

Joey de Mol

Breaking of ensemble equivalence in networks

Bachelor thesis, 11 juli 2014

Supervisors:

Dr. D. Garlaschelli (LION)

Prof. Dr. W.Th.F. den Hollander (MI)



Leiden Institute of Physics (LION)
Mathematical Institute (MI)
Universiteit Leiden

Contents

1	Abstract	3
2	Introduction	4
2.1	Statistical ensembles	4
2.2	Ensemble non-equivalence	4
2.3	Networks	5
3	Large deviations and ensemble equivalence in statistical physics	7
3.1	Basic principles	7
3.2	Entropy as a rate function	8
3.3	Ensemble equivalence in statistical physics	9
4	Statistical ensembles of networks	10
4.1	Motivation	10
4.2	Micro-canonical ensemble	11
4.3	Canonical ensemble	11
4.4	Grand-canonical ensemble	13
4.5	Ensemble equivalence	13
5	Results	15
5.1	Graphs with a fixed fraction of links	15
5.2	Regular graphs	16
5.3	Sparse graphs with given degree sequence	19
5.4	Star graphs	21
6	Conclusion	24
7	Appendix: An introduction to large deviation theory	25
7.1	Introduction	25
7.2	Cramér’s Theorem for the empirical average	26
7.3	The rate function	29
7.4	Sanov’s Theorem for the empirical measure	31

1 Abstract

This project addresses the problem of the breaking of ensemble equivalence between the canonical ensemble and the micro-canonical ensemble, motivated by examples in statistical physics. We propose complex networks as a novel candidate system and large deviation theory as a powerful mathematical tool. Several classes of networks are studied with the purpose of providing examples for the development of a general theory. Our results show the occurrence of ensemble non-equivalence in networks.

2 Introduction

2.1 Statistical ensembles

Statistical physics is concerned with the study of complex systems, i.e. systems for which we do not know the exact microstate since there are too many variables. For example, to completely describe the state of a gas we would need to know the position and the velocity of every individual atom, which is clearly impossible. However, the macroscopically observable variables can still be described by statistical ensembles, determined by only a couple of observable variables in equilibrium, and therefore the complexity of the problem is reduced. A statistical ensemble is a probability distribution for the microstate of the system. J. Willard Gibbs observed that the choice of ensemble required depends on the macroscopic constraints that are used. In 1902 Gibbs introduced the three following ensembles for physical systems [3]:

- **Micro-canonical ensemble**

The micro-canonical ensemble is a statistical expression of the conservation of energy for a closed system. It assigns a equal probability to all microstates with the specified energy. All other microstates are assigned probability zero.

- **Canonical ensemble**

The canonical ensemble models a system that can exchange energy, but no particles, with a heat reservoir having a constant temperature. The probability of a microstate with the specified amount of particles depends on its energy in such a way that the required temperature is obtained. Mathematically the temperature arises as the Lagrange multiplier enforcing the specified average energy.

- **Grand-canonical ensemble**

The grand-canonical ensemble models a system that can exchange both energy and particles with a reservoir having a constant temperature and chemical potential. The chemical potential replaces the constraint on the number of particles in the same way as the temperature replaces the constraint on the energy.

2.2 Ensemble non-equivalence

The micro-canonical ensemble had already been known since the work of Boltzmann in 1877 [4]. It is the most fundamental of three ensembles, since we often want to describe the equilibrium properties of a system for which the energy and the number of particles are fixed. The principle that the micro-canonical ensemble is built on is that nature does not favor any microstate over any other, so that each microstate must be assigned the same probability. In order to calculate this probability we must know the number of microstates with the prescribed energy and number of particles. Unfortunately, except for trivial cases like perfect gases or non-interacting systems, this task can be difficult. It is easier to work within the canonical ensemble, but we have to pay a price by also allowing microstates with the 'wrong' energy. Gibbs argued that in the limit where

the number of particles goes to infinity, the so-called thermodynamic limit, the micro-canonical and canonical ensembles are equivalent. His argument was that the energy fluctuations in the canonical ensemble become negligible compared to the total energy, so that the canonical ensemble essentially has a unique value of the energy, which is exactly what the micro-canonical ensemble prescribes. This would justify the use of the canonical ensemble instead of the more fundamental, but more difficult, micro-canonical ensemble. Equivalently, it would mean that energy and temperature can be used as two different parameters describing the equilibrium properties of the same system. One could similarly argue that the grand-canonical ensemble essentially has a unique number of particles in the thermodynamic limit.

For simple examples, like the ideal gas and non-interacting systems, the argument proposed by Gibbs can be proven. Later it was confirmed for many systems that indeed the micro-canonical and canonical equilibrium properties were the same. The approach of complex systems with statistical ensembles has been a great success, but it has led people to naively assume that the three ensembles are equivalent for *any* system. That, however, is not always the case.

Recently scientists have stumbled upon many-body systems where ensemble equivalence breaks down. These systems include models of fluid turbulence [5] [6], quantum phase separation [7] [8] [9], star formation [10] [11] and nuclear fragmentation [12]. This equivalence breaking manifests itself through the appearance of negative heat capacities and the system having micro-canonical equilibrium properties that are not predicted by the canonical ensemble. Due to its close relationship with statistical physics [13], *large deviation theory* provides a conceptual framework for studying the underlying mathematical principles of equivalence breaking. In particular, it has been proven for three different definitions of ensemble equivalence that there is non-equivalence if and only if the micro-canonical entropy function is nonconcave as a function of the energy density in the thermodynamic limit [14] [15]. It is still an open question what the physical cause is of ensemble non-equivalence, although it is expected to be related to the presence of long-range interactions.

2.3 Networks

There is need for new classes of systems in order to study the problem of ensemble non-equivalence at a more fundamental and more general level. One class that seems particularly suitable is that of *networks*.

A network is represented by a set of objects, called the nodes, and a set of relationships between these objects, called the links. In mathematics this is usually referred to as a graph $\mathbf{G} = (V, E)$, where V is a set of nodes and E is a set of links. The term network is commonly used when referring to something empirical, whereas a graph is the abstract mathematical object. Moreover, networks are typically without self-loops and multiple links. We only consider binary unweighted labeled graphs. Note that such graphs completely specified by the $n \times n$ adjacency matrix $\mathbf{A} = (a_{ij})_{i,j=1}^n$, where n is the number of nodes, $a_{ij} = 1$ if nodes i and j are connected and $a_{ij} = 0$ if they are not.

Examples of networks are: road networks, where cities are connected via highways, social networks, where people are linked depending on a specific social

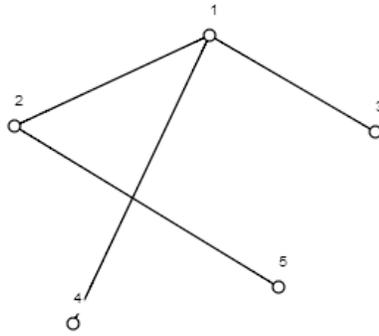


Figure 2.1: A graph with 5 nodes and 4 links.

relationship like friendship, and economic networks, where the nodes can be any economic agent and the links can be any financial interaction. One important economic network is the World Trade Network, the trade between countries. Due to the diversity and the vast amount of data available for them, networks have become a popular research topic.

Statistical physics and probability theory can be applied to networks by considering statistical ensembles of networks. The above ideas all carry over when, instead of the energy of a physical system, we consider a constraint on a network, like a specified number of links. We will go through this in more detail later. For networks with certain topological properties observed in empirical networks, ensemble non-equivalence is observed. Therefore real-world networks offer an exciting new way to study the breaking of ensemble equivalence. One of the criteria used is a non-vanishing specific relative entropy for the micro-canonical and canonical ensembles. The aim of the project was to study this equivalence breaking in networks and investigate whether the used criteria can be related to the criteria based on large deviation theory as used in [14] [15].

