Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering

Department of Physics
Study programme: Mathematical Physics

Complex network modelling using non-Shannonian entropies

MASTER'S THESIS

Author:	Bc. David Dobáš
Supervisor:	doc. Ing. Mgr. Petr Jizba, Ph.D.
Year:	2025

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Dobáš** Jméno: **David** Osobní číslo: **494713**

Fakulta/ústav: **Fakulta jaderná a fyzikálně inženýrská**

Zadávající katedra/ústav: **Katedra fyziky**

Studijní program: **Matematická fyzika**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Modelování komplexních sítí s využitím neshannonovských entropií

Název diplomové práce anglicky:

Complex network modelling using non-Shannonian entropies

Pokyny pro vypracování:

1. Seznamte se s modelováním komplexních sítí pomocí maximalizace Shannonovy entropie, např. dle [1].
2. Seznamte se s ne-shannonovskými entropiemi (Tsallis, Rényi, Sharma-Mittal, atd.) a jejich užitím ve statistické fyzice [2], [3].
3. Modelujte sítě pomocí MaxEnt přístupu založeném na vybraných ne-shannonovských entropiích. Odvoďte analytické vlastnosti modelu a srovnejte s vlastnostmi shannonovských modelů.
4. Předveďte výsledky také numericky, například v úloze rekonstrukce sítí.

Seznam doporučené literatury:

[1] Cimini, G., Squartini, T., Saracco, F. et al. The statistical physics of real-world networks. Nat Rev Phys 1, 58–71 (2019).
[2] Jizba, P. and Korbel, J. Maximum Entropy Principle in Statistical Inference: Case for Non-Shannonian Entropies, Phys. Rev. Lett. 122 (2019) 120601.
[3] Somazzi, A. and Garlaschelli, D. Learn your entropy from informative data: an axiom ensuring the consistent identification of generalized entropies: arXiv:2301.05660.
[4] Bianconi, G. Multilayer Networks: Structure and Function, Oxford University Press, Oxford, 2018.

Jméno a pracoviště vedoucí(ho) diplomové práce:

doc. Ing. Mgr. Petr Jizba, Ph.D. katedra fyziky FJFI

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **30.10.2024** Termín odevzdání diplomové práce: **09.05.2025**

Platnost zadání diplomové práce: **30.10.2026**

doc. Ing. Mgr. Petr Jizba, Ph.D.
podpis vedoucí(ho) práce

podpis vedoucí(ho) ústavu/katedry

doc. Ing. Václav Čuba, Ph.D.
podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

9. 12. 2024
Datum převzetí zadání

Podpis studenta

Declaration

I hereby declare, that I wrote this research project on my own and using the cited resources only.

I agree with the usage of this thesis in the purport of the §60 Act 121/2000 (Copyright Act).

Prague ....
Bc. David Dobáš

Acknowledgement

I gratefully acknowledge the support provided by the SGS ČVUT grants (SGS25/163/OHK4/3T/14 and SGS22/178/OHK4/3T/14), which were instrumental in making this research possible. I also extend my heartfelt thanks to my supervisor, doc. Petr Jizba, whose continuous support, scientific inspiration, and passion for complex systems have significantly shaped my academic journey.

Bc. David Dobáš

Název práce:
Modelování komplexních sítí s využitím neshannonovských entropií

Autor: Bc. David Dobáš

Studijní program: Aplikace přírodních věd
Obor: Matematická fyzika
Druh práce: Diplomová práce

Vedoucí práce: doc. Ing. Mgr. Petr Jizba, Ph.D.
Katedra fyziky, Fakulta jaderná a fyzikálně inženýrská

Abstrakt: TODO

Klíčová slova: TODO

Title:
Complex network modelling using non-Shannonian entropies

Author: Bc. David Dobáš

Abstract: TODO

Key words: TODO

Contents

Introduction

The emergence of complex network analysis over the past few decades has revealed profound patterns in systems ranging from biological interactions to social networks and technological infrastructures. Understanding these systems requires moving beyond traditional graph theory, with models based on maximizing entropy play a fundamental role. This thesis explores the application of non-extensive entropy frameworks to complex network modeling, with a particular focus on Tsallis entropy.

Complex networks differ fundamentally from regular graphs in their non-uniform topology and emergent properties like scale-free degree distributions, high clustering, and rich hierarchical structures. While Shannon entropy has been successfully applied to network modeling through exponential random graphs, it definitely does not constitute the only possible choice for network modelling.

Tsallis entropy, a generalization of Shannon entropy introduced in 1988, provides a promising framework for modeling systems with sub-exponential phase space growth, where standard statistical mechanics approaches may not apply. By incorporating a non-extensivity parameter q, Tsallis entropy can account for long-range correlations and power-law distributions commonly observed in complex networks.

The thesis is organized as follows: first, we briefly introduce complex networks and several network models. Then, we will talk about entropy, its importance in physical systems, information theory and statistical inference. We will introduce Shannon entropy and then extend it to the case of so called non-extensive entropies, with a focus on Tsallis entropy. Finally, we will apply the Tsallis entropy to define a statistical model on networks.

Chapter 1

Introduction to complex networks

This chapter introduces complex networks analysis by first defining the fundamentals of graph theory and key topological properties. We then distinguish complex networks from standard graphs, highlighting the emergent properties commonly found in real-world networks. The chapter concludes with a review of established random graph models. The main resources for this chapter are [1, 2].

1.1 Graphs and their properties

A **graph** (network) $\mathcal{G}$ is a set of **vertices** (nodes) V and a set of **edges** (links) E connecting the vertices (i.e. set of tuples of vertices). In the whole thesis, we consider labeled graphs, meaning that the vertices are distinguishable and we can explicitly label them $i = 1, \dots, N_V$ where N_V is the number of vertices. Then for every graph, we can define the **adjacency matrix** $\mathbb{A}$, whose entries are

$$a_{ij} = \begin{cases} 1 \text{ if there is an edge from node } i \text{ to } j\text{, } (i, j) \in E \\ 0 \text{ otherwise} \end{cases}$$

There is a one-to-one correspondence between labeled graphs and their adjacency matrices.

Graphs can be **directed** or **undirected**. In the undirected case, $a_{ij} = a_{ji}$, i.e., the adjacency matrix is symmetric. Directed graphs do not have such a constraint. Also, one can either allow or disallow **self-loops**, edges from a node to itself ($a_{ii} = 1$). A graph can be **weighted**, which means each edge present in the graph can be assigned a weight, a real number w_{ij}.

1.1.1 Network properties

The topological structure of a network can be characterized through various quantitative measures that reveal important structural patterns and organizational principles. The following properties, while not exhaustive, provide a mathematical foundation for analyzing network topology and comparing different network structures.

Number of links, link density

The simplest network property is the number of links. For an undirected graph with no self-loops, one can compute it using the adjacency matrix as

$$L(G) = \frac{1}{2} \sum_{i,j=1, i \neq j}^{N} a_{ij}$$

For a directed graph with self-loops, it is given by

$$L(G) = \sum_{i,j=1}^{N} a_{ij}$$

The **link density** is then the number of links divided by the total number of possible links $L_{max}(N_V)$, which depends on the number of vertices N_V and the type of graph. For undirected graphs with no self-loops, the total number of possible links is $\frac{N_V(N_V-1)}{2}$, for directed graphs with self-loops it is N_V^2, and for directed graphs without self-loops it is $N_V(N_V-1)$. Then one can write the link density as

$$c(G) = \frac{L(G)}{L_{max}(N_V)}$$

Degrees

Degree of a node corresponds to the number of its neighbors, i.e., the number of edges adjacent to it. In an undirected graph, there is only the total degree of each node,

$$k_i(G) = \sum_{i=1}^{N_V} a_{ij}$$

In a directed graph, we need to distinguish between **in-degree** and **out-degree**. Using the adjacency matrix, they can be written in a simple way,

$$k_i^{in}(G) = \sum_{i=1}^{N_V} a_{ji}$$

$$k_i^{out}(G) = \sum_{i=1}^{N_V} a_{ij}$$

Average nearest neighbor degree

Now we move on to the second-order properties. For these properties, we do not only take into account the neighbors of a node, but also neighbors of neighbors. The first example of such a property is the **average nearest neighbor degree** (ANND).

For undirected graphs, the average nearest neighbor degree of node i is

$$k_i^{nn} = \frac{\sum_{j \neq i} a_{ij} k_j}{\sum_{j \neq i} k_i}$$

In case of a directed graph, we can define the average out-degree of out-neighbors of node i as

$$k_i^{nn,out} = \frac{\sum_{j \neq i} a_{ij} k_j^{out}}{k_i^{out}} = \frac{\sum_{j \neq i} \sum_{k \neq j} a_{ij} a_{jk}}{\sum_{j \neq i} a_{ij}}$$

The average in-degree of in-neighbors of node i is

$$k_i^{nn,in} = \frac{\sum_{j \neq i} \sum_{k \neq j} a_{ji} a_{kj}}{\sum_{j \neq i} a_{ji}}$$

We could also define the average out-degree of in-neighbors and vice versa, but we will not use these.

Local clustering coefficient

The local clustering coefficient is defined for undirected graphs only. For directed graphs, we can define the local clustering coefficient on an undirected projection of our graph, whose adjacency matrix can be obtained as

$$b_{ij} = \max(a_{ij}, a_{ji})$$

The idea of the local clustering coefficient is to look at all neighbors of a selected node and find out how many of all possible links between these neighbors are realized. If neighbors of a node are highly connected among themselves, we assume such a node is in a group of highly connected nodes, which we call a cluster. Mathematically, the local clustering coefficient of the i-th node can be expressed as

$$C_i = \frac{\sum_{j \neq i} \sum_{k \neq i,j} b_{ij} b_{jk} b_{ki}}{\sum_{j \neq i} \sum_{k \neq i,j} b_{ij} b_{ki}}$$

1.2 Complex networks, real-world networks

What do we mean by complex networks? Typically, networks represent structures observed in the real world, while graphs serve as their mathematical representations. Complex networks are characterized by the absence of a uniform topology. It is often easier to define them by what they are not: for instance, a regular lattice would not be considered a complex network.

What distinguishes graph theory from the theory of complex networks? Newman et al. [3] argue that while graph theory has contributed many significant mathematical results, it has primarily focused on combinatorial properties of graphs rather than on the structure of networks that occur naturally. The advent of the internet and the availability of large datasets enabled empirical studies of network properties, revealing that real-world network structures cannot be adequately described by lattices or simple random graphs. Consequently, network science has emerged as a field dedicated to thoroughly investigating these empirical properties and developing models that can accurately reproduce them.

In the following, we will introduce few (yet definitely not all) important properties of real-world networks.

Sparsity

When real-world networks are compared, one finds that the number of links is usually not proportional to the number of all possible links, but rather to the number of nodes, i.e. $L(N_V) \in \mathcal{O}(N_V)$. One says that such networks are **sparse**. This also means that the average degree of a nodes does not depend on the number of nodes, $k(N_V) \in \mathcal{O}(1)$, and that the link density is inversely proportional to the number of nodes, $c(N_V) \in \mathcal{O}(1/N_V)$. Most real-world networks are sparse [1].

Power-law degree distribution

The degree distribution $P(k)$ represents the fraction of nodes in a graph with degree k. Empirical studies [4] have revealed that numerous real-world networks, including the internet, World Wide Web, and protein interaction networks, follow a power-law degree distribution $P(k) \sim k^{-\gamma}$ with exponent γ typically ranging between 2 and 3. Networks exhibiting this property are commonly referred to as **scale-free**.

This distribution has profound implications for network structure: a majority of nodes possess relatively few connections, while a small minority of nodes—often called hubs—maintain an extraordinarily high number of connections. Such structural organization significantly impacts dynamic processes on networks. For instance, in epidemic spreading, these highly-connected hubs can dramatically accelerate transmission rates compared to predictions from standard compartmental models like SIR, which typically assume homogeneous mixing without accounting for the underlying network topology [1].

Assortativity

One property of interest in real-world networks can be their **assortativity**. Assortativity means the tendency of nodes with certain features to be neighbors with nodes of similar features. For this we can use the average nearest neighbor degree (ANND) as defined in 1.1.1. Specifically, we can study the average ANND for nodes with degree k, denoted as $\overline{k}_{nn}(k)$. If $\overline{k}_{nn}(k)$ increases with k, meaning that nodes with high degree tend to connect to other nodes with high degree, we say that the network is assortative, and if it decreases, we say that the network is disassortative.

An interesting result is that social networks are usually assortative, while technological networks are disassortative [4]. Also, several networks exhibit a power-law behavior $\overline{k}_{nn}(k) \sim k^{\beta}$, for example the internet has been shown to have $\beta \approx -0.5$ [5].

Clustering

Similarly to the average nearest neighbor degree, we can study the average clustering coefficient, either for all nodes denoted by C, or depending on the node degree, $\bar{C}(k)$. There have been two significant observations. First, the average clustering coefficient of nodes with degree k behave as $\bar{C}(k) \sim k^{-\omega}$. For the World trade web, it was found that $\omega \approx 0.7$ [6], for the internet or network of English synonyms, ω was found to be approximately 1 [7]. This behavior has been called **hierarchical clustering**. Nodes with small degrees tend to be part of highly interconnected clusters, while the neighbors of hubs with high degree are usually less densely connected. Ravasz and Barabási [7] argue that this is the mechanism which allows for both high average clustering and power-law degree distributions.

Second important observation is that total average clustering does not depend on number of vertices N_V, $C(N_V) \in \mathcal{O}(1)$ [8]. This is interesting especially when we consider that the link density $c(N_V)$ is inversely proportional to N_V. So even though the density of links goes to zero, the clustering remains high. We say that real-world networks are **highly clustered**.

1.3 Random network models

The main aim of network modelling is to find such models which can reproduce and possibly explain properties of real-world networks. One possible approach is to use models which are not deterministic, but rather stochastic. Two approaches are usually taken [9]:

- **Microscopic**: In these models, some microscopic network formation mechanism is identified and used to reproduce properties of real-world systems. Such models include, for example, the Barabási-Albert preferential attachment model and the Watts-Strogatz small-world model.
- **Macroscopic**: Here, macroscopic properties of networks of interest are identified, and based on these properties, one finds models which reproduce them on average, but otherwise stay maximally random. This approach is similar to that of statistical mechanics. The framework used here is called Exponential Random Graphs.

Let us first introduce several model, which are based on the microscopic idea. Later on, we will be thoroughly discussing the macroscopic approach, which will be based on maximizing entropy.

1.3.1 Erdős-Rényi model

The Erdős-Rényi model [10] is the simplest case of a random graph model. Let us have n vertices. Let us fix a probability $0 \leq p \leq 1$. Then edges are drawn one by one independently and the probability of edge between node i and j is simply

$$p_{ij} = p \tag{1.1}$$

This model is usually not suitable for reproducing properties of real-world networks. Some examples to mention are:

- Degree distribution: Edges are drawn independently and each of them is a Bernoulli variable, therefore the degree of a node is a binomial random variable, which in the large N limit with pN bounded converges to a Poisson distribution, which decreases faster than exponentially. On the other hand, real-world networks usually have fat-tailed, power-law degree distributions.
- Average nearest neighbor degree: The expected average number of neighbors of each node in the Erdős-Rényi model is $p(N-1)$. Therefore, even the average nearest neighbor degree of each node is $p(N-1)$, a constant independent of the degree of the node. However, real-world networks are usually assortative or disassortative.
- Local clustering coefficient: The expected local clustering coefficient in the Erdős-Rényi model is the same for each node, $C_i = p$. In real-world networks, however, the local clustering coefficient is usually a decreasing function of the node degree. Also, if we assume sparsity, then p needs to behave as $p \sim 1/N$, and therefore even the average clustering coefficient is $C \in \mathcal{O}(1/N)$. However, we saw that real-world networks are highly clustered, $C(N) \in \mathcal{O}(1)$.

1.3.2 Barabási-Albert preferential attachment model

Since the Erdős-Rényi model does not yield a power-law degree distribution, Barabási and Albert [11] were interested in finding such a network formation mechanism, which could reproduce this property. They came up with two important observations:

- Growth: Real-world networks are usually not static, they grow with time.
- Preferential attachment: When a new node emerges, it is more likely to connect to nodes that already have many connections.

With these observations, they developed the following model [11]. They start with a few vertices, for example, two vertices connected with an edge. Then, in each step, they add one node. This node then connects to the already existing nodes with a probability proportional to their degree,

$$p(i) = \frac{k_i}{\sum_{j=1}^{N} k_j} \tag{1.2}$$

They were then able to show that such a model yields a power-law degree distribution with an exponent $\gamma = 3$. Generalizations allowing for a tunable scaling exponent were also found [12].

Even though the Barabási-Albert model is able to reproduce a power-law degree distribution, it has been shown, that the clustering behaves as $C(N) \sim N^{-0.75}$ [8], meaning that the model is not able to reproduce high clustering for large N.

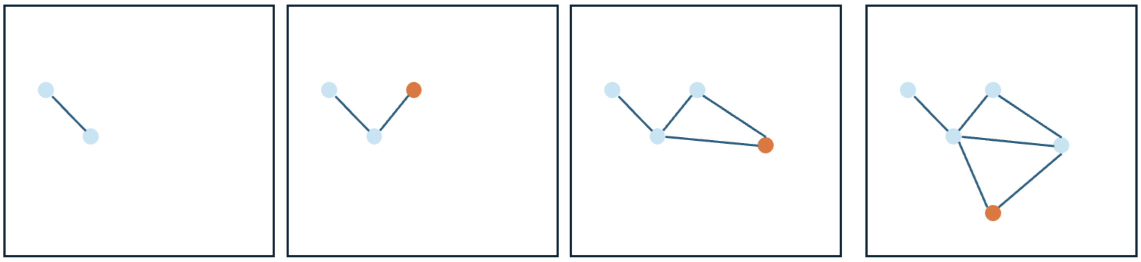


Figure 1.1: Illustration of the Barabási-Albert model. Each step adds one node and connects it to other nodes with a probability proportional to their degree. In the figure, we can see that a new node tends to connect to nodes with an already high degree.

1.3.3 Configuration model

The configuration model starts with a prescribed degree sequence k_i and tries to find graphs compatible with such that degree sequence. The simplest approach is to start with a desired number of vertices N and assign to each node a number of half-edges that corresponds to its prescribed degree. These half-edges can be connected together to form full edges. One way to do that is to connect them uniformly randomly. That gives a random network model, which yields graphs with a degree sequence precisely corresponding to the prescribed one.

However, this version of the configuration model cannot avoid self-loops. Therefore, graphs generated by this model can usually hardly be compared with real-world networks. An approach avoiding this issue is called link stub reconnection. It starts with a real-world network without self-loops and reconnects its edges in such a manner that the degree sequence is preserved. However, this method was shown to explore the space of all allowed graphs non-uniformly, staying near the original graph. This can be overcome by introducing an acceptance-probability [13], but that makes the whole problem more computationally demanding.

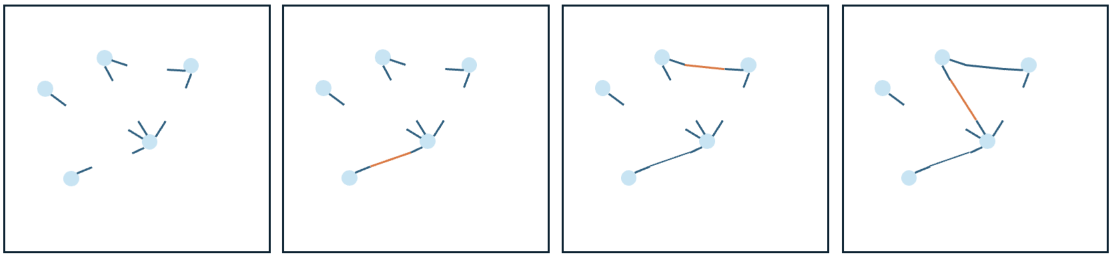


Figure 1.2: Illustration of the configuration model. Each node possesses the so-called half-stubs or half-edges. The number of half-stubs corresponds to the desired degree. Then these half stubs are connected uniformly randomly to form full edges.

1.3.4 Park-Newman model

The configuration model has several limitations because it requires the degree sequence to be followed exactly in each realization. This either produces graphs with self-loops and multiple edges, or makes uniform sampling difficult. Park and Newman [14] proposed a different approach. They assign each node a parameter x_i which they call fugacity (although later literature prefers the term fitness [15]). In their model, edges are independent random variables, and the probability of an edge between nodes i and j has the form:

$$p_{ij} = p(x_i, x_j)$$

They require that all graphs with the same degree sequence have equal probability. The parameters x_i then allow tuning of each node's degree, reproducing desired degree distributions on average rather than as a hard constraint in each graph. This model samples each edge only once and does not allow self-loops, thereby overcoming the limitations of the configuration model.

Under these assumptions, they demonstrated that the only possible link probability takes the form:

$$p_{ij} = \frac{g(x_i)g(x_j)}{1 + g(x_i)g(x_j)}$$

for some function g. Since they placed no specific requirements on the fitness parameters x_i, they argued that any reparametrization could be chosen, leading them to select the simple option $g(x_i) = \beta^{1/2}x_i$. The resulting model is therefore:

$$p_{ij} = \frac{\beta x_i x_j}{1 + \beta x_i x_j} \tag{1.3}$$

The parameter β is common to all nodes and determines the overall network density. The individual fitness parameters x_i indicate each node's tendency to connect with others, with higher x_i corresponding to higher connection probability. To ensure that p_{ij} is a valid probability, x_i must be positive for all i.

