delta+1}}, t_{j_\delta})}{\big(d+s\big)t_{j_{\delta+1}}} \\ &= c_\delta \frac{d-\delta}{d+s} t_{j_\delta}{}^{\frac{1+\delta}{d+s}-1} t_{j_{\delta+1}}^{-\frac{1+\delta}{d+s}}, \end{aligned} \tag{5.30}$$

where to get from the first line to the the second line we have inserted the mean-field expressions for $k_{d,\delta}(t_{j_{\delta+1}}, t_{j_\delta})$ given in Eq. (5.21). Finally, using Eq. (5.29) and Eq. (5.30) we get a closed expression for the probability p_{α_δ} of a δ-face as a function of the times $\{t_{j_1}, t_{j_2}, \ldots, t_{j_\delta}\}$ of arrival of its nodes in the simplicial complex, given by

$$p_{\alpha_\delta} = m\frac{d!}{(d-\delta)!(d+s)^\delta}\big(t_{j_0}t_{j_1}...t_{j_{\delta-1}}\big)^{\frac{1}{d+s}-1} t_{j_\delta}^{-\frac{\delta}{d+s}}. \tag{5.31}$$

This probability will prove useful in Section 5.3.4 for calculating the generalized strengths of faces conditioned on the arrival times of their nodes.

5.3.3 Mean-field solution for the weight of a simplex

In this section we derive a mean-field expression for the weight $w(t, t_\alpha)$ that the d-dimensional simplex α added to the simplicial complex at time t_α has at time t. In combination with the expression for the probability of a simplex given Eq. (5.31) of the previous section, we will use the results of this section to help us calculate the generalised strengths of faces in Section 5.3.4.

At each time step of our model we reinforce m' random simplices increasing their weight by w_0. The expected change in weight during $[t, t+1]$ of a simplex with weight $w(t, t_\alpha)$ at time t is $m'\tilde{\Pi}_d(\alpha)$, with $\tilde{\Pi}_d(\alpha)$ as given in equation (5.7). Similar to our approach to calculating the generalised degrees in the Section 5.3.1, we make a continuous time approximation, resulting in the following differential equation

$$\frac{\partial w(t, t_\alpha)}{\partial t} = w_0 m' \tilde{\Pi}_d(\alpha), \tag{5.32}$$

with initial condition

$$w(t_\alpha, t_\alpha) = w_0 \tag{5.33}$$

since each new simplex initially has weight w_0.

At each time step m new simplices, each of weight w_0, are added to the simplicial complex, and m' existing simplices increase their weight by w_0. Therefore the normalization constant $\tilde{\mathcal{Z}}_t$ can be approximated for $t \gg 1$ as

$$\tilde{\mathcal{Z}}_t = \sum_{\alpha \in S_d(t)} w_\alpha(t) = (m' + m)w_0 t + W_I \simeq (m' + m)w_0 t, \tag{5.34}$$

where W_I is the total weight of the initial simplicial complex and is negligible compared to $\tilde{\mathcal{Z}}_t$ for $t \gg 1$. Eq. (5.32) can then be written as

$$\frac{\partial w(t, t_\alpha)}{\partial t} = \lambda \frac{w(t, t_\alpha)}{t}, \tag{5.35}$$

with

$$\lambda = \frac{m'}{m + m'}. \tag{5.36}$$

This equation, together with the initial condition expressed in Eq. (5.33) is easily solved:

$$w(t, t_\alpha) = w_0 \left(\frac{t}{t_\alpha} \right)^\lambda. \tag{5.37}$$

This mean-field expression can be interpreted as the expected weight of a simplex α conditioned on its arrival time t_α and also conditioned on its *existence*. Together with the probability of existence of a simplex calculated in the previous section, this will be used in the following section to calculate the average contribution to the generalised strength of a face coming from simplices with given arrival times of their nodes.

5.3.4 Mean-field approach for the generalized strengths

In this section we evaluate the generalized strength of a δ-face within our mean-field approximation. Recall that the generalized strength of a face of dimension δ is the sum of the weights of the simplices of dimension d of which that face is a subset.

In the spirit of the mean-field approximation, i.e. neglecting fluctuations, we approximate the generalized strength $s^{\alpha}_{d,\delta}$ of a δ-face α by its expected value $s_{d,\delta}(t, t_\alpha)$ over different simplicial complex realizations and conditioned on the existence of the face α with arrival time t_α. This is given by

$$s_{d,\delta}(t, t_\alpha) = \frac{\sum_{\alpha' \in Q_d(N) | \alpha' \supset \alpha} p_{\alpha'} w(t, t_{\alpha'})}{p_\alpha}. \tag{5.38}$$

The sum in the numerator is over all possible d-simplices that could exist at time t and which would be incident to α. The numerator is thus the expected sum of the weights of each of these possible d-simplices. We divide by p_α in the denominator to get the expected generalized strength of α conditional on its existence. As we have shown in the previous two sections, within our mean-field continuous time approximation the probabilities p_α and $p_{\alpha'}$ and the weights $w(t, t_{\alpha'})$ are functions of the birth times of the nodes contained within α and α'. In this section we seek to approximate the sum in Eq. (5.38) with a set of integrals over the birth times of the nodes of α'.

To this end we indicate each δ-simplex α by the set of its nodes $[i_0, i_1, \ldots, i_\delta]$ ordered according to the arrival times in the simplicial complex $t_{i_0} < t_{i_1} < \ldots < t_{i_\delta} = t_\alpha$. Similarly we will indicate each d-simplex α' by the ordered set of its nodes $[j_0, j_1, \ldots, j_d]$ ordered according to the arrival times in the simplicial complex $t_{j_0} < t_{j_1} < \ldots < t_{j_d} = t_{\alpha'}$.

The transition from the sum in Eq. (5.38) to an integral over the birth times $t_{j_0}, t_{j_1}, \ldots, t_{j_d}$ is complicated slightly by the fact that α is a face of each α' (i.e. $[i_0, i_1, \ldots i_\delta] \subset [j_0, j_1, \ldots j_n]$) and so α fixes the birth times of exactly $\delta + 1$ of the nodes in each α'. As we will see, the structure of our integrals will in fact depend on

the 'position' that each of the birth times belonging to α takes in the the set of birth times belonging to α'. We indicate this position with the following notation. For the node $i_r \subset \alpha$ (i.e. the rth node in α) we indicate its *index* in the list $[j_0, j_1, \ldots j_d]$ with $q(r)$. We therefore have

$$i_r = j_{q(r)}. \tag{5.39}$$

In order to be concrete, consider the following example. In a simplicial complex of dimension $d = 4$ consider the 4-simplex α'

$$\alpha' = [j_0, j_1, j_2, j_3, j_4] = [5, 7, 11, 19, 25] \tag{5.40}$$

and the 1-face α

$$\alpha = [i_0, i_1] = [7, 19]. \tag{5.41}$$

Since $i_0 = j_1$ and $i_1 = j_3$ we have

$$q(0) = 1, \quad q(1) = 3. \tag{5.42}$$

Additionally, for a face α, the positions set by $\{q(r)\}_{r=0,1,\ldots\delta}$ fix not only the birth times of the nodes $t_{j_{q(0)}} = t_{i_0}, \ldots, t_{j_{q(\delta)}} = t_{i_\delta}$, but also tell us precisely how many nodes of α' are born in between each consecutive pair of nodes $i_r, i_{r+1} \in \alpha$. This will prove useful for constructing the integrals over the birth times, because for fixed $\{q(r)\}_{r=0,1,\ldots\delta}$ we know that there are $q(r+1) - q(r) - 1$ birth times to integrate over between $t_{j_{q(r)}}$ and $t_{j_{q(r+1)}}$ and that the integration limits of these integrals will be t_{i_r} and $t_{i_{r+1}}$. In particular, for a fixed $\{q(r)\}_{r=0,1,\ldots\delta}$ we will integrate over the birth times of the nodes in α' using integrals of the form

$$\int_{t_{i_r}}^{t_{i_{r+1}}} dt_{j_{q(r+1)-1}} \int_{t_{i_r}}^{t_{j_{q(r+1)-1}}} dt_{j_{q(r+1)-2}} \cdots \int_{t_{i_r}}^{t_{j_{q(r)+3}}} dt_{j_{q(r)+2}} \int_{t_{i_r}}^{t_{j_{q(r)+2}}} dt_{j_{q(r)+1}} \big(\cdot \big) \tag{5.43}$$

The above form integrates over all possible combinations of the (non-fixed) birth times $t_{j_{q(r)+1}} < \ldots < t_{j_{q(r+1)-1}}$ between the (fixed) birth times $t_{j_{q(r)}} = t_{i_r}$ and $t_{j_{q(r+1)}} = t_{i_{r+1}}$.

Having introduced all of the above concepts and notation we are now in a position to approximate Eq. (5.38) with a set of integrals over birth times. We sum over all possible positions $\{q(r)\}_{r=0,1,...\delta}$ that the nodes of α can take in α', and get the following expression for the expected generalized strength of α

$$s_{d,\delta}(t,t_\alpha) = \\ \frac{1}{p_{[i_0,\ldots,i_\delta]}} \sum_{\{q(r)\}_{r=0,1,...\delta}} \int_{t_{j_0}<\ldots<t_{j_d}} \left[\prod_{n=0}^{d} dt_{j_n}\right] \prod_{r=0}^{\delta} \hat{\delta}\left(t_{j_{q(r)}}, t_{i_r}\right) p_{[j_0,\ldots,j_d]} w(t,t_{j_d}), \tag{5.44}$$

where $\hat{\delta}(x,y)$ indicates the Kronecker delta. Using Eq. (5.31) for the probability $p_{[j_0,\ldots,j_d]}$ and Eq. (5.37) for the weight $w(t,t_d)$ we get

$$\begin{aligned} s_{d,\delta}(t,t_\alpha) = \; & \frac{1}{p_\alpha} w_0 m \frac{d!}{(d+s)^d} t^\lambda \left(t_{i_0} t_{i_1} ... t_{i_\delta}\right)^{\frac{1}{d+s}-1} \\ & \times \sum_{\{q\}} A_{q(\delta)} \left(\prod_{r=0}^{\delta-1} X_{q(r),q(r+1)}\right) B_{q(0)}, \end{aligned} \tag{5.45}$$

where $A_{q(\delta)}$, $X_{q(r),q(r+1)}$ and $B_{q(0)}$ are nested integrals taking similar forms to that in Eq. (5.43) and integrate over birth times greater than $t_{j_{q(\delta)}}$, between $t_{j_{q(r)}}$ and $t_{j_{q(r+1)}}$, and less than $t_{j_{q(0)}}$ respectively. To be clearer, $A_{q(\delta)}$ integrates over the non-fixed birth times $t_{j_{q(\delta)+1}} < \ldots < t_{j_d}$ between the final fixed birth time $t_{j_{q(\delta)}} = t_{i_\delta}$ and time t, while $X_{q(r),q(r+1)}$ integrates over the non-fixed birth times $t_{j_{q(r)+1}} < \ldots < t_{j_{q(r+1)-1}}$ between the fixed birth times $t_{j_{q(r)}} = t_{i_r}$ and $t_{j_{q(r+1)}} = t_{i_{r+1}}$, and $B_{q(0)}$ integrates over the non-fixed birth times $t_{j_0} < \ldots < t_{j_{q(0)-1}}$ between time 0 and the first fixed birth time $t_{j_{q(0)}} = t_{i_0}$.

It will be convenient to express all of these quantities in terms of the function $I(\tau_L, \tau_U)$ defined to be

$$I^n_{\tau_L,\tau_U} \; = \; \int_{\tau_L}^{\tau_U} dt_n t_n^{\frac{1}{d+s}-1} \int_{\tau_L}^{t_n} dt_{n-1} t_{n-1}^{\frac{1}{d+s}-1} ... \int_{\tau_L}^{t_2} dt_1 t_1^{\frac{1}{d+s}-1}.$$

In particular, by distinguishing between the cases in which there is at least one node whose arrival time is being integrated over, and the case where the allocation of positions specified by $\{q\}$ implies that there are no arrival times to integrate over we obtain

$$A_{q(\delta)} = \begin{cases} \int_{t_{i_\delta}}^{t} dt_{j_d} t_{j_d}^{-\lambda-\frac{d}{d+s}} I_{t_{i_\delta},t_{j_d}}^{d-q(\delta)-1} & \text{if} \quad 0 \leq q(\delta) \leq d-1, \\ t_{i_\delta}^{-\lambda+\frac{s-1}{d+s}} & \text{if} \qquad q(\delta) = d, \end{cases} \tag{5.46}$$

$$X_{q(r),q(r+1)} = \begin{cases} I_{t_{i_r},t_{i_{r+1}}}^{q(r+1)-q(r)-1} & \text{if} \quad q(r+1)-q(r) > 1, \\ 1 & \text{if} \quad q(r+1)-q(r) = 1, \end{cases} \tag{5.47}$$

$$B_{q(0)} = \begin{cases} I_{0,t_{i_0}}^{q(0)} & \text{if} \quad q(0) > 0, \\ 1 & \text{if} \quad q(0) = 0. \end{cases} \tag{5.48}$$

We also note that Eq. (5.45) may be simplified further, by substituting the expression for p_α given in Eq. (5.31):

$$s_{d,\delta}(t,t_\alpha) = w_0 \frac{(d-\delta)!}{(d+s)^{d-\delta}} t^\lambda t_{i_\delta}^{-\frac{d+s-\delta-1}{d+s}} \sum_{\{q\}} A_{q(\delta)} \left(\prod_{r=0}^{\delta-1} X_{q(r),q(r+1)} \right) B_{q(0)}. \tag{5.49}$$

Calculation of the strength thus requires the evaluation of the integrals in Eq.s (41-5.48) followed by the evaluation of the sum over the positions $\{q(r)\}_{r=0,1,\ldots\delta}$. In Appendix C we calculate these quantities exactly, finding that $s_{d,\delta}(t,t_\alpha)$ is given by

$$s_{d,\delta}(t,t_\alpha) = \begin{cases} w_0 \frac{d-\delta}{(d+s)(\lambda_\delta-\lambda)} \left(\frac{t}{t_\alpha}\right)^{\lambda_\delta} + w_0 \left[1 - \frac{d-\delta}{(d+s)(\lambda_\delta-\lambda)}\right] \left(\frac{t}{t_\alpha}\right)^{\lambda} & \text{if} \quad \lambda \neq \lambda_\delta, \\ w_0 \left(\frac{t}{t_\alpha}\right)^{\lambda} \left[1 + \frac{d-\delta}{d+s} \log\left(\frac{t}{t_\alpha}\right)\right] & \text{if} \quad \lambda = \lambda_\delta, \end{cases} \tag{5.50}$$

where λ_δ is given by Eq. (5.22) and λ is given by Eq. (5.36). We have therefore achieved our objective of calculating the generalized strength at time t of a δ-face α as a function of its birth times and the model parameters. Remarkably, Eq. (5.50) depends only on the ratio between the time t and the birth time of α $t_\alpha = t_{i_\delta}$, and

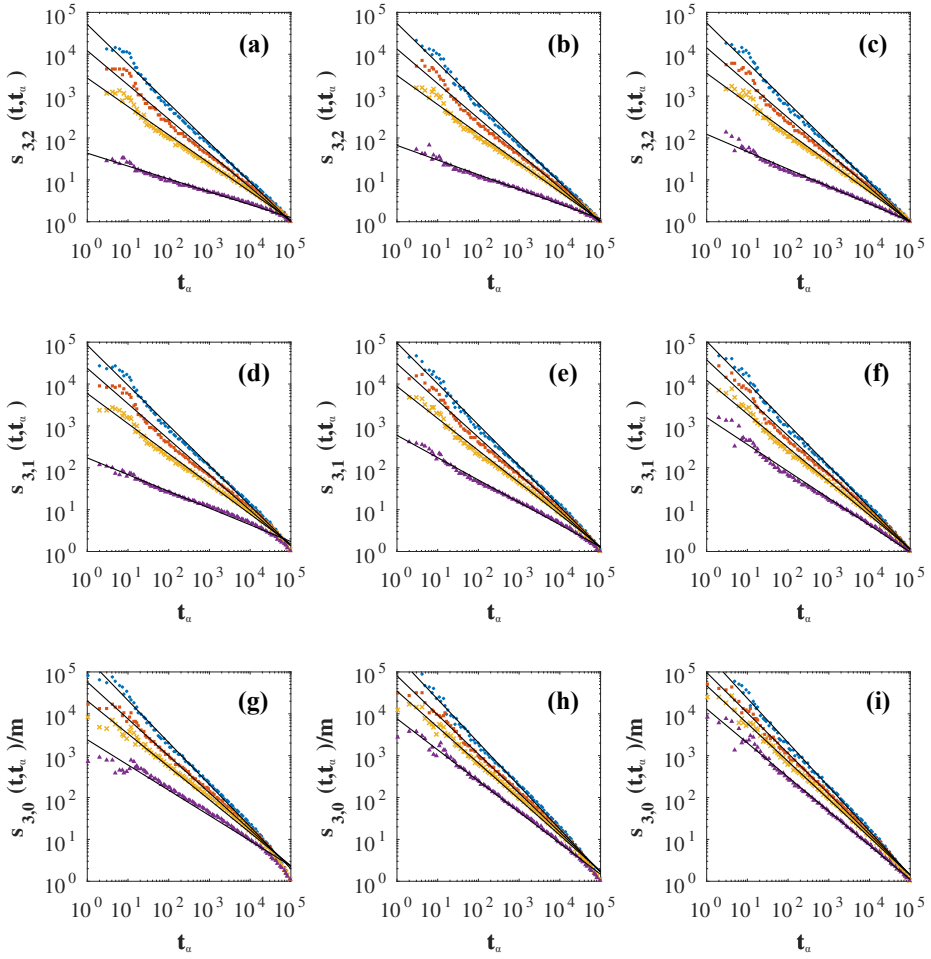

Figure 5.4: The average generalized strengths $s_{d,\delta}(t, t_\alpha)$ of the δ-faces arrived in the network at time t_α are shown for simplicial complexes of dimension $d = 3$, and flavor $s = -1$ (panels a, d, g), $s = 0$ (panels b, e, h) and $s = 1$ (panels c, f, i) and for faces of dimension $\delta = 2$ (panels a, b, c), $\delta = 1$ (panels d, e, f) and $\delta = 0$ (panels g, h, i). The results of simulations are shown for $(m = 1,\ m' = 5)$, $(m = 2,\ m' = 5)$, $(m = 2,\ m' = 3)$ and $(m = 3,\ m' = 1)$ (blue circles, red squares, yellow x's and purple triangles respectively). The simulated simplicial complexes have $N = 10^5$ nodes and the results are averaged over 10 simplicial complex realizations.

not on the birth times of other nodes in α.

As discussed in the introduction to this chapter, we are particularly interested in how the growth and reinforcement dynamics of our model (controlled by the model parameters) affect the distribution of weight across the simplicial complex. This distribution can be characterized by the scaling of the generalized strengths of the simplices with their generalized degrees: a linear scaling indicates a homogeneous distribution of weight, while a super-linear scaling would indicate an inhomogeneous distribution where faces with high generalized degrees are incident to d-simplices with high weight. We can use our expression for the generalized strength given in Eq. (5.50) in combination with the results we obtained about the generalized degree in Section 5.3.1 to find the generalized strength-degree scaling that emerges in our model.

To this end we keep only the leading terms for $t/t_\alpha \gg 1$ both in Eq. (5.50) for the average generalized strength $s_{d,\delta}(t, t_\alpha)$ and in Eq. (5.21) for the average generalized degree $k_{d,\delta}(t, t_\alpha)$ and we neglect the fluctuations of the generalized strengths ($s_{d,\delta}(t, t_\alpha) \simeq s^\alpha_{d,\delta}$) and generalized degrees ($k_{d,\delta}(t, t_\alpha) \simeq k^\alpha_{d,\delta}$).