We start with an introduction that covers the basic principles of large deviations and ensemble equivalence in statistical physics. We proceed by defining statistical ensembles for networks and discussing the methods that can be used to compute them. We also look at the different definitions of ensemble equivalence. Next we compute the specific relative entropy for different constraints. Finally we elaborate on what these examples tell us about the subject of ensemble breaking in networks and what the future outlook is. The appendix provides an introduction to large deviation theory.

3 Large deviations and ensemble equivalence in statistical physics

Large deviation theory is a part of probability theory that deals with the description of unlikely events. It is an important tool in other branches of probability theory and arises in statistical physics, chemistry, biology, computer science, information theory and financial mathematics as well. Most important for this project is the relation between large deviation theory and statistical physics. One could reformulate the theory of statistical physics, in particular, the theory of statistical ensembles, in such a way that it is just an application of the theory of large deviations. An introduction to large deviation theory describing the key ideas and definitions can be found in the appendix. In this section we borrow from [13] to show that the micro-canonical entropy is a negative rate function associated with a large deviation principle, up to an additive constant. Large deviation theory then explains why the entropy and the free energy are related by a Legendre transform, and why variational principles arise in statistical physics. Moreover, the mathematical framework of large deviation theory has provided a method to detect ensemble inequivalence in physics by checking the non-concavity of the entropy function. We will give an overview of these large deviation results for statistical ensembles as described in [14].

3.1 Basic principles

We sum up a list of definitions and postulates that enable the application of large deviation theory in (equilibrium) statistical physics.

- The objects of study are n particles (atoms, spins, molecules, etc.) with joint state denoted by $\omega = (\omega_1, \omega_2, \dots, \omega_n)$, where ω_i is the state of the i th particle. The joint state ω gives a complete microscopic description of the n -particle system and is called a *microstate*. The one-particle state space is denoted by Λ , so that the n -fold product Λ^n equals the space of all microstates Λ_n .
- The particles mutually interact through some forces or potentials which leads to a *Hamiltonian* $H_n(\omega)$ and *mean energy* $h_n(\omega) = H_n(\omega)/n$.
- The microstate ω of the n -particle system is a random variable, distributed according to a prior probability measure $P(d\omega)$ on Λ_n . Because of the assumption of uniformity of nature, the prior distribution is chosen to be the uniform measure $P(d\omega) = d\omega/|\Lambda_n|$, where $|\Lambda_n| = |\Lambda|^n$ is a constant. The need for statistical ensembles arises when specifying the constraints of the system and when having to select a new probability distribution on the space Λ_n taking into account the constraints and the prior distribution.
- A *macrostate* is a function $\omega \mapsto M_n(\omega)$ of the microstates that reflects the macroscopic behavior of the system. The *equilibrium states* are the most probable outcomes of the macrostates in the specified ensemble in the thermodynamic limit $n \rightarrow \infty$.

Large deviation theory comes into play by observing that most macrostates M_n satisfy a large deviation principle for the fluctuations around the equilibrium states. The Law of Large Numbers says that the probability to deviate from the equilibrium state tends to zero in the thermodynamic limit, and it turns out that these probabilities are often exponentially decaying in the number of particles n . This is the reason why statistical physics and large deviation theory are so closely related. We will now see how the micro-canonical entropy is actually a negative rate function.

3.2 Entropy as a rate function

The mean energy $h_n(\omega)$ is an important microstate and we will now consider its large deviations. Letting $du = [u, u + du]$ be an infinitesimal interval of energy values, we have the formula for the probability distribution of h_n with respect to the prior $P(d\omega)$ on Λ^n

$$P(h_n \in du) = \int_{\{\omega \in \Lambda_n: h_n(\omega) \in du\}} P(d\omega).$$

Assuming that the uniform prior measure is $P(d\omega) = d\omega/|\Lambda|^n$, we get

$$P(h_n \in du) = \frac{1}{|\Lambda|^n} \int_{\{\omega \in \Lambda_n: h_n(\omega) \in du\}} d\omega = \frac{1}{|\Lambda|^n} \Omega(h_n \in du),$$

where

$$\Omega(h_n \in du) = \int_{\{\omega \in \Lambda_n: h_n(\omega) \in du\}} d\omega$$

is the volume of the microstates ω with mean energy $h_n(\omega) \in du$. Then we calculate the rate function $I(u)$ of $P(h_n \in du)$, assuming it exists, as

$$I(u) \equiv - \lim_{n \rightarrow \infty} \frac{1}{n} \log P(h_n \in du) = \log |\Lambda| - s(u),$$

where

$$s(u) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \Omega(h_n \in du)$$

is the micro-canonical entropy. We observe that if the rate function $I(u)$ exists, then it equals the negative of the entropy $s(u)$ up to an additive constant. We may re-define the entropy as

$$s(u) = \lim_{n \rightarrow \infty} \frac{1}{n} \log P(h_n \in du),$$

so that $I(u) = -s(u)$, which is more manageable.

Before being able to state results about ensemble equivalence, we mention the existence of the canonical free energy function

$$\varphi(\beta) = - \lim_{n \rightarrow \infty} \log Z_n(\beta),$$

where

$$Z_n(\beta) = \int_{\Lambda_n} e^{-\beta H_n(\omega)} d\omega$$

is the n -particle *partition function* associated with H_n at inverse temperature β . The (canonical) free energy function is the basic thermodynamic function of the canonical ensemble, like the entropy is for the micro-canonical ensemble.

3.3 Ensemble equivalence in statistical physics

The micro-canonical and canonical ensembles are said to be *thermodynamically equivalent* when the entropy and the free energy are one-to-one related by a Legendre transform. A one-to-one relationship between entropy and the free energy us that the micro-canonical description of a physical system determined from the entropy as a function of its energy is equivalent to the canonical description determined from the free energy as a function of its temperature. With the help of large deviation theory, it has been proven that there is thermodynamical non-equivalence when the micro-canonical entropy function has one or more points of noncavity [15]. This is a interesting new insight, because for a long time physicists had thought that the entropy should always be a concave function. Indeed, the examples of models showing ensemble non-equivalence mentioned in the introduction all have a nonconcave entropy function.

A more natural way to define ensemble equivalence is to compare the equilibrium states for a certain macrostate (energy, magnetization, etc.) predicted by the ensembles. The micro-canonical and canonical ensembles are said to be *macrostate equivalent* when there exists a one-to-one relationship between the elements of the set of equilibrium values of the macrostate predicted by the micro-canonical ensemble as a function of energy and the elements of the set of equilibrium values predicted by the canonical ensemble. It has been proven that macrostate equivalence is equivalent to thermodynamical equivalence [15], again via the mathematical framework provided by large deviation theory, so that the concavity of the entropy function is all the information one needs for determining whether or not there is equivalence of ensembles.

Finally, there is the concept of *measure equivalence* when the canonical probability distribution converges to the micro-canonical probability distribution in the thermodynamic limit. We will make this more precise in the section about ensembles in networks, because measure equivalence is the type of equivalence that we study in this project. Measure equivalence for physical systems has recently been proven to be equivalent to thermodynamic and macrostate equivalence [15].

4 Statistical ensembles of networks

This section starts off with a motivation for to use statistical ensembles of networks. We discuss the most important ensembles, their limitations and the methods to compute them. The section is based on [2].

4.1 Motivation

When studying real networks, it is important to determine whether an observed pattern is caused by non-trivial structural features or it is the result of simple constraints on the network. The most important and simple constraint is the specification of the degree k_i , the number of nodes connected to node i , for each labeled node $i = 1, \dots, n$, where n is the total number of nodes. If \mathbf{A} is the adjacency matrix, then $k_i = \sum_{j \neq i} a_{ij}$ for all i .

Consider for example one instance of the World Trade Network where a large number of triangles is observed. That is, when country A trades with country B, which in turn trades with country C, there is a high probability that A trades with C. Perhaps this is the case because two countries are more inclined to start trading with each other when they have a common trading partner, for some economic reason. Or it could be due to purely mathematical reasons, namely, that in any graph with that exact same degree distribution there is a high probability of observing a large number of triangles.