Importantly, they showed that sampling fitnesses from a power-law distribution $P(x) \sim x^{-\alpha}$ produces a power-law degree distribution, while also explaining the disassortativity observed in networks such as the internet.

Chapter 2

Shannon entropy

Shannon entropy serves as a fundamental measure of information, uncertainty, and disorder across multiple disciplines. In physics, it provides the mathematical foundation for statistical mechanics and thermodynamics, establishing the microscopic basis for macroscopic phenomena. In information theory, it quantifies the theoretical limits of data compression and transmission, while in statistical inference, it offers a principled approach to model selection and parameter estimation under constraints. In this chapter, we aim for a brief yet comprehensive introduction to these concepts. For the general framework we follow mainly [16], while using other sources for further clarifications and details.

2.1 Shannon entropy in physics

From thermodynamics, we know that entropy needs to satisfy the following conditions:

1. $dS = \frac{\delta Q}{T}$ is an exact differential, i.e., its integral does not depend on the path in phasespace, if it is given by a reversible process. δQ is the heat added to the system.

2. It is additive, meaning that for two systems A, B, the entropy of the composed system $A \cup B$ is $S_{A\cup B} = S_A + S_B$.

3. In a closed (isolated) system, the entropy is non-decreasing, $dS \geq 0$.

Statistical mechanics aims to explain these macroscopic thermodynamic properties from a microscopic perspective, introducing the notions of *microstates* and *macrostates*. A microstate is a single configuration of the system at the microscopic level, for example, a specific arrangement of positions and momenta of all particles. A macrostate, on the other hand, is a description of the system in terms of macroscopic observables, such as temperature, pressure, volume, etc. Multiple microstates can correspond to the same macrostate. If all microstates are assumed to have equal probability, then the probability of a macrostate is proportional to the number of microstates corresponding to it, which is called multiplicity. As we will see, the entropy is the logarithm of this multiplicity multiplied by an appropriate constant.

Specifically, let us define the entropy in statistical mechanics at equilibrium as

$$\sigma = \ln \Delta\Gamma \tag{2.1}$$

where $\Delta\Gamma$ is the phasespace volume accessible to the system for a given energy range [17]. The logarithm appears precisely because of the additivity requirement. Assuming that we have two systems A, B with phasespace volumes $\Delta\Gamma_A$, $\Delta\Gamma_B$, and further assuming that the phasespace volume of the composite system is $\Delta\Gamma_{A\cup B} = \Delta\Gamma_A \Delta\Gamma_B$, we get

$$\sigma_{A\cup B} = \ln \Delta\Gamma_{A\cup B} = \ln(\Delta\Gamma_A \Delta\Gamma_B) = \sigma_A + \sigma_B \tag{2.2}$$

One can also show, that the entropy defined in eq. 2.1 satisfies the conditions to be exact differential and non-decreasing [17]. The value of σ achieves its maximum at equilibrium and is then defined by the internal energy U, number of particles N_i of different kinds, and external variables x_ν (e.g. volume, magnetic field, etc.). One can then write

$$\sigma = \sigma(U, N_i, x_\nu) \tag{2.3}$$

Assuming that we have two systems A, B in thermal equilibrium, such that they can only exchange energy, we can from additivity write

$$\sigma_{A\cup B} = \sigma_A + \sigma_B \tag{2.4}$$

If only energy is exchanged, we can write

$$d\sigma = d\sigma_A + d\sigma_B = \frac{\partial \sigma_A}{\partial U_A} dU_A + \frac{\partial \sigma_B}{\partial U_B} dU_B = 0 \tag{2.5}$$

Then, using the fact that total energy is conserved, $dU_A = -dU_B$, we get

$$\frac{\partial \sigma_A}{\partial U_A} = \frac{\partial \sigma_B}{\partial U_B} \tag{2.6}$$

which allows us to define temperature as

$$\frac{1}{\tau} = \frac{\partial \sigma}{\partial U} \tag{2.7}$$

and see that in equilibrium, the temperature is the same for both systems. One can, for example in the case of ideal gas, find, that the relation with absolute temperature T is given by [17]

$$T = k_B \tau \tag{2.8}$$

where k_B is the Boltzmann constant.

Now, assuming that volume V of our system is variable as well, we can write

$$d\sigma = \frac{\partial \sigma}{\partial U} dU + \frac{\partial \sigma}{\partial V} dV = \frac{1}{\tau} dU + \frac{1}{\tau} \frac{\partial U}{\partial V} dV \tag{2.9}$$

Identifying $-\frac{\partial U}{\partial V}$ as pressure p, we get

$$dU = \tau d\sigma - p dV \tag{2.10}$$

and therefore $\tau d\sigma = \delta Q$ is the heat added to the system. Since from thermodynamics we know that $\delta Q = T dS$ and we had $T = k_B \tau$, we can finally find the connection between the statistical mechanics entropy σ and the thermodynamic entropy S

$$S = k_B \sigma \tag{2.11}$$

2.1.1 Additivity vs. extensivity

Let us turn our focus on the additivity requirement once again. From $dU = TdS - pdV$, we see that since U and V are extensive (where by extensive we mean that they scale with the size of the system), while T and p are intensive, the entropy has to be extensive as well. Otherwise, if we enlarge the system and the term TdS does not scale accordingly, its contribution would vanish for large systems, which is not the case in physics.

Should the entropy then be additive or extensive? To avoid confusion, we need to define the terms *additivity* and *extensivity* properly. Let us suppose we have two separated systems A, B with finite sizes of discrete phasespaces W_A, W_B. Also let us denote by $W_{A\cup B}$ the size of the phasespace of the composite system $A \cup B$. A quantity X is called *extensive* if

$$X(W_{A\cup B}) = X(W_A) + X(W_B) \tag{2.12}$$

A quantity X is called *additive* if

$$X(W_A W_B) = X(W_A) + X(W_B) \tag{2.13}$$

For systems, where $W_{A\cup B} = W_A W_B$, the two definitions coincide, and we do not need to distinguish between them. The equation $W_{A\cup B} = W_A W_B$ means, that the phasespace of the system grows exponentially with the size of the system. By size of the system, we usually mean the number of particles, or further indivisible elements, which can be in different states.

Let us assume a microcanonical situation, where the system of N particles is perfectly isolated and has fixed energy E. The accessible phasespace is therefore only such states. All the states have the same probability, which is inversely proprotional to the multiplicity of such state, which we denote $W(N)$. Then the entropy is given by $S(N) = k_B \ln W(N)$. Supposing, that the multiplicity grows exponentially with the size of the system $W(N) \propto e^N$, we obtain $S(N) \propto N$, i.e. indeed the entropy scales with the system size.

However, the assumption $W_{A\cup B} = W_A W_B$ is not at all obvious for all systems. As is argued in [16], many complex systems actually follow a slower phasespace growth, $W_{A\cup B} < W_A W_B$. In this case, the entropy of the form $S(N) = k_B \ln \Delta\Gamma$ is not extensive and this will be one of the motivations to introduce generalized entropies.

2.1.2 Canonical ensemble

Microcanonical ensembles are difficult to work with, since the multiplicity is often hard to compute. We can instead use the canonical ensemble, introduced by Gibbs. In canonical ensemble, we have a system in thermal contact with a heat reservoir, with which it can exchange energy. For such system, one can find [17] that the probability distribution of energy is given by the famous Boltzmann distribution

$$P(E_i) \propto e^{-\frac{E_i}{k_B T}} \tag{2.14}$$

and using this knowledge, we can find that the entropy can be expressed as

$$S = -k_B \sum_i P(E_i) \ln P(E_i), \tag{2.15}$$

or using equivalent integral formulation in the case of continuous energy spectrum. This is the famous Boltzmann-Gibbs-Shannon formula for entropy. The other way around, one can first assume, that the entropy needs to be maximized, but since the energy of the system is not fixed anymore, we can only assume we observe some average energy U. Then, maximizing the corresponding Lagrange function

$$\mathcal{L} = -k_B \sum_i P(E_i) \ln P(E_i) - \beta \left(\sum_i P(E_i) E_i - U \right) - \alpha \left(\sum_i P(E_i) - 1 \right) \quad (2.16)$$

one can recover the Boltzmann distribution as the solution of the maximization problem. We will discuss this approach in more detail in section 2.3.

2.2 Shannon entropy in information theory

Even though we are not concerned with information theory in this thesis, it is still important to remind some of the important notions and results, which will be useful later for generalization of entropy. We will follow [18] and [19] here.

Let us consider a simple example. How many yes/no questions do we need to identify an integer between 0 and 63? There could be many possible ways, but we can imagine, that the optimal strategy is to first ask if the number is smaller than 32, then if it is smaller than 16, then 8, 4, 2, and finally 1. This way, we can identify the number with at most 6 questions. And this approach is equivalent to finding a binary code for the number. Therefore, the random number “contains” 6 bits of information.

To formalize the problem, let us define an **ensemble** X as a triple $(x, \mathcal{A}_X, P_X)$, where x is an outcome of a random variable, which comes from a discrete set $\mathcal{A}_X = \{a_1, \dots, a_n\}$ with probabilities $P_X = \{p_1, \dots, p_n\}$ s.t. $p(x = a_i) = p_i$ and $\sum_{i=1}^n p_i = 1$. One of the results of information theory is showing, that

- $h(x = a_i) = \log_2 \frac{1}{p_i}$ is a sensible measure of information content of the outcome $x = a_i$.
- $H(X) = \sum_{i=1}^n p_i h(x = a_i) = -\sum_{i=1}^n p_i \log_2 p_i$, is a sensible measure of average information content of the ensemble X.

The importance of Shannon entropy $H(X)$ as a measure of information is expressed by the *Shannon's source coding theorem*, which states

Theorem 2.2.1 (Shannon's source coding theorem - verbal version)**.** *N i.i.d. random variables $X_1, \dots, X_N$, each with entropy $H(X)$ can be compressed into $NH(X)$ bits with negligible error as $N \to \infty$. On the other hand, compression below $NH(X)$ will almost certainly lead to a loss of information.*

Shannon's source coding theorem

Let us delve a bit deeper into the meaning of these concepts. Having a source of information, we can say, that the information content is essentialy the number of bits needed to encode the information. If the source produces symbols and we show that we need L bits per source symbol on average to encode and decode the data reliably, we say that the source has average information content at most L bits per symbol.

If we want to be always guaranteed to identify an outcome from the ensemble, we need at least $H_0(X) = \log_2 |\mathcal{A}_X|$ bits to encode the outcome, where $|\mathcal{A}_X|$ is the number of all possible outcomes (the size of the alphabet). This is clear, because if we want to assign a unique code to each outcome, we need at least $|\mathcal{A}_X|$ different codes, and to encode the outcome in binary, we need at least $\log_2 |\mathcal{A}_X|$ bits.

What if, however, we say we are willing to accept the risk, that in small fraction δ of cases we will not be able to identify the outcome? Then, the situation becomes much more interesting. We can then only encode those symbols, which taken together appear with probability greater than $1 - \delta$. To encode the least symbols possible, we should encode the so-called **δ-sufficient subset** $S_\delta(X)$, which is the smallest subset of $\mathcal{A}_X$, such that $P(x \in S_\delta(X)) \geq 1 - \delta$. The **$\delta$-essential bit content** is then defined as $H_\delta(X) = \log_2 |S_\delta(X)|$. Essential bit content therefore computes, how many bits we need to encode all outcomes from the δ-sufficient subset.

Now, instead of single realizations, let us consider sequences $\mathbf{x} = (x_1, \ldots, x_N)$ of N i.i.d. realizations from the ensemble X. We denote the ensemble of sequences by $X^N = (X_1, \ldots, X_N)$ and the set of all possible sequences by $\mathcal{A}_X^N$. Number of bits needed to encode such sequences with error less than δ is $H_\delta(X^N)$. On the other hand, the Shannon entropy of the ensemble X^N is $H(X^N) = NH(X)$, since the entries are independent. The Shannon's source coding theorem then finds a relation between these two quantities.

Theorem 2.2.2 (Shannon's source coding theorem - formal version)**.** Let X be an ensemble with entropy $H(X) = H$. Then for any $\epsilon > 0$ and $0 < \delta < 1$, there exists $N_0(\epsilon, \delta)$ such that for all $N \geq N_0(\epsilon, \delta)$,

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon \tag{2.17}$$

The first part of the theorem, $\frac{1}{N} H_\delta(X^N) < H + \epsilon$, means, that for a sequence of length N, even if we allow only a small rate of error δ, we will only need $H + \epsilon$ bits per symbol to encode such sequences. On the other hand, the second part, $\frac{1}{N} H_\delta(X^N) > H - \epsilon$, means, that even if we allow large error δ, we will still need at least $H - \epsilon$ bits per symbol.

This remarkable result would be hard to prove using only the δ-sufficient subsets, since they are hard to enumerate. Instead, one first defines a so-called *typical set* $T_{N\beta}$, which is easier to "count" and then relates it to the δ-sufficient subsets.

Using the law of large numbers, one can prove the following theorem [19].

Theorem 2.2.3 (Asymptotic equipartition property)**.** If $X_1, \ldots, X_N$ are i.i.d. random variables with entropy $H(X)$, then for any $\epsilon > 0$ and $0 < \delta < 1$, there exists $N_0(\epsilon, \delta)$ such that for all $N \geq N_0(\epsilon, \delta)$,

$$\Pr \left\{ \left| -\frac{1}{N} \log_2 P(X_1, \ldots, X_N) - H(X) \right| < \epsilon \right\} \geq 1 - \delta \tag{2.18}$$

i.e. $-\frac{1}{N} \log_2 P(X_1, \ldots, X_N)$ converges to $H(X)$ in probability.

This motivates the definition of the typical set $T_{N\beta}$ as the set of all sequences $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{A}_X^N$ for which

$$\left| -\frac{1}{N} \log_2 P(x_1, \ldots, x_N) - H(X) \right| < \beta \tag{2.19}$$

or equivalently

$$2^{-N(H(X)-\beta)} \leq P(x_1, \dots, x_N) \leq 2^{-N(H(X)+\beta)} \tag{2.20}$$

One can then show, that

$$\Pr\{\mathbf{x} \in T_{N\beta}\} \geq 1 - \frac{\sigma^2}{\beta^2 N} \tag{2.21}$$

where $\sigma = \operatorname{var}(\log_2 1/P(x_n))$. Equation 2.21 shows, that for any fixed $\beta > 0$, the probability of the sequence being in the typical set converges to 1 as $N \to \infty$. This is sometimes called the *asymptotic equipartition principle* [18]. One can also show $|T_{N\beta}| < 2^{N(H(X)+\beta)}$ for large N, i.e. one needs at most $N(H(X)+\beta)$ bits to encode the sequence from the typical set. This is already similar to what we saw in the case of the Shannon's source coding theorem. Essentially, showing that the typical set and δ-sufficient are similar enough, one can finally prove the Shannon's source coding theorem 2.2.2. For details, we refer to [18].

The case of ergodic Markov chains

By a **discrete random process**, we mean a sequence of random variables $X := (X_t)_{t\in\mathbb{N}}$, where each X_i comes from a discrete set $\mathcal{A}_X$. Let us consider the process is **Markov**, meaning that for all $n > 1$ and all $a_{i_1}, \dots, a_{i_n} \in \mathcal{A}_X$,

$$P(X_n = a_{i_n} | X_1 = a_{i_1}, \dots, X_{n-1} = a_{i_{n-1}}) = P(X_n = a_{i_n} | X_{i_{n-1}} = a_{i_{n-1}}) \tag{2.22}$$

i.e. the probability of the next state depends only on the previous state and does not depend on the past. We assume that the process is **time-homogenous**, meaning that $P(X_{n+1} = a_j | X_n = a_i) = P(X_1 = a_j | X_0 = a_i)$ for all n. The process is then fully characterized by the transition probabilities $P(X_1 = a_j | X_0 = a_i)$ which we write shortly as $p(j|i)$ (for further details on the definitions, see [20]).

Hanel et al. [16] (based on the original treatment by Khinchin [21]) argue, that the Shannon entropy is a sensible measure of information not only for the ensembles of i.i.d. random variables, but also for those discrete random processes, which are Markov and ergodic. Markov process is **ergodic**, if it has a unique **stationary distribution**, i.e. such a distribution that solves the equation

$$p_j = \sum_{i\in\mathcal{A}_X} p_i p(j|i) \tag{2.23}$$

and if for each symbol $a_i \in \mathcal{A}_X$, its relative frequency of occurence $p_{i,rel}(N) = \frac{1}{N}\sum_{n=1}^{N} \delta_{a_i}(X_n)$ converges to $p(a_i)$ as $N \to \infty$, i.e. for $\epsilon > 0$ and $\delta > 0$ arbitrarily small, there exists $N_0(\epsilon, \delta)$ such that for all $N \geq N_0(\epsilon, \delta)$,

$$\Pr\{|p_{i,rel}(N) - p_i| > \delta\} < \epsilon \tag{2.24}$$

for any $\delta > 0$. The meaning is, that to observe the stationary distribution, we do not need to run several realizations of the process, but it suffices to run one and wait long enough. This is essentially the same as the ergodic hypothesis in physics, where we assume that we can exchange the time average with the ensemble average. Mathematically, the definition means, that the process observes the law of large numbers.

Suppose the Markov chain is in state a_i at time $t = 0$. Then we have the probabilities of transitions to all other states $p(j|i)$ for all $j \in 1, \dots n$. Khinchin [21] then defines the one-step entropy in the state a_i as

$$h(a_i) := -\sum_{j=1}^{n} p(j|i) \log_2 p(j|i) \tag{2.25}$$

and the average entropy of the Markov process X as

$$H(X) := \sum_{i \in \mathcal{A}_X} p_i h(a_i) \tag{2.26}$$

where p_i are the probabilities of the states in the stationary distribution.

Now assume that the ergodic Markov process X produced a sequence $C = (a_{i_1}, \dots, a_{i_N})$. The probability of observing such a sequence is given by

$$P(C) = p_{i_1} p(i_2|i_1) \dots p(i_N|i_{N-1}) \tag{2.27}$$

Khinchin then shows [21], that the asymptotic equipartition property holds even for the one-step entropy, i.e. for $\epsilon > 0$ and $\delta > 0$ arbitrarily small, there exists $N_0(\epsilon, \delta)$ such that for all $N \geq N_0(\epsilon, \delta)$,

$$\Pr\left\{\left|-\frac{1}{N}\log_2 P(C) - H(X)\right| < \epsilon\right\} \geq 1 - \delta \tag{2.28}$$

Shannon-Khinchin axioms

Before showing, how the entropy can be defined axiomatically, let us first clarify the notation, when we deal with two systems. Assume we have systems A with elementary states $\{a_1, \dots, a_n\}$ and probabilities $P_A = (p_{a_1}, \dots, p_{a_n})$ and system B with elementary states $\{b_1, \dots, b_m\}$ and probabilities $P_B = (p_{b_1}, \dots, p_{b_m})$. On the joint system, assume a probability distribution $P_{AB} = (p_{a_1 b_1}, \dots, p_{a_1 b_m}, \dots, p_{a_n b_1}, \dots, p_{a_n b_m})$, where $p_{a_i b_j} = P(A = a_i, B = b_j)$.

Now let us assume the situation, when observing realizations of system A, the probabilities of states of system B are

$$\begin{aligned} P_{B|A=a_1} &= (p_{b_1|a_1}, \dots, p_{b_m|a_1}), \\ P_{B|A=a_2} &= (p_{b_1|a_2}, \dots, p_{b_m|a_2}), \\ &\vdots \\ P_{B|A=a_n} &= (p_{b_1|a_n}, \dots, p_{b_m|a_n}) \end{aligned} \tag{2.29}$$

where $p_{b_j|a_i} := P(B = b_j | A = a_i)$. From the definition of conditional probability, we have $p_{a_i b_j} = p_{a_i} p_{b_j|a_i}$.

We can now also approach the problem of finding "sensible measures of information" from an axiomatic point of view. The axioms are usually called **Shannon-Khinchin's axioms** and demand the measure of information $S(P)$ as a functional of probability distribution P to satisfy the following axioms (following [22]):

- **SK1: Continuity**: $S(P) = S(p_1, p_2, \dots, p_n)$ is a continuous function of $p_1, p_2, \dots, p_n$.
- **SK2: Maximality**: $S(p_1, p_2, \dots, p_n)$ is maximal for the uniform distribution $p_i = \frac{1}{n}$ for all i.

- **SK3: Expandability**: $S(p_1, p_2, \ldots, p_n, 0) = S(p_1, p_2, \ldots, p_n)$ for any $n \geq 1$, i.e. adding an elementary state with probability 0 does not change the entropy.
- **SK4: Shannon additivity**: $S(P_{AB}) = S(P_A) + S(P_B|P_A)$, where $S(P_B|P_A) = \sum_{j=1}^{n} p_{a_j} S(P_{B|A=a_j})$ is the conditional entropy of system B given system A.

One can easily show, that the Shannon entropy $S(P) = -k \sum_{i=1}^{n} p_i \ln p_i$ satisfies all the axioms (k being any positive constant). Importantly, Shannon [23] showed the other direction, that the only measure satisfying all the axioms is the Shannon entropy. This means, that unless we have a reason to modify the axioms, we shall prefer the Shannon entropy over other measures. As we will see, the axiom which is mostly being modified is the axiom SK4. Hanel et al. [16] argue, that this axiom is strongly connected with the ergodicity of the system. Many complex systems, however, are not ergodic, for example, they get trapped in some part of the state space and can never return back. This will be the motivation to generalize the axiom SK4.