The respective scalings with respect to t/t_α are

$$s_{d,\delta}(t, t_\alpha) \propto \begin{cases} \left(\frac{t}{t_\alpha}\right)^{\lambda_\delta} & \text{if} \quad \lambda < \lambda_\delta, \\ \left(\frac{t}{t_\alpha}\right)^{\lambda} \log\left(\frac{t}{t_\alpha}\right) & \text{if} \quad \lambda = \lambda_\delta, \\ \left(\frac{t}{t_\alpha}\right)^{\lambda} & \text{if} \quad \lambda > \lambda_\delta, \end{cases} \tag{5.51}$$

for the generalized strengths and

$$k_{d,\delta}(t, t_\alpha) \propto \begin{cases} \left(\frac{t}{t_\alpha}\right)^{\lambda_\delta} & \text{if} \quad \lambda_\delta \neq 0, \\ \log\left(\frac{t}{t_\alpha}\right) & \text{if} \quad \lambda_\delta = 0, \end{cases} \tag{5.52}$$

for the generalized degrees. Recall that $\lambda = \frac{m'}{m+m'}$ is the proportion of the weight added at each time step that is added via the reinforcement mechanism, while

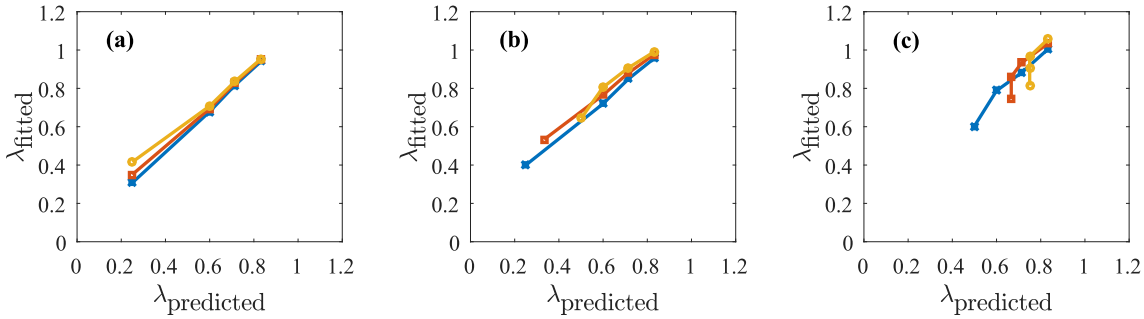

Figure 5.5: The exponents λ_{fitted} obtained by fitting Eq. (5.55) to the data in figure 5.4 are shown versus the predicted exponents $\lambda_{predicted}$ given by Eq. 5.56 for different δ-faces. The panels (a), (b) and (c) refer respectively to triangles ($\delta = 2$), links ($\delta = 1$) and nodes ($\delta = 0$). The blue stars, red squares, and yellow circles indicate the data obtained respectively for the the flavors $s = -1$, $s = 0$ and $s = 1$.

$\lambda_\delta = \frac{d+s-\delta-1}{d+s}$ governs the behaviour of the generalized degree of a face with given dimension δ. Using Eq.s (5.51) and (5.52) above we can find a scaling relation for the generalized strength of some face given its generalized degree. In fact, as long as $\lambda_\delta > 0$, we obtain

$$s_{d,\delta}(k_{d,\delta}) \propto \begin{cases} k_{d,\delta} & \text{for} \quad \lambda < \lambda_\delta, \\ k_{d,\delta} \ln k_{d,\delta} & \text{for} \quad \lambda = \lambda_\delta, \\ (k_{d,\delta})^{\lambda/\lambda_\delta} & \text{for} \quad \lambda > \lambda_\delta. \end{cases} \tag{5.53}$$

For $\lambda_\delta = 0$, instead we derive an exponential scaling of the average of the generalized strength versus the average of the generalized degree of the δ-faces, i.e.

$$s_{d,\delta}(k_{d,\delta}) \propto e^{\beta k_{d,\delta}}, \tag{5.54}$$

with $\beta = \lambda \frac{d+s}{(1-s)c_\delta}$.

These results predict that by tuning the parameter values (d,s,m,m') it is possible to observe either linear, superlinear or even exponential scalings of the generalized strengths versus the generalized degrees. Remarkably, for some choices of the parameter values (e.g. $d = 4$, $s = 0$, and $m = m'$) our predictions suggest it is possible to observe all three types of scaling for faces of different dimension δ within the same

simplicial complex. We stress here that the scaling relations Eqs. (5.53) and (5.54) are obtained in the limit $t/t_\alpha \gg 1$, and neglecting the fluctuations of the generalized degrees and the generalized strengths over different network realizations. Therefore these expressions need to be compared to numerical simulations for assessing the limits of the considered approximations. In the next section we do exactly this, and find that for $\lambda_\delta > 0$ Eq. (5.53) gives a good prediction of the numerically obtained scaling while for $\lambda_\delta = 0$ the prediction given by Eq. (5.54) differs markedly from the numerical scaling, suggesting that in this case the fluctuations of the generalized strengths and generalized degrees around their mean-field values cannot be neglected entirely.

5.4 Numerical simulations

In order to check the validity of our mean-field calculations we ran extensive simulations of our model. The code for these simulations is available online [85].

We use the code to generate simplicial complexes from our model and compare the generalised degree and generalised strength statistics to the predictions given in Section 5.3.4. In particular, for our simulations we use simplices of dimension $d = 3$ (tetrahedra) and all possible values of the flavor $s = -1, 0, 1$ as well as a variety of choices of m and m'. For each combination of the parameters we generate 10 realisations of the model and then average the statistics over these realisations. Our main goal is to characterize the limit of validity of the mean-field calculations performed in the Section 5.3.4.

In Figure 5.3 we show the simulation results for the generalized degree distribution $P_{d,\delta}(k_{d,\delta})$ and $N = 10^5$ averaged over 10 realizations of the model. We observe that the mean-field calculation accurately predicts for which dimension δ and for which flavor s we observe bimodal, exponential or power-law degree distributions.

In Figure 5.4 we show the average generalized strengths $s_{d,\delta}(t, t_\alpha)$ of δ-faces α as a function of their arrival time t_α. Here we observe a clear power-law scaling of

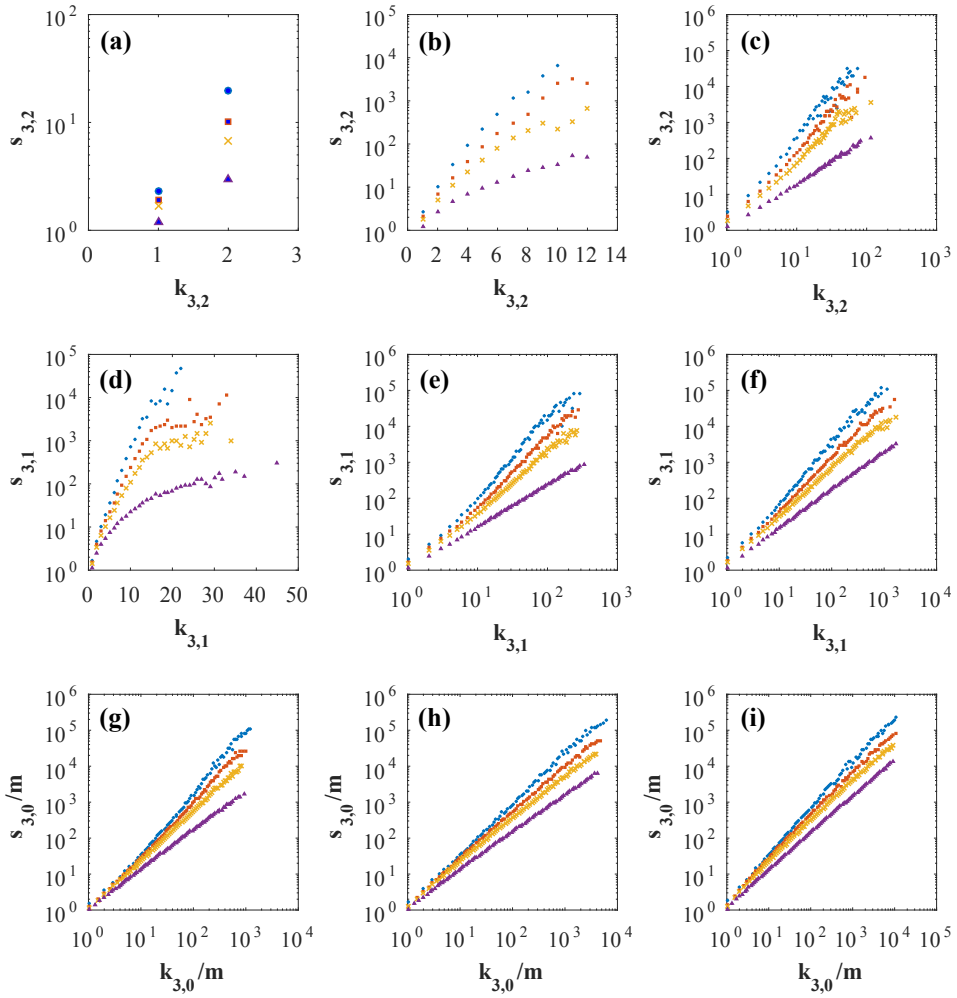

Figure 5.6: The average generalized strengths of δ-faces as a function of their corresponding generalized degree $s_{d,\delta}(k_{d,\delta})$ are shown for simplicial complexes of dimension $d = 3$, and flavor $s = -1$ (panels a, d, g), $s = 0$ (panels b, e, h) and $s = 1$ (panels c, f, i) and for faces of dimension $\delta = 2$ (panels a, b, c), $\delta = 1$ (panels d, e, f) and $\delta = 0$ (panels g, h, i). The results of simulations are shown for ($m = 1$, $m' = 5$), ($m = 2$, $m' = 5$), ($m = 2$, $m' = 3$) and ($m = 3$, $m' = 1$) (blue circles, red squares, yellow x's and purple triangles respectively). The simulated simplicial complexes have $N = 10^5$ nodes and the results are averaged over 10 simplicial complex realizations.

$s_{d,\delta}(t, t_\alpha)$ as a function of t/t_α for $t/t_\alpha \gg 1$ as predicted by the mean field approximation (Eq. 5.51).

In order to evaluate in closer detail the validity of our mean-field approach we have also checked whether the exponents of the power-law scalings observed in Figure 5.4 are well predicted by our mean-field equations. To this end we performed the following power-law fits

$$s_{d,\delta}(t, t_\alpha) = a \left(\frac{t}{t_\alpha} \right)^{\lambda_{fitted}}, \tag{5.55}$$

valid for $t/t_\alpha \gg 1$. We have then compared the fitted exponent λ_{fitted} with the predicted exponent obtained from Eq. (5.51)

$$\lambda_{predicted} \simeq \max(\lambda, \lambda_\delta). \tag{5.56}$$

The above prediction for the exponent $\lambda_{predicted}$ comes from Eq. (5.51) which itself gives the leading order terms from Eq. (5.50). This is valid with the assumption $t/t_\alpha \gg 1$, as the term with the larger exponent will dominate the term with the smaller exponent.

The comparison between the exponents λ_{fitted} and $\lambda_{predicted}$ is shown in Figure 5.5 for simplicial complexes of dimension $d = 3$ and different values of m and m' determining λ. We observe that while the overall trend of λ_{fitted} is captured by the mean-field result, some deviations are observed. These deviations become more significant for $\lambda \simeq \lambda_\delta$ where it is expected to be more difficult to observe the leading term in Eq. (5.50) starting for finite time simulations results.

Finally, as discussed in Section 5.3.4, for simplicial complexes with a large number of nodes the mean-field approximation predicts that the generalized strengths of the faces are related to their generalized degrees by the scaling relation given in Eqs. (5.53) and (5.54). As we have mentioned previously, this approximation neglects the role of fluctuations for the generalized degree and for the generalized strengths. It is therefore important to check to what extend the mean-field calculations capture

the simulation results. Figure 5.6 shows the average generalized strength versus the generalized degree $s_{d,\delta}(k_{d,\delta})$.

We observe that the role of fluctuations is particularly pronounced for δ-faces with exponential generalized degree distributions (i.e. $\delta = d - 2$ for $s = -1$ and $\delta = d - 1$ for $s = 0$). These fluctuations are more significant for values of the parameters m and m' corresponding to low-values of λ.

For the other combinations of the parameters and face dimension δ which do not correspond to exponential generalized degree distributions, we instead find the mean-field predictions provides a rather good prediction of the scaling of the average generalized strength $s_{d,\delta}(k_{d,\delta})$.

5.5 Conclusions

The research presented in this chapter, originally published in [34], follows in the tradition of non-equilibrium network models that seek to characterize the relations between growth mechanisms of networks and their structural properties. Our model is a non-equilibrium model of a weighted simplicial complex. In this model the simplicial complex evolves at each time by A) the addition of a new node belonging to m new d-dimensional simplices and B) the reinforcement of the weights of m' d-dimensional simplices. For certain choices of the parameters our model reduces to the weighted BA model of [43] or the NGF of [57] both of which were discussed in Chapter 3.

The 'growth' and 'reinforcement' dynamics of our model differ in a significant way from those in network models. They act on simplices of dimension $d - 1$ and d respectively, with probabilities that depend on properties belonging to those simplices. Thus the model dynamics is not limited to depending only on 'node properties' but instead considers the properties of groups of nodes represented by simplices.

In particular, we saw in Section 5.3.1 how a 'manifold' attachment mechanism at

dimension $d-1$ is felt as a uniform attachment mechanism at dimension $d-2$ and as a preferential attachment mechanism at dimension $d-3$, and similarly uniform attachment mechanisms at dimension $d-1$ are felt as preferential attachment at dimension $d-2$. These 'effective' attachment mechanisms are responsible for the distributions of the generalised degrees that we find for the lower dimensional faces, and so our model has uncovered new mechanisms for generating scale-free and exponential distributions of the generalised degree (and degree) via higher dimensional attachment.

Our model allows us to investigate the competing effects of growth and reinforcement on the generalized strengths of the simplices and on the distribution of weight throughout the simplicial complex. We found that depending on the choice of the model parameters that the evolution of the generalised strength of a simplex is either dominated by the accumulation of new simplices via the growth process or the accumulation of new weight via the reinforcement process. We also characterized the distribution of weight across the simplicial complex by finding scaling relations between the generalized strengths of the simplices and their strengths, finding that the model is capable of producing linear, super-linear or exponential scalings of generalised strength with generalised degree.

We believe that this model could be rather fruitful for modelling real-world simplicial complexes that are typically weighted such as collaboration networks. Additionally the model could be used a benchmark to test the wide range of topological and geometrical measures and computational techniques that have been proposed in recent years for the study of real datasets. These include different definitions of curvature, and the persistent homology conducted using a filtration based on the weights of the links or on the simplices.

Chapter 6

Dense networks and Simplicial Complexes

In this chapter we present a modelling framework for producing dense networks and simplicial complexes with power-law distributions of the degrees or generalised degrees. The original research presented in this chapter was published in our paper *Dense Power-law Networks and Simplicial Complexes* [35].

Power-law degree distributions $P(k) \propto k^{-\gamma}$ have been identified as a universal property of complex networks [1, 45] and have been found in areas as diverse as the world wide web [87], cell biology [88] and scientific co-authorship [89]. In the vast majority of these networks the power-law exponent γ is in the range $(2, 3)$, resulting in a finite average degree

$$\langle k \rangle \approx C^{-1} \int_1^N dk k^{1-\gamma} = \frac{C}{2-\gamma} \left(N^{2-\gamma} - 1 \right) \sim \mathcal{O}(1), \tag{6.1}$$

where $C = \int_1^N dk k^{-\gamma}$ normalises the distribution, and where in the above the degrees are approximated as continuous variables. We say that these networks are *sparse*, as their average degree remains constant as the system size increases. In contrast, for $\gamma < 2$ the average degree grows with the system size like $\langle k \rangle \sim \mathcal{O}(N^{2-\gamma})$, and so the networks in this case are *dense*. We say that a power-law network is *scale-free* if $\gamma \leq 3$ as the second moment diverges with the system size like $\langle k^2 \rangle \sim \mathcal{O}(N^{3-\gamma})$ and

so the degree distribution has no 'typical' or 'characteristic' scale. Networks with $\gamma \in (2, 3)$ are thus sparse and scale-free while networks with $\gamma \leq 2$ are dense and scale-free.

The ubiquitousness of power-law degree distributions and in particular sparse power-law degree distributions with $\gamma \in (2, 3)$ has motivated the development of numerous explanatory models. These models seek to explain the emergence of power-law distributions in complex networks through simple stochastic processes that may mirror the processes that occur in real networks.

The pivotal Barabási and Albert model [45] was developed to explain the emergence of power-law degree distributions by including two simple yet fundamental elements: a growth element in which new nodes are continuously added to the network, and a preferential attachment element where the new nodes form links with exisiting nodes with probabilities proportional to the degrees of the exisiting nodes. This work triggered the formulation of several other models for producing power-law networks, including the Bianconi-Barabási model [46, 47], the non-linear preferential attachment model [48] and the model with initial attractiveness of the nodes [58].

These models all produce sparse power-law networks with exponent $\gamma \in (2, \infty)$, and so are suited as explanatory models for the majority of complex networks. However, not all networks are sparse. There is increasing evidence that dense networks often occur in on-line social networks [38, 39], recommendation networks [38] and in the brain [40].

Furthermore the vast majority of *these* networks appear to be both dense and scale-free, and specifically have been found to have a power-law exponent $\gamma \in (1, 2]$. For this reason the development of new theoretical frameworks for modelling these networks is highly relevant.

However, producing networks that are both dense and scale-free is far from simple. In [36] it was shown that dense scale-free simple networks are not graphical unless some cutoff on the maximum degree is imposed. This means that for these

degree distributions no 'simple' network can be produced where there are no multiedges or self-loops (also called tadpoles). Casting the problem in terms of the configuration model discussed in Chapters 2 and 4, the essential problem is that it is not possible to find 'wirings' for all of the stubs of the high degree nodes without creating multiple links between two nodes, or without wiring some nodes to themselves.

This result reflects the fact that realizing dense power-law networks is rather challenging compared to realizing sparse networks. However it does not imply that dense power-law networks do not exist. The fact that dense scale-free networks do exist is demonstrated by the existence of a few modelling frameworks that extend the configuration model to dense scale-free networks [39] by imposing a suitable structural cutoff, or that generate dense power-law networks with specific values of the power-law exponents (i.e. $\gamma = 1$ [90] or $\gamma = 1.5$ [39]).

In the research presented in this chapter we propose a theoretical framework that is designed to generate dense growing power-law networks and simplicial complexes using preferential attachment and without imposing any ad hoc cutoff. Our approach is based on the Pitman-Yor process [91–93], also known as the Chinese Restaurant Process. This process was originally defined for generating exchangeable partitions or, in more physical terms it is defined as a ball-in-the-box process. Our primary aim was to generate growing dense power-law networks using a variation of the Pitman-Yor algorithm.

We have developed three distinct models of weighted networks and simplicial complexes. In each model, the total number of nodes and their strengths evolve analogously to the Pitman-Yor process, yielding dense power-law strength distributions with tunable exponent $\gamma \in (1, 2]$. A link (or simplex) is considered to be present in the network (or simplicial complex) as soon as its weight is greater than 0, and so the degree of a node is closely related to, but not the same as its strength. The central question we address in this work is what is the degree distribution of these networks and how does it relate to the strength distribution?

In the first of our models the networks produced are undirected. While the distribution of the strengths in this model is dense with tunable $\gamma \in (1, 2]$, the degree distribution is only marginally dense with $\gamma = 2$, regardless of the choice of the parameter of the model. Out of our three models, this is the only one producing the kind of networks under discussion in Ref. [36], and we believe that the inability of our model to produce dense networks with exponent $\gamma < 2$ supports the results found in [36].

In the second model we instead consider directed networks. Unlike in the undirected model, in this model both the out-strengths and out-degrees of the nodes (see Section 6.1 for a definition) have a tunable dense exponent $\gamma \in (1, 2]$, while the in-strengths and in-degrees follow homogeneous distributions. These networks are thus truly dense and scale-free in their out-degree distributions, and so this version has a use as a basic explanatory model for the dense scale-free directed networks observed in recommendation networks [38] and the brain [40].

Finally we further extend this directed model to include directed simplicial complexes formed not only by nodes and links but also by triangles. Similar to in the directed network model, the simplicial complexes produced have dense distributions of the generalised out-strengths and generalised out-degrees and homogeneous distributions of the generalised in-strengths and generalised in-degrees.

This chapter is structured as follows: in Section 6.1 we recall some basic concepts for networks that are weighted and possibly directed, and show how these concepts can be extended to simplicial complexes. In Section 6.2 we give an overview of the Yule-Simon and Pitman-Yor processes for generating power-law distributions with exponents $\gamma \in (2, \infty)$ and $\gamma \in (1, 2]$ respectively. In Section 6.3 we present a modelling framework which exploits the Pitman-Yor process in order to produce dense scale-free networks and simplicial complexes. In Section 6.4 we derive mean-field expressions for the total number of nodes and their strengths. We use these results in Section 6.5 to find the probability that a link or triangle is reinforced in any given time-step. In Section 6.6 we derive mean-field equations for the degrees, and show that the degree distributions are scale-free with dense exponent $\gamma \in (1, 2]$. In

Section 6.7 we explore the relation between the strengths and degrees of the nodes. In Section 6.8 we examine the clustering and degree correlations produced by the model. Finally, in Section 6.9 we give the chapter conclusions.

6.1 Definitions and notation

The networks and simplicial complexes that we discuss in this chapter are weighted and either undirected or directed.