The way to deal with this problem is to introduce statistical ensembles of graphs with specified constraints, which are otherwise random. Let \mathbf{G}^* be the empirical network in question, and X the topological property we want to investigate. Consider a set of constraints denoted by the vector \vec{C} . Then each graph \mathbf{G} in some family of graphs \mathcal{G} is assigned an occurrence probability $P(\mathbf{G})$, and evaluation values $X(\mathbf{G})$ and $\vec{C}(\mathbf{G})$. We must have that

$$\sum_{\mathbf{G} \in \mathcal{G}} P(\mathbf{G}) = 1.$$

The ensemble average of X is defined as

$$\langle X \rangle \equiv \sum_{\mathbf{G} \in \mathcal{G}} X(\mathbf{G})P(\mathbf{G})$$

We choose the probabilities $P(\mathbf{G})$ in a way that reflects the constraint $\vec{C}(\mathbf{G}^*)$ of the real network. The method we choose depends on which ensemble we are using. The idea is the same as for choosing probabilities that reflect the energy of physical systems in the ensembles used in statistical physics.

If \mathcal{G} consists of all graphs comparable with \mathbf{G}^* in some way, typically the family of graphs with the same number of nodes as \mathbf{G}^* , then we can simply compare $X(\mathbf{G}^*)$ with $\langle X \rangle$. If these are almost the same, then we may conclude that the observed topological property $X(\mathbf{G}^*)$ is the result of the constraints $\vec{C}(\mathbf{G}^*)$, whereas if they are very different, then the observed topological property $X(\mathbf{G}^*)$ must be related to the real-world meaning of \mathbf{G}^* .

Now that we have motivated the use of statistical ensembles in networks we will discuss the computation of probabilities in the different ensembles in detail.

4.2 Micro-canonical ensemble

The distinctive property of the micro-canonical ensemble is that it enforces the constraints *exactly*. It uniformly assigns a positive probability to all graphs $G \in \mathcal{G}$ satisfying $\vec{C}(G) = \vec{C}(G^*)$, i.e.,

$$P_{MC}(G) = \begin{cases} \frac{1}{\mathcal{N}[\vec{C}(G^*)]}, & \text{if } \vec{C}(G) = \vec{C}(G^*), \\ 0, & \text{else,} \end{cases}$$

where $\mathcal{N}[\vec{C}(G^*)] \equiv |\{G \in \mathcal{G} : \vec{C}(G) = \vec{C}(G^*)\}|$.

Recalling our motivation, we see that the micro-canonical ensemble is the ideal distribution to use, since it enforces the constraints exactly, enabling a comparison between the observed property and the average of the property on the family of graphs satisfying the same constraint as the observed graph G^* . However, the micro-canonical ensemble has its limitations in that it is difficult to use. Computing $\mathcal{N}[\vec{C}(G^*)]$ is a challenging task, except for the most trivial cases, just like it was for physical systems. Most often only approximations asymptotic in the graph size under some conditions on the constraints are known. An analytical computation of the ensemble average can be even more difficult. Therefore the micro-canonical graph ensembles typically have to be sampled computationally by a uniformly random generation of many graphs with the required constraints, so that the property in question can be measured on all these graphs and averaged. This is a time-consuming and computationally demanding task, and the uniformly random generation is not straightforward.

Consider for example the generation of graphs with the same degree sequence as the observed degree sequence $\vec{k}(G^*) = \{k_i(G^*)\}$. One cannot naively assign to each node i a number of k_i edge stubs and randomly match pairs of stubs while avoiding self-loops and multiple links, since this will sometimes leave us with no legal pairings before all stubs are used. The method used instead is where one starts with the real network G^* and randomly selects two links, (A, B) and (C, D) for example, and then replaces these links by (A, D) and (C, B) if they did not already exist. Repeating this a large number of times, we get a random graph with the desired degree sequence. This method is additionally time-consuming, since the rewiring step has to be done many times to generate just one random graph, and it requires full use of the real network G^* , which is more information than one initially wanted to use.

Therefore, despite the fact that the micro-canonical ensemble is the most fundamental ensemble, an ensemble that is analytically manageable is preferable.

4.3 Canonical ensemble

Instead of enforcing the constraints exactly, the canonical ensemble enforces the constraints *on average*. That is, it enforces the ensemble average $\langle \vec{C} \rangle$ of the constraints.

The canonical probabilities $P_C(G)$ are assigned so as to

$$\begin{aligned}
&\text{maximize} && S \equiv - \sum_{\mathbf{G} \in \mathcal{G}} P_C(\mathbf{G}) \log P_C(\mathbf{G}) \\
&\text{subject to} && \sum_{\mathbf{G} \in \mathcal{G}} P_C(\mathbf{G}) = 1 \text{ and } \langle \vec{C} \rangle \equiv \sum_{\mathbf{G} \in \mathcal{G}} P_C(\mathbf{G}) \vec{C}(\mathbf{G}) = \vec{C}(\mathbf{G}^*).
\end{aligned}$$

S is called the Shannon-Gibbs entropy. It equals 0 if and only if the probability distribution is deterministic, and is maximal for a uniform probability distribution. It makes sense to maximize the Shannon-Gibbs entropy, since this gives the most uniform probability distribution satisfying the constraints, which is what nature does as well.

We now give the maximum-likelihood method to compute the canonical ensemble, first described in this form in [2]. It is both analytically manageable and fast. Below \mathcal{G} will be the family of graphs with the same number of nodes as \mathbf{G}^* .

To find the maximum-entropy graph probabilities, we introduce a set of Lagrange multipliers $\vec{\theta} = \{\theta_a\}$ enforcing the constraints $\vec{C}(\mathbf{G}^*)$. Each graph $\mathbf{G} \in \mathcal{G}$ is assigned the following probability conditional on $\vec{\theta}$,

$$P_C(\mathbf{G}|\vec{\theta}) = \frac{e^{-H(\mathbf{G}, \vec{\theta})}}{Z(\vec{\theta})},$$

where $H(\mathbf{G}, \vec{\theta})$ is the *graph Hamiltonian* defined as the inner product of \vec{C} and $\vec{\theta}$,

$$H(\mathbf{G}, \vec{\theta}) \equiv \vec{\theta} \cdot \vec{C}(\mathbf{G}),$$

and $Z(\vec{\theta})$ is the normalizing constant called the *partition function*

$$Z(\vec{\theta}) \equiv \sum_{\mathbf{G} \in \mathcal{G}} e^{-H(\mathbf{G}, \vec{\theta})}.$$

It can be shown that the resulting log-likelihood of the empirical network \mathbf{G}^* ,

$$\mathcal{L}(\vec{\theta}) \equiv \log P_C(\mathbf{G}^*|\vec{\theta}) = -H(\mathbf{G}^*, \vec{\theta}) - \log Z(\vec{\theta}),$$

is maximized by the particular parameter choice $\vec{\theta}^*$ such that the ensemble average of the constraints equals the value of the constraint evaluated on the real network, i.e.,

$$\langle \vec{C} \rangle_{\vec{\theta}^*} \equiv \sum_{\mathbf{G} \in \mathcal{G}} \vec{C}(\mathbf{G}) P_C(\mathbf{G}|\vec{\theta}^*) = \vec{C}(\mathbf{G}^*).$$

This enables us to compute $\vec{\theta}^*$, which in turn determines the canonical probabilities. Graphs $\mathbf{G} \in \mathcal{G}$ satisfying $\vec{C}(\mathbf{G}) = \vec{C}(\mathbf{G}^*)$ are most likely. Other graphs also have a non-zero probability of occurrence in the canonical ensemble, but they pay exponentially in probability for their deviation in the graph Hamiltonian.

When the constraint is a specified number of links or degree sequence, this method results in assigning a probability of connectance p_{ij} to each pair of nodes i and j . The probability of a graph $\mathbf{G} \in \mathcal{G}$ occurring in the canonical ensemble is then

$$P_C(\mathbf{G}) = \prod_{i < j} p_{ij},$$

so each link is realised independently. Once the probabilities of connectance are known, one is able to analytically compute ensemble averages in the same order of time required to evaluate the property on the empirical network. We postpone the computations leading to the probabilities of connectance when the constraint is a specified number of links or a specified degree sequence until Section 5.