2.3 Shannon entropy in statistical inference

In statistical inference, the task is, given observations of some random process, to find the properties of the distribution which generated them. Consider the following task: having once again sequences of N i.i.d. random variables coming from a discrete set $\mathcal{A}_X = \{a_1, \ldots, a_n\}$ with probabilities $\mathbf{q} = \{q_1, \ldots, q_n\}$, we want to find the most likely histogram of the observations of the letters from $\mathcal{A}_X$. What we mean by that? For each letter $a_i \in \mathcal{A}_X$, let us denote the number of its occurences in the sequence by k_i and its relative frequency by $p_i = \frac{k_i}{N}$. Now the question is: what are the values of k_i (or p_i) which are the most likely to be observed? The idea is, that in practice, we might not know the exact values of q_i, but just some more coarse information about the distribution.

If we knew the distribution $P(X)$, we could write the probability of a histogram $\mathbf{k} = (k_1, \ldots, k_n)$ as

$$P(\mathbf{k}|\mathbf{q}) = \frac{N!}{k_1! \ldots k_n!} q_1^{k_1} \ldots q_n^{k_n} = M(\mathbf{k}) G(\mathbf{k}|\mathbf{q}) \tag{2.30}$$

We see, that the probability can be factorized in two terms, the multiplicity term $M(\mathbf{k}) := \frac{N!}{k_1! \ldots k_n!}$ and the constraint term $G(\mathbf{k}|\mathbf{q}) := q_1^{k_1} \ldots q_n^{k_n}$. Importantly, the multiplicity term does not depend on the underlying distribution $\mathbf{q}$, and all the information about the distribution is contained in the constraint term $G(\mathbf{k}|\mathbf{q})$.

Then, the most likely histogram is such a sequence $\mathbf{k}$ which maximizes the probability $P(\mathbf{k}|\mathbf{q})$. For the maximization problem, we can instead consider the logarithm of the probability and divide by N, which yields

$$\frac{1}{N} \ln P(\mathbf{k}|\mathbf{q}) = \frac{1}{N} \ln M(\mathbf{k}) + \frac{1}{N} \ln G(\mathbf{k}|\mathbf{q}) \tag{2.31}$$

The reason why we divided by N is, that the term $\frac{1}{N} \ln M(\mathbf{k})$ has a limit

$$\lim_{N \to \infty} \frac{1}{N} \ln M(\mathbf{k}) = -\sum_{i=1}^{n} p_i \ln p_i = S(\mathbf{p}) \tag{2.32}$$

which can be proven for example using the Stirling's approximation of factorials [16], or using bounds on the multinomial coefficient (we will see the case with binomial coefficient and binary entropy in section 4.2.1).

For the constraint part, we can employ a useful reparametrization

$$q_i = e^{-\alpha-\beta\epsilon_i} \tag{2.33}$$

in which α will be responsible for normalization and we extract β as a common factor for all probabilities in $\mathbf{q}$. This leads to

$$\frac{1}{N}\ln G(\mathbf{k}|\mathbf{q}) = -\frac{1}{N}\sum_{i=1}^{n} k_i\left(\alpha+\beta\epsilon_i\right) = -\alpha\sum_{i=1}^{n} p_i - \beta\sum_{i=1}^{n} p_i\epsilon_i \tag{2.34}$$

Finally, one wants to find such a sequence $\mathbf{p} = (p_1, \ldots, p_n)$ which maximizes the expression

$$\frac{1}{N}\ln P(\mathbf{k}|\mathbf{q}) = S(\mathbf{p}) - \alpha\sum_{i=1}^{n} p_i - \beta\sum_{i=1}^{n} p_i\epsilon_i \tag{2.35}$$

The maximization can be done by taking the derivative with respect to p_i and setting it to zero. This yields

$$\frac{\partial}{\partial p_i}\left(S(\mathbf{p}) - \alpha\sum_{i=1}^{n} p_i - \beta\sum_{i=1}^{n} p_i\epsilon_i\right) = 0 \tag{2.36}$$

This is the famous **maximum entropy principle** (MEP or MaxEnt). It corresponds to maximizing the Shannonian entropy $S(\mathbf{p})$ under the normalization constraint (using the Lagrange multiplier α) and the constraint on average values of properties ϵ_i (using the Lagrange multiplier β).

The maximum entropy principle was popularized by Jaynes [24]. He argued, that instead of first finding the probability distribution and then arguing, that Shannon entropy corresponds to thermodynamic entropy (as we saw in section 2.1), one should treat statistical mechanics as an inference problem from the beginning. The final goal is to assign a probability to each state, which has to be done in a minimally biased way. Therefore, we should only take in account the information we have, while keeping as much uncertainty as possible. Since the information theory gives us the measure of uncertainty, namely the Shannon entropy, the recipe is to maximize the Shannon entropy under the constraints given by the information we have. We will utilize this approach for modeling of complex networks in section 2.4.

Shore-Johnson axioms

Shore and Johnson [25] made a further point in favor of Shannon entropy. They proposed a set of axioms one should demand from an inference method involving maximization of some functional (maximum entropy principle) to avoid inconsistencies. We will state the axioms as they are given in [26]:

- **SJ1: Uniqueness axiom**: The result should be unique.
- **SJ2: Permutation invariance**: The permutation of states should not matter
- **SJ3: Subset independence**: It should not matter whether one treats disjoint subsets of system states in terms of separate conditional distributions or in terms of the full distribution

- **SJ4: System independence**: It should not matter whether one accounts for independent constraints related to independent systems separately in terms of marginal distributions or in terms of full-system constraints and joint distribution.
- **SJ5: Maximality**: In absence of any prior information, the uniform distribution should be the solution.

The fifth axiom is not explicitly mentioned in the original paper [25], since they assumed more general case, where one has some nontrivial prior knowledge about the distribution and minimizes cross-entropy. However, when we have no prior knowledge (and have therefore uniform prior distribution), the minimum cross-entropy inference is equivalent to the maximum entropy principle.

Shore and Johnson argue, that maximizing the Shannon entropy is the only inference procedure satisfying all their axioms. However, we will see in section 3.4 that they actually used a hidden assumption and the true family of functionals satisfying their axioms is actually richer.

2.4 Complex network modelling using Shannon entropy

As we have argued, the Shannon entropy can be utilized to find probability distributions which are least biased under given constraints. According to Jaynes [24], this approach can go well beyond modeling of physical systems. We can use it for essentially any measurable space. In this section, we will utilize the Shannon entropy for modelling of graphs. The resulting models are called *Exponential Random Graphs* (ERG). The derivation using maximization of Shannon entropy was first shown by Park and Newman [27].

Let us assume we have a set of all possible graphs with N nodes, denoted by $G \in \mathcal{G}_N$. This set is finite and therefore measurable. We want to find a probability distribution P, i.e. we want to assign to each graph $G' \in \mathcal{G}_N$ a probability $P(G')$ s.t. $\sum_{G\in\mathcal{G}} P(G) = 1$. Also, we want to satisfy canonical constraints

$$c_i^* = \sum_{G\in\mathcal{G}} c_i(G)P(G) \tag{2.37}$$

where $c_i(G)$ is a value of some property c_i on the graph G and c_i^* is the value of that property we want to recover on average. We use the subscript i to denote the possibility of more constraints at once.

For any probability distribution P, we can define the graph Shannon entropy as

$$S = -\sum_{G\in\mathcal{G}} P(G)\ln P(G) \tag{2.38}$$

We now want to maximize the Shannon entropy given the constraints in eq. 2.37 and the normalization condition $\sum_{G\in\mathcal{G}} P(G) = 1$. The corresponding Lagrange function to maximize is

$$\mathcal{L} = -\sum_{G\in\mathcal{G}_N} P(G)\ln P(G) + \alpha\left(\sum_{G\in\mathcal{G}_N} P(G) - 1\right) + \sum_i \theta_i \left(\sum_{G\in\mathcal{G}} c_i(G)P(G) - c_i^*\right)$$

Then for maximization, we impose for each graph $G' \in \mathcal{G}_N$

$$\frac{\partial \mathcal{L}}{\partial P(G')} = -\ln P(G') - 1 - \alpha - \sum_i \theta_i c_i(G') = 0$$

which leads to

$$P(G') = e^{-1-\alpha} e^{-\sum_i \theta_i c_i(G')}$$

We can see that the factor with the Lagrange multiplier α factorizes out and is the same for all graphs. This means we can set

$$e^{1+\alpha} = \sum_{G \in \mathcal{G}_N} e^{-\sum_i \theta_i c_i(G)} \equiv \mathcal{Z}(\vec{\theta}) \tag{2.39}$$

with $\mathcal{Z}(\vec{\theta})$ being the partition function. The resulting probability distribution then is

$$P(G) = \frac{e^{-H(G,\vec{\theta})}}{\mathcal{Z}(\vec{\theta})} \tag{2.40}$$

where $H(G,\theta) = \sum_i \theta_i c_i(G)$ is called graph Hamiltonian and $\mathcal{Z}(\vec{\theta}) = \sum_{G \in \mathcal{G}} e^{-H(G,\theta)}$ is the partition function.

We can also introduce the free energy $F(\vec{\theta}) = -\ln \mathcal{Z}(\vec{\theta})$ and similarly to statistical mechanics obtain

$$\frac{\partial F(\vec{\theta})}{\partial \theta_i} = -\frac{1}{\mathcal{Z}(\vec{\theta})} \frac{\partial \mathcal{Z}(\vec{\theta})}{\partial \theta_i} = -\frac{1}{\mathcal{Z}(\vec{\theta})} \sum_{G \in \mathcal{G}_N} -c_i(G) e^{-\sum_i \theta_i c_i(G)} = \langle c_i \rangle \tag{2.41}$$

We observe that the knowledge of partition function is central in our study. Fortunately, it is possible to compute the partition function analytically in several cases, which we show below.

Undirected case

Let us first consider the case of undirected graphs without self-loops. That means that the adjacency matrix is a symmetrical binary matrix with zeros on the diagonal. Then we can assign a distinct Lagrange multiplier to each possible edge in the following manner:

$$H(G,\vec{\theta}) = \sum_{i<j} \theta_{ij} a_{ij} \tag{2.42}$$

where a_{ij} are the elements of the adjacency matrix of G. Then one can compute the partition function analytically

$$\mathcal{Z}(\vec{\theta}) = \sum_{\{a_{kl}\}} e^{-\sum_{i<j} \theta_{ij} a_{ij}} = \sum_{\{a_{kl}\}} \prod_{i<j} e^{-\theta_{ij} a_{ij}} = \prod_{i<j} \sum_{\{a_{ij}=0,1\}} e^{-\theta_{ij} a_{ij}} = \prod_{i<j} (1 + e^{-\theta_{ij}}) \tag{2.43}$$

The notation $\sum_{\{a_{kl}\}}$ means summing over all possible adjacency matrices. The free energy then is

$$F(\vec{\theta}) = -\ln \mathcal{Z}(\vec{\theta}) = -\sum_{i<j} \ln(1 + e^{-\theta_{ij}}) \tag{2.44}$$

Let us note that we obtained a model with independent edges. This is because the probability of the whole graph factorizes into probabilities depending only on single edges:

$$P(G|\vec{\theta}) = \frac{e^{-\sum_{i<j}\theta_{ij}a_{ij}}}{\mathcal{Z}(\vec{\theta})} = \frac{\prod_{i<j} e^{-\theta_{ij}a_{ij}}}{\prod_{i<j}(1+e^{-\theta_{ij}})} = \prod_{i<j}\frac{e^{-\theta_{ij}a_{ij}}}{1+e^{-\theta_{ij}}} \tag{2.45}$$

Then, each edge is a Bernoulli random variable, and the probability that node i and j are connected is

$$p_{ij} = \langle a_{ij}\rangle = \frac{\partial F(\vec{\theta})}{\partial\theta_{ij}} = \frac{1}{1+e^{\theta_{ij}}} \tag{2.46}$$

Remarkably, this result reminds us of the Fermi-Dirac statistics. This is not a coincidence, since each edge is either present or not present, which is similar to a state that can be occupied by at most one particle, as in the fermionic case. Park and Newman [27] also show similarity with Bose-Einstein statistics in case where we allow for multiple edges between same two nodes.

Now let us consider two cases:

- Setting $\theta_{ij} = \theta$: Then the Hamiltonian is $H(G,\theta) = \sum_{i<j}\theta a_{ij} = \theta L_u(G)$, where $L_u(G)$ is the total number of links in G. This Hamiltonian corresponds to only constraining the total number of links. Then we have

$$p_{ij} = \frac{1}{1+e^{\theta}} \tag{2.47}$$

  i.e. the same for all nodes. Since we have a model with independent edges, each with the same probability of occurrence, this model is actually equivalent to the **Erdős-Rényi model**, as we have seen in section 1.3.1. The partition function in this case is

$$\mathcal{Z}(\vec{\theta}) = \prod_{i<j}(1+e^{-\theta}) = (1+e^{-\theta})^{N(N-1)/2} \tag{2.48}$$

- Setting $\theta_{ij} = \theta_i + \theta_j$: Then

$$H(G,\theta) = \sum_{i<j}(\theta_i+\theta_j)a_{ij} = \sum_{i\neq j}\theta_i a_{ij} = \sum_{i\neq j}\theta_i k_i(G) \tag{2.49}$$

  where $k_i(G)$ is the degree of i-th node of the graph G. This corresponds to constraining the degrees of nodes, therefore it is the canonical version of the configuration model. The link probabilities are

$$p_{ij} = \frac{1}{1+e^{\theta_i+\theta_j}} = \frac{1}{1+e^{\theta_i}e^{\theta_j}} = \frac{\beta x_i x_j}{1+\beta x_i x_j} \tag{2.50}$$

  where we used the reparametrization $\sqrt{\beta}x_i = e^{-\theta_i}$. This is exactly the **Park-Newman model** as we have defined it in section 1.3.4.

Directed case

In the directed case, the adjacency matrix does not have to be symmetric, and therefore we can assign a Lagrange multiplier to each of its entries. We allow self-loops, meaning that even the diagonal entries are allowed to be non-zero. The Hamiltonian is

$$H(G,\vec{\theta}) = \sum_{i,j} \theta_{ij} a_{ij} \tag{2.51}$$

Analogically, we have

$$Z(\vec{\theta}) = \prod_{i,j}(1+e^{-\theta_{ij}}) \tag{2.52}$$

$$F(\vec{\theta}) = -\sum_{i,j} \ln(1+e^{-\theta_{ij}}) \tag{2.53}$$

$$p_{ij} = \mathbb{E}(a_{ij}) = \frac{\partial F(\vec{\theta})}{\partial \theta_{ij}} = \frac{1}{1+e^{\theta_{ij}}} \tag{2.54}$$

Let us once again consider two important cases:

- Setting $\theta_{ij} = \theta$: The Hamiltonian is $H(G,\theta) = \sum_{i,j} \theta a_{ij} = \theta L(G)$, where $L(G)$ is this time the total number of directed links in G (self-loops are considered only once). The link probability is similarly to the undirected case

  $$p_{ij} = \frac{1}{1+e^{\theta}} \tag{2.55}$$

  and the partition function is

  $$\mathcal{Z}(\vec{\theta}) = \prod_{i,j}(1+e^{-\theta}) = (1+e^{-\theta})^{N^2} \tag{2.56}$$

  We call this model the **Directed Erdős-Rényi model**.

- Setting $\theta_{ij} = \alpha_i + \beta_j$ and not allowing self-loops: The Hamiltonian is

  $$H(G,\vec{\alpha},\vec{\beta}) = \sum_{i\neq j}(\alpha_i+\beta_j)a_{ij} = \sum_i(\alpha_i k_i^{out}(G) + \beta_i k_i^{in}(G)) \tag{2.57}$$

  This amounts to constraining both the out-degree and in-degree sequence. Then the edge probabilities are

  $$p_{ij} = \frac{1}{1+e^{\alpha_i+\beta_j}} = \frac{x_i y_j}{1+x_i y_j} \tag{2.58}$$

  where we reparametrized $x_i \equiv e^{-\alpha_i}$, $y_j \equiv e^{-\beta_j}$. A model where each node i is assigned values x_i, y_i and edges are drawn independently with probability 2.58 is called **Directed Binary configuration model** (DBCM) [28].

Chapter 3

Non-Shannonian entropies

In Section 2, we made a strong case for the omnipresence of Shannon entropy. However, we also argued that there are situations where a generalization is needed. For example, we saw that in physics, we require entropy to be extensive, ensuring that temperature remains a meaningful concept across all scales. Extensivity depends on the scaling behaviour of the system's phase space. For exponential scaling, Shannon entropy is suitable. However, for systems where the phase space grows sub-exponentially, a generalization is required.

Furthermore, we will see that considering the Shore-Johnson axioms for statistical inference reveals that a broader family of entropies, beyond just the Shannon form, is permissible. This broader family corresponds to a generalization of the Shannon-Khinchin axioms.

In this chapter, we will first examine several Rényi and Tsallis entropies individually. Subsequently, we will demonstrate how these entropies can be derived from generalized versions of the axioms presented in Section 2.

3.1 Rényi entropy

Rényi [29] studied the special case of axiom SK4, where we consider two independent systems A and B and we demand $S(P_{AB}) = S(P_A) + S(P_B)$. He found, that this requirement together with SK1-3 is actually satisfied with a larger family of functionals, namely

$$S_\alpha(p_1, \dots, p_n) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha \right) \tag{3.1}$$

for $\alpha > 0$ and $\alpha \neq 1$. Importantly, in the limit $\alpha \to 1$, we recover the Shannon entropy. However, due to the uniqueness theorem, we know that only Shannon entropy satisfies the full version of axiom SK4. Rényi realized, that the reason for this is, that a standard linear average is used. He then proposed, that instead of arithmetic mean, we could have used a more general average defined for real numbers $x_1, \dots, x_n$ as

$$f^{-1} \left(\sum_{i=1}^{n} p_i f(x_i) \right) \tag{3.2}$$

where f is a real function called Komlogorov-Nagumo function. If we then assume, that an information learned from an event with probability p is $\mathcal{I}(p) = -\log p$, then the average amount of information from a system with probabilities $P = (p_1, \ldots, p_n)$ using the general average is

$$\mathcal{S}_f(P) = f^{-1}\left(\sum_{i=1}^{n} p_i f(-\log(p_i))\right) \tag{3.3}$$

Now the question is, with which functions f we can still recover the additivity rule for independent events. It can be shown [30], that there are only two options

- The linear case $f(x) = cx$: then the corresponding entropy is the Shannon entropy $S(P) = -k\sum_{i=1}^{n} p_i \log p_i$
- The exponential case $f(x) = e^{(1-\alpha)x}$: then the corresponding entropy is the Rényi entropy $S_\alpha(P) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} p_i^\alpha\right)$

Rényi and Shannon entropies are therefore the only entropies, which are additive for independent events. This makes Rényi entropy particularly useful in information theory.

In MaxEnt context, one can consider canonical constraints

$$\langle\epsilon\rangle = \sum_{i=1}^{n} p_i \epsilon_i = U \tag{3.4}$$

and constrained maximization leads to the following distribution [31]

$$p_i = \frac{1}{\mathcal{Z}_\alpha}\left[1 - \beta\frac{\alpha-1}{\alpha}(\epsilon_i - U)\right]_+^{1/(\alpha-1)} \tag{3.5}$$

with partition function

$$\mathcal{Z}_\alpha = \sum_{i=1}^{n}\left[1 - \beta\frac{\alpha-1}{\alpha}(\epsilon_i - U)\right]_+^{1/(\alpha-1)} \tag{3.6}$$

where $[x]_+ = \max(x, 0)$ is the positive part. Note, that the distribution is power-law, and in the limit $\alpha \to 1$, we recover the Boltzmann distribution.

Rényi entropy has been successfully applied to a wide range of problems, including the thermodynamics of multifractals [30], measuring of entanglement entropy in quantum many-body systems [32] and others.

3.2 Tsallis entropy

We have seen, that in case of Rényi entropy, we have a family of entropic functionals, where the Shannon entropy is a special case. Tsallis[1] [36] proposed a different entropic family, parametrized by a real number q, defined as

$$S_q(P) = k\frac{1}{1-q}\left(\sum_{i=1}^{n} p_i^q - 1\right) \tag{3.7}$$

[1] Current literature [33] recognizes that this entropy form has been found even earlier, by Havrda and Charvát in 1967 [34] and independently by Daróczy in 1970 [35], so it is sometimes called Tsallis-Havrda-Charvát entropy or Havrda-Charvát-Daróczy-Tsallis entropy.

Once again, in the limit $q \to 1$, we recover the Shannon entropy. The question immediately arises, what is the behavior of this entropy when considering independent events. We have seen, that the Rényi entropy is additive. Keeping the notation we had for the Shannon-Khinchin axioms, we can for independent A and B write

$$\begin{aligned} \frac{1-q}{k} S_q(P_{AB}) &= \left(\sum_{i=1}^{n} \sum_{j=1}^{m} p_{a_i}^q p_{b_j}^q - 1 \right) \\ &= \left(\sum_{i=1}^{n} p_{a_i}^q - 1 \right) \left(\sum_{j=1}^{m} p_{b_j}^q - 1 \right) + \left(\sum_{i=1}^{n} p_{a_i}^q - 1 \right) + \left(\sum_{j=1}^{m} p_{b_j}^q - 1 \right) \\ &= \left(\frac{1-q}{k} \right)^2 S_q(P_A) S_q(P_B) + \frac{1-q}{k} S_q(P_A) + \frac{1-q}{k} S_q(P_B) \end{aligned} \tag{3.8}$$

which is then rewritten as

$$S_q(P_{AB}) = S_q(P_A) + S_q(P_B) + \frac{1}{k}(1-q) S_q(P_A) S_q(P_B) \tag{3.9}$$

This means Tsallis entropy is non-additive for $q \neq 1$.