In the case of a network, we write the weight of a link between two nodes i and j as w_{ij}, with $w_{ij} = w_{ji}$ if the network is undirected but not necessarily if the network is directed. In our models the weights take non-negative integer values and the presence of the links (or later the simplices) are in fact determined by the weights, with

$$a_{ij} = \begin{cases} 1 & \text{if} \quad w_{ij} > 0 \\ 0 & \text{if} \quad w_{ij} = 0, \end{cases}$$

for any pair of nodes i and j. As discussed in previous chapters, the strength of a node in a weighted network is the sum of the weights of its links. For an undirected network we write the strength of a node i as

$$s_i = \sum_{j=1}^{N} w_{ij}, \tag{6.2}$$

while for a directed network, we count the weight directed in to a node and the weight directed out from the node separately using the in-strength s_i^{in} and out-strength s_i^{out} respectively:

$$s_i^{in} = \textstyle\sum_{j=1}^{N} w_{ji} \qquad s_i^{out} = \textstyle\sum_{j=1}^{N} w_{ij}. \tag{6.3}$$

Similarly, for a directed network we distinguish between the in-degree k_i^{in} and out-

degree k_i^{out}:

$$k_i^{in} = \sum_{j=1}^{N} a_{ji} \qquad k_i^{out} = \sum_{j=1}^{N} a_{ij}. \tag{6.4}$$

In the simplicial complex version of our model, the simplicial complexes are pure 2-dimensional, meaning that they are constructed exclusively from triangles and their sub-faces (nodes and links). The simplicial complexes we consider are also weighted and directed.

Similar to in the networks versions of our model, in this version we associate weights w_{ijl} to each triangle (i, j, l). The triangles are directed in the sense that we map a differently 'directed' or 'oriented' triangle to each permutation of its three nodes, i.e. for three nodes labelled i, j and l we can create 6 distinct directed triangles: ijl, ilj, jil, jli, lij and lji. The first node in the triangle we call the 'source' node, the second we call the 'first target node' and the third we call the 'second target node'. As with an undirected simplicial complex, the triangles contain their faces of dimension 1 (links) and 0 (nodes). In the simplicial complexes we present in this chapter, these links are also directed, and their directions are determined by the direction of their parent triangle. The two links coming from the source node are directed outwards from the source node towards the two target nodes. The third link between the two target nodes is directed from the first target node towards the second target node. Figure 6.1 is a diagram showing the relation between the direction of a triangle and the direction of its links.

Directed triangles allow us to distinguish between the triangles a node has gained from distinct attachment mechanisms in our model through the node's 'generalized out-degree' and 'generalized in-degree'. Of course the specific relation between the direction of a triangle and the directions of its links that we use in this chapter is just a convention that we have chosen, and there are indeed other conventions that could be chosen instead. We have chosen ours as it produces simplices that have what could be called a 'temporal direction', where the links of the simplices produced are acyclic. Interestingly directed simplices of this type have recently been used [94] to analyse brain networks and coupling for topological information with the neuronal

network dynamics.

Similar to in the directed version of the model we define the generalised in-strength and generalised out-strength of a node i as

$$\tilde{s}_i^{in} = \sum_{j,l=1}^{N} (w_{jil} + w_{jli}) \qquad \tilde{s}_i^{out} = \sum_{j,l=1}^{N} w_{ijl}, \tag{6.5}$$

i.e. $\tilde{s}_i^{out}$ is the total weight of the triangles for which node i is the source node, and $\tilde{s}_i^{in}$ is the total weight of the triangles for which it is one of the target nodes. The in-degree and out-degree are defined similarly;

$$\tilde{k}_i^{in} = \sum_{j,l=1}^{N} (a_{jil} + a_{jli}) \qquad \tilde{k}_i^{out} = \sum_{j,l=1}^{N} a_{ijl}. \tag{6.6}$$

6.2 Dense scale-free distributions and the Pitman-Yor Process

In this section we discuss the relation between preferential attachment mechanisms in network models and so called 'ball-in-the-box' models. In Section 6.2.1 we show how the evolution of the degree distribution in the Barabási-Albert model [45] can be mapped to the Yule-Simon model [95], which is a discrete time stochastic process that produces power-law distributions with exponent $\gamma \in (2, \infty)$. As discussed earlier in this chapter, degree distributions with exponents in this range are necessarily sparse. In Section 6.2.2 we discuss the Pitman-Yor process [91] which similar to the Yule-Simon process is a discrete time stochastic process that utilises preferential attachment to produce power-law distributions. However, the distributions produced by Pitman-Yor are in the dense range with tunable exponent $\gamma \in (1, 2]$. Pitman-Yor thus provides us with the inspiration for the models of dense networks and simplicial complexes that we present in Section 6.3.

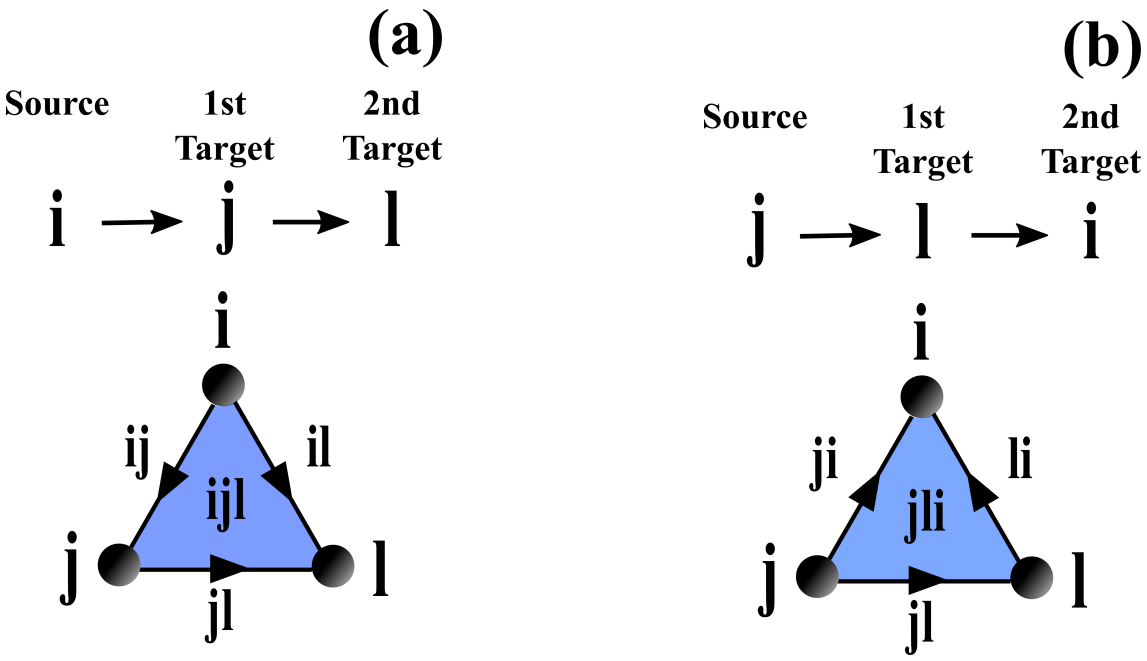

Figure 6.1: Diagram showing the relation between the direction of a triangle and the direction of its links. In panel (a) the triangle ijl has node i as its source node, node j as its first target node and node l as its second target node. There are three directed links present here: (i, j), (i, l) directed away from i towards the two target nodes and (j, l) directed from j to l. The generalized out-strength and generalized out-degree are equal to 1 for node i and 0 for nodes j and l. In panel (b) the triangle jli has node j as its source node, node l as its first target node and node i as its second target node. The three directed links here are instead (j, i), (j, l) and (l, i), while the generalized out-strength and generalized out-degree are now equal to 1 for node j and 0 for nodes i and l.

6.2.1 Mapping the Barabási-Albert model to the Yule-Simon model

The Barabási-Albert model [45] provides a fundamental mechanism for the emergence of scale-free networks with degree distribution $P(k) \sim k^{-\gamma}$ and diverging second moment $\langle k^2 \rangle$. In the Barabási-Albert model we start at time $t = 1$ from a finite network and at each time $t > 1$ we add one new node and m links connected to the new node and to a node i with degree k_i chosen with probability

$$\Pi_i^{BA} = \frac{k_i}{\sum_j k_j}. \tag{6.7}$$

This probability implements the preferential attachment mechanism according to which nodes which already have high degree k_i are more likely increase their degree further by acquiring new links. The number of nodes $N(t)$ and the number of links $L(t)$ at time t are deterministic variables growing linearly with time as we have $N(t) = t$ and $L(t) = mt$. Therefore the average degree of this network is independent on the network size and given by

$$\langle k \rangle = \frac{2L(t)}{N(t)} = 2m. \tag{6.8}$$

The degree distribution $P(k)$ of the Barabási-Albert model can be evaluated exactly in the large network limit $N(t) \gg 1$ and is given by [48, 58]

$$P(k) = \frac{2\Gamma(m+2)}{\Gamma(m)} \frac{\Gamma(k)}{\Gamma(k+3)} \simeq 2m(m+1)k^{-3}, \tag{6.9}$$

where the latter approximated expression describes the tail of the distribution where $k \gg 1$.

The mathematical origin of the power-law in the Barabási-Albert model can be rooted back to a ball-in-the-box model called the Yule-Simon model [95]. This is a discrete time stochastic process analogous to placing balls in a growing number of boxes with probabilities dependent on the number of balls already in the boxes.

The process starts at time $t = 1$ with a single box with one ball in it. At each subsequent time step $t > 1$ a new ball is introduced and is either placed in an existing box (with probability $\epsilon \in (0, 1)$) or a new box is created (with probability $1 - \epsilon$) and the ball is placed in it. The process may be thought of as producing a random partition of the set of balls introduced up to time t. For any given time t we indicate the total number of boxes by N, and the total number of balls by M. Additionally we indicate by s_i the number of balls in the ith box.

The reinforcement dynamics called preferential attachment in network models is implemented by assuming that the probability of placing a new ball in box i grows linearly with the number of balls s_i already in box i. Therefore in the Yule-Simon model the probability that a new ball is placed in box i is

$$\Pi_i^{YS} = \begin{cases} \epsilon s_i/t & \text{for} \quad 1 < i \leq N, \\ (1-\epsilon) & \text{for} \quad i = N+1. \end{cases} \tag{6.10}$$

Clearly in this model the average number of boxes $\langle N(t) \rangle$ increases linearly with time, i.e. $\langle N(t) \rangle = (1 - \epsilon)t$. Moreover, since a new ball is added at each time-step the number of balls at time t is a deterministic variable given by $M(t) = t$. It follows that the average number of balls per box is constant in time, i.e.

$$\left\langle \frac{M}{N} \right\rangle \simeq \frac{1}{1-\epsilon} = O(1). \tag{6.11}$$

This implies that if the distribution $P(s)$ of balls in the boxes decays as a power-law $P(s) \simeq s^{-\gamma}$ it must necessarily have the power-law exponent γ in the range $\gamma \in (2, \infty)$. In fact, as we have already seen, for $\gamma \in (1, 2]$ power-law distributions have a diverging average value. An exact expression for the distribution produced by the Yule-Simon process can be calculated exactly in the limit $t \to \infty$ finding

$$\begin{aligned} P(s) &= \frac{1}{\epsilon}\Gamma\left(1 + \frac{1}{\epsilon}\right) \frac{\Gamma(s)}{\Gamma(s + 1 + 1/\epsilon)} \\ &\simeq \frac{1}{\epsilon}\Gamma\left(1 + \frac{1}{\epsilon}\right) s^{-\gamma} \end{aligned} \tag{6.12}$$

where the last expression is derived in the limit $s \gg 1$ and where the power-law

exponent γ is given by

$$\gamma = 1 + \frac{1}{\epsilon} \in (2, \infty). \tag{6.13}$$

Therefore we see that using both growth and preferential attachment the Yule-Simon model generates power-law distributions $P(s) \simeq Cs^{-\gamma}$ with $\gamma > 2$ and with a finite average number of balls in the boxes.

The Barabási-Albert model can be mapped to a ball-in-the-box model by assuming that each node corresponds to a box and each half-edge attached to a given node corresponds to a ball in the box. Note that for the Barabási-Albert model the number of nodes is a deterministic variable $N(t) = t$ as is the number of half edges, which is given by twice the number of links $M(t) = 2L(t) \simeq 2mt$. However the Barabási-Albert model can be considered as being in the same *universality class* as the Yule-Simon process with $\epsilon = 1/2$.

Other variations of the Barabási-Albert model have been shown to yield scale-free networks with tunable exponent γ. However, network models in which the number of nodes and the number of links both increase linearly with time can only produce networks with power-law exponent $\gamma \in (2, \infty)$.

In order to produce dense scale-free networks with $\gamma \in (1, 2]$ we need a model in which the density of links increases in time. In the following section we introduce another ball-in-the-box process known as the Pitman-Yor process. Similar to the Yule-Simon process, this process uses growth and preferential attachment to produce power-law distributions of the number of balls in the boxes, however the density of balls in the boxes is not constant but instead increases in time, leading to power-law exponent $\gamma \in (1, 2]$.

6.2.2 The Pitman-Yor process

The Pitman-Yor process [91–93] is a discrete stochastic process similar to Yule-Simon that is known to yield dense power-law distributions.

The model starts at time $t = 1$ with one ball in a single box, and at each time $t > 1$ a new ball is added and either placed in an existing box i or placed in a new box $i = N(t)+1$. Specifically the probability Π_i^{PY} that the new ball goes in the box i is parametrized by the parameter $\alpha \in (0, 1)$ and given by

$$\Pi_i^{PY} = \begin{cases} \frac{s_i(t)-\alpha}{t} & \text{for} \quad 1 < i \leq N, \\ \frac{\alpha N}{t} & \text{for} \quad i = N + 1. \end{cases} \tag{6.14}$$

As in the Yule-Simon process there is a growth in both the number of balls, and the number of boxes, with a preferential attachment mechanism for placing the balls. Unlike with Yule-Simon however, the probability of adding a new box is not constant, but depends on the number of boxes already added while decaying with the total number of balls already added [91, 93].

The marginal distribution of a single box in the large t limit is [91]

$$P(s) = \frac{\alpha}{\Gamma(1-\alpha)} \frac{\Gamma(s-\alpha)}{\Gamma(s+1)} \simeq \frac{\alpha}{\Gamma(1-\alpha)} s^{-\gamma}, \tag{6.15}$$

where the latter expression is an approximation for the tail of the distribution (i.e. for $s \gg 1$). Here the power-law exponent γ is given by

$$\gamma = 1 + \alpha \in (1, 2), \tag{6.16}$$

which implies that the distribution has a diverging first moment. This is consistent with the fact that the number of balls increases superlinearly with the expected number of boxes $\langle N(t) \rangle$ [91] as we have

$$\begin{aligned} \langle N(t) \rangle &\simeq t^{\alpha} \\ M(t) &= t, \end{aligned} \tag{6.17}$$

and therefore

$$\left\langle \frac{M}{N} \right\rangle = O(t^{1-\alpha}). \tag{6.18}$$

In this chapter we explore whether the Pitman-Yor process can be exploited in order to formulate dense power-law network models, and we emphasize the challenges posed by the density of the resulting networks. In this endeavour our objective is to construct not only dense power-law networks formed by pairwise interactions but also dense simplicial complexes which allow one to go beyond the framework of pairwise interactions.

6.3 Evolution of dense scale-free networks and simplicial complexes

In this section we introduce our modelling framework in which the Pitman-Yor process is exploited in order to generate dense weighted power-law networks and 2-dimensional simplicial complexes.

These networks and simplicial complexes evolve by the subsequent addition of nodes and the establishment of new links (or 2-simplices) or reinforcement of already existing links (or 2-simplices). For the network case we consider both a version of the model where the links are directed and a version with undirected links. In the simplicial complex case the simplices are 'directed' in the sense given in Section 6.1.

Unlike in many other models of growing networks, the number of nodes in the network at a given time is not a deterministic function of time but instead depends on the stochastic growth dynamics of the network. The relative probabilities of nodes being created or selected for reinforcement are analogous to a Pitman-Yor process, with an equivalent parameter $\alpha \in (0, 1)$. Below we give the dynamics for each of the three versions of the model.

6.3.1 Undirected network growth dynamics

In the undirected network version of the model, we write the total number of nodes in the network at time t as $N(t)$. Every pair of nodes $i, j \in \{1, 2, ..., N(t)\}$ has an associated weight $w_{ij}(t)$ taking non-negative integer values. A link exists between two nodes i and j if their weight is non-zero, i.e. if $w_{i,j}(t) > 0$.

In this version of the model we start at $t = 1$ with an undirected link between node 1 and node 2. At each time step $t \geq 1$ we select a node i with probability

$$\Pi_i^U = \begin{cases} \frac{s_i(t)-\alpha}{2t} & \text{for} \quad 1 \leq i \leq N \\ \frac{\alpha N}{2t} & \text{for} \quad i = N+1. \end{cases} \tag{6.19}$$

We then update the value of N and select a second node j using the same algorithm. If the two selected nodes are not already linked we add a link between them, if they are already linked we reinforce the weight of the links. In other words the adjacency matrix element a_{ij} and the weight w_{ij} of the link (i, j) are updated according to

$$\begin{aligned} a_{ij}(t+1) &= 1, \\ w_{ij}(t+1) &= w_{ij}(t) + 1. \end{aligned} \tag{6.20}$$

Moreover, since the network is undirected we have

$$\begin{aligned} a_{ji}(t) &= a_{ij}(t), \\ w_{ji}(t) &= w_{ij}(t). \end{aligned} \tag{6.21}$$

Therefore in this model we treat half-edges as the balls of the Pitman-Yor process and we treat the nodes as the boxes of the Pitman-Yor process.

Therefore it is to be expected that the strength distribution will follow a power-law exponent with exponent $\gamma = 1+\alpha$. However, given that the network is weighted, the degree distribution could potentially be significantly different because if a new link is placed between nodes that are already connected the strengths of these nodes will increase but their degrees will remain unchanged.

6.3.2 Directed network growth dynamics

The directed network growth dynamics assumes that links are directed and that only the source node of the links is chosen according to the Pitman-Yor reinforcement dynamics, while the target node is chosen uniformly at random among all the existing nodes of the network. In this way we expect that the density of links in the network will grow more rapidly than in the undirected case. In fact by choosing the target node with uniform probability we are more likely to add new links because we are not biasing the target node to be a node of high degree.

In the directed version, we start at $t = 1$ with a directed link from node 2 to node 1.
At each time step $t \geq 1$ a pair of nodes is selected and its weight is incremented by one. The source node is either an existing node or a new node, while the target node is chosen uniformly at random from the remaining existing nodes. The probability at time t of selecting node i as the source node is given by:

$$\Pi_i^D = \begin{cases} \frac{s_i^{out}(t)-\alpha}{t} & \text{for} \quad 1 < i \leq N \\ \frac{\alpha(N-1)}{t} & \text{for} \quad i = N+1. \end{cases} \tag{6.22}$$

The probability at time t of selecting node j as the target node is given by:

$$\hat{\Pi}_j^N = \frac{1}{N(t)}. \tag{6.23}$$

When both source node i and target node j have been selected we update the adjacency matrix element and weight of the link, i.e.

$$\begin{aligned} a_{ij}(t+1) &= 1, \\ w_{ij}(t+1) &= w_{ij}(t) + 1. \end{aligned} \tag{6.24}$$

Note that here we have chosen to select the source node of the link according to the Pitman-Yor dynamics while the target node is chosen with uniform probability. However it is also possible to consider a directed network model in which the target node is chosen according to the Pitman-Yor dynamics and the source node is chosen

uniformly at random. Since the two versions of the model are simply related by the inversion of the direction of the links here we omit the explicit treatment of the latter possible definition.

6.3.3 Directed simplicial complex growth dynamics

Here we consider a directed 2-dimensional simplicial complex formed only by 'directed triangles'.

The triangles are directed in the sense that each permutation of three nodes is associated with a different triangle. We say that the first node in the triangle is the "source node", the second node is the "first target node" and the third node is the "second target node". For the triangles in this version of the model the links are also directed, and we have chosen the convention that the two links coming from the source node are directed away from the source node towards the target nodes and the third link is directed from the first target node to the second.

In this version of the model the triangles ijl also have an associated weight $w_{ijl}(t)$ taking non-negative integer values. These weights are associated specifically to the directed triangles and thus are also directed in the same sense. We define the *generalized out-strength* $\tilde{s}_i^{out}$ of a node i to be the total weight of triangles for which i is the source node, i.e. $\tilde{s}_i^{out} = \sum_{j,l=1}^{N} w_{ijl}$.