It is also worth noting that the canonical ensemble is more resistant to errors in the presence of links in the empirical network \mathbf{G}^* than the micro-canonical ensemble. In the micro-canonical ensemble, \mathbf{G}^* would be assigned a probability zero since it does not match the constraints exactly, whereas in the canonical ensemble it would have a near-maximal probability since the constraints will still be close to the exact constraints. Moreover, the canonical ensemble does not require full information about \mathbf{G}^* , like was needed for the rewiring algorithm in the micro-canonical ensemble.

4.4 Grand-canonical ensemble

Also enforcing the number of nodes only on average, one gets the grand-canonical ensemble. To do so we add the number of nodes as a constraint and a corresponding Lagrange multiplier in the method of computing the canonical ensemble.

4.5 Ensemble equivalence

The use of the computationally easier canonical ensemble instead of the more fundamental micro-canonical ensemble needs to be justified. For physical systems Gibbs proposed that the ensembles become equivalent in the thermodynamic limit, but this turned out to be not always true. Long-range interactions are thought to play a central role in the breaking of equivalence, and networks are full of long-range interactions. We need to define equivalence of ensembles before we can study this problem in more detail.

We will say that the micro-canonical and canonical ensembles are equivalent if their *specific relative entropy* is zero. That is, we compute the *relative entropy*, also called the *Kullback - Leibler divergence*,

$$S(P_{MC}||P_C) \equiv \sum_{\mathbf{G} \in \mathcal{G}} P_{MC}(\mathbf{G}) \log \left(\frac{P_{MC}(\mathbf{G})}{P_C(\mathbf{G})} \right).$$

Noting that

$$P_{MC}(\mathbf{G}) = \begin{cases} \frac{1}{\mathcal{N}[\vec{C}(\mathbf{G}^x)]}, & \text{if } \vec{C}(\mathbf{G}) = \vec{C}(\mathbf{G}^*), \\ 0, & \text{else,} \end{cases}$$

where $\mathcal{N}[\vec{C}(\mathbf{G}^x)] \equiv |\{\mathbf{G} \in \mathcal{G} : \vec{C}(\mathbf{G}) = \vec{C}(\mathbf{G}^*)\}|$, and noting that the canonical probability is the same for all graphs having the same values of the constraints, we conclude that the sum cancels against the micro-canonical probabilities, so that we remain with

$$S(P_{MC}||P_C) = \log\left(\frac{P_{MC}(\mathbf{G}^*)}{P_C(\mathbf{G}^*)}\right) = \log(P_{MC}(\mathbf{G}^*)) - \log(P_C(\mathbf{G}^*)).$$

We then take the thermodynamic limit of the empirical graph \mathbf{G}^* in some natural way. For example, when the constraint is a specified number of links we must express this number as the *fraction* of occurring links, and if the constraint is a degree sequence we must know how to draw new degrees when the graph size increases. The *specific relative entropy* is then

$$\lim_{N \rightarrow \infty} \frac{1}{N} S(P_{MC}||P_C),$$

where N is the leading order scale of the micro-canonical and canonical ensembles when taking the thermodynamic limit of the empirical graph, so that we get a converging limit. This scale most often simply is the number of nodes. We say that the ensembles are equivalent if and only if the specific relative entropy of the ensembles vanishes.

The above is called *measure equivalence* of ensembles. Other criteria for ensemble equivalence used in physics are *thermodynamic equivalence* and *macrostate equivalence*. In physics the ensembles are called thermodynamically equivalent if the micro-canonical entropy and the free energy are one-to-one related by a Legendre transform, and macrostate equivalent if the sets of equilibrium values predicted by the ensembles are the same. For physical systems it has been proven that the ensembles are non-equivalent on all three levels when the micro-canonical entropy function is nonconcave as a function of the energy density in the thermodynamic limit [15]. It remains to be seen whether this holds for networks as well. With that question in mind we have computed the specific relative entropy for different constraints. We will present these computations in the next section.

5 Results

For different examples of constraints on the empirical graph \mathbf{G}^* we determine whether or not there is ensemble equivalence by computing the specific relative entropy of the micro-canonical distribution with respect to the canonical distribution. In each case we first compute the micro-canonical and canonical probability distributions.

5.1 Graphs with a fixed fraction of links

As a first example we consider the case where the constraint of the empirical graph is the number of realised links, i.e.,

$$C(\mathbf{G}^*) = L(\mathbf{G}^*) = \sum_{i < j} a_{ij}(\mathbf{G}) = \lambda \binom{n}{2},$$

where λ is the fraction of realised links on the set of $\binom{n}{2}$ different pairs of nodes. The number of graphs with n nodes and a fraction λ of realised links is

$$\Omega_n(\lambda) = \binom{M}{\lambda M},$$

where $M = \binom{n}{2}$. The micro-canonical probabilities for a graph \mathbf{G} with n nodes is $P_{MC}(\mathbf{G}) = \frac{1}{\Omega_n(\lambda)}$ if $C(\mathbf{G}) = \lambda M$ and zero otherwise.

We use the maximum-likelihood method described in Section 4.3 to obtain the canonical probabilities. The partition function is

$$Z(\theta) = \sum_{\mathbf{G}} e^{-\theta \cdot \sum_{i < j} a_{ij}(\mathbf{G})} = \sum_{\mathbf{G}} \prod_{i < j} e^{-\theta \cdot a_{ij}(\mathbf{G})},$$

where we sum over all graphs with n nodes. We may rewrite this as

$$Z(\theta) = \prod_{i < j} \sum_{\{a_{ij}\}} e^{-\theta \cdot a_{ij}(\mathbf{G})},$$

where the sum is over all possible values of a_{ij} , i.e., $a_{ij} = 0$ and $a_{ij} = 1$. The interchange of the order of the sum and the product is allowed because, for each graph first multiplying over each pair of nodes and then summing over all graphs is equivalent to, for each pair of nodes, first summing over all possible values of $a_{ij}(\mathbf{G})$ and then multiplying over all pairs of nodes (which is proven by induction). We then have

$$Z(\theta) = \prod_{i < j} \sum_{\{a_{ij}\}} e^{-\theta \cdot a_{ij}(\mathbf{G})} = \prod_{i < j} (1 + e^{-\theta}) = (1 + e^{-\theta})^M,$$

so that the canonical probability for each graph \mathbf{G} having n nodes becomes

$$P_C(\mathbf{G}|\theta) = \frac{e^{-\theta \cdot \sum_{i < j} a_{ij}(\mathbf{G})}}{(1 + e^{-\theta})^M} = \prod_{i < j} p_{ij}^{a_{ij}(\mathbf{G})} (1 - p_{ij})^{1 - a_{ij}(\mathbf{G})},$$

where

$$p_{ij} = \frac{e^{-\theta}}{1 + e^{-\theta}}.$$

The required Lagrange multiplier is found by maximizing the log-likelihood of \mathbf{G}^* ,

$$\mathcal{L}(\theta) = -\theta \cdot L(\mathbf{G}^*) - M \log(1 + e^{-\theta}),$$

which is solved by

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 \implies M \cdot \frac{e^{-\theta}}{1 + e^{-\theta}} = L(\mathbf{G}^*).$$

These considerations tell us that the canonical distribution is given by the Erdős-Rényi model for generating random graphs, where each pair of nodes has a probability $p = \frac{e^{-\theta}}{1 + e^{-\theta}}$ of being realised, so that the expected number of realised links is that of the empirical graph.