Tsallis then studies two cases. First, in the *microcanonical* case, he considers W states and shows that even here, the entropy is maximized by the uniform distribution $p_i = \frac{1}{W}$. Then the entropy is

$$S_q(P) = k \frac{W^{1-q} - 1}{1-q} \tag{3.10}$$

This equation is important, because it shows the scaling of Tsallis entropy. If we consider that the state space of the system grows exponentially, i.e. $W \propto e^N$, then we have $S_q(P) \propto e^{(1-q)N}$, meaning it is non-extensive for $q \neq 1$. However, if we consider a state space growing as $W \propto N^{1/(1-q)}$, then we have $S_q(P) \propto N$. Therefore we see, that event though the Tsallis entropy is non-additive for $q \neq 1$, it can be extensive for state-spaces, which scale as a power-law. Since being proposed, Tsallis collected immense amount of evidence, that the Tsallis entropy is relevant for many systems [37].

3.2.1 First version of thermostatistics

Tsallis also considered a *canonical* case, with a constraint

$$\langle \epsilon \rangle = \sum_{i=1}^{n} p_i \epsilon_i = U_q \tag{3.11}$$

and claims that the entropy is maximized by the distribution

$$p_i = \frac{1}{\mathcal{Z}_q} \left(1 - \beta(q-1)\epsilon_i\right)^{1/(q-1)} \tag{3.12}$$

with β being the Lagrange multiplier associated with the constraint on energy and the partition function

$$\mathcal{Z}_q = \sum_{i=1}^{n} \left(1 - \beta(q-1)\epsilon_i\right)^{1/(q-1)} \tag{3.13}$$

However, later on it has been realized, that it is not that simple to eliminate the Lagrange multiplier α responsible for normalization. Ferri et al. [38] showed, that it is indeed possible to eliminate the multiplier α (we will see this procedure in section 4.1, where we even consider more than one constraint), with the inevitable consequence of introducing self-referentiality. Specifically, the resulting distribution is

$$p_i = \frac{1}{\mathcal{Z}_q}\left[1-\beta\frac{q-1}{q}\frac{1}{\mathcal{Z}_q^{1-q}}(\epsilon_i - U_q)\right]_+^{1/(q-1)} \equiv \frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-\beta\frac{1}{q\mathcal{Z}_q^{1-q}}(\epsilon_i - U_q)\right) \tag{3.14}$$

with

$$\mathcal{Z}_q = \sum_{i=1}^{n}\left[1-\beta\frac{q-1}{q}\frac{1}{\mathcal{Z}_q^{1-q}}(\epsilon_i - U_q)\right]_+^{1/(q-1)} = \left(\sum_{i=1}^{n} p_i^q\right)^{1/(q-1)} \tag{3.15}$$

where $\exp_q$ is the q-exponential function defined as

$$\exp_q(x) := [1+(1-q)x]_+^{1/(1-q)} \tag{3.16}$$

Compare this result with eq. 3.5. Here, the significant difference is the self-referentiality caused by the term $\frac{1}{\mathcal{Z}_q^{1-q}}$. Ferri et. al state, that because of this, we need to iterate the equation for $\mathcal{Z}_q$ until convergence. More recently, Jizba et al. [31] argued, that if we define β_{THC} as

$$\beta_{THC} = \beta\frac{1-q}{q}\frac{1}{\mathcal{Z}_q^{1-q}} \tag{3.17}$$

we can use the self-referentiality to find an universal parameter. Since in physics we only measure the energy differences and not the absolute energy, the dependence of probability on the energy U_q is undesirable. But since β_{THC} depends U_q through $\mathcal{Z}_q$, Jizba et al. [31] were able to show that it transforms through the group called *Möbius parabolic group* under the shifts of energy levels. They then propose to use the invariant of this group as an observable. In section 4.2, we will explicitly show this redefinition of parameters.

3.2.2 Third version of thermostatistics

Soon after Tsallis's seminal work, it was realized that the constraint in eq. 3.11 causes certain difficulties. Therefore, later on, Tsallis et al. [39] proposed, that for constraints, one should use the so-called escort distribution

$$P_i = \frac{p_i^q}{\sum_{j=1}^n p_j^q} \tag{3.18}$$

so that the constraints then are

$$\langle\epsilon\rangle_q := \sum_{i=1}^{n} P_i\epsilon_i = \frac{\sum_{i=1}^n p_i^q\epsilon_i}{\sum_{j=1}^n p_j^q} = U_q \tag{3.19}$$

The quantities $\langle\dots\rangle_q$ are called **q-averages**. Maximization of the Tsallis entropy under this constraint leads to

$$p_i = \frac{1}{\bar{\mathcal{Z}}_q}\left[1-\beta(1-q)\frac{1}{\bar{\mathcal{Z}}_q^{1-q}}(\epsilon_i - U_q)\right]_+^{1/(1-q)} = \frac{1}{\bar{\mathcal{Z}}_q}\exp_q\left(-\beta\frac{1}{\bar{\mathcal{Z}}_q}(\epsilon_i - U_q)\right) \tag{3.20}$$

with the partition function

$$\bar{\mathcal{Z}}_q = \sum_{i=1}^{n} \left[1 - \beta(1-q)\frac{1}{\bar{\mathcal{Z}}_q^{1-q}}(\epsilon_i - U_q)\right]_{+}^{1/(1-q)} = \left(\sum_{i=1}^{n} p_i^q\right)^{1/(1-q)} \tag{3.21}$$

To motivate the use of escort distributions, let us define the q-logarithm as

$$\ln_q(x) := \frac{1}{1-q}(x^{1-q} - 1) \tag{3.22}$$

and introduce the following useful identities [37]

$$\frac{\mathrm{d}\ln_q(x)}{\mathrm{d}x} = \frac{1}{x^q} \tag{3.23}$$

$$\frac{\mathrm{d}\exp_q(x)}{\mathrm{d}x} = (\exp_q(x))^q \tag{3.24}$$

Using these identities, it is easy to show that

$$\frac{\partial}{\partial\beta}\ln_q \bar{\mathcal{Z}}_q(\tilde{\beta}) = 0 \tag{3.25}$$

This motivates the definition of new partition function

$$\ln_q \mathcal{Z}_q(\beta) := \ln_q \bar{\mathcal{Z}}_q(\beta) - \beta U_q \tag{3.26}$$

for which we trivially have $-\frac{\partial}{\partial\beta}\ln_q \mathcal{Z}_q(\beta) = U_q$. Also, from the definition of Tsallis entropy (eq. 3.7), and using the identity for partition function (eq. 3.21), can find that $S_q(P) = k\ln_q \bar{\mathcal{Z}}_q(\beta)$. Now if we define the free energy as

$$F_q(\beta) := -\frac{1}{\beta}\ln_q \mathcal{Z}_q(\beta) \tag{3.27}$$

we can recover the familiar relation (using $T = \frac{1}{k\beta}$)

$$F_q = U_q - TS_q \tag{3.28}$$

and other thermodynamic relations as well [39]. For the first version of thermostatistics, such relations have not been found.

3.3 Hanel-Thurner entropies

Hanel and Thurner [40, 16] identified scaling as a fundamental property of entropy and examined how it relates to the Shannon-Khinchin axioms. They determined that entropy scaling is directly connected to axiom SK4 (see section 2.2), and proposed replacing this axiom with the condition that entropy must have a trace-form, meaning there exists a function g such that

$$S(P) = \sum_{i=1}^{W} g(p_i) \tag{3.29}$$

Under this requirement, they found [16] that all such entropies can be written as

$$S(P) = \frac{r}{c}A^{-d}e^{A}\left[\sum_{i=1}^{W}\Gamma(1+d, A - c\log p_i) - rp_i\right] \tag{3.30}$$

where $\Gamma(a,b) = \int_b^\infty t^{a-1}e^{-t}dt$ is the incomplete gamma function, $A = \frac{cdr}{1-(1-c)r}$ and r needs to satisfy further conditions (see [16]).

The parameters c, d are of a great importance, since they determine the scaling of the entropy. Specifically, they considered the microcanonical case with W states and probabilities $p_i = \frac{1}{W}$ and found that

$$\lim_{W\to\infty} \frac{S(\lambda W)}{S(W)} = \lambda^{1-c} \qquad \text{for } 0 \le c < 1 \tag{3.31}$$

$$\lim_{W\to\infty} \frac{S(W^{1+a})}{S(W)W^{a(1-c)}} = (1+a)^d \qquad \text{for } d \text{ real valued} \tag{3.32}$$

They argue that this determines the equivalence classes of all systems for which SK1-3 hold. Importantly, for $(c,d) = (1,1)$ we recover the Shannon entropy, and for $(c,d) = (q,0)$ we find the Tsallis entropy. From this we can also see, that in the limit $q \to 1$, even though Tsallis entropy converges to the Shanon entropy, the asymptotic properties do not change continuously, meaning that the thermodynamic limit and the limit $q \to 1$ do not commute. This fact will be important later, when we will consider the termodynamic limit for network models.

Hanel and Thurner also showed, that the corresponding probability distribution functions are

- $(c,d) = (1,1)$: the Boltzmann distribution $p_{(1,1)} \propto e^{-x}$
- $(c,d) = (q,0)$: power-laws (q-exponentials) $p_{(q,0)} \propto (1+(1-q)x)^{1/(1-q)}$
- $(c,d) = (1,d)$: stretched exponentials $p_{(1,d,r)} \propto e^{-dr\left((1+x/r)^{1/d}-1\right)}$

They also considered the situation, when we know the scaling of the state space, i.e. we know the function $W(N)$ depending on the size of the system N and require the entropy to be extensive. They find that the corresponding scaling coefficients have to be

$$\frac{1}{1-c} = \lim_{N\to\infty} N\frac{W'(N)}{W(N)} \tag{3.33}$$

$$d = \lim_{N\to\infty} \log W(N) \left(\frac{1}{N}\frac{W(N)}{W'(N)} + c - 1\right) \tag{3.34}$$

Using this, one can distinguish the following cases:

- $W(N) = e^N$: the corresponding coefficients are $(c,d) = (1,1)$, which corresponds to the Shannon entropy
- $W(N) = N^b$: the corresponding coefficients are $(c,d) = (1-\frac{1}{b}, 0)$, which corresponds to the Tsallis entropy with $q = 1 - \frac{1}{b}$
- $W(N) = \exp(\lambda N^\gamma)$: the corresponding coefficients are $(c,d) = (1, 1/\gamma)$, which corresponds to Anteonodo-Plastino entropy.

We shall note that Rényi entropy is not of the trace form, so it does not belong, to the family of entropies considered by Hanel and Thurner. However, its scaling properties can be studied as well and it can be shown [16], that the corresponding coefficients are $(c,d) = (1,1)$. On the other hand, as we saw, the maximizing distribution is a q-exponential, which does not correspond to the results shown here.

3.4 Entropies satisfying Shore-Johnson axioms

Now let us turn our attention to statistical inference again. We have introduced the Shore-Johnson axioms in section 2.3, which are supposed to ensure consistence of inference methods based on maximization of some functional. Shore and Johnson claimed, that the only solution to their axioms is the Shannon entropy. However, as was pointed out by Uffink [41], they made an assumption, which went further then what was originally intended. Let us recall the axiom SJ4 once again:

> **SJ4: System independence:** It should not matter whether one accounts for independent constraints related to independent systems separately in terms of marginal distributions or in terms of full-system constraints and joint distribution.

However, in their derivation of the uniqueness, they treat the systems as independent (using a probability distribution, which factorizes), even if the independence is not guaranteed. Uffink showed, that if we consider only truly independent systems in axiom SJ4, the family of functionals which can be maximized under constraints is actually

$$\mathcal{U}_q = \left(\sum_{i=1}^{n} p_i^q\right)^{1/(1-q)} \tag{3.35}$$

modulo equivalency condition [26]. By equivalence, we mean that we can actually maximize $f(\mathcal{U}_q)$ for some strictly monotonic function f, since the maximization procedure will have the same functional form of the resulting distribution (however, the role of constraints can differ). We call the functionals of this form **Uffink functionals**. Note that for $f(x) = \ln x$, we recover the Rényi entropy and for $f(x) = \ln_q x$, we recover the Tsallis entropy. On the other hand, Hanel-Thurner (c,d) entropies generally do not belong to this family.

Jizba and Korbel [26] then argue, that the original result proving the uniqueness of Shannon entropy can only hold, if we assume a strong system independence.

> **Strong system independence (SSI):** Whenever two subsystems of a system are disjoint, we can treat the subsystems in terms of independent distributions.

This is, however, a very strong assumption. Systems can, for example, start building correlations even on long distances. Jizba and Korbel [26] take the specific example of 2-qubit quantum system, where they show that using maximum-entropy principle with Shannon entropy can lead to an expectation of entanglement, which might not be actually present in the system, and therefore argue, that one needs to use more general entropies. In quantum systems, the SSI can easily be violated, since even disjoint systems can have correlated measurements.

In another paper, Jizba and Korbel [22] make a final unification step with the information theory, when they find such a modification of Shannon-Khinchin axioms, which yields the same family of entropies as the Uffink functionals. Similarly to Hanel and Thurner, they modify the axiom SK4, but this time, they just modify the way how the entropy of composed system is computed, using generalized arithmetics.

Chapter 4

Tsallis network models

As we have seen, non-Shannonian entropies emerge as natural candidates for modelling systems with slower then exponential state-space growth or whenever we can not assume strong independence of its subsystems. In case of networks, we are mostly interested in statistical inference, and there, according to SJ axioms, we can use any entropy equivalent to the Uffink functional. Such entropies include the Tsallis and Rényi entropy.

Both Tsallis and Rényi entropy lead to power-law distributions, which, as we have seen, are omnipresent in real networks. The natural question is, whether we could extend the procedure of ensemble network modelling introduced in section 2.4 to these entropies. This question seems to have not been answered yet. Squartini et al. [28] explicitly mention:

> "Remarkably, S_q (the Tsallis entropy) can be employed to define a non-extensive version of the ERG formalism, whose derivation proceeds along similar lines. For example, imposing only the normalization condition leads to the functional $L_q[P] = S_q - \lambda_0 \sum_{G \in \mathcal{G}_N} P(G) - 1$ which is maximized by the uniform distribution $P(G) = 1/|\mathcal{G}_N|$. Imposing less trivial constraints, however, has not been attempted yet: as a consequence, a thorough comparison between the goodness of the reconstruction performances induced by extensive and non-extensive entropies is still missing."

With this motivation in mind, let us study the case of Tsallis entropy.

4.1 General Tsallis network model

Let $\mathcal{G}$ be the set of graphs, for example determined by the number of nodes. Then we define the graph Tsallis entropy as

$$S_q(P) = \frac{1}{q-1}\left(1 - \sum_{G \in \mathcal{G}} P(G)^q\right) \tag{4.1}$$

Next important step is the choice of constraints. As we have seen, using the q-averages has advantages, like the connection to thermodynamics. On the other hand, it is harder to interpret them. In network science, one is usually considering standard averages over an ensemble when studying network properties. We will therefore use the standard average.

Let us therefore have $c_i^* = \sum_{G\in\mathcal{G}} P(G)c_i(G) = \langle c_i\rangle$, where $c_i(G)$ are the network properties we are constraining, and let us demand normalization $\sum_{G\in\mathcal{G}} P(G) = 1$. Then we can write the corresponding Lagrange function as

$$\mathcal{L} = \frac{1}{q-1}\left(1 - \sum_{G\in\mathcal{G}} P(G)^q\right) - \sum_{i=1}^k \theta_i \left(\sum_{G\in\mathcal{G}} c_i(G)P(G) - c_i^*\right) - \alpha\left(\sum_{G\in\mathcal{G}} P(G) - 1\right)$$

Computing the derivative of Lagrange function and demanding it to be zero, we get

$$\frac{\partial\mathcal{L}}{\partial P(G')} = \frac{q}{1-q}P(G')^{q-1} - \sum_{i=1}^k \theta_i c_i(G') - \alpha = 0 \qquad \forall G' \in \mathcal{G} \tag{4.2}$$

Now we need to eliminate the multiplier α. This is done via a simple trick presented for example in [38]. We multiply the equation above by $P(G')$ and sum over all $G' \in \mathcal{G}$. This leads to

$$\alpha = \frac{q}{1-q}\sum_{G\in\mathcal{G}} P(G)^q - \sum_{i=1}^k \theta_i\langle c_i\rangle$$

Finally, one can obtain

$$\begin{aligned} P(G')^{q-1} &= \sum_{G\in\mathcal{G}} P(G)^q + \frac{1-q}{q}\sum_{i=1}^k \theta_i(c_i(G') - \langle c_i\rangle) \\ P(G') &= \left(\sum_{G\in\mathcal{G}} P(G)^q\right)^{1/(q-1)} \left[1 - \frac{q-1}{q}\frac{1}{\sum_{G\in\mathcal{G}} P(G)^q}\left(\sum_{i=1}^k \theta_i(c_i(G') - \langle c_i\rangle)\right)\right]_+^{1/(q-1)} \end{aligned}$$

The factor $\left(\sum_{G\in\mathcal{G}} P(G)^q\right)^{1/(q-1)}$ must as well be the normalizing factor of the distribution (i.e. partition function), therefore we have the identity

$$\mathcal{Z}_q(\vec{\theta}) \equiv \left(\sum_{G\in\mathcal{G}} P(G)^q\right)^{1/(1-q)} \tag{4.3}$$

$$= \sum_{G'\in\mathcal{G}} \left[1 - \frac{q-1}{q}\frac{1}{\sum_{G\in\mathcal{G}} P(G)^q}\left(\sum_{i=1}^k \theta_i(c_i(G') - \langle c_i\rangle)\right)\right]_+^{1/(q-1)} \tag{4.4}$$

The general Tsallis network model can be therefore written as

$$\begin{aligned} P_q(G) &= \frac{1}{\mathcal{Z}_q}\left[1 - \frac{q-1}{q}\frac{1}{\mathcal{Z}_q^{1-q}}\left(\sum_{i=1}^k \theta_i(c_i(G) - \langle c_i\rangle)\right)\right]_+^{1/(q-1)} \\ &= \frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-\frac{1}{q\mathcal{Z}_q^{1-q}}\sum_{i=1}^k \theta_i(c_i(G) - \langle c_i\rangle)\right) \end{aligned} \tag{4.5}$$

The model suffers from self-referentiality, as we discussed in section 3.2.1. However, as shown in [31], we will use this fact to find an universal parameter, for which the model will not explicitly depend on the constraints $\langle c_i\rangle$. Let us first realize, that the following useful formula holds

$$\exp_{2-q}(a+b) = \exp_{2-q}(a)\exp_{2-q}\left(\frac{b}{1+(q-1)a}\right) \tag{4.6}$$

Then we can write

$$\exp_{2-q}\Big(-\frac{1}{q\mathcal{Z}_q^{1-q}}\sum_{i=1}^{k}\theta_i(c_i(G)-\langle c_i\rangle)\Big)$$
$$=\exp_{2-q}\left(\frac{1}{q\mathcal{Z}_q^{1-q}}\sum_{i=1}^{k}\theta_i\langle c_i\rangle\right)\exp_{2-q}\left(-\frac{1}{q\mathcal{Z}_q^{1-q}+(q-1)\sum_{i=1}^{k}\theta_i\langle c_i\rangle}\sum_{i=1}^{k}\theta_i c_i(G)\right) \tag{4.7}$$

The first bracket does not depend on G and thus factorizes in both numerator and denominator of eq. 4.5. Finally, we can define the universal parameters $\tilde{\theta}_i$ as

$$\tilde{\theta}_i \equiv \frac{\theta_i}{q\mathcal{Z}_q^{1-q}+(q-1)\sum_{i=1}^{k}\theta_i\langle c_i\rangle} \tag{4.8}$$

and obtain the simplified form of the probability distribution

$$P_q(G)=\frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-\sum_{i=1}^{k}\tilde{\theta}_i c_i(G)\right) \tag{4.9}$$

The parameters $\tilde{\theta}_i$ are not the Lagrange multipliers of the original problem anymore, but they are still supposed to be fitted according to the constraints. In further considerations, we will drop the tilde notation and write just θ_i. In section 2.4, we have defined the graph Hamiltonian $H(G,\vec{\theta})$. Here, we can use the same notation and write

$$P_q(G)=\frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-H(G,\vec{\theta})\right) \tag{4.10}$$

where $H(G,\vec{\theta})=\sum_{i=1}^{k}\theta_i c_i(G)$ contains all the constraints c_i, each coupled with its own parameter θ_i.

4.2 Tsallis-Erdös-Rényi model

Now let us once again consider directed graphs with self-loops and N_V nodes. We denote the set of all such graphs as $\mathcal{G}_{N_V}$. Let us also denote the maximum number of links in the ensemble as N. We will consider the case of constraining the average number of links in the ensemble, i.e.

$$L^*=\sum_{G\in\mathcal{G}_{N_V}}P_q(G)L(G)=\langle L\rangle \tag{4.11}$$

The corresponding model Hamiltonian is

$$H(G)=\theta L(G) \tag{4.12}$$

and the probability distribution is given by

$$P_q(G)=\frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-\theta L(G)\right)=\frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-\theta\sum_{i,j=1}^{N}a_{ij}\right) \tag{4.13}$$

The partition function is given by

$$\mathcal{Z}_q=\sum_{G\in\mathcal{G}_N}\exp_{2-q}\left(-\theta L(G)\right)=\sum_{\{a_{ij}\}}\exp_{2-q}\left(-\theta\sum_{i,j=1}^{N}a_{ij}\right) \tag{4.14}$$

where the sum $\sum_{\{a_{ij}\}}$ runs over all possible adjacency matrices.