In this version of the model we start at $t = 1$ with three nodes labelled $1, 2, 3$ and the single directed triangle 123. At each time step $t \geq 1$ we select a triangle to be created or reinforced. The source node i of this triangle is selected with probability

$$\tilde{\Pi}_i^{SC} = \begin{cases} \frac{\tilde{s}_i^{out}(t)-\alpha}{t} & \text{for} \quad 1 \leq i \leq N, \\ \frac{\alpha(N-2)}{t} & \text{for} \quad i = N+1. \end{cases} \tag{6.25}$$

Once this source node i has been selected, a link (j, l) is selected uniformly at random

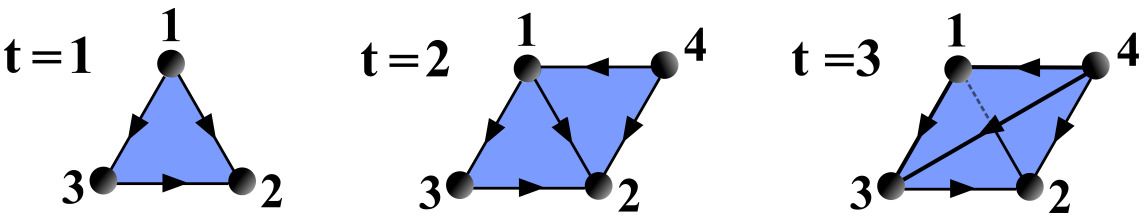

Figure 6.2: Diagram showing one possible way the simplicial complex could grow in the first three time-steps. At $t = 1$ all simplicial complexes in the model start as the single triangle 123. In this example, at $t = 2$ a new node (4) is created and randomly selects the link $(1, 2)$ to form the triangle 412. At $t = 3$ no new node is created. Instead, node 4 is selected for reinforcement and randomly selects the link $(1, 3)$ to form the triangle 413.

from the set of existing links with probability

$$\hat{\Pi}^{L}_{jl} = \frac{1}{L(t)}, \tag{6.26}$$

where $L(t)$ is the number of links in the simplicial complex at time t. If this triangle already exists then its weight is reinforced according to

$$w_{ijl}(t+1) = w_{ijl}(t) + 1, \tag{6.27}$$

while if it doesn't exist yet then it is created with initial weight one:

$$\begin{aligned} a_{ijl}(t+1) &= 1, \\ w_{ijl}(t+1) &= 1. \end{aligned} \tag{6.28}$$

Figure 6.2 is a diagram illustrating one possible way the simplicial complex could grow in its first three time-steps. In this example we start with a single triangle at $t = 1$. At $t = 2$ a new node labelled 4 is added and forms a triangle with the link $(1, 2)$. At time $t = 3$ no new nodes are added, but instead node 4 gains an additional triangle formed with the link $(1, 3)$.

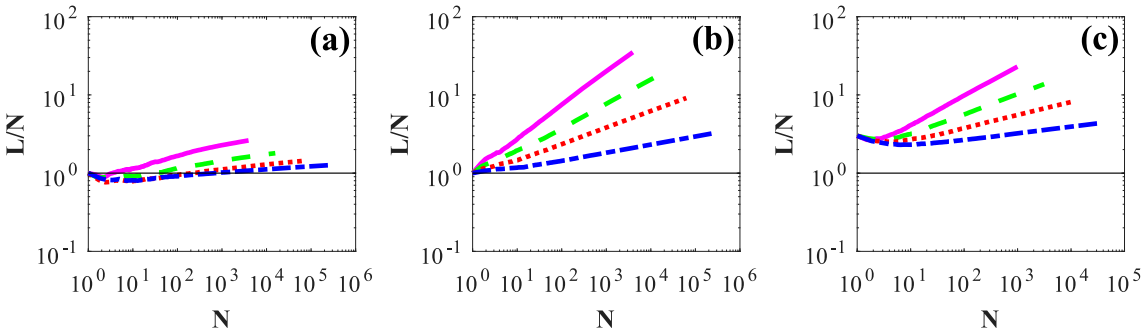

Figure 6.3: Total number of links over the total number of nodes, as a function of the total number of nodes, taken from simulation data. Panels (a), (b) and (c) show the results for the undirected network, directed network and simplicial complex respectively. For each model version and each choice of parameter α, 50 realizations of the network were generated and averaged over. The results were obtained for $t = 10^6$ and for $\alpha = 0.6$, (purple solid line), $\alpha = 0.7$ (green dashed line), $\alpha = 0.8$ (red dotted line), and $\alpha = 0.9$ (blue dot-dashed line).

6.3.4 Number of links as a function of the number of nodes

In all three versions of the model, the distributions of the strengths (or out-strengths) of the nodes are generated by a Pitman-Yor process and so therefore have 'dense' power-law exponents $\delta \in (1, 2)$. However, as mentioned before, this does not guarantee that the networks themselves are dense as the links can be weighted multiple times.

Therefore we have run extensive simulations of the three versions of the model to investigate whether the total number of links grows super-linearly with the number of nodes. Figure 6.3 shows how the number of links grows with the total number of nodes for a range of values of parameter α and for the three different versions of the model. We see that for all three versions the total number of links grows faster than the total number of nodes, indicating that the model produces dense networks and simplicial complexes. Moreover, as expected, the directed version of the model allows for the exploration of cases in which the ratio between the number of links and the number of nodes increases more rapidly than for the undirected version of the model.

6.4 Strengths of the nodes

The mean-field approximation is known to give very reliable results in the context of sparse growing network models. Therefore it is natural to approach the study of dense growing networks and simplicial complexes with the same techniques. Specifically here our goal is to derive the distribution of the strength s in the undirected network model, the distribution of the out-strength s^{out} in the directed network model and the distribution of the generalized out-strength $\tilde{s}^{out}$ for the simplicial complex model using the mean-field approximation.

6.4.1 Evolution of the number of nodes and of the strengths

The mean-field differential equations for the three cases differ only trivially, and therefore we will treat them using a unified set of equations that apply to all three cases. To this end we use the symbol $\hat{s}_i$ that indicates $s_i, s_i^{out}, \tilde{s}_i^{out}$ for the undirected network, directed network and directed simplicial complex versions of the model respectively.

Similarly the Pitman-Yor probabilities Π_i^U, Π_i^D and Π_i^{SC} can be unified in a single expression

$$\Pi_i = \begin{cases} \frac{\hat{s}_i(t)-\alpha}{(2-a)t} & \text{for} \quad 1 \leq i \leq N \\ \frac{\alpha(N-a-b)}{(2-a)t} & \text{for} \quad i = N+1. \end{cases} \tag{6.29}$$

where we have introduced the parameters a and b taking values: $a = b = 0$ for the undirected network case; $a = 1, b = 0$ for the directed network case and $a = 1, b = 1$ for the directed simplicial complex case. Eq. (6.29) thus subsumes Eq.s (6.19), (6.22) and (6.25) for the probabilities of a node i being selected (or created) at time $t + 1$ in the undirected, directed and simplicial complex versions of the model respectively.

As usual in the mean-field approximation, we will treat our stochastic variables $N(t)$, $\hat{s}_i(t)$ as deterministic continuous variables equal to their expected value over different realizations of the network or simplicial complex evolution. Since at each

time the number of nodes to be chosen according to the Pitman-Yor probability is $(2-a)$ the mean-field equation determining the growth of the number of nodes in the network is given by

$$\frac{dN}{dt} = (2-a)\Pi_{N+1} = \frac{\alpha(N-a-b)}{t}, \tag{6.30}$$

with the initial condition

$$N(t=1) = 2+b. \tag{6.31}$$

The solution is then

$$N(t) = (2-a)t^{\alpha} + a + b. \tag{6.32}$$

The differential equations for $\hat{s}_i$ of a node i born at time $t_i > 1$ is given by

$$\frac{d\hat{s}_i}{dt} = (2-a)\Pi_i = \frac{\hat{s}_i - \alpha}{t}, \tag{6.33}$$

with initial condition

$$\hat{s}_i(t_i) = 1. \tag{6.34}$$

Therefore $\hat{s}_i$ increases linearly with time, and is given by

$$\hat{s}_i = (1-\alpha)\left(\frac{t}{t_i}\right) + \alpha. \tag{6.35}$$

6.4.2 Strength distribution

The strength distribution can easily be derived within the mean-field approximation by using the mean-field expressions for the number of nodes $N(t)$ (Eq. 6.32) and the strength $\hat{s}(t)$ (Eq. (6.35)) as a function of time t.

To this end by using Eq. (6.35) we first note that the cumulative strength distribution $P(\hat{s}_i > \hat{s})$ indicating the probability that a random node i has a strength

$\hat{s}_i(t) > \hat{s}$ can be written as

$$P(\hat{s}_i \geq \hat{s}) = P\left(t_i \leq t^\star(\hat{s})\right), \tag{6.36}$$

where $P\left(t_i \leq t^\star(\hat{s})\right)$ is the probability that a random node i arrives in the network at time $t_i \leq t^\star(\hat{s})$ and where $t^\star(\hat{s})$ satisfies

$$\hat{s} = (1-\alpha)\left(\frac{t}{t^\star}\right) + \alpha. \tag{6.37}$$

Moreover we observe that the probability $P\left(t_i \leq t^\star(\hat{s})\right)$ is simply given by the fraction of nodes that arrived in the network before time $t^\star(\hat{s})$, i.e.

$$P\left(t_i \leq t^\star(\hat{s})\right) = \frac{N(t^\star(\hat{s}))}{N(t)} = \frac{(2-a)\left[t^\star(\hat{s})\right]^\alpha + a + b}{(2-a)t^\alpha + a + b}. \tag{6.38}$$

For $\hat{s} \gg 1$ and $t^\star(\hat{s}) \gg 1$ we can write Eq. (6.38) as

$$P\left(t_i \leq t^\star(\hat{s})\right) = \left(\frac{t^\star(\hat{s})}{t}\right)^\alpha = \left(\frac{1-\alpha}{s}\right)^\alpha. \tag{6.39}$$

The strength distribution $\tilde{P}(\hat{s})$ is thus given by

$$\begin{aligned} \tilde{P}(\hat{s}) &= \frac{d}{d\hat{s}}\left[1 - P(\hat{s}_i \geq \hat{s})\right] \\ &\simeq \frac{\alpha}{1-\alpha}\left(\frac{1-\alpha}{\hat{s}}\right)^{\alpha+1}. \end{aligned} \tag{6.40}$$

Therefore the strength distribution is power-law distributed with exponent $1+\alpha \in (1,2]$. Figure 6.4 shows the strength distributions arising from simulations of the three models. We see that in all three versions of the model, and for all values of α used, that the strength distributions follow a power-law. In the insets of each panel we see that the exponents fitted to the distributions are very close to $1+\alpha$ as predicted by Eq. (6.40).

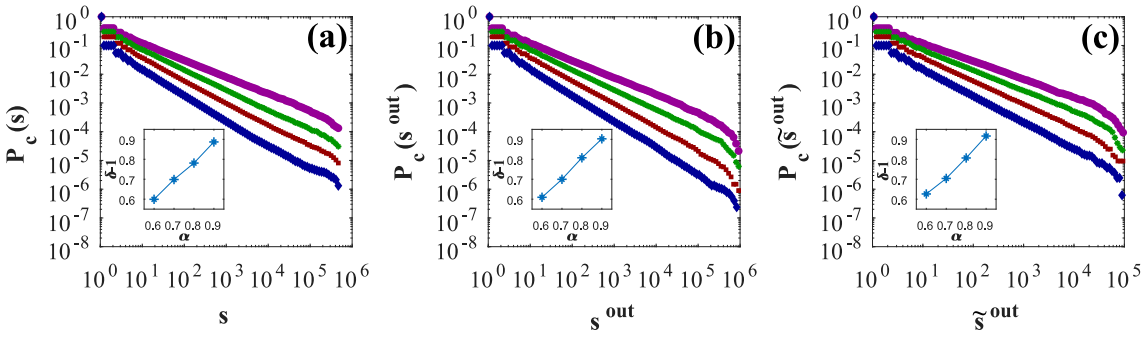

Figure 6.4: Cumulative strength, out-strength and generalized out-strength distributions for the three versions of the model. For each version of the model and for each choice of parameter α, 50 realizations of the network were generated and averaged over. The simulated results were obtained for $t = 10^6$ and are represented by purple circles ($\alpha = 0.6$), green stars ($\alpha = 0.7$), red squares ($\alpha = 0.8$) and blue diamonds ($\alpha = 0.9$). The insets show the fitted exponents of the distributions for the four values of α.

6.5 Reinforcement probabilities

By reinforcement probability we mean the probability that at time t we either add a new link (or a new triangle) or in the case where the link (or triangle) already exists that we increase its weight. In this section we show how this probability can be directly calculated within the mean-field approximation using Eqs. (6.32) and (6.35).

In the undirected case we indicate with $\pi_{ij}(t, t_i, t_j)$ the probability that at time t we reinforce or add a link between node i and node j given that node i and node j have been added to the network at time t_i and t_j respectively. The reinforcement probability is therefore given by

$$\pi_{ij}^{U}(t, t_i, t_j) = 2\Pi_i^U \Pi_j^U \tag{6.41}$$

where Π_i^U, Π_j^U are calculated at time t. Therefore we have

$$\pi_{ij}^{U}(t, t_i, t_j) = \frac{(s_i(t) - \alpha)\,(s_j(t) - \alpha)}{2t^2}, \tag{6.42}$$

which by inserting the mean-field expression for $s_i(t)$ gives

$$\pi_{ij}^{U}(t, t_i, t_j) = \frac{(1-\alpha)^2}{2t_i t_j}. \tag{6.43}$$

Similarly it can be shown that in the directed case the probability $\pi_{ij}^{D}(t, t_i, t_j)$ that a link from node i to node j is reinforced at time t given that nodes i iand j are arrived in the network at time t_i and t_j respectively can be expressed as

$$\pi_{ij}^{D}(t, t_i, t_j) = \frac{1}{N(t)} \frac{s_i^{out}(t) - \alpha}{t}, \tag{6.44}$$

which using the mean-field solution of s_i^{out} and $N(t)$ gives

$$\pi_{ij}^{D}(t, t_i, t_j) = \frac{1-\alpha}{(t^{\alpha}-1)t_i}. \tag{6.45}$$

For the simplicial complex we indicate with $\tilde{\pi}_{i,\ell}(t, t_i, \tau_\ell)$ the probability that a triangle with source node i and target link ℓ is reinforced, conditioned on their respective birth times t_i and τ_ℓ. We write this as

$$\tilde{\pi}_{i,\ell}(t, t_i, \tau_\ell) \;=\; \frac{\tilde{s}_i^{out} - \alpha}{t} \frac{1}{L(t)}, \tag{6.46}$$

where $L(t)$ is the total number of (directed) links at time t. Using Eq. (6.35) we obtain

$$\tilde{\pi}_{i,\ell}(t, t_i, \tau_\ell) \;=\; \frac{1-\alpha}{t_i} \frac{1}{L(t)}. \tag{6.47}$$

Ideally, we would like to have obtained a mean-field prediction for $L(t)$, but this proved too difficult. However, the above equation still proves useful when investigating the distribution of the generalised out-degrees later in Section 6.6.

6.6 Degree distribution

In this section we use the mean-field results of Sections 6.4 and 6.5 to derive equations for the degrees of the nodes conditioned on their birth times. We evaluate these equations numerically for a range of values of the parameter α and find power-law scalings with $\gamma = 2$ for k_i in the undirected case, and $\gamma < 2$ for k_i^{out} and $\tilde{k}_i^{out}$ in the directed and simplicial complex cases. We also compare our numerically obtained predictions with simulation results, validating our mean-field approach.

6.6.1 Undirected network case

We write the degree of a node i born at time t_i in terms of the link probabilities $p_{ij}(t, t_i, t_j)$:

$$k_i(t, t_i) = \int_1^t dt_j \dot{N}(t_j) p_{ij}(t, t_i, t_j), \tag{6.48}$$

where $\dot{N}(t_j)$ indicates the derivative of $N(t)$ with respect to t, evaluated at t_j, and also corresponds to the probability that a new node is born at time t_j. The link probability $p_{ij}(t, t_i, t_j)$ is the probability that a link exists between nodes i and j conditioned on their birth times t_i and t_j (and also conditional on the event that a new node j is born at t_j). We can write this probability as

$$p_{ij}(t, t_i, t_j) = 1 - \prod_{t'=\tau}^{t} \left[1 - \pi_{ij}(t', t_i, t_j)\right], \tag{6.49}$$

where $\pi_{ij}(t', t_i, t_j)$ is the reinforcement probability given in (6.43) and $\tau = \max\{t_i, t_j\}$ is the first time that both i and j are present in the network, i.e. (6.49) is 1 minus the probability that the pair (i, j) is not reinforced in the time interval $[\tau, t]$. The π_{ij} reinforcement probabilities are very small for almost all pairs of nodes, so we make the approximation

$$\prod_{t'=\tau}^{t} \left[1 - \pi_{ij}(t', t_i, t_j)\right] \simeq \exp\left(-\sum_{t'=\tau}^{t} \pi_{ij}(t', t_i, t_j)\right). \tag{6.50}$$

Taking t to be very large, we approximate the sum in the above equation with an integral, and write (6.49) as

$$p_{ij}(t,t_i,t_j) \simeq 1 - \exp\left(-\int_\tau^t dt'\ \pi_{ij}(t',t_i,t_j)\right). \tag{6.51}$$

It is then straight-forward to obtain the following expression for the degree from Eq.s (6.51) and (6.43):

$$\begin{aligned} k_i(t,t_i) &\simeq 2\alpha \int_1^{t_i} dt_j t_j^{\alpha-1}\left[1 - e^{-\frac{(1-\alpha)^2}{2t_i t_j}(t-t_i)}\right] \\ &\quad +2\alpha \int_{t_i}^{t} dt_j t_j^{\alpha-1}\left[1 - e^{-\frac{(1-\alpha)^2}{2t_i t_j}(t-t_j)}\right]. \end{aligned} \tag{6.52}$$

The first integral in the above is the contribution to the generalised degree from nodes born before t_i, i.e. the contribution when $\tau = \max\{t_i, t_j\} = t_i$. The second integral is instead the contribution when $\tau = \max\{t_i, t_j\} = t_j$. We perform changes of variables on each of the integrals in Eq. (6.52), obtaining

$$\begin{aligned} k_i(t,t_i) &\simeq 2\alpha A^\alpha \left(\frac{t}{t_i} - 1\right)^\alpha \int_{A(\frac{t}{t_i}-1)t_i^{-1}}^{A(\frac{t}{t_i}-1)} dx x^{-(1+\alpha)}\left[1 - e^{-x}\right] \\ &\quad +2\alpha A^\alpha \left(\frac{t}{t_i}\right)^\alpha \int_0^{A(\frac{t}{t_i}-1)t_i^{-1}} dx (At_i^{-1} + x)^{-(1+\alpha)}\left[1 - e^{-x}\right], \end{aligned} \tag{6.53}$$

where $A = \frac{(1-\alpha)^2}{2}$ and in the first integral we have used the change in variable $x = A(t - t_i)t_i^{-1}t_j^{-1}$ while in the second integral we have used $x = At_j^{-1}(\frac{t}{t_j} - 1)$. We wish to extract a scaling between the degree $k_i(t,t_i)$ and the birth time t_i so that we can eventually calculate the degree distribution similarly to how we calculated the strength distribution in Section 6.4. In order to do this we take the limit $t, t_i \to \infty$ while at the same time fixing the ratio $t/t_i = y$. In the second integral $(At_i^{-1} + x)^{-(1+\alpha)}$ becomes simply $x^{-(1+\alpha)}$ and we may write Eq. (6.53) as

$$\begin{aligned} k_i(t,t_i) &\simeq 2\alpha A^\alpha (y-1)^\alpha \int_{A(y-1)yt^{-1}}^{A(y-1)} dx x^{-(1+\alpha)}\left[1 - e^{-x}\right] \\ &\quad +2\alpha A^\alpha y^\alpha \int_0^{A(y-1)yt^{-1}} dx x^{-(1+\alpha)}\left[1 - e^{-x}\right]. \end{aligned} \tag{6.54}$$

Now, by considering $y = \frac{t}{t_i} \gg 1$ we can make the approximation $y - 1 \simeq y$ and write Eq. (6.54) as a single integral:

$$k_i(t, t_i) \simeq 2\alpha A^\alpha y^\alpha \int_0^{Ay} dx x^{-(1+\alpha)} \left[1 - e^{-x}\right]. \tag{6.55}$$

Evaluating the integral then gives

$$k_i(t, t_i) \simeq 2\alpha A^\alpha y^\alpha \left[-\alpha^{-1} A^{-\alpha} y^{-\alpha} - \Gamma(-\alpha) - \Gamma(-\alpha, Ay)\right], \tag{6.56}$$

where $\Gamma(\cdot)$ indicates the gamma function and $\Gamma(\cdot, \cdot)$ indicates the upper incomplete gamma function defined by

$$\Gamma(s, z) = \int_z^\infty dt t^{s-1} e^{-t}. \tag{6.57}$$

For $-1 < s < 0$ and $z \to \infty$ the incomplete gamma function converges to 0. Considering this fact, and also considering our earlier assumption that $y \gg 1$ we find that we can approximate Eq. (6.56) further as

$$k_i(t, t_i) \simeq -2\alpha A^\alpha y^\alpha \Gamma(-\alpha), \tag{6.58}$$

and so obtain the following scaling relation between $k_i(t, t_i)$ and t_i

$$k_i(t, t_i) \propto \left(\frac{t}{t_i}\right)^\alpha. \tag{6.59}$$

We are now finally in a position where we can calculate the cumulative degree distribution $P(k_i > k)$ within the mean-field approximation using

$$P(k_i > k) = \frac{N\left(t^\star(k)\right)}{2t^\alpha} = \left(\frac{t^\star(k)}{t}\right)^\alpha, \tag{6.60}$$

where $t^\star(k)$ is the birth time such that

$$k_i\big(t, t^\star(k)\big) = k. \tag{6.61}$$

From Eq.s (6.59) and (6.61) we obtain the scaling

$$P(k_i > k) \propto k^{-1}, \tag{6.62}$$

indicating that for large t, the degree distribution $P(k)$ has a power-law tail with exponent $\gamma = 2$. We confirm these results by comparing them to simulations of the process for a range of values of α. Figure 6.5 shows the average cumulative degree distributions given by the simulations. Also included are theoretical predictions obtained by evaluating Eq. (6.52) numerically. The inset plot shows the values of γ obtained from fitting power-laws to the tails of the simulation data. We see that for all choices of the parameter α the degrees of the nodes follow power-laws with values of γ close to the theoretical prediction of 2.