The canonical probability for a graph \mathbf{G} with n nodes is therefore

$$P_C = p^{L(\mathbf{G})} (1 - p)^{M - L(\mathbf{G})}, \quad p = \lambda.$$

We are now able to compute the relative entropy

$$\begin{aligned} S(P_{MC} || P_C) &= \log(P_{MC}(\mathbf{G}^*)) - \log(P_C(\mathbf{G}^*)) \\ &= -\log \Omega_n(\lambda) - L(\mathbf{G}^*) \log p - (M - L(\mathbf{G}^*)) \log(1 - p) \\ &= -\log \binom{M}{\lambda M} - \lambda M \log \lambda - (1 - \lambda) M \log(1 - \lambda). \end{aligned}$$

We may apply Stirling's formula on $\binom{M}{\lambda M}$ when computing the specific relative entropy. Instead, we note that $2^{-M} \binom{M}{\lambda M}$ is the probability that a random variable binomially distributed with probability $\frac{1}{2}$ has value λM . Then [1],

$$\lim_{M \rightarrow \infty} \frac{1}{M} 2^{-M} \binom{M}{\lambda M} = -\log 2 - \lambda \log \lambda - (1 - \lambda) \log(1 - \lambda).$$

We note that the leading order of both probabilities is $M = \binom{n}{2}$, so that the specific relative entropy is

$$\begin{aligned} \lim_{M \rightarrow \infty} \frac{1}{M} S(P_{MC} || P_C) &= \lim_{M \rightarrow \infty} \frac{1}{M} \binom{M}{\lambda M} - \lambda \log \lambda - (1 - \lambda) \log(1 - \lambda) \\ &= 0. \end{aligned}$$

The micro-canonical and canonical probabilities cancel each other in the limit $M \rightarrow \infty$. We conclude that when the constraint is the fraction of realised links the ensembles are equivalent.

5.2 Regular graphs

As a second example we consider empirical graphs where the constraint is the specified degree sequence, i.e., $\vec{C}(\mathbf{G}^*) = \vec{k}(\mathbf{G}^*)$. Not for every degree sequence a

closed-form expression of the micro-canonical and canonical probabilities exists, as far as we know, so we restrict ourselves to some special classes of degree sequences for which we can find an explicit solution. First we consider graphs with a homogeneous degree sequence. Graphs for which each node has the same degree $k \in \mathbb{N}_0$ are called k -regular graphs.

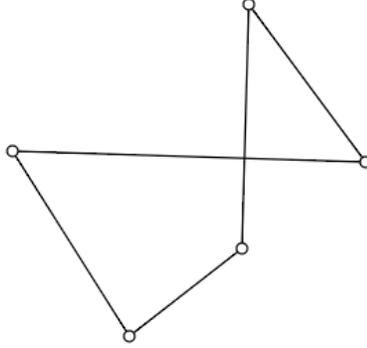


Figure 5.1: A 2-regular graph with 5 nodes.

There are results on the asymptotic number of labeled graphs with given degree sequence such that the degrees are $o(\sqrt{n})$ [16] [17]. A constant homogeneous degree sequence has this property. When we apply the results described in those papers to k -regular graphs, we find that the number of k -regular graphs with n nodes is asymptotically

$$\begin{aligned} \Omega_n(k) &= \frac{(nk)!}{\left(\frac{nk}{2}\right)! 2^{nk/2} (k!)^n} \exp\left(-\frac{k^2-1}{4} - \frac{k^3}{12n} + O(k^2/n)\right) \\ &= \frac{\sqrt{2} \left(\frac{nk}{e}\right)^{\frac{nk}{2}}}{(k!)^n} \exp\left(-\frac{k^2-1}{4} - \frac{k^3}{12n} + O(k^2/n)\right), \end{aligned}$$

where we applied Stirling's formula to rewrite the expression.

In the maximum-entropy method for the canonical ensemble there is now a Lagrange multiplier θ_i for each node $1 \leq i \leq n$. The partition function reads

$$Z(\vec{\theta}) = \sum_{\mathbf{G}} e^{-\sum_i \theta_i k_i(\mathbf{G})} = \sum_{\mathbf{G}} e^{-\sum_{i<j} (\theta_i + \theta_j) a_{ij}(\mathbf{G})} = \sum_{\mathbf{G}} \prod_{i<j} e^{-(\theta_i + \theta_j) a_{ij}(\mathbf{G})},$$

where the sum is over all graphs with n nodes. With the same interchange of sum and product as in the previous example, we get

$$Z(\vec{\theta}) = \prod_{i<j} (1 + e^{-\theta_i - \theta_j}).$$

Again, the canonical graph probability for a graph \mathbf{G} with n nodes can be written as

$$P_C(\mathbf{G}) = \prod_{i<j} p_{ij}^{a_{ij}(\mathbf{G})} (1 - p_{ij})^{1 - a_{ij}(\mathbf{G})},$$

where now

$$p_{ij} = \frac{e^{-\theta_i - \theta_j}}{1 + e^{-\theta_i - \theta_j}} = \frac{x_i x_j}{1 + x_i x_j},$$

where we abbreviate $x_i = e^{-\theta_i}$. Since all degrees must be k on average, all Lagrange multipliers must be the same. Hence $x_i = x$ for all i , where x satisfies

$$k = (n-1) \frac{x^2}{1+x^2}.$$

Therefore

$$p_{ij} = \frac{k}{n-1} \quad \forall i < j.$$

We conclude that, again, the canonical distribution is given by the Erdős-Rényi model, now with probability of connectance $p = \frac{k}{n-1}$. The canonical probability for a graph \mathbf{G} with n nodes is therefore

$$P_C = p^{L(\mathbf{G})} (1-p)^{M-L(\mathbf{G})}, \quad p = \frac{k}{n-1}.$$

So in the case of k -regular graphs the relative entropy becomes

$$\begin{aligned} S(P_{MC}||P_C) &= \log(P_{MC}(\mathbf{G}^*)) - \log(P_C(\mathbf{G}^*)) \\ &= -\log \Omega_n(\lambda) - L(\mathbf{G}^*) \log p - (M - L(\mathbf{G}^*)) \log(1-p) \\ &= -\left(\frac{nk}{2} \log(nk) - \frac{nk}{2} - n \log(k!) + O(k^2)\right) \\ &\quad - \frac{nk}{2} \log\left(\frac{k}{n-1}\right) - \left(\frac{n(n-1)}{2} - \frac{nk}{2}\right) \log\left(1 - \frac{k}{n-1}\right). \end{aligned}$$

In the limit $n \rightarrow \infty$ we have that $\log(nk) + \log(\frac{k}{n-1}) \rightarrow \log(k^2)$, so that

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} S(P_{MC}||P_C) &= \lim_{n \rightarrow \infty} \frac{1}{n} \left(-\left(\frac{nk}{2} \log(k^2) - \frac{nk}{2} - n \log(k!) + O(k^2)\right) \right. \\ &\quad \left. - \frac{n(n-1)}{2} \log\left(1 - \frac{k}{n-1}\right) + \frac{nk}{2} \log\left(1 - \frac{k}{n-1}\right) \right) \\ &= -\frac{k}{2} \log(k^2) + \frac{k}{2} + \log(k!) \\ &\quad + \lim_{n \rightarrow \infty} \left[-\frac{n-1}{2} \log\left(1 - \frac{k}{n}\right) + \frac{k}{2} \log\left(1 - \frac{k}{n}\right) \right] \end{aligned}$$

Noting that $\lim_{n \rightarrow \infty} n \log(1 - \frac{k}{n}) = -k$ and $\lim_{n \rightarrow \infty} \log(1 - \frac{k}{n}) = 0$, we find the specific relative entropy

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} S(P_{MC}||P_C) &= -k \log(k) + \frac{k}{2} + \log(k!) + \frac{k}{2} + 0 \\ &= \log(k!) - k \log k + k, \end{aligned}$$

for every constant k . Since there is a non-zero specific relative entropy, we say that there is non-equivalence of ensembles for the class of regular graphs. We note that the leading order in the thermodynamic limit for both ensembles is n , which is the number of constraints that occur by specifying a degree sequence.