The significant difference from the Shannonian case is that the probability distribution can not be factorized into a product of probabilities of individual edges. This means the edges are not independent random variables anymore. We will study the dependence between edges later.

Also note, that the probability does not depend on the ordering of edges. For the sake of simplicity, let us therefore drop the j index and just assume N random binary variables a_i. We will still call a_i links, to keep connection with our graph motivation.

In that case, we have a probability of a sequence $\mathbf{a} = (a_1, \dots, a_N)$ of N random binary variables a_i given by

$$P(\mathbf{a}) = \frac{1}{\mathcal{Z}_q(\theta)} \left(1 + (1-q)\theta \sum_{i=1}^{N} a_i\right)^{1/(q-1)} = \frac{1}{\mathcal{Z}_q(\theta)} \exp_{2-q}\left(-\theta \sum_{i=1}^{N} a_i\right) \tag{4.15}$$

4.2.1 Large N behavior

Let us now study the behavior of the probability and of the partition function when N is large. For these purposes, we will follow the approach similar to the one given in [42] in the case of Curie-Weiss model.

Our strategy will be, similarly to the asymptotic equipartition property, to determine the large N behavior of $\frac{1}{N} \ln P$. For that, it will be essential to study the behavior of quantity called **free entropy density**

$$\Phi_q(\theta, N) = \frac{1}{N} \ln \mathcal{Z}_{q,N}(\theta) \tag{4.16}$$

with the hope, that it converges for large N to a non-zero value. We will show, that its value for large N is determined by the minimum of another quantity called *effective free entropy density*. Note, that in statistical physics, one usually uses *free energy density*, but we avoid the term "free energy", because we do not have the connection to thermodynamic identities (which are available only when q-average constraints are considered).

From now on, we will be using the notation $\mathcal{Z}_{q,N}(\theta)$, since we are interested in the large N behavior and the partition function is dependent on N as well. To proceed, let us first realize, that the number of possible sequences where k links are present is given by $\binom{N}{k}$. That means that the probability of obtaining a sequence with k links under our model is

$$P\left(\sum_{i=1}^{N} a_i = k\right) = \binom{N}{k} \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-\theta k) \tag{4.17}$$

We could use a Stirling formula to approximate the binomial coefficient for large N. Or we can utilize bounds, often used in the information theory literature (see proof for example at [19])

$$\frac{1}{N+1} e^{NH(k/N)} \leq \binom{N}{k} \leq e^{NH(k/N)}$$

where $H(x) = -x \log x - (1-x)\log(1-x)$ is the Shannon binary entropy function. We can then write

$$\frac{1}{N+1} e^{NH(k/N)} \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-\theta k) \leq$$
$$\leq P\left(\sum_{i=1}^{N} a_i = k\right) \leq e^{NH(k/N)} \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-\theta k)$$

or in terms of the link-density $c = k/N$ and average number of links $\bar{a}_i = \frac{1}{N}\sum_{i=1}^{N} a_i$

$$\frac{1}{N+1} e^{NH(c)} \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-\theta N c) \leq P(\bar{a}_i = c) \leq e^{NH(c)} \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-\theta N c) \tag{4.18}$$

Note, that even though we used the Tsallis entropy, Shannon entropy naturally emerged because of the binomial multiplicity. Now let us rewrite the q-exponential in terms of the standard exponential and merge with the entropy term.

$$e^{NH(c)} \exp_{2-q}(-\theta N c) = e^{NH(c)} \left(1 + (1-q)\theta N c\right)^{1/(q-1)} = e^{N\left(H(c) - \frac{1}{N}\frac{1}{1-q}\ln(1+(1-q)\theta N c)\right)}$$

This allows us to define the **effective free entropy density** as

$$\phi_q(c, \theta, N) = H(c) - \frac{1}{N}\frac{1}{1-q}\ln\left(1 + (1-q)\theta N c\right) \tag{4.19}$$

Using it, the bounds are rewritten as

$$\frac{1}{N+1}\frac{1}{\mathcal{Z}_{q,N}(\theta)} e^{N\phi_q(c,\theta,N)} \leq P(\bar{a}_i = c) \leq \frac{1}{\mathcal{Z}_{q,N}(\theta)} e^{N\phi_q(c,\theta,N)} \tag{4.20}$$

Let us examine both bounds.

- For the lower bound, we use that $P(\bar{a}_i = c) \leq 1$, meaning that

  $$\frac{1}{N+1} e^{N\phi_q(c,\theta,N)} \leq \mathcal{Z}_{q,N}(\theta) \tag{4.21}$$

  This holds especially for $c = c_{max}(\theta, N) \equiv \arg\max_c \phi_q(c, \theta, N)$. Taking logarithms on both sides and dividing by N, we get

  $$-\frac{\ln(N+1)}{N} + \phi_q(c_{max}, \theta, N) \leq \frac{\ln \mathcal{Z}_{q,N}(\theta)}{N} \tag{4.22}$$

- For the upper bound, we sum over all possible c (there are $(N+1)$ of them, since there are $(N+1)$ possible values for k), getting

  $$1 \leq \sum_c \frac{e^{N\phi_q(c,\theta,N)}}{\mathcal{Z}_{q,N}(\theta)} \leq (N+1)\frac{e^{N\phi_q(c_{max},\theta,N)}}{\mathcal{Z}_{q,N}(\theta)} \tag{4.23}$$

  Taking logarithms on both sides and dividing by N, we get

  $$\frac{1}{N}\ln \mathcal{Z}_{q,N}(\theta) \leq \frac{\ln(N+1)}{N} + \phi_q(c_{max}, \theta, N) \tag{4.24}$$

Since $\lim_{N\to\infty}\frac{\ln(N+1)}{N}=0$, we find that[1]

$$\begin{aligned}\Phi_q(\theta) := \lim_{N\to\infty}\Phi_q(\theta,N) = \lim_{N\to\infty}\frac{1}{N}\ln\mathcal{Z}_{q,N}(\theta) &= \lim_{N\to\infty}\phi_q(c_{max}(\theta,N),\theta,N)\\ &= \lim_{N\to\infty}\max_{c\in[0,1]}\phi_q(c,\theta,N)\end{aligned} \tag{4.25}$$

meaning, that the free entropy density converges to a non-zero value, which is given by the maximum of the effective free entropy density, if this maximum converges as well. From the definition of ϕ_q in eq. 4.19 we see, that its behavior is determined by the interplay of the entropy term $H(c)$ and the constraint term $\frac{1}{N}\frac{1}{1-q}\ln(1+(1-q)\theta Nc)$. However, for fixed θ and q, we have

$$\lim_{N\to\infty}\frac{1}{N}\frac{1}{1-q}\ln(1+(1-q)\theta Nc)=0 \tag{4.26}$$

This means the energy term is negligible for large N and all the behavior of Φ_q (and therefore of $\mathcal{Z}_{q,N}(\theta)$) is determined by the Shannon entropy term.

In the limit, we therefore for fixed θ and q obtain

$$\lim_{N\to\infty}\Phi_q(\theta,N)=\max_{c\in[0,1]}H(c)=H(1/2) \tag{4.27}$$

where we used the fact that the Shannon binary entropy $H(c)$ is maximized for $c_{max}=\arg\max_c H(c)=\frac{1}{2}$. Therefore, we have

$$\lim_{N\to\infty}\frac{\ln\mathcal{Z}_{q,N}(\theta)}{N}=H(1/2) \quad\Leftrightarrow\quad \mathcal{Z}_{q,N}(\theta)\asymp e^{NH(1/2)}=2^N \tag{4.28}$$

where we use the notation of equality up to the first order of exponent defined as $a_n\asymp b_n \iff \lim_{N\to\infty}\frac{1}{N}\ln a_n=\lim_{N\to\infty}\frac{1}{N}\ln b_n$, similarly to [43].

Combining the bounds for the probability distribution (eq. 4.20) and bounds for the partition function (eq. 4.22 and eq. 4.24), we get

$$\begin{aligned}-2\frac{\ln(N+1)}{N}&+\phi_q(c,\theta,N)-\phi_q(c_{max},\theta,N)\le\\ &\le\frac{1}{N}\ln P(\bar{a}_i=c)\le\phi_q(c,\theta,N)-\phi_q(c_{max},\theta,N)+\frac{\ln(N+1)}{N}\end{aligned} \tag{4.29}$$

We saw, that $\lim_{N\to\infty}\phi_q(c,\theta,N)=H(c)$ and $\lim_{N\to\infty}\phi_q(c_{max},\theta,N)=H(1/2)$, therefore

$$\lim_{N\to\infty}\frac{1}{N}\ln P(\bar{a}_i=c)=H(c)-H(1/2) \quad\Leftrightarrow\quad P(\bar{a}_i=c)\asymp e^{-N(H(1/2)-H(c))} \tag{4.30}$$

This means that the probability distribution is maximal for the link density $c=1/2$ and decays exponentially for other values of c. This behavior is called *large deviation principle* [43]. The function $I(c):=H(1/2)-H(c)$ is called *rate function* and determines the rate of exponential decay of the probability. We can see it has a unique minimum at $c=1/2$, where $I(1/2)=0$. This means that the states with $c=1/2$ are the only states, whose probability does not decay exponentially and all the probability is being concentrated around $c=1/2$ as N becomes large. This behavior is called *concentration of measure.*

[1] Note that to be precise, we should for each N find a maximum among all possible values of c, i.e. from the set $\{0,1/N,2/N,\dots,1\}$. However, for N large enough, one can show that the difference between the value of ϕ_q at the discrete maximum and the maximum among $c\in[0,1]$ is of the order of $1/N$, i.e. it vanishes in the limit $N\to\infty$ [42].

Comparison with the Shannonian case

In the Shannonian case, we have

$$P\left(\bar{a}_i = c\right) = \binom{N}{Nc} \frac{1}{\mathcal{Z}^N(\theta)} e^{-\theta N c} \tag{4.31}$$

Therefore, the effective free entropy density ϕ_1 function is $\phi_1(c,\theta) = H(c) - \theta c$. Notably, it does not depend on N and the contribution of the constraint term θc does not vanish. If we extremize with respect to c, we get one and only extremum at $c_{max} = \frac{1}{1+e^\theta}$. The probability then is

$$P\left(\bar{a}_i = c\right) \asymp e^{-N(\phi_1(c_{max},\theta)-\phi_1(c,\theta))} \tag{4.32}$$

Note that if we choose $\theta = 0$, we get $\phi_1(c,\theta,N) = H(c)$ and $c_{max} = 1/2$, which corresponds to our results for the Tsallis model. This means that in large N limit, we expect the Tsallis model to behave the same as Erdös-Rényi model with $\theta = 0$ or equivalently with a link probability $p = 1/2$.

How to overcome the degeneracy of the Tsallis model?

The whole large N behavior of the Tsallis model is encoded in the $\phi_q(c,\theta,N)$ function, as a result of the concentration of measure. Looking at its definition in eq. 4.19, we see, that to overcome the vanishing energy term, we have two options:

- $\theta \leftarrow e^{\theta N}$: If θ grows exponentially with N, the logarithm will not vanish and we will get nontrivial contribution from the energy term. However, remember that in case of graphs, N is the number of all possible links, which is of the order of N_V^2, for N_V the number of vertices. Already for 100 nodes, we would have to set $\theta \approx e^{10^5}$, which is not numerically tractable.
- $\lim_{N\to+\infty}(1-q)N = \mathit{const.}$: If q behaves as $1/N$, the energy term does not vanish and we will obtain nontrivial contribution from the energy term.

We identify the second case as the more interesting one and we will study it later.

4.2.2 Means and covariances

Now let us study whether and how the Tsallis entropy introduces nontrivial correlations between edges. First, let us examine the average number of links. In Shannonian case we know that

$$\langle L \rangle = -\frac{\partial \ln \mathcal{Z}(\theta)}{\partial \theta} \tag{4.33}$$

This equation, however, does not hold in the Tsallis case. In appendix A, we show that

$$\langle L \rangle = N \exp_{2-q}(-\theta) \frac{\mathcal{Z}_{q,N-1}\left(\frac{\theta}{1-(q-1)\theta}\right)}{\mathcal{Z}_{q,N}(\theta)} \tag{4.34}$$

Note the occurence of $\mathcal{Z}_{q,N-1}\left(\frac{\theta}{1-(q-1)\theta}\right)$, partition function for system with size $N-1$ and modified parameter. Remarkably, this result holds even in the Shannonian case (setting $q = 1$). Therefore, we observe that the general relationship is not expressed through the derivative, but rather through this direct formula connecting the partition functions of different system sizes.

The marginal probability of single link is due to permutation symmetry of the probability just $\langle a_i \rangle = \langle L \rangle / N = \langle c \rangle \, \forall i$, i.e. it is the same as the average link density. We therefore have

$$\langle a_i \rangle = \langle c \rangle = \exp_{2-q}(-\theta) \frac{\mathcal{Z}_{q,N-1}\left(\frac{\theta}{1-(q-1)\theta}\right)}{\mathcal{Z}_{q,N}(\theta)} \tag{4.35}$$

Instead of computing $\langle a_i \rangle$ using $\langle L \rangle$, there is actually an elegant way how to compute it directly. This is because one can easily sum over other degrees of freedom. Indeed, we can write

$$\begin{aligned}
\langle a_i \rangle &= \frac{1}{\mathcal{Z}_{q,N}(\theta)} \sum_{\{a_j\}} a_i \exp_{2-q}\left(-\theta \sum_j a_j\right) = \frac{1}{\mathcal{Z}_{q,N}(\theta)} \sum_{a_i=0,1} a_i \sum_{\{a_j\}_{j\neq i}} \exp_{2-q}\left(-\theta \sum_j a_j\right) \\
&= \frac{1}{\mathcal{Z}_{q,N}(\theta)} \sum_{\{a_j\}_{j\neq i}} \exp_{2-q}\left(-\theta - \theta \sum_{j\neq i} a_j\right) \\
&\overset{(*)}{=} \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-\theta) \sum_{\{a_j\}_{j\neq i}} \exp_{2-q}\left(-\frac{\theta}{1-(q-1)\theta} \sum_{j\neq i} a_j\right) \\
&= \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-\theta) \mathcal{Z}_{q,N-1}\left(\frac{\theta}{1-(q-1)\theta}\right)
\end{aligned} \tag{4.36}$$

where in the $\overset{(*)}{=}$ equation, we utilized formula 4.6.

Similar approach can be used for computation of higher joint moments. Having k distinct indices $i_1 \neq \ldots \neq i_k$, we show in appendix B the following identity

$$\langle a_{i_1} \ldots a_{i_k} \rangle = \exp_{2-q}(-k\theta) \frac{\mathcal{Z}_{q,N-k}\left(\frac{\theta}{1-k(q-1)\theta}\right)}{\mathcal{Z}_{q,N}(\theta)} \tag{4.37}$$

Now the covariance between links $a_i, a_j, i \neq j$ is

$$Cov(a_i, a_j) = \langle a_i a_j \rangle - \langle a_i \rangle \langle a_j \rangle \tag{4.38}$$

Using eq. 4.37, we find that for two distinct indices $i \neq j$, we have

$$\langle a_i a_j \rangle = \exp_{2-q}(-2\theta) \frac{\mathcal{Z}_{q,N-2}\left(\frac{\theta}{1-2(q-1)\theta}\right)}{\mathcal{Z}_{q,N}(\theta)} \tag{4.39}$$

The relation with moments of number of links L is shown in appendix C. Finally, we can write the covariance as

$$Cov(a_i, a_j) = \exp_{2-q}(-2\theta) \frac{\mathcal{Z}_{q,N-2}\left(\frac{\theta}{1-2(q-1)\theta}\right)}{\mathcal{Z}_{q,N}(\theta)} - \left(\exp_{2-q}(-\theta) \frac{\mathcal{Z}_{q,N-1}\left(\frac{\theta}{1-(q-1)\theta}\right)}{\mathcal{Z}_{q,N}(\theta)}\right)^2 \tag{4.40}$$

The Shannonian case

In the Shannonian case, we computed the partition function in eq. 2.56 (here, we have N links instead of N^2). For the covariance, we therefore get

$$Cov(a_i, a_j) = \exp(-2\theta)\frac{(1+e^{-\theta})^{N-2}}{(1+e^{-\theta})^N} - \left(\exp(-\theta)\frac{(1+e^{-\theta})^{N-1}}{(1+e^{-\theta})^N}\right)^2 = 0 \tag{4.41}$$

Since we saw that in the Shannonian case, the links are independent, this is an expected result.

Perturbative expansion

The problem with the expression in eq. 4.40 is that it is hard to compute the partition function. Even though we know the asymptotic behavior of the partition function, we do not know its limit. Since we saw, that one interesting case is when q is close to one, we will analyze this limit perturbatively.

The q-exponential can be expanded in $(q-1)$ as

$$\exp_{2-q}(x) = e^x\left(1 - \frac{1}{2}(q-1)x^2 + \frac{1}{3}(q-1)^2x^3\left(1+\frac{3}{8}x\right) \ + \ \mathcal{O}((q-1)^3)\right) \tag{4.42}$$

Since the q-exponential is expanded as an exponential multiplied by a correction polynomial, let us defined the residual function as

$$\mathrm{res}_q(x) = \frac{\exp_{2-q}(x)}{e^x} \tag{4.43}$$

Then the partition function can be written as

$$\begin{aligned}\mathcal{Z}_{q,N}(\theta) &= \sum_{G\in\mathcal{G}_N}\exp_{2-q}(-\theta L(G)) = \sum_{G\in\mathcal{G}_N} e^{-\theta L(G)}\mathrm{res}_{2-q}(-\theta L(G)) &(4.44)\\ &= \mathcal{Z}_{1,N}(\theta)\sum_{G\in\mathcal{G}_N}\frac{e^{-\theta L(G)}}{\mathcal{Z}_{1,N}(\theta)}\mathrm{res}_{2-q}(-\theta L(G)) = \mathcal{Z}_{1,N}(\theta)\langle\mathrm{res}_{2-q}(-\theta L(G))\rangle_{\mathrm{Sh}} &(4.45)\end{aligned}$$

where $\langle\ldots\rangle_{\mathrm{Sh}}$ denotes the average using the probability given by the Shannonian model and $\mathcal{Z}_{1,N}(\theta) = (1+e^{-\theta})^N$ is the partition function of the Shannonian model. Note that this approach is equivalent to perturbation method, studied in the original paper on statistical mechanics of networks [27]. Expanding the residual function in $(q-1)$, we can write

$$\begin{aligned}\mathcal{Z}_{q,N}(\theta) = \mathcal{Z}_{1,N}(\theta)\Big(1 &- \frac{1}{2}(q-1)\theta^2\langle L^2\rangle_{\mathrm{Sh}}\\ &- \frac{1}{3}(q-1)^2\theta^3\left(\langle L^3\rangle_{\mathrm{Sh}} - \frac{3}{8}\theta\langle L^4\rangle_{\mathrm{Sh}}\right) + \mathcal{O}((q-1)^3)\Big)\end{aligned} \tag{4.46}$$

To compute higher orders of the partition function, we need to compute terms $\langle L^k\rangle_{\mathrm{Sh}}$. However, in the Shannonian case, each link is a Bernoulli random variable with $p = 1/(1+e^\theta)$, and therefore $L \sim B(N, 1/(1+e^\theta))$, i.e. L is a binomial random variable with $p = 1/(1+e^\theta)$. The calculation of higher moments for the binomial distribution is well-documented in statistical literature.

Using eq. 4.35 and the expansion of the partition function, we show in appendix D, that to the first order in $(q-1)$, we have

$$\langle a_i \rangle = \frac{1}{1+e^\theta}\left[1+(1-q)\frac{1}{2}\frac{\theta^2}{(1+e^\theta)^2}e^\theta(2N-1+e^\theta)+\mathcal{O}((q-1)^2)\right] \tag{4.47}$$

$$\langle a_i a_j \rangle = \frac{1}{(1+e^\theta)^2}\left[1+(1-q)\frac{\theta^2}{(1+e^\theta)^2}e^\theta(2N-1+2e^\theta)+\mathcal{O}((q-1)^2)\right] \tag{4.48}$$

Therefore, the covariance is

$$Cov(a_i,a_j) = \langle a_i a_j \rangle - \langle a_i \rangle \langle a_j \rangle = (1-q)\frac{\theta^2 e^{2\theta}}{(1+e^\theta)^4}+\mathcal{O}((q-1)^2) \tag{4.49}$$

Remarkably, with q further from 1, we get positive correlations between links. Also, the covariance is to the first order in $(q-1)$ independent of the network size N. However, our experiments did not prove the system size independence for higher orders.