Additionally we have studied the scaling of the maximum degree (cutoff) k_{max} as a function of the network size N. This study reveals that the cutoff scales with the network size N with a proportionality constant depending on the value of α (see Figure 6.6).

6.6.2 Directed network case

We can derive the distribution of the out-degrees in the directed network case using a similar approach to in the previous section. The out-degree $k_i^{out}(t, t_i)$ at time t of node i born at time t_i is

$$k_i^{out}(t, t_i) \simeq \int_1^t dt_j \dot{N}(t_j) p_{ij}(t, t_i, t_j), \tag{6.63}$$

where $p_{ij}(t, t_i, t_j)$ is the probability that at time t there is a link between i and j conditioned on their birth times t_i and t_j. As in the undirected case we write this probability in terms of the reinforcement probabilities:

$$p_{ij}(t, t_i, t_j) \;\; = \;\; 1 - \prod_{t'=\tau}^{t} \left[1 - \pi_{ij}(t', t_i, t_j)\right], \tag{6.64}$$

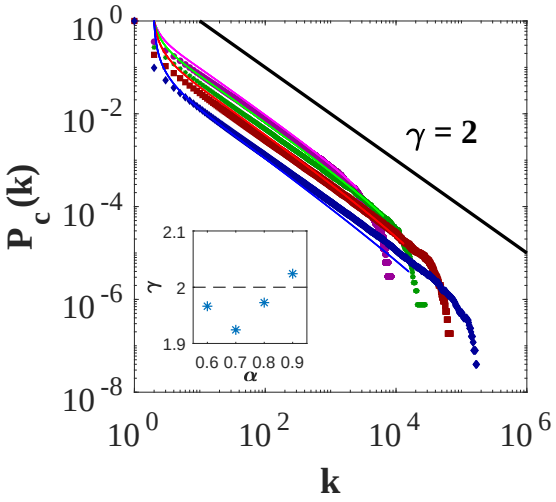

Figure 6.5: Cumulative degree distributions in the undirected case. For each choice of parameter α, 50 realizations of the network were generated and averaged over. The simulated results were obtained for $t = 10^6$ and are represented by purple circles ($\alpha = 0.6$), green stars ($\alpha = 0.7$), red squares ($\alpha = 0.8$) and blue diamonds ($\alpha = 0.9$). The numerical results are represented by the purple solid line ($\alpha = 0.6$), green dashed line ($\alpha = 0.7$), red dotted line ($\alpha = 0.8$) and blue dot-dashed line ($\alpha = 0.9$). The solid black line shows an exact power-law with exponent $\gamma = 2$ for comparison. The inset shows the fitted exponents of the simulated distributions for the four values of α.

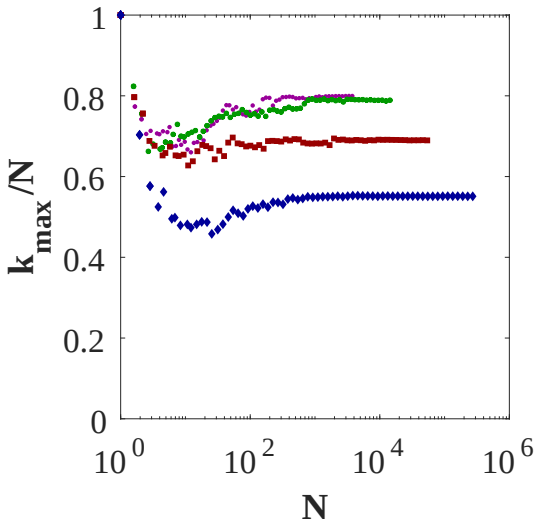

Figure 6.6: Evolution of the average normalized degree cut-off for the undirected case as a function of average total number of nodes in the network. For each choice of parameter α, 50 realizations of the network were generated and averaged over. The simulated results were obtained for $t = 10^6$ and are represented by purple circles ($\alpha = 0.6$), green stars ($\alpha = 0.7$), red squares ($\alpha = 0.8$) and blue diamonds ($\alpha = 0.9$).

where τ is again the maximum of t_i and t_j, and where $\pi_{ij}(t', t_i, t_j)$ is now given by Eq. (6.45). The reinforcement probabilites are small, allowing us to again make the approximation $1 - \pi_{ij}(t', t_i, t_j) \simeq e^{-\pi_{ij}(t', t_i, t_j)}$, leading to

$$p_{ij}(t, t_i, t_j) \simeq 1 - \exp\left(- \int_{\tau}^{t} dt' \pi_{ij}(t', t_i, t_j) \right), \tag{6.65}$$

where in the above equation we have again used a continuous time approximation allowing us to replace the sum over t' with an integral. For $t, \tau \gg 1$ the reinforcement probability becomes $\pi_{ij}(t', t_i, t_j) = \frac{1-\alpha}{(t'^{\alpha}-1)t_i} \simeq \frac{1-\alpha}{t'^{\alpha} t_i}$, and so we evaluate the integral in Eq. (6.66) to get

$$p_{ij}(t, t_i, t_j) \simeq 1 - \exp\left(-t^{-\alpha} y \left[1 - \left(\frac{\tau}{t} \right)^{1-\alpha} \right] \right), \tag{6.66}$$

where $y = \frac{t}{t_i}$. Unlike in the undirected network case, the link probability is constant with respect to t_j when $t_j \leq t_i$ (i.e. when $\tau = t_i$). Specifically, in this range we have

$$p_{ij}(t, t_i, t_j) = p_{ij}(t, t_i, t_j = t_i) = 1 - \exp\left(-t^{-\alpha} y \left[1 - y^{\alpha - 1} \right] \right). \tag{6.67}$$

The mean-field out-degree given in Eq. (6.63) can now be written as

$$\begin{aligned} k_i^{out}(t, t_i) \simeq\ & t_i^{\alpha} \left[1 - e^{-t^{-\alpha} y \left[1 - y^{\alpha - 1} \right]} \right] \\ & + \alpha \int_{t_i}^{t} dt_j t_j^{\alpha - 1} \left[1 - e^{-t^{-\alpha} y \left[1 - \left(\frac{t_j}{t} \right)^{1-\alpha} \right]} \right]. \end{aligned} \tag{6.68}$$

We perform integration by parts on the remaining integral in Eq. (6.68), using $u = 1 - e^{-t^{-\alpha} y \left[1 - \left(\frac{t_j}{t} \right)^{1-\alpha} \right]}$ and $\frac{dv}{dt_j} = \alpha t_j^{\alpha - 1}$ with the standard formula for integration by parts:

$$\int_a^b dt \left[u \frac{dv}{dt} \right] = uv|_{t=a}^{t=b} - \int_a^b dt \left[\frac{du}{dt} v \right], \tag{6.69}$$

which leads to

$$\begin{aligned} k_i^{out}(t,t_i) &= (1-\alpha)t^{-1}y\int_{t_i}^{t} dt_j e^{-t^{-\alpha}y\left[1-\left(\frac{t_j}{t}\right)^{1-\alpha}\right]} \\ &= (1-\alpha)y\int_{y^{-1}}^{1} dx e^{-t^{-\alpha}y\left[1-x^{1-\alpha}\right]}, \end{aligned} \tag{6.70}$$

where in the second line we have used the change of variable $x = \frac{t_j}{t}$. We now let $t, t_i \to \infty$ with $y = \frac{t}{t_i}$ fixed. In this regime, the exponent of the exponential function is very small, $-t^{-\alpha}y\left[1-x^{1-\alpha}\right] \ll 1$, and so we can make the approximation $e^{-t^{-\alpha}y\left[1-x^{1-\alpha}\right]} \simeq 1 - t^{-\alpha}y\left[1-x^{1-\alpha}\right]$. This leads to

$$\begin{aligned} k_i^{out}(t,t_i) &= (1-\alpha)y\int_{y^{-1}}^{1} dx \left(1 - t^{-\alpha}y\left[1-x^{1-\alpha}\right]\right) \\ &= (1-\alpha)y\left[1 - y^{-1} + t^{-\alpha}\left(-y + 1 + \frac{1}{2-\alpha}y - \frac{1}{2-\alpha}y^{\alpha-1}\right)\right] \end{aligned} \tag{6.71}$$

which for large t tends to

$$k_i^{out}(t,t_i) = (1-\alpha)\left[y-1\right] = (1-\alpha)\left[\frac{t}{t_i} - 1\right]. \tag{6.72}$$

Therefore the cumulative degree distribution may be found from

$$P(k_i(t) > k) = P(t_i < t^\star(k))) \tag{6.73}$$

with

$$t^\star(k) = t\left(\frac{1-\alpha}{k+1}\right) \tag{6.74}$$

and

$$P(t_i < t^\star(k))) = \frac{N(t^\star(k))}{N(t)}. \tag{6.75}$$

Using the mean-field expression for the number of nodes given by Eq. (6.32) we find the cumulative out-degree distribution to be

$$P(k_i > k) = \left(\frac{1-\alpha}{k+1}\right)^{\alpha}, \tag{6.76}$$

which implies the out-degree distribution is

$$P(k) = \alpha \frac{(1-\alpha)^{\alpha}}{(k+1)^{\alpha+1}} \propto k^{-\alpha-1}. \tag{6.77}$$

Therefore, we see that within our mean-field approximation the distribution of out-degrees has a power-law tail with exponent $\gamma = 1 + \alpha$. Figure 6.7 shows theoretical predictions for the full cumulative out-degree distribution, obtained from numerical evaluation of Eq. (6.63) for a selection of values of α. Also included in the figure are the results of simulations for the same values of α. The inset plot shows the values of γ obtained from fitting power-laws to the tails of the simulation data. We see that the out-degrees of the nodes follow power-laws with increasing values of γ for larger α, and with all values of γ between $\gamma = 1$ and $\gamma = 2$. From the inset plot it is clear that the exponents of tails of the distributions closely agree with the theoretical prediction of $\gamma = 1 + \alpha$.

6.6.3 Directed simplicial complex case

In the case of the simplicial complex, the generalized out-degree $\tilde{k}_i^{out}(t, t_i)$ of a node i is the number of triangles for which i is the source node. In the mean-field approximation we may write this as

$$\tilde{k}_i^{out}(t, t_i) \simeq \int_1^t d\tau_\ell \dot{L}(\tau_\ell) \hat{p}(t, t_i, \tau_\ell) \tag{6.78}$$

where

$$\hat{p}(t, t_i, \tau_\ell) = 1 - e^{-(1-\alpha)/t_i \int_{\max(t_i,\tau_\ell)}^{t} dt' [L(t')]^{-1}} \tag{6.79}$$

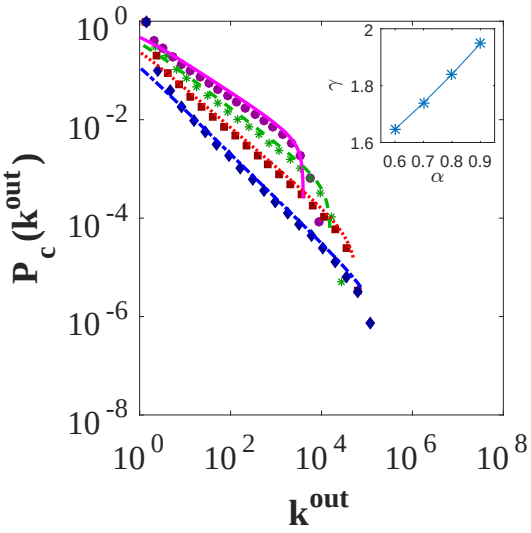

Figure 6.7: Cumulative distributions of the out-degrees. For each choice of parameter α, 50 realizations of the network were generated and averaged over. The simulated results were obtained for $t = 10^6$ and are represented by purple circles ($\alpha = 0.6$), green stars ($\alpha = 0.7$), red squares ($\alpha = 0.8$) and blue diamonds ($\alpha = 0.9$). The numerical results are represented by the purple solid line ($\alpha = 0.6$), green dashed line ($\alpha = 0.7$), red dotted line ($\alpha = 0.8$) and blue dot-dashed line ($\alpha = 0.9$). The inset shows the fitted exponents of the simulated distributions for the four values of α.

is the probability of a triangle with source node i and target link l, and was derived using the reinforcement probability given in Eq. (6.47) in the same way as in the network versions of the model. Figure 6.3 provides strong evidence that for large t, $L(t)$ grows like a power of t. We therefore assume

$$L(t) = ct^b. \tag{6.80}$$

The values c and b could be obtained for each choice of α by fitting Eq. (6.80) to the data shown in figure 6.3(c), however we will see that in fact this isn't necessary for obtaining the scaling of the generalised degree. Substituting Eq. (6.80) in to (6.79) we obtain the following for the generalized out-degree of a node i born at time t_i,

$$\tilde{k}_i^{out}(t,t_i) = cb\int_{t_i}^{t} d\tau_\ell \tau_\ell^{b-1}\left[1 - e^{-\frac{1-\alpha}{c(1-b)t_i}\left(t^{1-b} - \max(t_i,\tau_\ell)^{1-b}\right)}\right]. \tag{6.81}$$

The above integral has a similar form to the integral in Eq. (6.68), and we take a very similar approach. In particular we note that for $\tau_l \leq t_i$ the probability of a link given in Eq. (6.79) is constant with respect to τ_l. We write Eq. (6.81) as

$$\begin{aligned}\tilde{k}_i^{out}(t,t_i) &= ct_i^b\left[1 - e^{-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-y^{b-1}\right]}\right] \\ &\quad + cb\int_{t_i}^{t} d\tau_l \tau_l^{b-1}\left[1 - e^{-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-\left(\frac{\tau_l}{t}\right)^{1-b}\right]}\right],\end{aligned} \tag{6.82}$$

where $y = \frac{t}{t_i}$. Applying integration by parts to the integral in the second line we find

$$\begin{aligned}\tilde{k}_i^{out}(t,t_i) &= (1-\alpha)yt^{-1}\int_{t_i}^{t} d\tau_l e^{-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-\left(\frac{\tau_l}{t}\right)^{1-b}\right]} \\ &= (1-\alpha)y\int_{y^{-1}}^{1} dx e^{-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-x^{1-b}\right]},\end{aligned} \tag{6.83}$$

where in the second line we have applied the change of variable $x = \frac{\tau_l}{t}$. We now let $t, t_i \to \infty$ while keeping $y = \frac{t}{t_i}$ fixed, giving us $-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-x^{1-b}\right] \ll 1$ for all x within the the range being integrated over. This allows us to make the approximation $e^{-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-x^{1-b}\right]} \simeq 1 - t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-x^{1-b}\right]$ and so we can write

Eq. (6.83) as

$$\begin{aligned}
\tilde{k}_i^{out}(t,t_i) &= (1-\alpha)y\int_{y^{-1}}^{1} dx\left(1-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\left[1-x^{1-b}\right]\right) \\
&= (1-\alpha)y\left[\left(1-t^{-b}\frac{(1-\alpha)}{c(1-b)}y\right)x+t^{-b}\frac{(1-\alpha)}{c(1-b)(2-b)}yx^{2-b}\right]\Bigg|_{x=y^{-1}}^{x=1} \\
&\to (1-\alpha)\left[y-1\right]=(1-\alpha)\left[\frac{t}{t_i}-1\right], \qquad (6.84)
\end{aligned}$$

where in the last line we use the fact that $t^{-b} \to 0$ as $t \to \infty$ since we must have $b > 0$ the total number of links cannot decrease in time. Remarkably, the mean-field expression for the generalised out-degree that we give in Eq. (6.84) has no direct dependence on c or b, so assuming that $L(t)$ is indeed a power-law, there is no need to find fitted values for c and b in order to predict the $\tilde{k}_i^{out}(t,t_i)$.

The approach to calculating the distribution of the generalised out-degrees mirrors the approach to calculating the distribution of the out-degrees in the directed network version of the model. We find the distribution to be given by

$$P(k) = \alpha\frac{(1-\alpha)^{\alpha}}{(k+1)^{\alpha+1}} \propto k^{-\alpha-1}. \qquad (6.85)$$

Therefore, we see that within our mean-field approximation the distribution of generalized out-degrees has a power-law tail with exponent $\gamma = 1+\alpha$.

Figure 6.8 shows theoretical predictions for the full cumulative generalized out-degree distribution, obtained from numerical evaluation of Eq. (6.81) for a selection of values of α. Also included in the figure are the results of simulations for the same values of α. The inset plot shows the values of γ obtained from fitting power-laws to the tails of the simulation data.

We see that the generalized out-degrees of the nodes follow power-laws with increasing values of γ for larger α, and with all values of γ between $\gamma = 1$ and $\gamma = 2$. From the inset plot we see that the exponents of the tails of the distributions are

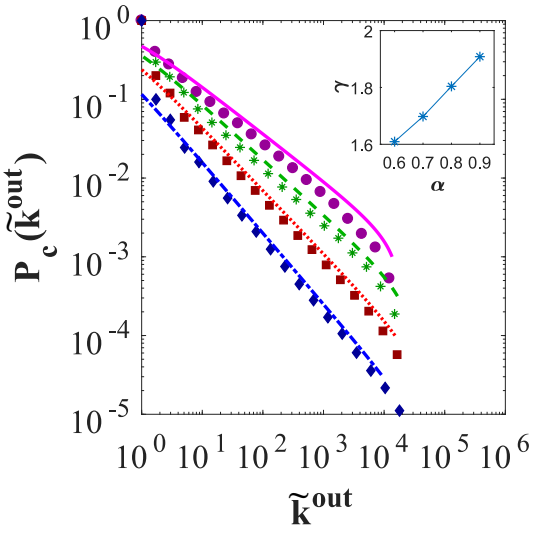

Figure 6.8: Cumulative distributions of the generalized out-degrees. For each choice of parameter α, 50 realizations of the simplicial complex were generated and averaged over. The simulated results were obtained for $t = 10^5$ and are represented by purple circles ($\alpha = 0.6$), green stars ($\alpha = 0.7$), red squares ($\alpha = 0.8$) and blue diamonds ($\alpha = 0.9$). The numerical results are represented by the purple solid line ($\alpha = 0.6$), green dashed line ($\alpha = 0.7$), red dotted line ($\alpha = 0.8$) and blue dot-dashed line ($\alpha = 0.9$). The inset shows the fitted exponents of the simulated distributions for the four values of α.

quite close to the exponents of the generalized out-strengths $\delta = 1 + \alpha$.