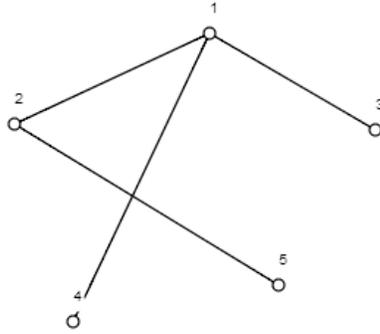


Figure 5.2: A graph with heterogeneous degree sequence $\vec{k} = (3, 2, 1, 1, 1)$.

5.3 Sparse graphs with given degree sequence

For the next example we consider heterogeneous degree sequences $\vec{k} = (k_1, \dots, k_n)$. We already mentioned that we cannot compute the ensembles analytically in general, but we can for *sparse* graphs. We say a graph is sparse if

$$k_i < \sqrt{n} \quad \text{for all } 1 \leq i \leq n.$$

The asymptotic number of sparse graphs $\Omega_n(\vec{k})$ with n nodes and degree sequence \vec{k} is found by again applying the results in [16] [17],

$$\Omega_n(\vec{k}) = \frac{\sqrt{2} \left(\frac{2L}{e}\right)^L}{\prod_{i=1}^n k_i!} \exp\left(-b^2 - b + o\left(\frac{\bar{k}^3}{n}\right)\right),$$

where L is the number of links,

$$L = \frac{1}{2} \sum_{i=1}^n k_i = \frac{n\bar{k}}{2},$$

and where b is given by

$$b = \frac{\overline{k^2} - \bar{k}}{2\bar{k}},$$

where we used overlines to denote averages.

So

$$\log P_{MC}(\mathbf{G}^*) = -L \log(2L) + L + \sum_{i=1}^n \log(k_i!) + b^2 + b + o\left(\frac{\bar{k}^3}{n}\right)$$

In the example of regular graphs we have seen that the maximum-entropy method gives the following canonical probability for a graph \mathbf{G} with n nodes

$$P_C(\mathbf{G}) = \prod_{i < j} p_{ij}^{a_{ij}(\mathbf{G})} (1 - p_{ij})^{1 - a_{ij}(\mathbf{G})},$$

where

$$p_{ij} = \frac{e^{-\theta_i - \theta_j}}{1 + e^{-\theta_i - \theta_j}} = \frac{x_i x_j}{1 + x_i x_j},$$

and where we abbreviated $x_i = e^{-\theta_i}$.

For sparse graphs with large n the probabilities of connectance must satisfy $P_{ij} \ll 1$ for all $i < j$, so that asymptotically $p_{ij} = x_i x_j$. Noting that asymptotically

$$k_i = \sum_{i \neq j} p_{ij} = \sum_{i \neq j} x_i x_j = x_i x_{tot} \quad \text{for all } i,$$

where $x_{tot} = \sum_i x_i$, we get

$$p_{ij} = \frac{k_i k_j}{x_{tot}^2} \text{ for all } i < j.$$

To determinate x_{tot} we further note that

$$\sum_{i,j} p_{ij} = 2L \text{ and } \sum_{i,j} p_{ij} = \sum_{i,j} \frac{k_i k_j}{x_{tot}^2} = \frac{(2L)^2}{x_{tot}^2},$$

which taken together gives $x_{tot}^2 = 2L$, and therefore asymptotically for sparse graphs

$$p_{ij} = \frac{k_i k_j}{2L} \text{ for all } i < j.$$

For computing the canonical probability of \mathbf{G}^* , we have the following lemma:

Lemma 5.1.

$$\log P_C(\mathbf{G}^*) = \sum_{\mathbf{G}} P_C(\mathbf{G}) \log P_C(\mathbf{G})$$

Proof.

$$\begin{aligned} \sum_{\mathbf{G}} P_C(\mathbf{G}) \log P_C(\mathbf{G}) &= \sum_{\mathbf{G}} P_C(\mathbf{G}) \log P_C(\mathbf{G}) [-H(\mathbf{G}, \vec{\theta}) - \log Z(\vec{\theta})] \\ &= \sum_{\mathbf{G}} P_C(\mathbf{G}) \log P_C(\mathbf{G}) [-\vec{\theta} \cdot \vec{C}(\mathbf{G}) - \log Z(\vec{\theta})] \\ &= -\vec{\theta} \cdot \langle \vec{C}(\mathbf{G}) \rangle - \log Z(\vec{\theta}), \end{aligned}$$

and $\vec{\theta}$ was chosen such that $\langle \vec{C}(\mathbf{G}) \rangle = \vec{C}(\mathbf{G}^*)$. □

Applying the lemma, we have

$$\begin{aligned} \log P_C(\mathbf{G}^*) &= \sum_{\mathbf{G}} P_C(\mathbf{G}) \log P_C(\mathbf{G}) \\ &= \sum_{i < j} p_{ij} \log p_{ij} + (1 - p_{ij}) \log(1 - p_{ij}) \\ &= \sum_{i < j} \frac{k_i k_j}{2L} \log \left(\frac{k_i k_j}{2L} \right) + \left(1 - \frac{k_i k_j}{2L} \right) \log \left(1 - \frac{k_i k_j}{2L} \right) \\ &= \sum_{i=1}^n k_i \log k_i - L \log(2L) + \sum_{i < j} \left(1 - \frac{k_i k_j}{2L} \right) \log \left(1 - \frac{k_i k_j}{2L} \right), \end{aligned}$$

where in the second expression we again applied the trick of switching sums and products. For the second sum in the final expression, note that $p_{ij} = \frac{k_i k_j}{2L} \ll 1$ so that by a Taylor expansion we asymptotically have

$$\log P_C(\mathbf{G}^*) = \sum_{i=1}^n k_i \log k_i - L \log(2L) - L$$

We conclude that the relative entropy of P_{MC} with respect to P_C is

$$\begin{aligned} S(P_{MC}||P_C) &= \log(P_{MC}(\mathbf{G}^*)) - \log(P_C(\mathbf{G}^*)) \\ &= -L \log(2L) + L + \sum_{i=1}^n \log(k_i!) + b^2 + b + o\left(\frac{\bar{k}^3}{n}\right) \\ &\quad - \left(\sum_{i=1}^n k_i \log k_i - L \log(2L) - L \right) \\ &= \sum_{i=1}^n \log(k_i!) - k_i \log k_i + k_i \\ &\quad + b^2 + b + o\left(\frac{\bar{k}^3}{n}\right), \end{aligned}$$

where we used that $2L = \sum_i k_i$. If we assume that the mean and variances of the degrees are finite, then we see that the specific relative entropy becomes

$$\lim_{n \rightarrow \infty} S(P_{MC}||P_C) = \overline{\log(k!)} - \bar{k} \log \bar{k} + \bar{k}.$$

So, we say that we have ensemble non-equivalence when the constraint is a sparse degree sequence. The correct leading order is again n , the number of nodes or the number of constraints. The specific relative entropy reminds us of that of regular graphs, except now we have to take averages over all the degrees since they are not all the same. It remains to see what happens when the degrees do not have finite mean or variance, like when the degrees are taken from certain power-law distribution, which is an important setting for network theory.

We expect that the concept of a *dual* graph, obtained by removing all realised links and adding all links that were not realised, may be useful for studying ensemble equivalence of *dense* graphs, defined as having a sparse dual graph. The micro-canonical probabilities will stay the same since for each sparse graph there is a unique dual graph. The computations for the canonical ensemble should also hold for the sparse dual graph, but one has to transform the probabilities of connectance back to the dense graph, which gives a more complicated expression for the canonical probability, but which possibly can be solved with some more work.

5.4 Star graphs

For our final example we consider a heterogeneous degree sequence that belongs neither to a sparse graph nor a dense graph. Let us first consider a *star* graph

of n nodes, where node 1 is called the hub, i.e., $k_1 = n - 1$, and the rest are called leaves, i.e., $k_i = 1$ for all $1 < i \leq n$. The micro-canonical entropy of this graph is zero because only one graph satisfies this degree sequence.