4.2.3 Saddle point equation

In equation 4.35, we have found an expression for the average link density. This expression is rigorous and holds for all values of N. However, we have also seen that with N large, the probability distribution is concentrated around the state with maximal free entropy density $\Phi_q(\theta, N)$. The typical approach in physics is then to replace the average by value at the maximum configuration, which is the so-called *saddle point approximation.*

To that end, it is important to be able to find the maximum configuration c_{max}. Let us first define a new parameter $r=(1-q)N$. With such reparametrization, we can write the effective free entropy density as

$$\phi_r(c,\theta) = H(c) - \frac{1}{r}\ln(1+r\theta c) \tag{4.50}$$

To find the maxima, let us compute the derivative of $\phi_r(c,\theta)$ with respect to c is

$$\frac{\partial \phi_r}{\partial c} = \ln\left(\frac{1-c}{c}\right) - \frac{\theta}{1+r\theta c} \tag{4.51}$$

Setting to zero, we obtain the *saddle point equation*

$$c = \frac{1}{1+\exp\left(\frac{\theta}{1+r\theta c}\right)} \tag{4.52}$$

We have seen that in Shannonian case, there was only one solution at $c=\frac{1}{1+e^\theta}$. For $q \neq 1$, the situation can differ. See figure 4.1, where we plot the effective free entropy density $\phi_r(c,\theta)$ for $r=2.7$ for $\theta \in \{4, 5, 5.5\}$. For $\theta = 4$, we have a unique maximum, however, for $\theta = 5$ and $\theta = 5.5$, we have two maxima. Also, we can see that the location of the global maxima changes abruptly between $\theta = 5$ and $\theta = 5.5$, which is a signature of phase transition.

To see, whether a phase transition actually occurs, let us compute the local maxima and determine the global maximum, depending on θ. On figure 4.2 on the right, we show the locations of the local maxima as a function of θ. For $r = 2.7$, we

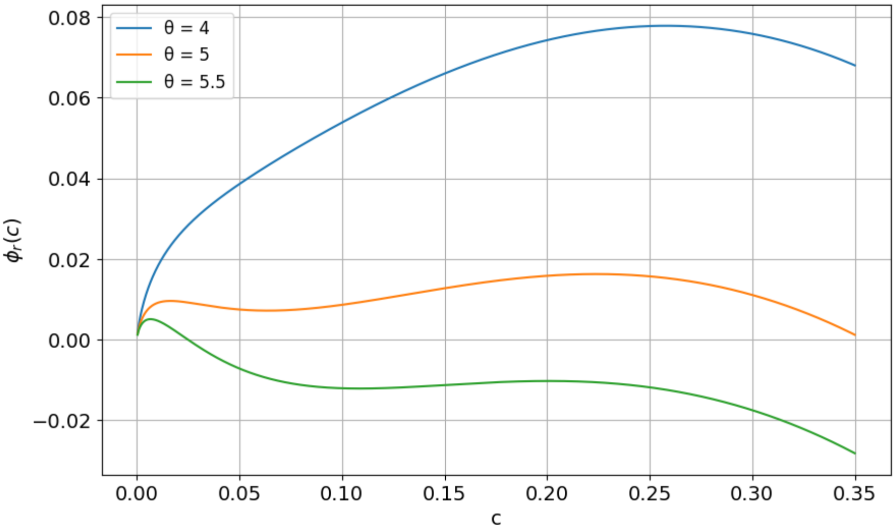


Figure 4.1: The effective free entropy density $\phi_r(c, \theta)$ for $r = 2.7$ as a function of link density c for different values of θ. The global maxima of these curves determine the typical configuration.

find that in range $\theta \in [4.831, 5.846]$ there are two local maxima c_{low} and c_{high}, where one is stable and one metastable. $\theta_{sp-} = 4.831$ and $\theta_{sp+} = 5.846$ are the so-called *spinodal points*. The phase transition happens at $\theta_{crit} = 5.154$, below which the high density maximum is stable and above which the low density maximum is stable. On the left, we show situation for $r = 2$, where no phase transition occurs. We also show the Shannonian network density $c_{Shannon} = \frac{1}{1+e^\theta}$ for reference.

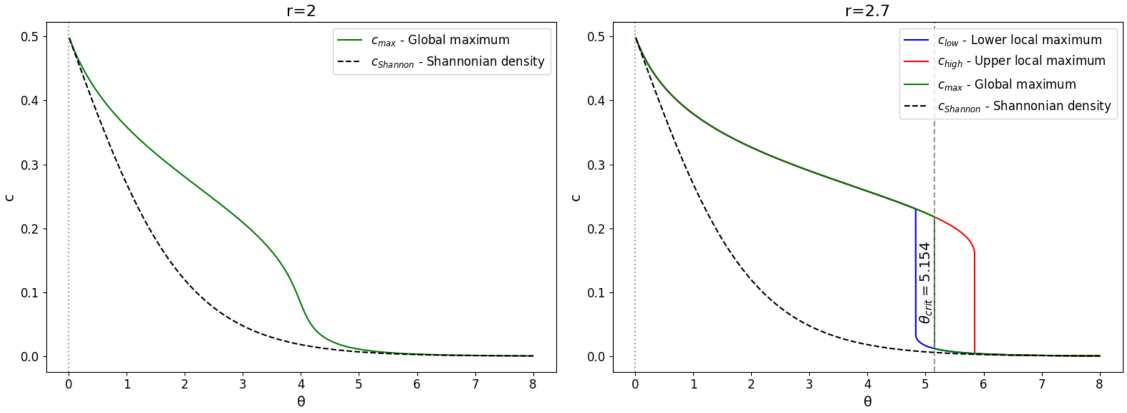


Figure 4.2: On the left, we show the behavior of the typical configuration link density c_{max} for $r = 2$ as a function of θ. On the right, for $r = 2.7$, we plot the local and global maxima of the effective free entropy density $\phi_r(c, \theta)$ as a function of θ. In the range $\theta \in [4.831, 5.846]$, there are two local maxima c_{low} and c_{high}, where one is stable and one metastable. The phase transition happens at $\theta_{crit} = 5.154$. For reference, we also show the Shannonian network density $c_{Shannon} = \frac{1}{1+e^\theta}$.

4.2.4 Phase transition

To understand the aforementioned phase transition, let us study the saddle point equation 4.52 in more detail. The right-hand side can be rewritten as

$$R(c) = \left(1 + \exp\left(\frac{1}{\frac{1}{\theta} + rc}\right)\right)^{-1} \tag{4.53}$$

It is bounded between 0 and 1 and is discontinuous at $c = -\frac{1}{r\theta}$, which is negative, since $r, \theta > 0$. It can be easily seen, that on the interval $[-\frac{1}{r\theta}, +\infty)$, $R(c)$ is an increasing differentiable function of c approaching $\frac{1}{2}$ for $c \to +\infty$. As such, the only possibility for $R(c)$ to have more than one intersection with the left-hand side $L(c) = c$ is that the derivative of $R(c)$ with respect to c must be larger than 1 (since the derivative of $L(c)$ is 1) on some interval. And if this condition is satisfied, we will be able to find such θ, for which $R(c)$ will have more than one intersection with $L(c)$. For an illustration, see figure 4.3.

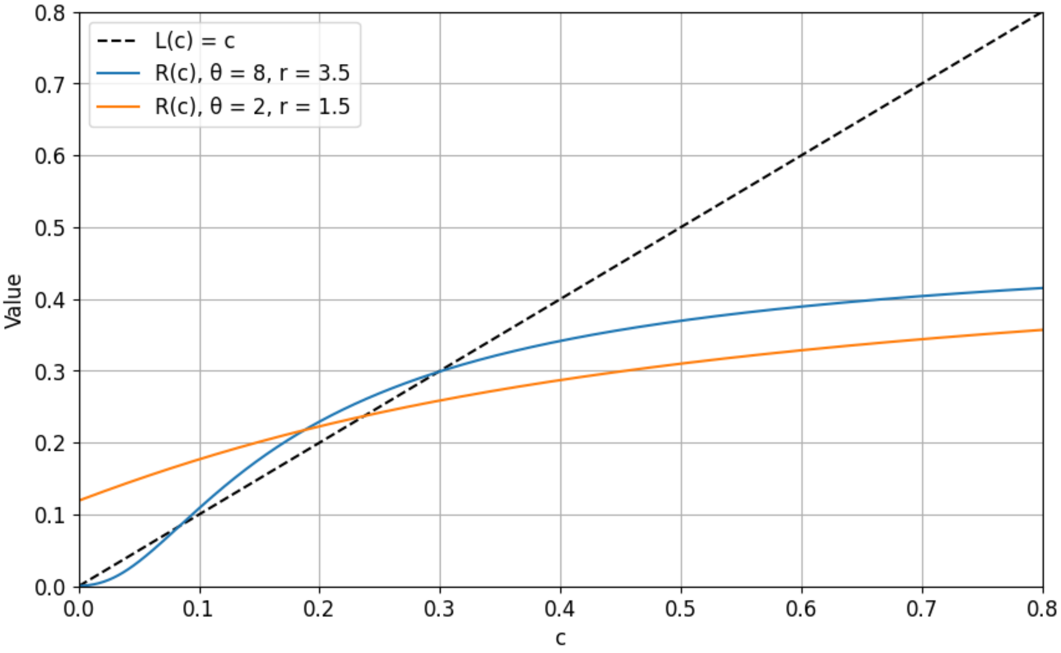


Figure 4.3: The left-hand side $L(c) = c$ and the right-hand side $R(c)$ (eq. 4.53) of the saddle point equation 4.52 for two distinct combinations of r and θ. We can see that to have more than one intersection (solutions to the saddle point equation), the necessary condition is that the derivative of $R(c)$ must be larger than 1 on some interval.

To examine this condition, let us realize that in eq. 4.53, θ plays a role of shift along the c-axis and therefore it does not change the shape of the function $R(c)$. So this condition is purely determined by the value of r and it suffices to study the condition:

$$\frac{\partial}{\partial c}\left(1 + \exp\left(\frac{1}{rc}\right)\right)^{-1} > 1 \tag{4.54}$$

In appendix E, we prove the following proposition:

Proposition 4.2.1. Let x^* be a positive solution of equation

$$\tanh\left(\frac{1}{x^*}\right) = x^* \tag{4.55}$$

Then the phase transition in Tsallis-Erdős-Rényi model can occur if and only if $r \equiv (1-q)N > \frac{(x^*)^2}{1-(x^*)^2} \approx 2.2767$.

Let us now characterize the critical point. The r value is given by $r_{cp} = \frac{(x^*)^2}{1-(x^*)^2} \approx 2.2768$. The necessary condition for the phase transition is, that the effective free energy density $\phi_r(c,\theta)$ is convex on some interval. The critical value θ_{cp} will be found as the first value of θ, for which $\phi_{r_{cp}}(c,\theta)$ has an inflection point. In appendix F, we show the following proposition:

Proposition 4.2.2. The critical point θ_{cp} is given by

$$\theta_{cp} = 2 + \sqrt{\frac{4(r_{cp}+1)}{r_{cp}}} \approx 4.3994 \tag{4.56}$$

To better characterize the phase transition, let us find the phase diagram. We solve the saddle point equation 4.52 numerically different values of r, and for each of them we find the spinodal points θ_{sp-} and θ_{sp+} as well as the phase transition boundary θ_{crit}. We also show the computed critical point $(r_{cp}, \theta_{cp}) \approx (2.2767, 4.3994)$. The resulting phase diagram is shown on figure 4.4.

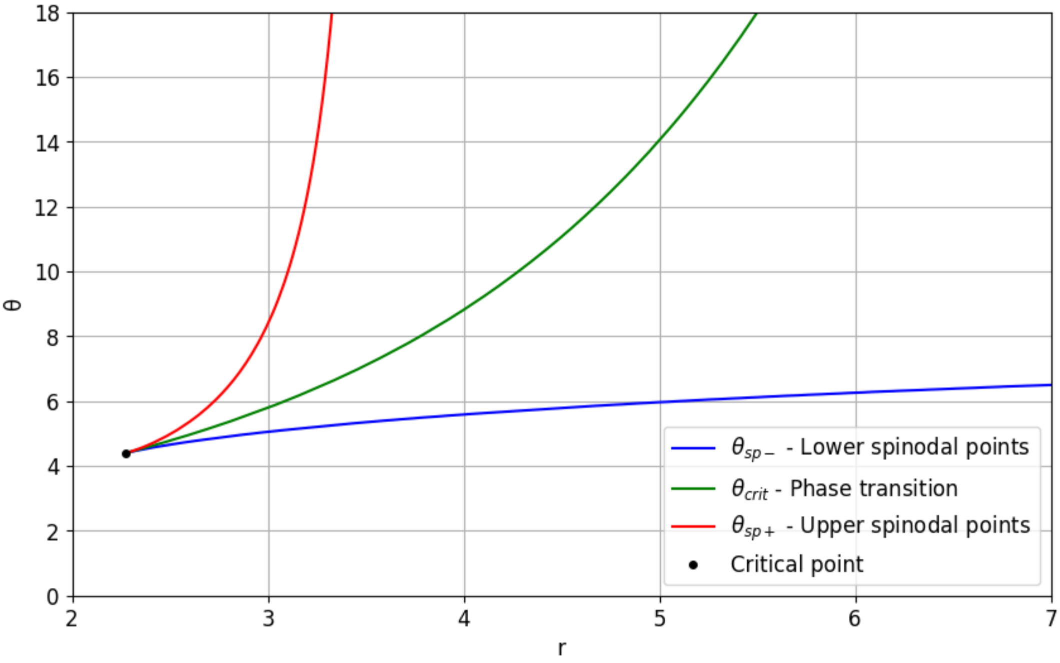


Figure 4.4: The phase diagram of the Tsallis-Erdős-Rényi model. For each value of r, we show the spinodal points θ_{sp-} and θ_{sp+} as well as the phase transition θ_{crit}. The critical point $(r_{cp}, \theta_{cp}) \approx (2.2767, 4.3994)$ is shown as the black dot.

Finally, we shall note that the phase transition has an important consequence on applications like network reconstruction. Specifically, there might be no combination of r and θ, which would be able to reproduce given target network density. In the metastable regime, we have two phases, one with lower density c_{low} and one with higher density c_{high}. Then it is not possible to reconstruct densities between the lowest possible value of c_{high} and highest possible value of c_{low}. This can be already seen on figure 4.2. For different values of r, we numerically computed these boundaries and highlighted the region, where the density reconstruction is not possible.

The results are shown on figure 4.5. We assume that the irreconstructible region grows in size and in the limit only allows either $c = 0$ or $c = 0.5$.

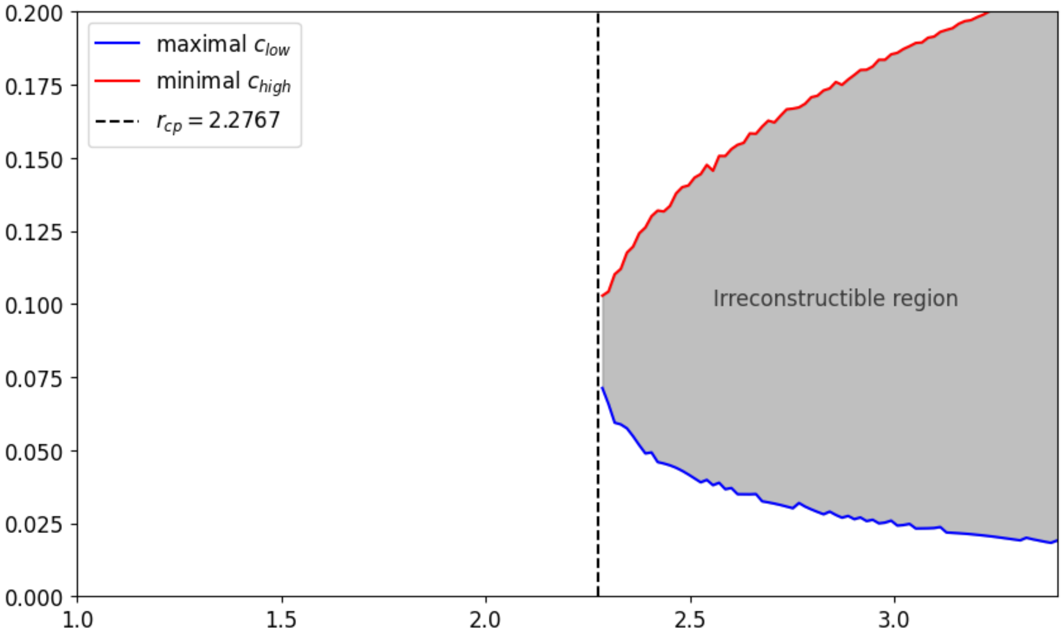


Figure 4.5: For different values of r, we numerically approximated boundaries of the region, where density reconstruction is not possible. Such a region is determined by the lowest possible value of c_{high} and the highest possible value of c_{low} in the metastable regime. We also show r value of the critical point, $r_{cp} = 2.2767$.

Note on the order of limits

Phase transitions are usually discussed in terms of non-analytical behavior of the free energy or free entropy. However, the non-analyticity only appears in the thermodynamic limit $N \to \infty$. Consider again the equation 4.25. The non-analyticity is caused by the maximization over c in the limit $N \to \infty$. However, we also claimed that for $N \to \infty$, the model converges to a Shannonian case with $\theta = 0$, where no phase transition occurs.

This seeming contradiction is due to the fact, that we imposed $r = (1 - q)N$ and consider it fixed. This, however, imposes a different characteristics of the limits. If we first perform the limit $N \to \infty$ with q fixed, we indeed obtain the degenerated Shannonian model, where the influece of the parameter θ diminishes, as discussed. Performing the $q \to 1$ limit after that does not change this fact. On the other hand, if we first performed the $q \to 1$ limit, we would obtain obtain the original Erdős-Rényi model. This noncommutative behavior of limits was already noticed by Hanel and Thurner (section 3.3), where we have seen, that Tsallis entropy belongs to $(c, d) = (q, 0)$ class but Shannon entropy belongs to $(c, d) = (1, 1)$ class, i.e. they have different scaling behavior in the $N \to \infty$ limit.

Our claim is that performing the $N \to \infty$ limit while keeping $(1 - q)N$ fixed gives interesting results, and only under such conditions we obtain the phase transition. On the other hand, we claim that even for finite N, the phase transition will have observable effects, like the existence of metastable states. Especially in case of networks, we usually consider N (number of all possible links) fixed but large, so we assume that the saddle-point approximation is going to be valid.

Now let us study, what implications $r = (1-q)N$ has on the entropy. One could define the Tsallis entropy using the parameter r as

$$S_r(P) = \frac{N}{r}\left(1 - \sum_{i=1}^{N} p_i^{r/N+1}\right) \tag{4.57}$$

Considering the microcanonical ensemble with W states, each with probability $p_i = \frac{1}{W}$, we obtain

$$S_r\left(\frac{1}{W}, \dots, \frac{1}{W}\right) = \frac{N}{r}\left(1 - W^{-r/N}\right) \tag{4.58}$$

If we assume exponential scaling of the state space $W \propto e^N$, we obtain

$$S_r\left(\frac{1}{W}, \dots, \frac{1}{W}\right) \propto N\frac{1-e^{-r}}{r} \propto N \tag{4.59}$$

meaning, that the entropy is extensive! So by the specific choice of limiting behavior $\lim_{N\to\infty}(1-q)N = \mathit{const.}$, we actually made the Tsallis entropy extensive again. However, remarkably, the behavior is much richer than in the Shannonian case, which is extensive as well.

4.2.5 Network properties

Let us now return back to our original model

$$P_q(G) = \frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-\theta L(G)\right) = \frac{1}{\mathcal{Z}_q}\exp_{2-q}\left(-\theta\sum_{i,j=1}^{N} a_{ij}\right) \tag{4.60}$$

and study its network properties. Since the probability does not factorize, we can not sample the links independently. This means we always have to sample the whole graph. We choose to use Metropolis-Hastings algorithm [44], because of its simplicity. Especially, we do not have to compute the partition function in it.

The Metropolis-Hastings algorithm is a Monte Carlo Markov chain method, which generates a Markov chain with a given stationary distribution. Given a current configuration graph G_t, we propose a new graph G' using a proposal distribution $Q(G'|G_t)$. Then we accept the new graph with probability

$$\alpha = \min\left(1, \frac{P_q(G')}{P_q(G_t)}\frac{Q(G_t|G')}{Q(G'|G_t)}\right) \tag{4.61}$$

or keep the current graph with probability $1-\alpha$. A sufficient condition for the algorithm to converge to the correct stationary distribution is one can reach any graph from any other graph in a finite number of steps using the proposal distribution.

Our proposal distribution is to flip a uniformly randomly chosen link. Such a proposal is symmetric, i.e. $Q(G'|G) = Q(G|G')$. This simplifies the acceptance probability to

$$\alpha = \min\left(1, \frac{P_q(G')}{P_q(G_t)}\right) = \min\left(1, \frac{\exp_{2-q}\left(-\theta L(G')\right)}{\exp_{2-q}\left(-\theta L(G_t)\right)}\right) \tag{4.62}$$

Note that the acceptance probability depends only on a ratio of the graph probabilities and therefore the partition function cancels out.

Using the Metropolis-Hastings algorithm, we will generate ensembles of networks. Since we always flip one edge at time, we will always perform numerous steps (of the order of number of all possible edges), before we append next graph to the ensemble. Like this, we hope to avoid the auto-correlations caused by the sampling.

Number of links, link density

For the link density, we have found two ways of approximating it. One was the perturbative approach (eq. 4.47), the other one was solving the saddle-point equation (eq. 4.52). On figure 4.6, we compare these two approaches. For 10 values of $(1-q)$ evenly spaced between 10^{-6} and 10^{-3} and for $\theta \in \{1, 3, 5\}$, we generated ensembles of 1000 graphs with 100 nodes, computed their link densities and compared with the theoretical results. We can see that the linear perturbative approximation looses relevance rather quickly, on the other hand, the densities obtained using the saddle-point equation follow the experimental results perfectly. We also see that with q getting further from 1, the link density for all the values of θ converge to 0.5, as we predicted.