6.7 Strength versus degree

The models we have introduced all produce networks or simplicial complexes where the strengths of the nodes and total number of nodes follow the statistics of the Pitman-Yor model of balls in boxes. In contrast, the degree statistics do not follow the Pitman-Yor model, as links (or triangles) between nodes may be weighted multiple times without altering the degrees (or generalized degrees) of the nodes. In Section 6.6 we saw that in the directed network and directed simplicial complex versions, the exponents of the degree and generalized degree distributions closely match the exponents of the strength distributions, while in the undirected version

the exponent of the degree distribution is always equal to 2. An interesting question therefore is what is the relation between the degrees and generalized degrees of the nodes and their strengths and generalized strengths? In particular, in the directed network and directed simplicial complex versions we would like to know to what extent the strengths and generalized strengths can act as proxies for the degrees and generalized degrees or whether the stregth increases super-linearly with the degree of the nodes [34, 96]. To this end we have run simulations of the models for various values of α with the aim of extracting the average relations over all of the realizations. The total number of nodes in a realization is non-deterministic and in fact can vary widely over a set of realizations. This is important as the probability that a source node either gains a new link (or triangle) or has one of its existing ones reinforced depends on the ratio of the relative size of the degree of the node with respect to N. Therefore, to see the effect of this 'crowding' of the links of high degree nodes we have normalized the strength and degree data by dividing by N for each realization before averaging over all realizations. Figure 6.9 shows the relation between the normalized strengths and degrees of the nodes for the three versions of the model. We see from panel a) that the strengths of the nodes in the undirected version are significantly higher than their degrees, with the effect being greater for nodes with higher degree. This is expected, as the probability of a link being reinforced more than once is larger when the strengths of its two nodes are larger. In contrast, we see from panels b) and c) that the average out-strengths and average generalized out-strengths are very close to equal to the out-degrees and generalized out-degrees of the nodes in the directed network and simplicial complex versions respectively, suggesting the strengths may indeed act as proxies for the degrees.

6.8 Clustering and Degree Correlations

In this section we explore using simulations the clustering and degree correlations of the undirected (and unweighted) networks produced by the three versions of our model. In order to compare the results for the undirected network model with the results of the directed network and directed simplicial complex versions, we decided to discard the information about the direction of the links in the directed network and directed simplicial complex versions.

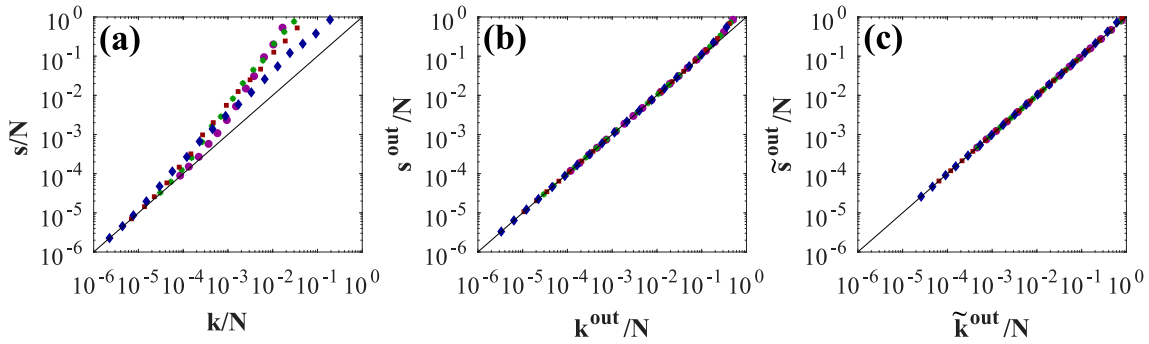

Figure 6.9: Plot of the normalized strengths versus the normalized degrees. Panels (a), (b) and (c) show the results for the undirected, directed and simplicial complex versions respectively. For each choice of parameter α, 50 realizations of the simplicial complex were generated and averaged over. The results were obtained for $t = 10^5$ and are represented by purple circles ($\alpha = 0.6$), green stars ($\alpha = 0.7$), red squares ($\alpha = 0.8$) and blue diamonds ($\alpha = 0.9$). The solid black line is the function $f(\frac{k}{N}) = \frac{k}{N}$, and is there as a guide to see how closely the strengths match the degrees.

Figure 6.10 shows the average degree $knn(k)$ of the neighbours of nodes with given degree k for the three versions. We see that for all three versions of the model, the networks produced are strongly disassortative. Interestingly in the undirected version, despite the fact that in Section 6.6 we found that the degree distribution has the same power-law exponent for different values of α, the strength of the disassortativity appears to be greater for increasing values of α. A likely explanation for this trend is that for larger α there is bias away from adding links between existing high degree nodes and towards creating links between a new node and a high degree node.

Figure 6.11 shows the average clustering of all nodes in the networks against the model parameter α for the three versions of the model. We see that the average clustering decreases with increasing values of α for all three versions.

6.9 Conclusions

In the research presented in this chapter, originally published in [35], we have presented three similar models for producing dense networks and simplicial complexes

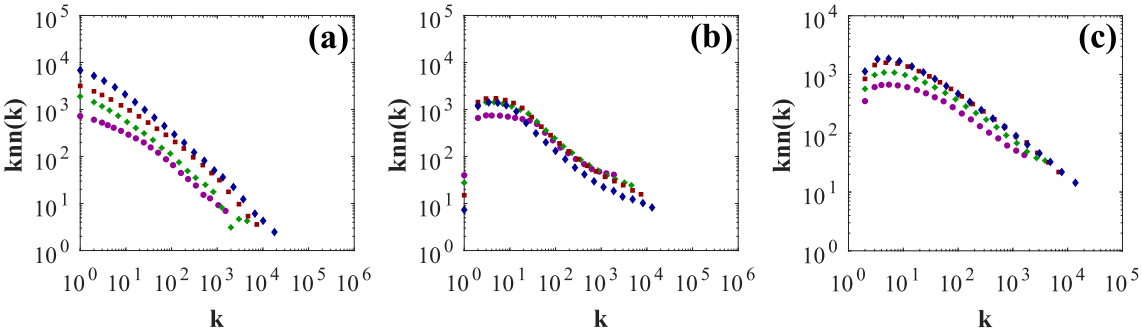

Figure 6.10: Average degree of the neighbours of nodes of degree k, taken from simulation data. Panels (a), (b) and (c) show the results for the undirected network, directed network and simplicial complex respectively. For each model version and each choice of parameter α, 50 realizations of the network were generated and averaged over. The results were obtained for $t = 10^5$ and for $\alpha = 0.6$, (purple circles), $\alpha = 0.7$ (green stars), $\alpha = 0.8$ (red squares), and $\alpha = 0.9$ (blue diamonds).

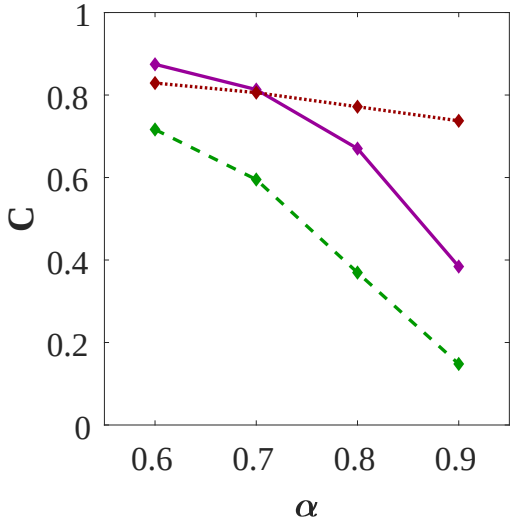

Figure 6.11: Average clustering of the nodes at different values of α, taken from simulation data with $t = 10^5$. For each model version and each choice of parameter α, 50 realizations of the network were generated and averaged over. Results from the three model versions are represented by the purple solid line (undirected), green dashed line (directed) and red dotted line (simplicial complex).

with power-law distributions of the strengths and degrees. The growth mechanisms of these models are analogous to the Pitman-Yor process, a stochastic process well-known among probability theorists for generating random partitions with power-law distributions of block sizes. Our undirected model can in one sense be thought of as a network with multiedges and a power-law degree distribution with tunable dense exponent $\gamma = 1 + \alpha \in (1, 2)$ or in a different sense as a weighted network with a power-law degree distribution with the border-line dense exponent $\gamma = 2$. Our directed network model produces dense directed networks with out-degree distributions that follow a power-law with tunable exponent $\gamma = 1 + \alpha$, and homogeneous distributions of the in-degrees. Our simplicial complex version extends the concept of a scale-free network to a scale-free simplicial complex with power-law distribution of the generalized out-degrees. These models demonstrate the difficulty in producing networks that are both dense and scale-free. However, they give insight into the possible mechanisms by which real-world networks densify, and may have a use as null-models for the growth of on-line social networks, recommendation networks or the brain. We show that the models are amenable to analytical calculations through our mean-field approach, and through simulation we verify the accuracy of our mean-field calculations.

Chapter 7

Conclusions

Simplicial complexes are a generalisation of networks that can encode the many-body interactions between the parts of a complex system. They allow for a more sophisticated characterisation of the structure of such systems that is not limited to node properties, but also provides a language to describe the higher order properties of simplices with dimension $d \geq 1$.

New measures of structure such as our generalised degrees and generalised strengths, or those based on algebraic topology provide a language that can describe a far richer variety of structures than what has previously been possible within traditional network science. On the one hand this opens up new possibilities: the discovery of new 'universal' properties, the development of more realistic models based on group-properties belonging to higher dimensional simplices rather than just nodes, or more subtle classifications of networks in terms of their homology. On the other hand this big increase in the types of structure that might be possible and the myriad measures that could potentially be introduced to characterise this structure should also be daunting. How do different measures relate to each other? What kind of structure would be interesting when we observe a real simplicial complex? What can we learn from this structure about the evolution or function of the simplicial complex?

Models help us answer these kinds of questions. They allow us to understand the dependences and relations between the structure of simplicial complexes, their evolution and processes occurring upon them. They help us to form hypotheses about

the origins of simplicial complexes, to make informed guesses at their structure in the absence of complete data, and to make generalisations and predictions about simplicial complexes sampled from the same source.

The models in this thesis fall into two categories: *explanatory models* in which the motivation is to find plausible hypotheses about the rules governing the evolution of simplicial complexes and to understand how these rules effect their structure, and *null models* which allow us to probe the relations between structural properties and help us answer questions such as, how can we 'best' model a simplicial complex given it has a given set of properties? In both types of model the aim is not necessarily to construct the most realistic model, but to isolate the effects of structural properties and growth mechanisms on the simplicial complexes produced.

The models presented in Chapter 4 are maximum entropy ensembles of d-dimensional simplicial complexes, based on the generalised degrees of the nodes. As such, our models have a clear use as null models for simplicial complexes. They are the least biased models for simplicial complexes given knowledge of these generalized degrees. Importantly, they allow for a statistically rigorous understanding of the implications of particular choices of the generalized degrees for the structure of simplicial complexes and dynamics taking place upon them. We believe that these models constitute an important first step in modelling simplicial complexes using the tools of equilibrium statistical physics and that this work will open up new perspectives for investigating a new generation of maximum entropy models of simplicial complexes.

This belief is already being vindicated, as since the publication of our work, a direct adaptation of our configuration model was used in [97] to quantify the significance of homological structure in simplicial complexes constructed from real data representing relations between pollinating insects and plants, human diseases and genes, and criminals, victims and witnesses in crimes. In the paper, a version of the configuration model was used to quantify the typical Betti numbers expected given only the generalised degrees in the real data. It was found that the Betti numbers found for the insect-plant data set appeared to be very similar to in the configuration model suggesting an absence of significant structure beyond that described

by the generalised degrees. For the disease and crime datasets it was instead found that the Betti numbers differed significantly from the configuration model indicating the presence of a significant structure not well explained in terms of the generalised degrees. This demonstrates the use of models designed along similar lines to ours for identifying interesting or non-trivial structure in complex data sets.

In Chapter 5 we presented a model of a d-dimensional simplicial complex that is both weighted and growing. This model follows in the tradition of growing network models that seek to characterize the relations between simple growth mechanisms of networks and their structural properties. The model allowed us to investigate the competing effects of growth, reinforcement and dimension on the generalized strengths of the simplices and on the distribution of weight across the simplicial complex. We found that the model could exhibit a rich variety of topologies and weight distributions, namely power-law, exponential and bimodal distributions of the generalised degrees and linear, super-linear or exponential scalings of generalised strength with generalised degree. Remarkably each of these distributions and scalings could be exhibited simultaneously within a single simplicial complex for faces of different dimension. This was due to the effective attachment and reinforcement mechanisms 'felt' by the lower dimension simplices, and highlights the importance that growth mechanisms which function based on clique or simplex properties may have in the evolution of networks and simplicial complexes.

In Chapter 6 we presented a modelling framework for producing networks and simplicial complexes which were both dense and scale-free. The growth mechanisms of the models contained within our framework are analogous to the Pitman-Yor process, a 'ball-in-the-box' process well-known among probability theorists for producing power-laws with exponent $\gamma \in (1, 2]$. Our undirected model demonstrated the difficulty of producing a simple network which is both dense and scale-free. In this model the network clearly densified over time and was also scale-free with a borderline dense exponent $\gamma = 2$. By relaxing the requirement for the network to be simple, either by considering only the out-degree or by reinterpreting the weight of a link as the number of multilinks between two nodes, we found that it was easy to create scale-free networks and simplicial complexes with tunable dense exponent

$\gamma \in (1, 2]$.

The models presented in this thesis are simple models of simplicial complexes that seek to isolate the effects either of structural constraints or of growth processes on the properties of the simplicial complexes produced. Our results contribute to a young but expanding literature on models of random simplicial complexes, and together with this literature our work will hopefully inform the development of more sophisticated models that successfully replicate the properties of real simplicial complexes while remaining as simple and transparent as possible.

The use of simplicial complexes to represent real, complex systems is still in its infancy and we do not know what novel measures of structure will be developed in the coming years, nor what properties will be observed that we will need to replicate. Developments in these areas will inform the creation of new models, while the insight gained from study of the models will in turn inform the analysis of real simplicial complexes.

Appendix A: Derivation of Eq. (4.68) for Ω

In this appendix we derive Eq. (4.68) for Ω in the presence of the structural cutoff. The quantity Ω indicates the logarithm of the probability that in the canonical ensemble of simplicial complexes enforcing the sequence of expected degree of the nodes $\{\overline{k_r}\}$ we observe a simplicial complex realization in which the sequence of the generalized degree of the nodes is exactly $\{\overline{k_r}\}$. This is written as

$$\begin{aligned}\Omega \quad &= -\ln \sum_{G} P_{CE}(G) \prod_{r} \delta\left(k_r, k_{d,0}(r)\right) \\ &= -\ln \sum_{G} \prod_{\alpha} p_{\alpha}^{a_\alpha}(1-p_\alpha)^{1-a_\alpha} \prod_{r} \delta\left(\overline{k_r}, k_{2,0}(r)\right) \end{aligned} \tag{1}$$

where, in presence of the structural cutoff, the probabilities p_α are given by Eq. (4.26). In order to evaluate Ω, we use the integral representation of the Kroenecker delta

$$\delta(x,y) = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega x - i\omega y}, \tag{2}$$

getting

$$\begin{aligned}\Omega &= -\ln \sum_{G} \prod_{\alpha} p_{\alpha}^{a_\alpha} (1-p_\alpha)^{1-a_\alpha} \prod_{r} \int_{-\pi}^{\pi} \frac{d\omega_r}{2\pi} e^{i\omega_r \overline{k_r} - i\omega_r \sum_{\alpha' | r\subset\alpha'} a_{\alpha'}} \\ &= -\ln \int_{-\pi}^{\pi} \prod_{r} \frac{d\omega_r}{2\pi} e^{\mathcal{G}[\{\omega_r\}]} \end{aligned} \tag{3}$$

where

$$\mathcal{G}[\{\omega_r\}] = \quad i \sum_{i} \omega_r \overline{k_r} + \sum_{\alpha} \ln \left[1 + p_\alpha \left(e^{-i\sum_{r\subset\alpha}\omega_r} - 1\right)\right].$$

For an uncorrelated simplicial complex ensemble with structural cutoff and with p_α given by Eq. (4.26) and $p_\alpha \ll 1$ we can approximate $\mathcal{G}[\{\omega_r\}]$ as

$$\mathcal{G}[\{\omega_r\}] = i \sum_{r} \omega_r \overline{k_r} + \sum_{\alpha} p_\alpha \left(e^{-i\sum_{r\subset\alpha}\omega_r} - 1\right). \tag{4}$$

Using the explicit factorized expression for p_α in the presence of the structural cutoff we observe that we can write

$$\sum_{\alpha} p_\alpha \left(e^{-i\sum_{r\subset\alpha}\omega_r} - 1\right) = \frac{d!}{(d+1)!} \langle \overline{k} \rangle N \left(\nu^{d+1} - 1\right) \tag{5}$$

where

$$\nu = \sum_{r} \frac{\overline{k_r}}{\langle \overline{k} \rangle N} e^{-i\omega_r}. \tag{6}$$

We now introduce the density

$$c(\omega|k) = \frac{1}{N_k} \sum_{r} \delta(\omega - \omega_r) \delta\left(k, \overline{k_r}\right) \tag{7}$$

where

$$N_k = NP_{d,0}(k) \tag{8}$$

indicates the number of nodes with generalized degree of the nodes $\overline{k_r} = k$ and $P_{d,0}(k)$ indicates the distribution of the generalized degree of the nodes. We can therefore express ν given by Eq. (6) in terms of $c(\omega|k)$ obtaining

$$\nu = \sum_k \frac{k}{\langle \overline{k} \rangle} P_{d,0}(k) \int d\omega e^{-i\omega} c(\omega|k). \tag{9}$$

Using the delta functions

$$\delta\left(c_{\omega,k}, c(\omega|k)\right) = \int_{-\pi}^{\pi} \frac{d\hat{c}_{\omega,k}}{2\pi N_k} e^{i\hat{c}_{\omega,k}[N_k c_{\omega,k} - \sum_r \delta(\omega-\omega_r)\delta\left(k,\overline{k_r}\right)]} \tag{10}$$

we can now express Ω as

$$\Omega = -\ln \int \mathcal{D}c_{\omega,k} \mathcal{D}\hat{c}_{\omega,k} e^{NF[c_{\omega,k},\hat{c}_{\omega,k}]},$$

where $F[c_{\omega,k}, \hat{c}_{\omega,k}]$ is given by

$$\begin{aligned} F[c_{\omega,k}, \hat{c}_{\omega,k}] = \quad & i \sum_k P_{d,0}(k) \int d\omega \hat{c}_{\omega,k} c_{\omega,k} + \frac{d!}{(d+1)!} \langle \overline{k} \rangle \left(\nu^{d+1} - 1\right) \\ & + \sum_k P_{d,0}(k) \ln \int \frac{d\omega}{2\pi} e^{i\omega k - i\hat{c}_{\omega,k}}. \end{aligned} \tag{11}$$

We evaluate the integral (11) with the saddle point method. The saddle point equations read,

$$\begin{aligned} \frac{\partial F[c_{\omega,k}, \hat{c}_{\omega,k}]}{\partial c_{\omega,k}} &= 0, \\ \frac{\partial F[c_{\omega,k}, \hat{c}_{\omega,k}]}{\partial \hat{c}_{\omega,k}} &= 0. \end{aligned}$$

Which gives us

$$
\begin{aligned}
-i\hat{c}_{\omega,k} &= k\nu^d e^{-i\omega}, && (12)\\
c_{\omega,k} &= \frac{\frac{1}{2\pi}e^{i\omega k - i\hat{c}_{\omega,k}}}{\int \frac{d\omega}{2\pi} e^{i\omega k - i\hat{c}_{\omega,k}}}. && (13)
\end{aligned}
$$

Using Eq. (12) we observe that the integral appearing in Eq. (13) can be expressed in terms of ν, obtaining

$$
\int \frac{d\omega}{2\pi} e^{i\omega k - i\hat{c}_{\omega,k}} = \int \frac{d\omega}{2\pi} e^{i\omega k + k\nu^d e^{-i\omega}} = \int \frac{d\omega}{2\pi} e^{i\omega k} \sum_h (\nu^d k)^h e^{-i\omega h} \frac{1}{h!} = \frac{(\nu^d k)^k}{k!}. \quad (14)
$$

Substituting this result in to Eq. (13), we get

$$
\begin{aligned}
c_{\omega,k} &= \frac{k!}{2\pi(\nu^d k)^k} e^{i\omega k - i\hat{c}_{\omega,k}} \\
&= \frac{k!}{2\pi(\nu^d k)^k} e^{i\omega k + k\nu^d e^{-i\omega}}. && (15)
\end{aligned}
$$

Finally, we can substitute this expression into the definition of ν given by Eq. (9) obtaining

$$
\begin{aligned}
\nu &= \sum_k \frac{k}{\langle \bar{k} \rangle} P_{d,0}(k) \int d\omega e^{-i\omega} c_{\omega,k} \\
&= \sum_k \frac{k}{\langle \bar{k} \rangle} P_{d,0}(k) \frac{k!}{(\nu^d k)^k} \int \frac{d\omega}{2\pi} e^{i\omega(k-1) + k\nu^d e^{-i\omega}} \\
&= \sum_k \frac{k}{\langle \bar{k} \rangle} P_{d,0}(k) \frac{k!}{(\nu^d k)^k} \int \frac{d\omega}{2\pi} e^{i\omega(k-1)} \sum_h (\nu^d k)^h e^{-i\omega h} \frac{1}{h!} \\
&= \sum_k \frac{k}{\langle \bar{k} \rangle} P_{d,0}(k) \frac{k!}{(\nu^d k)^k} \frac{(\nu^d k)^{k-1}}{(k-1)!} = \nu^{-d}. && (16)
\end{aligned}
$$