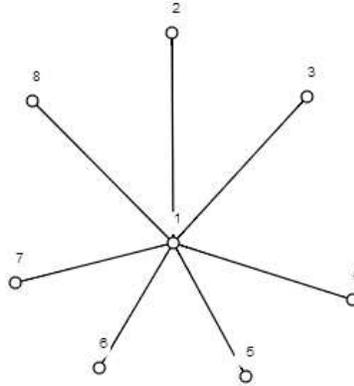


Figure 5.3: A star graph with 8 nodes.

However, things become more interesting when we consider M mutually connected hubs with $n - M$ leaves equally divided over the hubs. So the degree sequence is given by $k_i = \frac{n-M}{M} + (M - 1)$ if $1 \leq i \leq M$ and $k_i = 1$ if $i > M$.

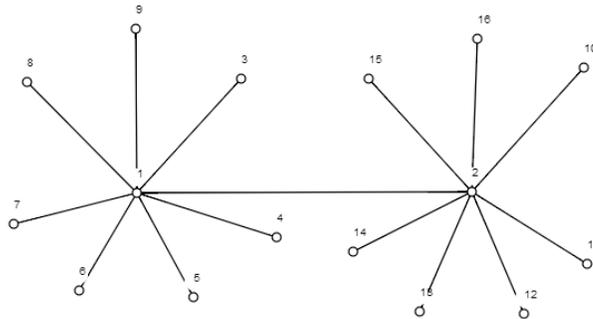


Figure 5.4: A “multi”-star graph with 16 nodes and 2 hubs.

The number of such graphs is a multinomial coefficient because the leaves are equally divided over the hubs

$$\Omega_n(M) = \binom{n - M}{\frac{n-M}{M} \dots \frac{n-M}{M}}.$$

Applying Stirling’s approximation, the asymptotic micro-canonical probability becomes

$$P_{MC}(\mathbf{G}^*) = M^{(M-n)} (2\pi(n - M))^{\frac{1}{2}(k-1)} M^{-\frac{M}{2}}$$

For the canonical ensemble we again have the configuration model

$$p_{ij} = \frac{x_i x_j}{1 + x_i x_j} \quad \text{for all } i < j.$$

If node i is a hub we write $x_i = x_H$ and if it is a leave we write $x_i = x_L$, which is allowed because the value of x_i should only depend on k_i . The system of equations fixing the degrees

$$\begin{cases} k_1 &= (M-1) + \frac{n-M}{M} = \sum_{i \neq j} p_{ij} = (M-1) \frac{x_H^2}{1+x_H^2} + (n-M) \frac{x_H x_L}{1+x_H x_L}, \\ k_n &= 1 = M \frac{x_H x_L}{1+x_H x_L} + (n-M-1) \frac{x_L^2}{1+x_L^2}, \end{cases}$$

has the (degenerate) solution

$$\frac{x_H^2}{1+x_H^2} = 1, \quad \frac{x_L^2}{1+x_L^2} = 0, \quad \frac{x_H x_L}{1+x_H x_L} = \frac{1}{M},$$

i.e. hubs are always mutually connected, leaves are never mutually connected and a leaf is connected with exactly one hub on average. Therefore the canonical probability is

$$P_C(\mathbf{G}^*) = \prod_{i < j} p_{ij}^{a_{ij}(\mathbf{G}^*)} (1-p_{ij})^{1-a_{ij}(\mathbf{G}^*)} = \left(\frac{1}{M}\right)^{n-M} \left(1 - \frac{1}{M}\right)^{(M-1)(n-M)}$$

Then the relative entropy is

$$\begin{aligned} S(P_{MC}||P_C) &= \log(P_{MC}(\mathbf{G}^*)) - \log(P_C(\mathbf{G}^*)) \\ &= (M-n) \log M + \left(\frac{1}{2}(k-1) \log(2\pi(n-M)) - \frac{M}{2} \log M\right) \\ &\quad - \left((n-M) \log\left(\frac{1}{M}\right) + (M-1)(n-M) \log\left(1 - \frac{1}{M}\right)\right) \\ &= (n-M)(M-1) \log\left(\frac{M}{M-1}\right) \\ &\quad + \left(\frac{1}{2}(k-1) \log(2\pi(n-M)) - \frac{M}{2} \log M\right), \end{aligned}$$

which gives a specific relative entropy of

$$\lim_{n \rightarrow \infty} \frac{1}{n} S(P_{MC}||P_C) = (M-1) \log\left(\frac{M}{M-1}\right).$$

So we again observe ensemble non-equivalence. The leading order in the limit is again n , the number of nodes or constraints.

6 Conclusion

Our results are the first to show the occurrence of breaking of ensemble equivalence in networks. It remains to develop general principles and a general theory. Our result indicate that breaking of ensemble equivalence occurs when the constraints imposed are extensive, i.e., proportional to the size of the network. The term N in the definition of specific relative entropy must be studied further, perhaps its correct value can be predicted by the type of specified constraints. It often appears to be equal to the number of constraints, but our first result on the specification of the fraction of realised links is paradoxical in that sense. We have proposed dual graphs as a possible useful concept for studying dense graphs. Recent papers using advanced combinatorial results in discrete mathematics could be used to study more difficult classes of constraints. For this project we focused on computations that could be done by hand, but the maximum-entropy method for the canonical ensemble makes it possible to also do numerical calculations on ensemble breaking. Furthermore, it could be investigated whether there exist analogues in statistical physics of the systems we studied in this project.

The aim of future research will be to completely characterise the onset of ensemble non-equivalence in complex networks using large deviation theory. This is a challenging task, but the framework provided by large deviation theory is likely to be very useful for studying ensemble breaking in networks like it has proven to be for statistical physics. One open question is whether measure equivalence, macrostate equivalence and thermodynamic equivalence are also mutually equivalent for networks, like in statistical physics [15]. Moreover, we have not studied the grand-canonical ensemble, so also that will be another possible direction of future research. Once there is more progress in the development of a theory and general principles, one could study breaking of ensemble equivalence in more general structures like matrices, and how ensemble non-equivalence is related to so-called “Gibbs-non-Gibbs” transitions.

Our research has confirmed that there is clear potential in studying ensemble non-equivalence in complex networks, and we look forward to further research on the subject.

7 Appendix: An introduction to large deviation theory

This section gives an introduction to the theory of large deviations. It closely follows chapters 1 and 2 of [1]. We formulate a basic result that describes the large deviation behavior of the empirical average of i.i.d. random variables satisfying a certain condition on the tail of the distribution. Although large deviation theory is certainly not limited to i.i.d. random variables, restricting to this setting is a good way to introduce the reader to some key principles and typical statements of large deviation theory. The rate function plays an important role in large deviation theory and is related to many of its key principles. We end with a large deviation result for the empirical measure, where the relative entropy appears as a rate function.

7.1 Introduction

Let X_1, X_2, \dots be i.i.d. \mathbb{R} -valued random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$. For $n \in \mathbb{N}$, let $S_n = \sum_{i=1}^n X_i$ be the partial sums.

There are two powerful theorems in probability theory that deal with such sums for large n :

Strong law of Large Numbers (SLLN)

$$\frac{1}{n} \xrightarrow{a.s.} \mu \text{ as } n \rightarrow \infty.$$

Central Limit Theorem (CLT)

$$\frac{1}{\sigma\sqrt{n}}(S_n - \mu n) \xrightarrow{d} Z \text{ as } n \rightarrow \infty,$$

where Z is a standard normal random variable.

In words, the SLLN tells us that the empirical average $\frac{1}{n}S_n$ converges to μ as $n \rightarrow \infty$, while the CLT gives us the probability that the partial sum S_n deviates from μn by an amount of order \sqrt{n} . Large deviation theory comes into play when we consider events where the deviation of S_n from μn is not “normal”, but of order n , i.e., “large”, so that the CLT does not suffice.

Consider the event

$$\left\{ \frac{1}{n}S_n \geq a \right\}, \quad a > \mu.$$

Then, because of the SLLN,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{1}{n}S_n \geq a \right) = 0.$$

We expect that the larger a is, the faster this probability decreases with n . The aim of large deviation theory is to quantify the *rate* of decays which typically is exponential in n .