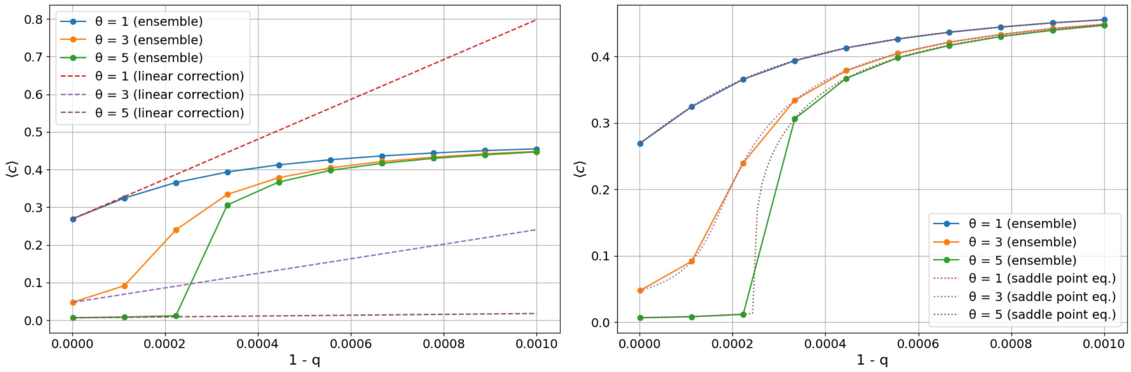
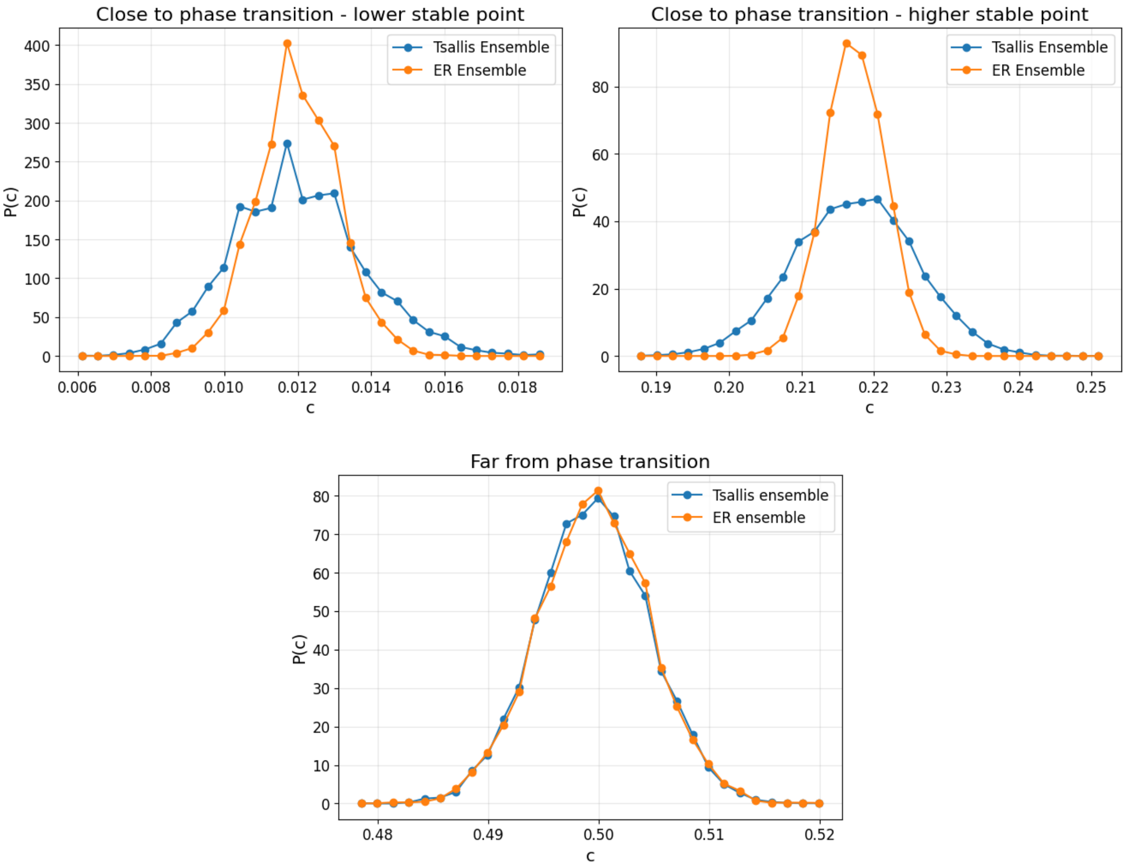


Figure 4.6: For ensemble of 1000 graphs with 100 nodes, we compute the average link density $\langle c \rangle$ and compare with the perturbative linear correction (eq. 4.47, on the left) and with the solutions of the saddle-point equation (eq. 4.52, on the right).

Let us also study, how the link densities are distributed among the ensemble. We study two situations: close to the phase transition and far from it. We will take the situation we have already seen: for $r = (1-q)N = 2.7$ we have found the phase transition happening at $\theta_{crit} = 5.154$ and the corresponding solutions of the saddle-point equations are $c_{low} = 0.0118$ and $c_{high} = 0.2177$. Let us therefore create two ensembles, each with a different initial condition. First ensemble will start with a graph with density $c = 0.01$ and the other one with $c = 0.3$. We generate 10000 graphs for each ensemble and plot the histogram of the link densities. For comparison, we also generate an ensemble with $r = 1000$ and $\theta = 1$. Finally, we generate Shannonian ensembles, each fitted to reproduce the average link densities of the Tsallis ensembles.

The results are shown on figure 4.7. The two initial conditions actually converged to two different fixed-points of the saddle-point equation, as we can see from the different distributions. The measured means are $\langle c \rangle_{low} = 0.0120$ and $\langle c \rangle_{high} = 0.2173$, which corresponds well with the theoretical values. Interestingly, for both ensembles near the phase transition, the distributions are much wider for the Tsallis ensembles than for the Shannonian ones. The measured standard deviations are $\sigma_{low} = 0.0018$ and $\sigma_{high} = 0.0011$ for the Tsallis ensembles and $\sigma_{low} = 0.0011$ and $\sigma_{high} = 0.0041$ for the Shannonian ensembles. On the other hand, the distributions for the ensemble far from the critical point are almost identical, both with $\langle c \rangle_{low} = 0.4995$ and $\sigma = 0.005$. This confirms our prediction, that for q far from 1, the model will be identical to the Erdős-Rényi model with link probability 0.5, while with q close to 1, the models differ significantly.


Close to phase transition - lower stable point

Close to phase transition - higher stable point

Far from phase transition

Figure 4.7: Comparison of distributions link densities for two situations: close to the phase transition (top two graphs) with $r = 2.7$ and $\theta = 5.154$ and far from it (bottom graph) with $r = 1000$ and $\theta = 1$. We also plot the corresponding Erdős-Rényi ensembles (fitted to reproduce the same average link density). For the top two graphs, we used two different initial conditions and converged to two different fixed-points of the saddle-point equation. Used ensembles of 10000 graphs with 100 nodes.

ANND, clustering

Let us compare other network properties as well. We report that we haven't found significant differences in degree distributions between the Tsallis and the Shannonian ensembles. However, for the average nearest neighbor degree and the clustering coefficient, we found an interesting difference between the Tsallis model close to the phase transition and far from it. See figure 4.8, where we compare the same ensembles as in figure 4.7 and compute the average ANND and clustering coefficient depending on node degree[2]. For the average nearest neighbor degree, we found that the Tsallis model develops positive assortativity and increasing clustering with degree. In the case $r = 1000$, $\theta = 1$, the model is close to the Shannonian case and the both assortativity and clustering are constant (as expected for the Erdős-Rényi model).

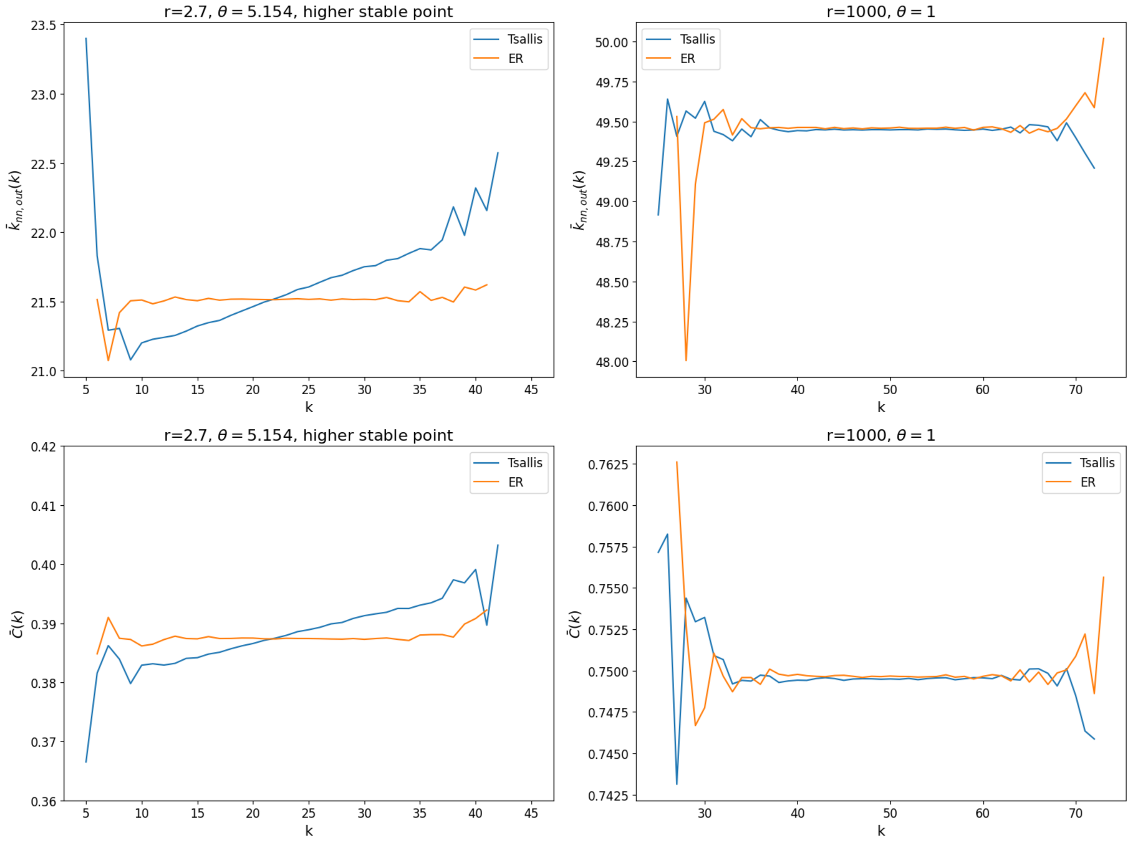


Figure 4.8: Comparison of average nearest neighbor out-degree and clustering coefficient (of the undirected projections) depending on node degree for the Tsallis ensembles close to the phase transition (left) and far from it (right). Interestingly, near the phase transition, the Tsallis model develops positive assortativity and increasing clustering with degree. For extreme values of degrees, there are less available data and therefore the computation is more noisy. Used the same ensembles as for figure 4.7.

4.3 Tsallis-Park-Newman model

Let us now study the generalization of the Park-Newman model using the Tsallis entropy. This time we consider undirected graphs with no self-loops. Park-Newman model corresponds to constraining the degree of each node, and the corresponding Hamiltonian is (see eq. 2.49)

$$H = \sum_{i<j} (\theta_i + \theta_j) a_{ij} = \sum_i \theta_i k_i(G) \tag{4.63}$$

[2] We first compute the ANND and clustering coefficient for each node in each graph. Then among all graphs in the ensemble, we group nodes with a same degree and compute the average ANND or clustering coefficient of the group.

The cgraph probability is then

$$P_q(G) = \frac{1}{\mathcal{Z}_q(\theta)} \exp_{2-q}\left(-\sum_{i<j} (\theta_i + \theta_j) a_{ij} \right) \tag{4.64}$$

with the partition function

$$\mathcal{Z}_q(\theta_1, \ldots, \theta_N) = \sum_{\{a_{ij}\}} \exp_{2-q}\left(-\sum_{i<j} (\theta_i + \theta_j) a_{ij} \right) \tag{4.65}$$

4.3.1 Average node degree

In the Shannonian case, the average node degree is according to eq. 2.50 given by

$$\langle k_i \rangle = \sum_{j \neq i} \langle a_{ij} \rangle = \sum_{j \neq i} p_{ij} = \sum_{j \neq i} \frac{1}{1 + e^{\theta_i + \theta_j}} \tag{4.66}$$

Let us therefore compute $\langle a_{kl} \rangle$ for the Tsallis-Park-Newman model.

$$\begin{aligned}
\langle a_{kl} \rangle \mathcal{Z}_q(\theta_1, \ldots, \theta_N) &= \sum_{\{a_{ij}\}} a_{kl} \exp_{2-q}\left(-\sum_{i<j} (\theta_i + \theta_j) a_{ij} \right) \\
&= \sum_{a_{kl}=0,1} a_{kl} \sum_{\substack{\{a_{ij}\} \\ (i,j) \neq (k,l)}} \exp_{2-q}\left(-(\theta_k + \theta_l) a_{kl} - \sum_{\substack{i<j \\ (i,j) \neq (k,l)}} (\theta_i + \theta_j) a_{ij} \right) \\
&= \sum_{\substack{\{a_{ij}\} \\ (i,j) \neq (k,l)}} \exp_{2-q}\left(-(\theta_k + \theta_l) - \sum_{\substack{i<j \\ (i,j) \neq (k,l)}} (\theta_i + \theta_j) a_{ij} \right) \\
&\overset{(*)}{=} \exp_{2-q}(-(\theta_k + \theta_l)) \sum_{\substack{\{a_{ij}\} \\ (i,j) \neq (k,l)}} \exp_{2-q}\left(- \sum_{\substack{i<j \\ (i,j) \neq (k,l)}} \frac{\theta_i + \theta_j}{1 - (q-1)(\theta_k + \theta_l)} a_{ij} \right) \\
&=: \exp_{2-q}(-(\theta_k + \theta_l)) \tilde{\mathcal{Z}}_q^{(k,l)}(\tilde{\theta}_1^{(k,l)}, \ldots, \tilde{\theta}_N^{(k,l)})
\end{aligned} \tag{4.67}$$

where we defined the partition function with the link (k, l) not present, and the new parameters as

$$\tilde{\mathcal{Z}}_q^{(k,l)}(\theta_1, \ldots, \theta_N) := \sum_{\substack{\{a_{ij}\} \\ (i,j) \neq (k,l)}} \exp_{2-q}\left(- \sum_{\substack{i<j \\ (i,j) \neq (k,l)}} (\theta_i + \theta_j) a_{ij} \right) \tag{4.68}$$

$$\tilde{\theta}_i := \frac{\theta_i}{1 - (q-1)(\theta_k + \theta_l)} \tag{4.69}$$

Then we obtain

$$\langle a_{kl} \rangle = \exp_{2-q}\left(-(\theta_k + \theta_l) \right) \frac{\tilde{\mathcal{Z}}_q^{(k,l)}(\tilde{\theta}_1, \ldots, \tilde{\theta}_N)}{\mathcal{Z}_q(\theta_1, \ldots, \theta_N)} \tag{4.70}$$

which reminds us of the expression we got in Tsallis-Erdős-Rényi model, eq. 4.35. Although we would like to further simplify the numerator $\tilde{\mathcal{Z}}_q^{(k,l)}(\tilde{\theta}_1, \dots, \tilde{\theta}_N)$, since the q-exponential of sum does not factorize in a simple way, this is probably as far as we can get.

The expression 4.70 would once again allow for a perturbative treatment, similar to the one we did in section 4.2.2. We will be, however, rather interested in the behavior of network properties and compute them numerically in the following section.

4.3.2 Network properties

Similarly to the Tsallis-Erdős-Rényi model (section 4.2.5), we will use the Metropolis-Hastings algorithm to sample the graph. This time, the acceptance probability is given by

$$\alpha = \min\left(1, \frac{P_q(G')}{P_q(G_t)}\right) = \min\left(1, \frac{\exp_{2-q}\left(-\sum_i \theta_i k_i(G')\right)}{\exp_{2-q}\left(-\sum_i \theta_i k_i(G_t)\right)}\right) \tag{4.71}$$

There is an unlimited number of experiments one could do, as we have freedom in choice of the model parameters. Let us consider one particular example. Gabrielli et al. [45] studied the interbank-loan networks and used the model, where edges are sampled independently with probability:

$$p_{ij} = \frac{z s_i s_j}{1 + z s_i s_j} \tag{4.72}$$

Here, s_i are the so-called strength of the network to be reconstructed, and parameter z tunes the overall density. Note that this is exactly the Park-Newman model, as seen in eq. 2.50, with $s_i = x_i$ and $z = \beta$. They found, that the interbank-loan networks are fit well by strengths sampled from log-normal distribution with $\sigma = 2.28$, $\mu = -\sigma^2/2$ and z fitted according to the target network density. We saw, that the relationship to the parameters θ_i is given by $\sqrt{\beta} x_i = e^{-\theta_i}$. This means we sample θ_i from normal distribution with $\sigma = 2.28$ and $\mu = \sigma^2/2$. Since parameter β is multiplicative, in terms of θ_i, we have to find an additive shift denoted by θ_{shift}, which will recover the target density of the reconstructed network.

Our goal now is to see, whether the Tsallis model gives significant corrections in terms of network properties and whether we will also observe signs of a phase transition.

Conclusion

TODO

Bibliography

1. BARABÁSI, Albert-László. *Network science*. Ed. by PÓSFAI, Márton. Cambridge: Cambridge University Press, 2016. Hier auch später erschienene, unveränderte Nachdrucke.
2. SQUARTINI, Tiziano; GARLASCHELLI, Diego. *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics*. Springer International Publishing, 2017.
3. NEWMAN, Mark E. J. *The Structure and Dynamics of Networks*. Ed. by WATTS, Duncan J.; BARABÁSI, Albert-László. Princeton: Princeton University Press, 2006. Princeton Studies in Complexity Ser, no. v.19.
4. NEWMAN, M. E. J. The Structure and Function of Complex Networks. *SIAM Review*. 2003, vol. 45, no. 2, pp. 167–256.
5. PASTOR-SATORRAS, Romualdo; VÁZQUEZ, Alexei; VESPIGNANI, Alessandro. Dynamical and Correlation Properties of the Internet. *Physical Review Letters*. 2001, vol. 87, no. 25, p. 258701.
6. SERRANO, Ma Ángeles; BOGUÑÁ, Marián. Topology of the world trade web. *Physical Review E*. 2003, vol. 68, no. 1, p. 015101.
7. RAVASZ, Erzsébet; BARABÁSI, Albert-László. Hierarchical organization in complex networks. *Physical Review E*. 2003, vol. 67, no. 2, p. 026112.
8. ALBERT, Réka; BARABÁSI, Albert-László. Statistical mechanics of complex networks. *Reviews of Modern Physics*. 2002, vol. 74, no. 1, pp. 47–97.
9. CIMINI, Giulio; SQUARTINI, Tiziano; SARACCO, Fabio; GARLASCHELLI, Diego; GABRIELLI, Andrea; CALDARELLI, Guido. The statistical physics of real-world networks. *Nature Reviews Physics*. 2019, vol. 1, no. 1, pp. 58–71.
10. ERDŐS, Paul; RÉNYI, Alfréd. On random graphs. *Publicationes Mathematicae Debrecen*. 1959, vol. 6, pp. 290–297.
11. BARABÁSI, Albert-László; ALBERT, Réka. Emergence of Scaling in Random Networks. *Science*. 1999, vol. 286, no. 5439, pp. 509–512.
12. HOFSTAD, Remco van der. *Random Graphs and Complex Networks*. Cambridge University Press, 2016.
13. COOLEN, A. C. C.; DE MARTINO, A.; ANNIBALE, A. Constrained Markovian Dynamics of Random Graphs. *Journal of Statistical Physics*. 2009, vol. 136, no. 6, pp. 1035–1067.
14. PARK, Juyong; NEWMAN, M. E. J. Origin of degree correlations in the Internet and other networks. *Physical Review E*. 2003, vol. 68, no. 2, p. 026112.

15. GARLASCHELLI, Diego; LOFFREDO, Maria I. Fitness-Dependent Topological Properties of the World Trade Web. *Physical Review Letters.* 2004, vol. 93, no. 18, p. 188701.

16. THURNER, Stefan. *Introduction to the theory of complex systems.* First edition. Ed. by KLIMEK, Peter; HANEL, R. A. Oxford: Oxford University Press, 2018. Includes bibliographical references (pages 407-424) and index.

17. KITTEL, Charles. *Elementary statistical physics.* Mineola, NY: Dover Publ., 2004. Dover books on physics. Originally published: New York : Wiley, 1958 - Includes index.

18. MACKAY, David J. C. *Information theory, inference, and learning algorithms.* 22nd printing. Cambridge [u.a.]: Cambridge University Press, 2019.

19. COVER, Thomas M. *Elements of information theory.* Second edition. Ed. by THOMAS, Joy A. Hoboken, N.J: Wiley-Interscience, 2006. Includes bibliographical references (pages 689-721) and index.