Therefore, ν is the solution of the equation $\nu = \nu^{-d}$, and so we have

$$
\nu = 1. \quad (17)
$$

Using this result, and Eq. (14) it is immediate to show that the value of the functional $F[c_{\omega,k}, \hat{c}_{\omega,k}]$ (Eq. (11)) at the saddle point is given by

$$F[c_{\omega,k}, \hat{c}_{\omega,k}] = i \sum_k P_{d,0}(k) \int d\omega \hat{c}_{\omega,k} c_{\omega,k} + \sum_k P_{d,0}(k) \ln \left[\frac{k^k}{k!} \right]. \tag{18}$$

Proceeding as in Eq. (16) it can be easily shown that

$$\begin{aligned} i \sum_k P_{d,0}(k) \int d\omega \hat{c}_{\omega,k} c_{\omega,k} &= -\sum_k P_{d,0}(k) k \frac{k!}{k^k} \int d\omega 2\pi e^{i\omega(k-1)+ke^{-i\omega}} \\ &= -\langle \overline{k} \rangle = -\sum_k P_{d,0}(k) k. \end{aligned} \tag{19}$$

Finally, evaluating the integral (11) at the saddle point we obtain the simple expression for Ω given by

$$\begin{aligned} \Omega &= -\ln \left(e^{NF[c_{\omega,k}, \hat{c}_{\omega,k}]} \right) \\ &= -N \sum_k P_{d,0}(k) \ln \left(\frac{k^k}{k!} e^{-k} \right) \\ &= -\sum_r \ln \left(\pi_{\overline{k_r}}(\overline{k_r}) \right), \end{aligned} \tag{20}$$

where $\pi_{\overline{k_r}}(\overline{k_r})$ indicated the Poisson distribution with average $\overline{k_r}$ calculated at $\overline{k_r}$, i.e.

$$\pi_{\overline{k_r}}(\overline{k_r}) = \frac{\overline{k_r}^{\overline{k_r}}}{\overline{k_r}!} e^{-\overline{k_r}}. \tag{21}$$

Appendix B: Derivation of the probability distributions for the generalized degrees of the nodes in the canonical ensemble with structural cutoff

In this appendix we show that in the canonical ensemble with structural cut-off, the probability $\rho_r(k)$ that a node r has generalized degree $k_{d,0}(r) = k$ follows a Poisson distribution with expected value $\overline{k_r}$. In keeping with the notation in Section 4.3.4, we show that $\rho_r(k) = \pi_{\overline{k_r}}(k)$, where

$$\pi_{\overline{k_r}}(k) = e^{-\overline{k_r}} \frac{\overline{k_r}^k}{k!}. \tag{22}$$

To prove this, we first observe that the probability that $k_{d,0}(r) = k$ in the canonical ensemble is

$$\rho_r(k) = \sum_G P(G)\delta\left(k_{d,0}(r), k\right), \tag{23}$$

i.e. the sum of the probabilities of all simplicial complexes in the ensemble for which the node r has generalized degree k. We use the representation of $P(G)$ as a product of simplex probabilities p_α given in Eq. (4.17) together with the integral representation of the Kronecker delta:

$$\delta(x,y) = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega x - i\omega y}, \tag{24}$$

to get

$$\rho_r(k) = \sum_G \prod_\alpha p_\alpha^{a_\alpha}(1-p_\alpha)^{1-a_\alpha} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k - i\omega \sum_{\alpha|r\subset\alpha} a_\alpha}. \tag{25}$$

Evaluating the sum over all simplicial complexes gives

$$\begin{aligned}\rho_r(k) &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_{\alpha|r\subset\alpha} \left[p_\alpha e^{-i\omega} + 1 - p_\alpha\right] \\ &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \sum_{\alpha|r\subset\alpha} \log\left(p_\alpha e^{-i\omega} + 1 - p_\alpha\right)}.\end{aligned} \tag{26}$$

Below the structural cut-off, we have $p_\alpha \ll 1$, so we may make the approximation

$$\log\left(p_\alpha e^{-i\omega} + 1 - p_\alpha\right) \approx p_\alpha\left(e^{-i\omega} - 1\right). \tag{27}$$

Eq. (26) becomes

$$\rho_r(k) = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + e^{-i\omega}\overline{k_r} - \overline{k_r}}, \tag{28}$$

where we have used $\overline{k_r} = \sum_{\alpha|r\subset\alpha} p_\alpha$. We make the expansion:

$$e^{e^{-i\omega}\overline{k_r}} = \sum_{\kappa=0}^{\infty} \frac{e^{-i\omega\kappa}\overline{k_r}^{\kappa}}{\kappa!}. \tag{29}$$

Using this expansion in Eq (28) gives

$$\rho_r(k) = \sum_{\kappa=0}^{\infty} e^{-\overline{k_r}} \frac{\overline{k_r}^{\kappa}}{\kappa!} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k - i\omega\kappa}, \tag{30}$$

which is nothing more than

$$\rho_r(k) = \sum_{\kappa=0}^{\infty} e^{-\overline{k_r}} \frac{\overline{k_r}^{\kappa}}{\kappa!}(\kappa)\delta(k,\kappa) = e^{-\overline{k_r}} \frac{\overline{k_r}^{k}}{k!}, \tag{31}$$

where we have identified the integral in Eq (30) as being the integral representation of $\delta(k, \kappa)$.

Appendix C: Main steps of the derivation of the Eq.(55)

In Section 5.3.4 of the main body of the thesis we derived the following equation (Eq. (5.50)) for the average generalized strength $s_{d,\delta}(t,t_\alpha)$ of the δ-face α with arrival time t_α

$$s_{d,\delta}(t,t_\alpha)=w_0\frac{(d-\delta)!}{(d+s)^{d-\delta}}t^{\lambda}t_{i_\delta}^{-\frac{d+s-\delta-1}{d+s}}\sum_{\{q\}}A_{q(\delta)}\left(\prod_{r=0}^{\delta-1}X_{q(r),q(r+1)}\right)B_{q(0)}, \tag{32}$$

where $A_{q(\delta)}$, $B_{q(\delta)}$ and $X_{q(r),q(r+1)}$ are defined respectively by Eqs. (41), (5.48) and (5.47) of the main text. These integrals are expressed in terms of $I(\tau_L,\tau_U)$ defined in Eq. (5.46), so evaluation of $I(\tau_L,\tau_U)$ is required for the evaluation of $A_{q(\delta)}$, $B_{q(\delta)}$ and $X_{q(r),q(r+1)}$. To do this we note that in the integral $I(\tau_L,\tau_U)$, the factors corresponding to each of the birth times takes the same form: $t_r^{\frac{1}{d+s}-1}$. We

exploit this symmetry in the following way:

$$
\begin{aligned}
I^n_{\tau_L,\tau_U} &= \int_{\tau_L}^{\tau_U} dt_n t_n^{\frac{1}{d+s}-1} \int_{\tau_L}^{t_n} dt_{n-1} t_{n-1}^{\frac{1}{d+s}-1} ... \int_{\tau_L}^{t_2} dt_1 t_1^{\frac{1}{d+s}-1} && (33)\\
&= \frac{1}{n!} \int_{\tau_L}^{\tau_U} dt_n t_n^{\frac{1}{d+s}-1} \int_{\tau_L}^{\tau_U} dt_{n-1} t_{n-1}^{\frac{1}{d+s}-1} ... \int_{\tau_L}^{\tau_U} dt_1 t_1^{\frac{1}{d+s}-1} && (34)\\
&= \frac{1}{n!} \left(\int_{\tau_L}^{\tau_U} d\tau \tau^{\frac{1}{d+s}-1} \right)^n && (35)\\
&= \frac{(d+s)^n}{n!} \left(t^{\frac{1}{d+s}} - \tau^{\frac{1}{d+s}} \right)^n && (36)\\
&= \frac{(d+s)^n}{n!} \sum_{r=0}^{n} \binom{n}{r} (-1)^r t^{\frac{n-r}{d+s}} \tau^{\frac{r}{d+s}}. && (37)
\end{aligned}
$$

This result allows us to express $X_{q(r),q(r+1)}$ as

$$
\begin{aligned}
X_{q(r),q(r+1)} &= \begin{cases} I^{q(r+1)-q(r)-1}_{t_{i_r},t_{i_{r+1}}} & \text{if} \quad q(r+1)-q(r) > 1, \\ 1 & \text{if} \quad q(r+1)-q(r) = 1, \end{cases} \\
&= \begin{cases} \frac{(d+s)^{q(r+1)-q(r)-1}}{(q(r+1)-q(r)-1)!} \left(t_{i_{r+1}}^{\frac{1}{d+s}} - t_{i_r}^{\frac{1}{d+s}} \right)^{q(r+1)-q(r)-1} & \text{if} \quad q(r+1)-q(r) > 1, \\ 1 & \text{if} \quad q(r+1)-q(r) = 1. \end{cases}
\end{aligned} \tag{38}
$$

Similarly $B_{q(0)}$ can be expressed as

$$
\begin{aligned}
B_{q(0)} &= \begin{cases} I^{q(0)}_{0,t_{i_0}} & \text{if} \quad q(0) > 0, \\ 1 & \text{if} \quad q(0) = 0, \end{cases} \\
&= \begin{cases} \frac{(d+s)^{q(0)}}{q(0)!} t_{i_0}^{\frac{q(0)}{d+s}} & \text{if} \quad q(0) > 0, \\ 1 & \text{if} \quad q(0) = 0. \end{cases}
\end{aligned} \tag{39}
$$

Finally using Eq. (37) and the definition of $A_{q(\delta)}$ that we rewrite here for conve-

nience,

$$A_{q(\delta)} = \begin{cases} \int_{t_{i_\delta}}^{t} dt_{j_d} t_{j_d}^{-\lambda-\frac{d}{d+s}} I_{t_{i_\delta},t_{j_d}}^{d-q(\delta)-1} & \text{if } 0 \le q(\delta) \le d-1, \\ t_{i_\delta}^{-\lambda+\frac{s-1}{d+s}} & \text{if } q(\delta) = d, \end{cases} \tag{40}$$

we obtain

$$A_{q(\delta)} = \begin{cases} \frac{(d+s)^{d-q(\delta)-1}}{(d-q(\delta)-1)!} \sum_{r=0}^{d-q(\delta)-1} \left[\binom{d-q(\delta)-1}{r} (-1)^r t_{i_\delta}^{\frac{r}{d+s}} \right. & \\ \left. \times \int_{t_{i_\delta}}^{t} dt_{j_d} t_{j_d}^{-\lambda+\frac{d+s-q(\delta)-r-1}{d+s}-1} \right] & \text{if } 0 \le q(\delta) \le d-1, \\ t_{i_\delta}^{-\lambda+\frac{s-1}{d+s}} & \text{if } q(\delta) = d. \end{cases} \tag{41}$$

In the case $q(\delta) \le d-1$ the integral present in Eq. (41) has two separate expressions for $\lambda = \frac{d+s-q(\delta)-1-r}{d+s}$ and $\lambda \neq \frac{d+s-q(\delta)-1-r}{d+s}$.

By performing the integral we find the following expression for $A_{q(\delta)}$

$$A_{q(\delta)} = \begin{cases} (d+s)^{d-q(\delta)-1} \sum_{r=0}^{d-q(\delta)-1} & \\ \times D_{q(\delta)+r} t_{i_\delta}^{\frac{r}{d+s}} \frac{(-1)^r}{r!} & \text{if } 0 \le q(\delta) \le d-1, \\ t_{i_\delta}^{-\lambda+\frac{s-1}{d+s}} & \text{if } q(\delta) = d, \end{cases} \tag{42}$$

where in the quantity $D_{q(\delta)+r}$ we have gathered together all of the factors in Eq. (42) that depend on $q(\delta)$ and r only through their sum (the reason why this is useful will become apparent later) and is given by

$$D_{q(\delta)+r} = \begin{cases} \frac{1}{(d-q(\delta)-r-1)!}\left(\frac{d+s-q(\delta)-1-r}{d+s} - \lambda\right)^{-1} & \text{if} \quad \lambda \neq \frac{d+s-q(\delta)-1-r}{d+s}, \\ \times\left(t^{\frac{d+s-q(\delta)-1-r}{d+s}-\lambda} - t_{i_\delta}^{\frac{d+s-q(\delta)-1-r}{d+s}-\lambda}\right) & \\ \frac{1}{(d-q(\delta)-r-1)!}\log\left(\frac{t}{t_{i_\delta}}\right) & \text{if} \quad \lambda = \frac{d+s-q(\delta)-1-r}{d+s}. \end{cases} \tag{43}$$

Having calculated $A_{q(\delta)}$, $B_{q(\delta)}$ and $X_{q(r),q(r+1)}$, we now wish to evaluate the sum over the positions $\{q(r)\}_{r=0,1,\ldots\delta}$ that the birth times of the nodes of α can take in the sequence of birth times $[t_{j_0}, \ldots, t_{j_d}]$. This sum has the form

$$\sum_{\{q\}} = \sum_{q(\delta)=\delta}^{d} \sum_{q(\delta-1)=\delta-1}^{q(\delta)-1} \cdots \sum_{q(1)=1}^{q(2)-1} \sum_{q(0)=0}^{q(1)-1} \tag{44}$$

Similar to the nested integrals we encountered earlier, we have here a set of nested sums. We now write the sum over $\{q(r)\}_{r=0,1,\ldots\delta}$ in Eq. (32) using the form given above

$$\begin{array}{l} \sum_{\{q\}} A_{q(\delta)} \left(\prod_{r=0}^{\delta-1} X_{q(r),q(r+1)}\right) B_{q(0)} = \\ \sum_{q(\delta)=\delta}^{d} A_{q(\delta)} \sum_{q(\delta-1)=\delta-1}^{q(\delta)-1} X_{q(\delta-1),q(\delta)} \cdots \sum_{q(1)=1}^{q(2)-1} X_{q(1),q(2)} \sum_{q(0)=0}^{q(1)-1} X_{q(0),q(1)} B_{q(0)}. \end{array} \tag{45}$$

In order to evaluate these sums, we rewrite the above in terms of a set of recursively defined functions as follows

$$\sum_{\{q\}} A_{q(\delta)} \left(\prod_{r=0}^{\delta-1} X_{q(r),q(r+1)}\right) B_{q(0)} = \sum_{q(\delta)=\delta}^{d} A_{q(\delta)} R_{q(\delta),\delta}, \tag{46}$$

where $R_{q(r),r}$ are functions defined recursively by the following pair of equations,

$$R_{q(1),1} = \sum_{q(0)=0}^{q(1)-1} X_{q(0),q(1)} B_{q(0)}, \tag{47}$$

$$R_{q(\beta),\beta} = \sum_{q(\beta-1)=\beta-1}^{q(\beta)-1} X_{q(\beta-1),q(\beta)} R_{q(\beta-1),\beta-1}. \tag{48}$$

The solution of these equations is calculated in Appendix D and reads

$$R_{q(\beta),\beta} = (d+s)^{q(\beta)-\beta} \frac{t_{i_\beta}^{\frac{q(\beta)-\beta}{d+s}}}{(q(\beta)-\beta)!}. \tag{49}$$

We will shortly use the above in our generalised strength equation

$$s_{d,\delta}(t,t_\alpha) \;=\; w_0 \frac{(d-\delta)!}{(d+s)^{d-\delta}} t^\lambda t_{i_\delta}^{-\frac{d+s-\delta-1}{d+s}} \sum_{q(\delta)=\delta}^{d} A_{q(\delta)} R_{q(\delta),\delta}. \tag{50}$$

First we note that that the solution we found for $A_{q(\delta)}$ differs when $q(\delta) = d$ from when $q(\delta) \leq d-1$. When $q(\delta) = d$, the contribution to the sum in Eq. (50) is

$$A_d R_{d,\delta} = \frac{(d+s)^{d-\delta}}{(d-\delta)!} t_{i_\delta}^{-\lambda+\frac{d+s-\delta-1}{d+s}}, \tag{51}$$

where we have used Eq.s (42) and (49) for A_d and $R_{d,\delta}$ respectively. Using Eq.s (42) and (49) for $q(\delta) \leq d-1$, the contribution to the sum is

$$\begin{aligned}
&\textstyle\sum_{q(\delta)=\delta}^{d-1} A_{q(\delta)} R_{q(\delta),\delta} = \\
&(d+s)^{d-\delta-1} \textstyle\sum_{q(\delta)=\delta}^{d-1} \sum_{r=0}^{d-q(\delta)-1} D_{q(\delta)+r} t_{i_\delta}^{\frac{q(\delta)+r-\delta}{d+s}} \frac{(-1)^r}{r!(q(\delta)-\delta)!} \\
&= (d+s)^{d-\delta-1} D_{q(\delta)+r} t_{i_\delta}^{\frac{q(\delta)+r-\delta}{d+s}} \Big|_{q(\delta)+r=\delta} \\
&= (d+s)^{d-\delta-1} D_\delta.
\end{aligned} \tag{52}$$

Note that in deriving the Eq.(52) we have used the following mathematical identity

$$\sum_{x=a}^{b}\sum_{y=0}^{b-x} f(x+y)\frac{(-1)^y}{y!(x-a)!} = f(a), \tag{53}$$

valid for integers $a, b > 0$ with $a < b$ (in Eq. (52) the substitution is $f(q(\delta)+r) = D_{q(\delta)+r}t_{i_\delta}^{\frac{q(\delta)+r-\delta}{d+s}}$). Therefore the average generalized strength given by Eq. (50) can be written as

$$\begin{aligned} s_{d,\delta}(t,t_\alpha) &= w_0\tfrac{(d-\delta)!}{(d+s)^{d-\delta}}t^\lambda t_{i_\delta}^{-\frac{d+s-\delta-1}{d+s}}\left[A_dR_{d,\delta} + \textstyle\sum_{q(\delta)=\delta}^{d-1}A_{q(\delta)}R_{q(\delta)}\right] \\ &= w_0\tfrac{(d-\delta)!}{(d+s)^{d-\delta}}t^\lambda t_{i_\delta}^{-\frac{d+s-\delta-1}{d+s}}\left[\tfrac{(d+s)^{d-\delta}}{(d-\delta)!}t_{i_\delta}^{-\lambda+\frac{d+s-\delta-1}{d+s}} + (d+s)^{d-\delta-1}D_\delta\right], \end{aligned} \tag{54}$$

which simplifies to

$$s_{d,\delta}(t,t_\alpha) = w_0\Big(\frac{t}{t_{i_\delta}}\Big)^\lambda + w_0\frac{(d-\delta)!}{d+s}t^\lambda t_{i_\delta}^{-\frac{d+s-\delta-1}{d+s}}D_\delta. \tag{55}$$

As noted earlier, D_δ takes different forms in the cases $\lambda \neq \lambda_\delta = \frac{d+s-\delta-1}{d+s}$ and $\lambda = \lambda_\delta = \frac{d+s-\delta-1}{d+s}$. Inserting Eq. (43) into Eq. (55) leads to our final expression for the generalized strength:

$$s^\alpha_{d,\delta}(t) = \begin{cases} w_0\frac{d-\delta}{(d+s)(\lambda_\delta-\lambda)}\left(\frac{t}{t_{i_\delta}}\right)^{\lambda_\delta} + w_0\left[1-\frac{d-\delta}{(d+s)(\lambda_\delta-\lambda)}\right]\left(\frac{t}{t_{i_\delta}}\right)^{\lambda} & \text{if} \quad \lambda \neq \lambda_\delta, \\ w_0\left(\frac{t}{t_{i_\delta}}\right)^{\lambda}\left[1+\frac{d-\delta}{d+s}\log\left(\frac{t}{t_{i_\delta}}\right)\right] & \text{if} \quad \lambda = \lambda_\delta. \end{cases} \tag{56}$$

Since this equation is the same as Eq. (5.50) of the main text, this concludes here our discussion.

Appendix D: Derivation of Eq. (49)

In this appendix our goal is to show that Eq. (49) holds. This equation is given by

$$R_{q(\beta),\beta} = (d+s)^{q(\beta)-\beta} \frac{t_{i_\beta}^{\frac{q(\beta)-\beta}{d+s}}}{(q(\beta)-\beta)!}, \tag{57}$$

where $R_{q(r),r}$ are functions defined recursively by the following pair of equations

$$R_{q(1),1} = \sum_{q(0)=0}^{q(1)-1} X_{q(0),q(1)} B_{q(0)}, \tag{58}$$

$$R_{q(\beta),\beta} = \sum_{q(\beta-1)=\beta-1}^{q(\beta)-1} X_{q(\beta-1),q(\beta)} R_{q(\beta-1),\beta-1}. \tag{59}$$

To this end we first check that (57) holds for $\beta = 1$. Inserting Eq. (38) for $X_{q(0),q(1)}$ and Eq. (39) for $B_{q(0)}$ into Eq. (58) gives

$$\begin{aligned} R_{q(1),1} &= \sum_{q(0)=0}^{q(1)-1} \sum_{l_0=0}^{q(1)-q(0)-1} \frac{(d+s)^{q(1)-1}(-1)^{l_0}}{(q(1)-q(0)-1-l_0)!(l_0)!q(0)!} \\ &\times t_{i_0}^{\frac{q(0)+l_0}{d+s}} t_{i_1}^{\frac{q(1)-q(0)-1-l_0}{d+s}}. \end{aligned} \tag{60}$$

We note that the expression being summed over factorises into a term depending on $q(0)$ and l_0 only through their sum and a term depending on $q(0)$ and l_0 otherwise:

$$R_{q(1),1} = \sum_{q(0)=0}^{q(1)-1} \sum_{l_0=0}^{q(1)-q(0)-1} f(q(0)+l_0) \frac{(-1)^{l_0}}{l_0! q(0)!}, \tag{61}$$

where

$$f(q(0)+l_0) = (d+s)^{q(1)-1} \frac{t_{i_0}^{\frac{q(0)+l_0}{d+s}} t_{i_1}^{\frac{q(1)-q(0)-1-l_0}{d+s}}}{(q(1)-q(0)-1-l_0)!}. \tag{62}$$

Using the mathematical identity Eq.(53), Eq. (61) simplifies to

$$R_{q(1),1} = f(0) = (d+s)^{q(1)-1} \frac{t_{i_1}^{\frac{q(1)-1}{d+s}}}{(q(1)-1)!}. \tag{63}$$

So (57) holds in the case $\beta = 1$.

We now show that in general if Eq. (57) holds for some β then it must also hold for $\beta+1$. Substituting in Eq. (38) and Eq. (57) into Eq. (59) gives

$$R_{q(\beta+1),\beta+1} = \sum_{q(\beta)=\beta}^{q(\beta+1)-1} \sum_{l_\beta=0}^{q(\beta+1)-q(\beta)-1} \left[\frac{(d+s)^{q(\beta+1)-\beta-1}(-1)^{l_\beta}}{(q(\beta+1)-q(\beta)-1-l_\beta)!(l_\beta)!(q(\beta)-\beta)!} \right.$$
$$\left. \times t_{i_\beta}^{\frac{q(\beta)+l_\beta-\beta}{d+s}} t_{i_{\beta+1}}^{\frac{q(\beta+1)-q(\beta)-1-l_\beta}{d+s}} \right]. \tag{64}$$

Similar to the $\beta = 1$ case we may write (64) in the form

$$R_{q(\beta+1),\beta+1} = \sum_{q(\beta)=\beta}^{q(\beta+1)-1} \sum_{l_\beta=0}^{q(\beta+1)-q(\beta)-1} f(q(\beta)+l_\beta) \frac{(-1)^{l_\beta}}{l_\beta! q(\beta)!}, \tag{65}$$

where in this case the term depending only on $q(\beta)$ and l_β through the sum of the two is

$$f(q(\beta)+l_\beta) = (d+s)^{q(\beta+1)-\beta-1} \frac{t_{i_\beta}^{\frac{q(\beta)+l_\beta-\beta}{d+s}} t_{i_{\beta+1}}^{\frac{q(\beta+1)-q(\beta)-1-l_\beta}{d+s}}}{(q(\beta+1)-q(\beta)-1-l_\beta)!}. \tag{66}$$

Using the identity (53) allows us to make the simplification

$$R_{q(\beta+1),\beta+1} = f(\beta) = (d+s)^{q(\beta+1)-\beta-1} \frac{t_{i_{\beta+1}}^{\frac{q(\beta+1)-\beta-1}{d+s}}}{(q(\beta+1)-\beta-1)!}, \tag{67}$$

which confirms Eq.(57) or equivalently, Eq.(49).

References

[1] A. L. Barabasi, "Network science," *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, vol. 371, no. 1987, 2013. 13, 30, 32, 33, 49, 50, 53, 110, 135

[2] M. J. Newman, *Networks: An introduction*. Oxford University Press, 2010. 13, 30, 32, 33, 110

[3] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," *Advances in Physics*, vol. 51, no. 4, pp. 1079–1187, 2002. 13, 30, 32, 33, 110

[4] M. A. Serrano and M. Boguna, "Topology of the world trade web," *Physical Review E*, vol. 68, no. 1, 2003. 13

[5] S. H. Yook, Z. N. Oltvai, and A. L. Barabasi, "Functional and topological characterization of protein interaction networks," *Proteomics*, vol. 4, no. 4, pp. 928–942, 2004. 13

[6] A. G. Haldane and R. M. May, "Systemic risk in banking ecosystems," *Nature*, vol. 469, no. 7330, pp. 351–355, 2011. 13

[7] G. Bianconi, "Interdisciplinary and physics challenges of network theory," *Epl*, vol. 111, no. 5, 2015. 14

[8] M. Kahle, *Topology of random simplicial complexes: a survey*, vol. 620 of *Contemporary Mathematics*, pp. 201–221. 2014. 14, 17

[9] O. Eisenberg, D. Krioukov, and K. Zuev *J. Phys. A*, vol. 48, 2015. 14, 17

[10] A. Patania, G. Petri, and F. Vaccarino, "The shape of collaborations," *Epj Data Science*, vol. 6, 2017. 14, 101

[11] E. N. Ciftcioglu, R. Ramanathan, and P. Basu, "Generative models for global collaboration relationships," *Scientific Reports*, vol. 7, 2017. 14, 17, 101

[12] C. J. Carstens and K. J. Horadam, "Persistent homology of collaboration networks," *Mathematical Problems in Engineering*, 2013. 14, 101

[13] G. Petri, M. Scolamiero, I. Donato, and F. Vaccarino, "Topological strata of weighted complex networks," *Plos One*, vol. 8, no. 6, 2013. 14, 15

[14] K. F. Kee, L. Sparks, D. C. Struppa, and M. Mannucci, "Social groups, social media, and higher dimensional social structures: A simplicial model of social aggregation for computational communication research," *Communication Quarterly*, vol. 61, no. 1, pp. 35–58, 2013, https://doi.org/10.1080/01463373.2012.719566. 14

[15] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, no. 3, 2003. 14, 84

[16] C. Giusti, R. Ghrist, and D. S. Bassett, "Twos company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data," *arXiv preprint arXiv: 1601. 01704*, 2016. 15, 101

[17] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino, "Homological scaffolds of brain functional networks," *Journal of the Royal Society Interface*, vol. 11, no. 101, 2014. 15, 101

[18] P. Dotko, "Topological analysis of the connectome of digital reconstructions of neural microcircuits," *arXiv preprint arXiv: 1601.01580*, 2016. 15

[19] C. Curto and V. Itskov, "Cell groups reveal structure of stimulus space," *Plos Computational Biology*, vol. 4, no. 10, 2008. 15

[20] D. Oriti, "Spacetime geometry from algebra: spin foam models for non-perturbative quantum gravity," *Reports on Progress in Physics*, vol. 64, no. 12, pp. 1703–1757, 2001. 16

[21] S. Gielen, D. Oriti, and L. Sindoni, "Cosmology from group field theory formalism for quantum gravity," *Physical Review Letters*, vol. 111, no. 3, 2013. 16

[22] J. Ambjorn, J. Jurkiewicz, and R. Loll, "Reconstructing the universe," *Physical Review D*, vol. 72, no. 6, 2005. 16

[23] J. Ambjorn, J. Jurkiewicz, and R. Loll, "Emergence of a 4d world from causal quantum gravity," *Physical Review Letters*, vol. 93, no. 13, 2004. 16

[24] M. A. Serrano, D. Krioukov, and M. Boguna, "Self-similarity of complex networks and hidden metric spaces," *Physical Review Letters*, vol. 100, no. 7, 2008. 16, 30, 36

[25] F. Papadopoulos, M. Kitsak, M. Angeles Serrano, M. Boguna, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012. 16, 30, 36

[26] G. Petri and A. Barrat, "Simplicial activity driven model," *Physical Review Letters*, vol. 121, no. 22, 2018. 17

[27] A. Costa and M. Farber *arXiv:1412.5805*, 2014. 17

[28] D. Cohen, A. Costa, M. Farber, and T. Kappeler, "Topology of random 2-complexes," *Discrete & Computational Geometry*, vol. 47, no. 1, pp. 117–149, 2012. 17

[29] G. Bianconi and C. Rahmede, "Emergent hyperbolic network geometry," *Scientific Reports*, vol. 7, 2017. 17

[30] G. Bianconi and C. Rahmede, "Emergent hyperbolic network geometry," *Scientific Reports*, vol. 7, 2017. 17

[31] G. Bianconi and C. Rahmede, "Complex quantum network manifolds in dimension $d > 2$ are scale-free," *Scientific Reports*, vol. 5, 2015. 17, 22

[32] G. Ghoshal, V. Zlatic, G. Caldarelli, and M. E. J. Newman, "Random hypergraphs and their applications," *Physical Review E*, vol. 79, no. 6, 2009. 17, 75

[33] O. T. Courtney and G. Bianconi, "Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes," *Physical Review E*, vol. 93, no. 6, 2016. 17, 24, 25, 26, 29, 59, 60, 99, 112

[34] O. T. Courtney and G. Bianconi, "Weighted growing simplicial complexes," *Physical Review E*, vol. 95, no. 6, 2017. 18, 27, 28, 29, 30, 42, 101, 133, 171

[35] O. T. Courtney and G. Bianconi, "Dense power-law networks and simplicial complexes," *Physical Review E*, vol. 97, no. 5, 2018. 18, 135, 172

[36] C. I. Del Genio, T. Gross, and K. E. Bassler, "All scale-free networks are sparse," *Physical Review Letters*, vol. 107, no. 17, 2011. 19, 136, 138

[37] M. Boguna, R. Pastor-Satorras, and A. Vespignani, "Cut-offs and finite size effects in scale-free networks," *European Physical Journal B*, vol. 38, no. 2, pp. 205–209, 2004. 25

[38] L. Lu, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems," *Physics Reports-Review Section of Physics Letters*, vol. 519, no. 1, pp. 1–49, 2012. 25, 136, 138

[39] H. Seyed-allaei, G. Bianconi, and M. Marsili, "Scale-free networks with an exponent less than two," *Physical Review E*, vol. 73, no. 4, 2006. 25, 136, 137

[40] P. Bonifazi, M. Goldin, M. A. Picardo, I. Jorquera, A. Cattani, G. Bianconi, A. Represa, Y. Ben-Ari, and R. Cossart, "Gabaergic hub neurons orchestrate synchrony in developing hippocampal networks," *Science*, vol. 326, no. 5958, pp. 1419–1424, 2009. 25, 136, 138

[41] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2004. 27, 103

[42] A. Barrat, M. Barthelemy, and A. Vespignani, "Weighted evolving networks: Coupling topology and weight dynamics," *Physical Review Letters*, vol. 92, no. 22, 2004. 27, 103

[43] G. Bianconi, "Emergence of weight-topology correlations in complex scale-free networks," *Europhysics Letters*, vol. 71, no. 6, pp. 1029–1035, 2005. 27, 28, 33, 36, 37, 38, 39, 40, 102, 103, 106, 133

[44] A. Allard, M. A. Serrano, G. Garcia-Perez, and M. Boguna, "The geometric nature of weights in real complex networks," *Nature Communications*, vol. 8, 2017. 27, 103

[45] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. 30, 33, 135, 136, 141, 143

[46] G. Bianconi and A. L. Barabasi, "Competition and multiscaling in evolving networks," *Europhysics Letters*, vol. 54, no. 4, pp. 436–442, 2001. 30, 36, 136

[47] G. Bianconi and A. L. Barabasi, "Bose-einstein condensation in complex networks," *Physical Review Letters*, vol. 86, no. 24, pp. 5632–5635, 2001. 30, 36, 136

[48] P. L. Krapivsky and S. Redner, "Organization of growing random networks," *Physical Review E*, vol. 63, no. 6, 2001. 30, 35, 36, 38, 136, 143

[49] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Pseudofractal scale-free web," *Physical Review E*, vol. 65, no. 6, 2002. 30, 36

[50] E. Ravasz and A. L. Barabasi, "Hierarchical organization in complex networks," *Physical Review E*, vol. 67, no. 2, 2003. 30, 36

[51] S. J. Wang, L. F. Xi, H. Xu, and L. H. Wang, "Scale-free and small-world properties of sierpinski networks," *Physica a-Statistical Mechanics and Its Applications*, vol. 465, pp. 690–700, 2017. 30, 36

[52] J. Kim, P. L. Krapivsky, B. Kahng, and S. Redner, "Infinite-order percolation and giant fluctuations in a protein interaction network," *Physical Review E*, vol. 66, no. 5, 2002. 30, 35

[53] P. L. Krapivsky and S. Redner, *Rate equation approach for growing networks*, vol. 625 of *Lecture Notes in Physics*, pp. 3–22. 2003. 30, 35

[54] I. Ispolatov, P. L. Krapivsky, and A. Yuryev, "Duplication-divergence model of protein interaction network," *Physical Review E*, vol. 71, no. 6, 2005. 30, 35

[55] R. Lambiotte, P. L. Krapivsky, U. Bhat, and S. Redner, "Structural transitions in densifying networks," *Physical Review Letters*, vol. 117, no. 21, 2016. 30, 35

[56] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998. 33

[57] G. Bianconi and C. Rahmede, "Network geometry with flavor: From complexity to quantum geometry," *Physical Review E*, vol. 93, no. 3, 2016. 34, 41, 42, 43, 102, 107, 116, 133

[58] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Structure of growing networks with preferential linking," *Physical Review Letters*, vol. 85, no. 21, pp. 4633–4636, 2000. 35, 36, 38, 136, 143

[59] A. Bhan, D. J. Galas, and T. G. Dewey, "A duplication growth model of gene expression networks," *Bioinformatics*, vol. 18, no. 11, pp. 1486–1493, 2002. 35

[60] *Maximum-Entropy Networks*. Springer, 2017. 45, 49, 50, 51, 52

[61] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, "An introduction to exponential random graph (p*) models for social networks," *Social Networks*, vol. 29, no. 2, pp. 173–191, 2007. 45, 49, 51

[62] T. P. Peixoto, "Entropy of stochastic blockmodel ensembles," *Physical Review E*, vol. 85, no. 5, 2012. 45, 49

[63] A. Annibale, A. C. C. Coolen, L. P. Fernandes, F. Fraternali, and J. Kleinjung, "Tailored graph ensembles as proxies or null models for real networks i: tools for quantifying structure," *Journal of Physics a-Mathematical and Theoretical*, vol. 42, no. 48, 2009. 45, 49

[64] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948, https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x. 46, 47, 56

[65] C. I. Del Genio, H. Kim, Z. Toroczkai, and K. E. Bassler, "Efficient and exact sampling of simple graphs with given arbitrary degree sequence," *Plos One*, vol. 5, no. 4, 2010. 50, 78, 81

[66] T. Snijders, "Markov chain monte carlo estimation of exponential random graph models," *Journal of Social Structure*, vol. 3, 06 2002. 51

[67] K. Anand and G. Bianconi, "Gibbs entropy of network ensembles by cavity methods," *Physical Review E*, vol. 82, no. 1, 2010. 52, 53, 54, 56, 57, 63, 76, 85, 86, 90

[68] S. Fortunato, "Community detection in graphs," *Physics Reports-Review Section of Physics Letters*, vol. 486, no. 3-5, pp. 75–174, 2010. 52

[69] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203, 2001. 52

[70] G. Bianconi, A. C. C. Coolen, and C. J. P. Vicente, "Entropies of complex networks with hierarchically constrained topologies," *Physical Review E*, vol. 78, no. 1, 2008. 53, 56, 57, 86, 89, 90

[71] K. Anand and G. Bianconi, "Entropy measures for networks: Toward an information theory of complex topologies," *Physical Review E*, vol. 80, no. 4, 2009. 54, 55, 56, 57, 63, 85, 90

[72] T. Squartini, J. de Mol, F. den Hollander, and D. Garlaschelli, "Breaking of ensemble equivalence in networks," *Physical Review Letters*, vol. 115, no. 26, 2015. 54, 56, 63, 85, 90

[73] G. Bianconi, "Statistical mechanics of multiplex networks: Entropy and overlap," *Phys. Rev. E*, vol. 87, p. 062806, Jun 2013. 57, 63, 85, 86, 90

[74] J. Park and M. E. J. Newman, "Statistical mechanics of networks," *Physical Review E*, vol. 70, no. 6, 2004. 72

[75] M. Boguna, R. Pastor-Satorras, and A. Vespignani, "Cut-offs and finite size effects in scale-free networks," *European Physical Journal B*, vol. 38, no. 2, pp. 205–209, 2004. 73

[76] V. Zlatic, G. Ghoshal, and G. Caldarelli, "Hypergraph topological quantities for tagged social networks," *Physical Review E*, vol. 80, no. 3, 2009. 75

[77] E. A. Bender and E. R. Canfield, "Asymptotic number of labeled graphs with given degree sequences," *Journal of Combinatorial Theory Series A*, vol. 24, no. 3, pp. 296–307, 1978. 63, 76, 91, 100

[78] P. Erdos and T. Gallai *Mat. Lapok*, vol. 11, pp. 477–477, 1960. 78

[79] H. Kim, C. I. Del Genio, K. E. Bassler, and Z. Toroczkai, "Constructing and sampling directed graphs with given degree sequences," *New Journal of Physics*, vol. 14, 2012. 81

[80] E. Agliari, A. Annibale, A. Barra, A. C. C. Coolen, and D. Tantari, "Immune networks: multitasking capabilities near saturation," *Journal of physics. A, Mathematical and theoretical*, vol. 46, no. 41. 84

[81] G. Bianconi, "The entropy of randomized network ensembles," *Epl*, vol. 81, no. 2, 2008. 91

[82] G. Bianconi, "Entropy of network ensembles," *Physical Review E*, vol. 79, no. 3, 2009. 92

[83] G. Bianconi, P. Pin, and M. Marsili, "Assessing the relevance of node features for network structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 28, pp. 11433–11438, 2009. 100

[84] J. Iacovacci, Z. Wu, and G. Bianconi, "Mesoscopic structures reveal the network between the layers of multiplex data sets," *Physical Review E*, vol. 92, no. 4, 2015. 100

[85] https://github.com/owencourtney/Weighted-Growing-Simplicial-Complexes. 104, 130

[86] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics-Theory and Experiment*, 2008. 109

[87] R. Albert, H. Jeong, and A. L. Barabasi, "Internet - diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999. 135

[88] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005. 135

[89] S. Redner, "How popular is your paper? an empirical study of the citation distribution," *European Physical Journal B*, vol. 4, no. 2, pp. 131–134, 1998. 135

[90] G. Timar, S. N. Dorogovtsev, and J. F. F. Mendes, "Scale-free networks with exponent one," *Physical Review E*, vol. 94, no. 2, 2016. 137

[91] J. Pitman, "Lecture notes in mathematics," *Springer*, vol. 1875, 2006. 137, 141, 146

[92] B. H. Alexander Gnedin and J. Pitman, "Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws," *Probability Surveys*, vol. 4, pp. 146–171, 2007. 137, 146

[93] B. Bassetti, M. Zarei, M. C. Lagomarsino, and G. Bianconi, "Statistical mechanics of the "chinese restaurant process": Lack of self-averaging, anomalous finite-size effects, and condensation," *Physical Review E*, vol. 80, no. 6, 2009. 137, 146

[94] M. W. Reimann, M. Nolte, M. Scolamiero, K. Turner, R. Perin, G. Chindemi, P. Dlotko, R. Levi, K. Hess, and H. Markram, "Cliques of neurons bound into cavities provide a missing link between structure and function," *Frontiers in Computational Neuroscience*, vol. 11, 2017. 140

[95] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, pp. 425–440, 1955. 141, 143

[96] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer, "Folks in folksonomies: Social link prediction from shared metadata," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, (New York, NY, USA), pp. 271–280, ACM, 2010. 171

[97] J.-G. Young, G. Petri, F. Vaccarino, and A. Patania, "Construction of and efficient sampling from the simplicial configuration model," *Physical Review E*, vol. 96, no. 3, 2017. 176