20. PRIVAULT, Nicolas. *Understanding Markov chains: Examples and applications.* Second edition. Singapore: Springer Nature, 2018. Springer undergraduate mathematics series.

21. KHINCHIN, Aleksandr Ja. *Mathematical foundations of information theory.* Dover ed., new transl. New York: Dover Publ., 1970. Dover books on intermediate and advanced mathematics. Enth.: The entropy concept in probability theory. - On the fundamental theorems of information theory. - Aus dem Russ. übers.

22. JIZBA, Petr; KORBEL, Jan. When Shannon and Khinchin meet Shore and Johnson: Equivalence of information theory and statistical inference axiomatics. *Physical Review E.* 2020, vol. 101, no. 4, p. 042126.

23. SHANNON, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal.* 1948, vol. 27, no. 3, pp. 379–423.

24. JAYNES, E. T. Information Theory and Statistical Mechanics. *Physical Review.* 1957, vol. 106, no. 4, pp. 620–630.

25. SHORE, J.; JOHNSON, R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory.* 1980, vol. 26, no. 1, pp. 26–37.

26. JIZBA, Petr; KORBEL, Jan. Maximum Entropy Principle in Statistical Inference: Case for Non-Shannonian Entropies. *Physical Review Letters.* 2019, vol. 122, no. 12, p. 120601.

27. PARK, Juyong; NEWMAN, Mark E. J. Statistical mechanics of networks. *Physical Review E.* 2004, vol. 70, no. 6, p. 066117.

28. SQUARTINI, Tiziano; CALDARELLI, Guido; CIMINI, Giulio; GABRIELLI, Andrea; GARLASCHELLI, Diego. Reconstruction methods for networks: The case of economic and financial systems. *Physics Reports.* 2018, vol. 757, pp. 1–47.

29. RÉNYI, Alfréd. On measures of information and entropy. *Berkeley Symposium on Mathematical Statistics and Probability.* 1961, pp. 547–561.

30. JIZBA, Petr; ARIMITSU, Toshihico. The world according to Rényi: thermodynamics of multifractal systems. *Annals of Physics*. 2004, vol. 312, no. 1, pp. 17–59.

31. JIZBA, Petr; KORBEL, Jan; ZATLOUKAL, Václav. Tsallis thermostatics as a statistical physics of random chains. *Physical Review E*. 2017, vol. 95, no. 2, p. 022103.

32. ISLAM, Rajibul; MA, Ruichao; PREISS, Philipp M.; ERIC TAI, M.; LUKIN, Alexander; RISPOLI, Matthew; GREINER, Markus. Measuring entanglement entropy in a quantum many-body system. *Nature*. 2015, vol. 528, no. 7580, pp. 77–83.

33. ILIĆ, V. M.; KORBEL, J.; GUPTA, S.; SCARFONE, A. M. An overview of generalized entropic forms(a). *Europhysics Letters*. 2021, vol. 133, no. 5, p. 50005.

34. HAVRDA, J.; F., Charvát. Quantification Method of Classification Processes. *Kybernetika, 3*. 1967, pp. 30–35.

35. DARÓCZY, Zoltán. Generalized information functions. *Information and Control*. 1970, vol. 16, no. 1, pp. 36–51.

36. TSALLIS, Constantino. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*. 1988, vol. 52, no. 1–2, pp. 479–487.

37. TSALLIS, Constantino (ed.). *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*. New York, NY: Springer New York, 2009. SpringerLink. Description based upon print version of record.

38. FERRI, G L; MARTÍNEZ, S; PLASTINO, A. Equivalence of the four versions of Tsallis's statistics. *Journal of Statistical Mechanics: Theory and Experiment*. 2005, vol. 2005, no. 04, P04009.

39. TSALLIS, Constantino; MENDES, RenioS.; PLASTINO, A.R. The role of constraints within generalized nonextensive statistics. *Physica A: Statistical Mechanics and its Applications*. 1998, vol. 261, no. 3–4, pp. 534–554.

40. HANEL, R.; THURNER, S. A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions. *EPL (Europhysics Letters)*. 2011, vol. 93, no. 2, p. 20006.

41. UFFINK, Jos. Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*. 1995, vol. 26, no. 3, pp. 223–261.

42. DEMBO, Amir; MONTANARI, Andrea. Gibbs measures and phase transitions on sparse random graphs. *Brazilian Journal of Probability and Statistics*. 2010, vol. 24, no. 2.

43. TOUCHETTE, Hugo. The large deviation approach to statistical mechanics. *Physics Reports*. 2009, vol. 478, no. 1–3, pp. 1–69.

44. ROBERT, Christian P.; CASELLA, George. The Metropolis—Hastings Algorithm. In: *Monte Carlo Statistical Methods*. Springer New York, 1999, pp. 231–283.

45. GABRIELLI, Andrea; MACCHIATI, Valentina; GARLASCHELLI, Diego. Critical Density for Network Reconstruction. In: *From Computational Logic to Computational Biology.* Springer Nature Switzerland, 2024, pp. 223–249.

Appendix

A The average number of links

The average number of links is computed as follows:

$$\langle L \rangle = \frac{\sum_{k=0}^{N} k\binom{N}{k} \exp_{2-q}(-\theta k)}{\mathcal{Z}_q^N(\theta)} \tag{73}$$

Next, we can use the identity $k\binom{N}{k} = N\binom{N-1}{k-1}$, and the fact that we can write the sum from $k=1$ as the term with $k=0$ does not contribute. The numerator then is

$$\begin{aligned}
\sum_{k=0}^{N} k\binom{N}{k} \exp_{2-q}(-\theta k) &= N\sum_{k=1}^{N}\binom{N-1}{k-1}(1-(q-1)\theta k)^{1/(q-1)} &(74)\\
&= N\sum_{k=0}^{N-1}\binom{N-1}{k}(1-(q-1)\theta(k+1))^{1/(q-1)} &(75)\\
&= N\sum_{k=0}^{N-1}\binom{N-1}{k}(1-(q-1)\theta-(q-1)\theta k)^{1/(q-1)} &(76)\\
&= N(1-(q-1)\theta)^{1/(q-1)}\sum_{k=0}^{N-1}\binom{N-1}{k}\left(1-(q-1)\frac{\theta}{1-(q-1)\theta}k\right)^{1/(q-1)} &(77)\\
&= N\exp_{2-q}(-\theta)\,\mathcal{Z}_q^{N-1}\left(\frac{\theta}{1-(q-1)\theta}\right) &(78)
\end{aligned}$$

The average number of links can be then written as

$$\langle L \rangle = N\exp_{2-q}(-\theta)\,\frac{\mathcal{Z}_q^{N-1}\left(\frac{\theta}{1-(q-1)\theta}\right)}{\mathcal{Z}_q^N(\theta)} \tag{79}$$

B Computation of higher moments $\langle a_{i_1} \dots a_{i_k} \rangle$

Similarly to how we computed $\langle a_i \rangle$ in 4.36, we can compute higher moments. First, let us realize, that $a_i = a_i^2 = a_i^n$ for any $n \in \mathbb{N}$, since we are dealing with binary variables. Therefore, it suffices to consider distinct indices.

Let us have k distinct indices $i_1, \dots, i_k$. Then we can write

$$\begin{aligned}
\langle a_{i_1} \dots a_{i_k} \rangle &= \frac{1}{\mathcal{Z}_{q,N}(\theta)}\sum_{\{a_k\}} a_{i_1}\dots a_{i_k}\exp_{2-q}\left(-\theta\sum_i a_i\right)\\
&= \sum_{a_{i_1}=0,1}\cdots\sum_{a_{i_k}=0,1} a_{i_1}\dots a_{i_k}\sum_{\{a_j\}_{j\neq i_1,\dots,i_k}}\exp_{2-q}\left(-\theta\sum_i a_i\right)
\end{aligned} \tag{80}$$

Now the only summands contributing to the sum are those for which $a_{i_1} = \cdots = a_{i_k} = 1$. Therefore, we can rewrite the sum as

$$\begin{aligned}\langle a_{i_1} \dots a_{i_k} \rangle &= \frac{1}{\mathcal{Z}_{q,N}(\theta)} \sum_{\{a_j\}_{j \neq i_1,\dots,i_k}} \exp_{2-q}\left(-k\theta - \theta \sum_{i \neq i_1,\dots,i_k} a_i\right) \\ &\overset{(*)}{=} \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-k\theta) \sum_{\{a_j\}_{j \neq i_1,\dots,i_k}} \exp_{2-q}\left(-\frac{\theta}{1-k(q-1)\theta} \sum_{i \neq i_1,\dots,i_k} a_i\right) \\ &= \frac{1}{\mathcal{Z}_{q,N}(\theta)} \exp_{2-q}(-k\theta)\, \mathcal{Z}_{q,N-k}\left(\frac{\theta}{1-k(q-1)\theta}\right)\end{aligned} \tag{81}$$

where in the $\overset{(*)}{=}$ equation we used the identity 4.6.

C The computation of $\langle a_i a_j \rangle$ using $\langle L^2 \rangle$ and $\langle L \rangle$

Now for two distinct double-indices $kl \neq mn$, we can compute

$$\begin{aligned}\langle a_k a_l \rangle &= \frac{1}{Z_q^N(\theta)} \sum_{\{a_i\}} a_k a_l \exp_{2-q}\left(-\theta \sum_i a_i\right) = \frac{1}{Z_q^N(\theta)} \frac{1}{N-1} \sum_{\{a_i\}} \sum_{m \neq l} a_m a_l \exp_{2-q}\left(-\theta \sum_i a_i\right) \\ &= \frac{1}{Z_q^N(\theta)} \frac{1}{N-1} \left[\sum_{\{a_i\}} \sum_m a_m a_l \exp_{2-q}\left(-\theta \sum_i a_i\right) - \sum_{\{a_i\}} a_l^2 \exp_{2-q}\left(-\theta \sum_i a_i\right)\right] \\ &= \frac{1}{Z_q^N(\theta)} \frac{1}{N-1} \left[\sum_{\{a_i\}} L(\mathbb{A}) a_l \exp_{2-q}\left(-\theta \sum_i a_i\right) - \sum_{\{a_i\}} a_l \exp_{2-q}\left(-\theta \sum_i a_i\right)\right] \\ &= \frac{1}{Z_q^N(\theta)} \frac{1}{N-1} \left[\frac{1}{N} \sum_{\{a_i\}} L(\mathbb{A})^2 \exp_{2-q}\left(-\theta \sum_i a_i\right) - \frac{1}{N} \sum_{\{a_i\}} L(\mathbb{A}) \exp_{2-q}\left(-\theta \sum_i a_i\right)\right] \\ &= \frac{1}{N(N-1)} \left(\langle L^2 \rangle - \langle L \rangle\right)\end{aligned}$$

Now, we can use $\binom{N}{k} k(k-1) = N\binom{N-1}{k-1}(k-1) = N(N-1)\binom{N-2}{k-2}$

$$\begin{aligned}\langle L^2 \rangle - \langle L \rangle &= \frac{1}{Z_q^N(\theta)} \sum_{k=2}^N k(k-1) \binom{N}{k} \exp_{2-q}(-\theta k) \\ &= \frac{1}{Z_q^N(\theta)} N(N-1) \sum_{k=2}^N \binom{N-2}{k-2} (1-(q-1)\theta k)^{1/(q-1)} \\ &= \frac{1}{Z_q^N(\theta)} N(N-1) \sum_{k=0}^{N-2} \binom{N-2}{k} (1-(q-1)\theta(k+2))^{1/(q-1)} \\ &= \frac{1}{Z_q^N(\theta)} N(N-1)(1-(q-1)2\theta)^{1/(q-1)} \sum_{k=0}^{N-2} \binom{N-2}{k} \frac{1}{[1-(q-1)\frac{\theta}{1-(q-1)2\theta}k]^{1/(q-1)}} \\ &= N(N-1) \exp_{2-q}(-2\theta) \frac{Z_q^{N-2}\left(\frac{\theta}{1-2(q-1)\theta}\right)}{Z_q^N(\theta)}\end{aligned}$$

This means

$$\langle a_k a_l \rangle = \exp_{2-q}(-2\theta) \frac{Z_q^{N-2}\left(\frac{\theta}{1-2(q-1)\theta}\right)}{Z_q^N(\theta)} \tag{82}$$

D The $q \approx 1$ expansion of $\langle a_i \rangle$ and $\langle a_i a_j \rangle$

Denoting $\theta' = \frac{\theta}{1-(q-1)\theta}$, we can rewrite the average number of links as

$$\langle a_i \rangle = \langle c \rangle = \exp_{2-q}(-\theta) \frac{\mathcal{Z}_q^{N-1}(\theta')}{\mathcal{Z}_q^N(\theta)} = \exp_{2-q}(-\theta) \frac{\left(1+e^{-\theta'}\right)^{N-1}}{(1+e^{-\theta})^N} \frac{\langle \mathrm{res}_{2-q}(-\theta' L(G)) \rangle_{N-1}^{\theta'}}{\langle \mathrm{res}_{2-q}(-\theta L(G)) \rangle_N^{\theta}}$$

where by $\langle \dots \rangle_N^\theta$ we denote the average over Shannonian ensemble with N links and probabilities defined by parameter θ. Now, expanding the residual to the first order in $(q-1)$, we get

$$\begin{aligned} \frac{\langle \mathrm{res}_{2-q}(-\theta' L(G)) \rangle_{N-1}^{\theta'}}{\langle \mathrm{res}_{2-q}(-\theta L(G)) \rangle_N^{\theta}} &= \frac{1 - \frac{1}{2}(q-1)\theta'^2 \langle L^2 \rangle_{N-1}^{\theta'} + \mathcal{O}((q-1)^2)}{1 - \frac{1}{2}(q-1)\theta^2 \langle L^2 \rangle_N^{\theta} + \mathcal{O}((q-1)^2)} \\ &= 1 + \frac{1}{2}(q-1)\theta^2 (\langle L^2 \rangle_N^\theta - \langle L \rangle_{N-1}^\theta) + \mathcal{O}((q-1)^2) \end{aligned}$$

where in the second equation we only kept θ' to the 0-th order.

L is a binomial random variable with $p = \frac{1}{1+e^\theta}$ and N trials. Therefore, we can write

$$\langle L^2 \rangle_N^\theta = Np(1-p) + Np^2 \tag{83}$$

$$\langle L^2 \rangle_N^\theta - \langle L \rangle_{N-1}^\theta = p(1-p+(2N-1)p) = \frac{1}{(1+e^\theta)^2}(e^\theta + 2N - 1) \tag{84}$$

We need to expand the other terms as well

$$\left(1+e^{-\theta'}\right)^{N-1} = (1+e^{-\theta})^{N-1} \left(1 - (q-1)(N-1)\frac{\theta^2}{1+e^\theta} + \mathcal{O}((q-1)^2)\right) \tag{85}$$

$$\exp_{2-q}(-\theta) = e^{-\theta}\left(1 - \frac{1}{2}(q-1)\theta^2 + \mathcal{O}((q-1)^2)\right) \tag{86}$$

Putting all expansions together, we get

$$\langle a_i \rangle = \frac{1}{1+e^\theta}\left[1 - (q-1)\frac{1}{2}\frac{\theta^2}{(1+e^\theta)^2} e^\theta (2N - 1 + e^\theta)\right] \tag{87}$$

For the term $\langle a_i a_j \rangle$, we define $\theta'' = \frac{\theta}{1-2(q-1)\theta}$ and have

$$\langle a_i a_j \rangle = \exp_{2-q}(-2\theta) \frac{\mathcal{Z}_q^{N-2}(\theta'')}{\mathcal{Z}_q^N(\theta)} \tag{88}$$

$$= \exp_{2-q}(-2\theta) \frac{\left(1+e^{-\theta''}\right)^{N-2}}{(1+e^{-\theta})^N} \frac{\langle \mathrm{res}_{2-q}(-\theta'' L(G)) \rangle_{N-2}^{\theta''}}{\langle \mathrm{res}_{2-q}(-\theta L(G)) \rangle_N^{\theta}} \tag{89}$$

It holds, that

$$\frac{\langle \mathrm{res}_{2-q}(-\theta'' L(G)) \rangle_{N-2}^{\theta''}}{\langle \mathrm{res}_{2-q}(-\theta L(G)) \rangle_N^{\theta}} = \frac{1 - \frac{1}{2}(q-1)\theta''^2 \langle L^2 \rangle_{N-2}^{\theta''} + \mathcal{O}((q-1)^2)}{1 - \frac{1}{2}(q-1)\theta^2 \langle L^2 \rangle_N^{\theta} + \mathcal{O}((q-1)^2)} \tag{90}$$

$$= 1 + \frac{1}{2}(q-1)\theta^2 (\langle L^2 \rangle_N^\theta - \langle L \rangle_{N-2}^\theta) + \mathcal{O}((q-1)^2) \tag{91}$$

$$\langle L^2 \rangle_N^\theta - \langle L \rangle_{N-2}^\theta = 2p(1-p+2(N-1)p) = \frac{2}{(1+e^\theta)^2}(e^\theta + 2N - 2) \tag{92}$$

The other terms are expanded as

$$\left(1+e^{-\theta''}\right)^{N-2} = (1+e^{-\theta})^{N-2}\left(1-(q-1)2(N-2)\frac{\theta^2}{1+e^{\theta}}+\mathcal{O}((q-1)^2)\right) \tag{93}$$

$$\exp_{2-q}(-2\theta) = e^{-2\theta}\left(1-(q-1)2\theta^2+\mathcal{O}((q-1)^2)\right) \tag{94}$$

Combining, we get

$$\langle a_i a_j\rangle = \frac{1}{(1+e^{\theta})^2}\left[1-(q-1)\frac{\theta^2}{(1+e^{\theta})^2}e^{\theta}(2N-1+2e^{\theta})\right] \tag{95}$$

E The phase transition condition

We have stated, that the phase transition is possible only for such r, for which there exists some interval of c values, on which

$$\frac{\partial}{\partial c}\left(1+\exp\left(\frac{1}{rc}\right)\right)^{-1} > 1 \tag{96}$$

Let us denote $f(c,r) = \left(1+\exp\left(\frac{1}{rc}\right)\right)^{-1}$ and recognise, that it can be written using a logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$ as $f(c,r) = \sigma\left(-\frac{1}{rc}\right)$. For logistic function, we have the following useful identities:

$$\sigma(-x) = 1-\sigma(x) \tag{97}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x)) \tag{98}$$

$$2\sigma(x) = 1+\tanh\left(\frac{x}{2}\right) \tag{99}$$

$$\tanh\left(\frac{x}{2}\right) = \sigma(x)-\sigma(-x) \tag{100}$$

Using 97 and 98, we rewrite our condition as

$$\frac{\partial}{\partial c}f(c,r) = \frac{\partial}{\partial c}\sigma\left(-\frac{1}{rc}\right) = \sigma\left(-\frac{1}{rc}\right)\sigma\left(\frac{1}{rc}\right)\frac{1}{rc^2} > 1 \tag{101}$$

We recognise that this expression is odd in c, and therefore it is enough to check the condition for $c > 0$. Since we consider $r > 0$, we have $\frac{\partial}{\partial c}f(c,r) > 0$ and it is easy to check that $\lim_{c\to 0^+}\frac{\partial}{\partial c}f(c,r) = \lim_{c\to+\infty}\frac{\partial}{\partial c}f(c,r) = 0$. Since $\frac{\partial}{\partial c}f(c,r)$ is differentiable (i.e. continuous), it suffices to check the condition at its maximum.

The condition for the maximum of $\frac{\partial}{\partial c}f(c,r)$ is given by

$$\frac{\partial^2}{\partial c^2}f(c,r) = 0 \tag{102}$$

Evaluating the derivative, we can obtain

$$\sigma\left(\frac{1}{rc}\right)-\sigma\left(-\frac{1}{rc}\right)-2rc = 0 \tag{103}$$

and using the identity 100

$$\tanh\left(\frac{1}{2rc}\right) = 2rc \tag{104}$$

The maximum of $\frac{\partial}{\partial c}f(c,r)$ is thus attained for such c, for which 104 holds. The condition 101 can be rewritten as

$$r > \frac{(2rc)^2}{4\sigma\left(-\frac{1}{rc}\right)\sigma\left(\frac{1}{rc}\right)} \tag{105}$$

which can be further rewritten using the identity 99 and the fact that $\tanh(x)$ is an odd function as

$$r > \frac{(2rc)^2}{1 - \tanh^2\left(\frac{1}{2rc}\right)} \tag{106}$$

Combining 104 and 106, we realize, that it suffices to find such a x^*, which is a solution of $\tanh(1/x^*) = x^*$ and condition 106 yields $r > \frac{(x^*)^2}{1-(x^*)^2}$, which finalizes the proof.

F The critical point

The second derivative of $\phi_r(c, \theta)$ is given by

$$\frac{\partial^2}{\partial c^2}\phi_r(c, \theta) = \frac{1}{c(c-1)} + \frac{r}{\frac{1}{\theta} + rc} \tag{107}$$

Setting to zero and rearranging, we get

$$1 + cr\theta(2 - \theta) - c^2 r\theta^2(r + 1) = 0 \tag{108}$$

We want to find the smallest θ, for which this equation has a solution. That is equivalent to finding such θ, for which the discriminant of the quadratic equation is zero. This yields

$$r^2\theta^2(2 - \theta)^2 - 4r\theta^2(r + 1) = 0 \tag{109}$$

Assuming $\theta \neq 0$ and using the r value of the critical point r_{cp}, we find

$$\theta_{cp} = 2 + \sqrt{\frac{4(r_{cp} + 1)}{r_{cp}}} \tag{110}$$