7.2 Cramér's Theorem for the empirical average

The following result by Cramér tells us that, under a certain condition on the moment-generating function of X_1 , the previous event has an exponentially decaying probability, and it allows us to explicitly calculate the rate of this decay.

Theorem 7.1. *Let (X_i) be i.i.d. \mathbb{R} -valued random variables satisfying*

$$\varphi(t) = \mathbb{E}[e^{tX_1}] < \infty \quad \forall t \in \mathbb{R}. \quad (1)$$

Let $S_n = \sum_{i=1}^n X_i$. Then, $\forall a > \mathbb{E}[X_1]$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}S_n \geq a\right) = -I(a), \quad (2)$$

where the function I is the Legendre transform of $\log \varphi$, the cumulant generating function of X_1 :

$$I(z) = \sup_{t \in \mathbb{R}} [zt - \log \varphi(t)], \quad z \in \mathbb{R}.$$

Proof. If $\mathbb{P}(X_1 = a)$ for some $a \in \mathbb{R}$, then $\varphi(t) = e^{ta}$, so that

$$I(z) = \sup_{t \in \mathbb{R}} [(z - a)t].$$

Therefore $I(a) = 0$ and $I(z) = \infty$ for $z \neq a$, so we may assume that X_1 is non-degenerate. We may suppose without loss of generality that $a = 0$ and $\mathbb{E}[X_1] < 0$, since the substitution $X_1 \rightarrow X_1 + a$ gives $\varphi(t) \rightarrow e^{at}\varphi(t)$ and therefore $I(a) \rightarrow I(0)$. We introduce

$$\varphi = \inf_{t \in \mathbb{R}} \varphi(t),$$

so that $I(0) = -\log \varphi$ (with $I(0) = \infty$ if $\varphi = 0$). The claim can now be written as

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) = \log \varphi.$$

Let $F(x) = \mathbb{P}(X_1 \leq x)$ be the cumulant distribution function of X_1 . It follows from equation (1) that φ is a smooth function, i.e., $\varphi \in C^\infty(\mathbb{R})$, and

$$\begin{aligned} \varphi(t) &= \int_{\mathbb{R}} e^{tx} dF(x), \\ \varphi'(t) &= \int_{\mathbb{R}} xe^{tx} dF(x), \\ \varphi''(t) &= \int_{\mathbb{R}} x^2 e^{tx} dF(x). \end{aligned}$$

Since $\varphi''(t) \geq 0$ for all $t \in \mathbb{R}$, φ is strictly convex. We also note that $\varphi'(0) = \mathbb{E}[X_1] < 0$. For proving the upper bound, we distinguish three cases:

1. $\mathbb{P}(X_1 < 0) = 1$.

Then it follows from $\phi'(t) < 0$ for all $t \in \mathbb{R}$ that $\rho = \lim_{t \rightarrow \infty} \varphi(t) = 0$. Also $\mathbb{P}(S_n \geq 0) = 0$, which proves the claim.

2. $\mathbb{P}(X_1 \leq 0) = 1$ and $\mathbb{P}(X_1 = 0) > 0$.

Then ϕ is still strictly decreasing and again $\rho = \lim_{t \rightarrow \infty} \varphi(t)$ so that now $\rho = \mathbb{P}(X_1 = 0) > 0$. The claim then follows from $\mathbb{P}(S_n \geq 0) = \mathbb{P}(X_1 = \dots = X_n = 0) = \mathbb{P}(X_1 = 0)^n = \rho^n$.

3. $\mathbb{P}(X_1 < 0) > 0$ and $\mathbb{P}(X_1 > 0) > 0$.

Now $\lim_{t \rightarrow \pm\infty} \varphi(t) = \infty$. Since φ is strictly convex, there exists a unique $\tau \in \mathbb{R}$ for which φ attains its minimum. So we have

$$\varphi(\tau) = \rho, \quad \varphi'(\tau) = 0.$$

Also $\tau > 0$ since $\varphi'(0) < 0$. By the exponential Chebyshev inequality, we then find the upper bound

$$\mathbb{P}(S_n \geq 0) = \mathbb{P}(e^{\tau S_n} \geq 1) \leq \mathbb{E}[e^{\tau S_n}] = [\varphi(\tau)]^n = \rho^n.$$

We conclude that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) \leq \log \rho.$$

To get the lower bound we introduce the *Cramér-transform* \hat{F} of the cumulative distribution function F , which is

$$\hat{F}(x) = \frac{1}{\rho} \int_{(-\infty, x]} e^{\tau y} dF(y).$$

Since $\rho = \varphi(\tau) = \int_{\mathbb{R}} e^{\tau y} dF(y)$, the function \hat{F} is a cumulant distribution function as well. The idea now is to consider a new i.i.d. sequence (\hat{X}_i) of random variables distributed according to \hat{F} . As the following lemmas will show, the large deviation event $\{S_n \geq 0\}$ becomes “typical” under the tilted probability measure.

Lemma 7.2. $\mathbb{E}[\hat{X}_1] = \hat{\mu} = 0$ and $\text{Var}[\hat{X}_1] = \hat{\sigma}^2 \in (0, \infty)$.

Proof. Let $\hat{\varphi}(t) = \mathbb{E}[e^{t\hat{X}_1}]$. Then

$$\hat{\varphi}(t) = \int_{\mathbb{R}} e^{tx} d\hat{F}(x) = \frac{1}{\rho} \int_{\mathbb{R}} e^{(t+\tau)x} dF(x) = \frac{1}{\rho} \varphi(t + \tau) < \infty \quad \forall t \in \mathbb{R}.$$

Like before, it follows that $\hat{\varphi} \in C^\infty(\mathbb{R})$ and

$$\begin{aligned} \mathbb{E}[\hat{X}_1] &= \hat{\varphi}'(0) = \frac{1}{\rho} \varphi'(\tau) = 0, \\ \text{Var}[\hat{X}_1] &= \hat{\varphi}''(0) = \frac{1}{\rho} \varphi''(\tau) \in (0, \infty). \end{aligned}$$

□

Lemma 7.3. Let $\hat{S}_n = \sum_{i=1}^n \hat{X}_i$. Then $\mathbb{P}(S_n \geq 0) = \rho^n \mathbb{E}[e^{-\tau \hat{S}_n} 1_{\{\hat{S}_n \geq 0\}}]$.

Proof. Write

$$\begin{aligned}\mathbb{P}(S_n \geq 0) &= \int_{\{x_1 + \dots + x_n \geq 0\}} dF(x_1) \dots dF(x_n) \\ &= \int_{\{x_1 + \dots + x_n \geq 0\}} [\rho e^{-\tau x_1} d\hat{F}(x_1)] \dots [\rho e^{-\tau x_n} d\hat{F}(x_n)] \\ &= \rho^n \mathbb{E}[e^{-\tau \hat{S}_n} \mathbf{1}_{\{\hat{S}_n \geq 0\}}].\end{aligned}$$

□

Lemma 7.4. $\liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}[e^{-\tau \hat{S}_n} \mathbf{1}_{\{\hat{S}_n \geq 0\}}] \geq 0$.

Proof. By Lemma 7.2, \hat{S}_n satisfies the conditions of the CLT. Moreover, since $\int_0^\infty e^{-x^2/2}$ we can pick a $C > 0$ such that $\frac{1}{\sqrt{2\pi}} \int_0^C e^{-x^2/2} > \frac{1}{4}$. This gives us the estimate

$$\mathbb{E}[e^{-\tau \hat{S}_n} \mathbf{1}_{\{\hat{S}_n \geq 0\}}] \geq e^{-\tau C \hat{\sigma} \sqrt{n}} \mathbb{P}\left(\frac{\hat{S}_n}{\hat{\sigma} \sqrt{n}} \in [0, C)\right).$$

For large enough n the probability in the RHS is at least $\frac{1}{4}$, which proves the claim. □

Combining Lemmas 7.3 and 7.4, we get the lower bound required to complete the proof of Cramér's Theorem:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) \geq \log \rho.$$

□

Remarks

1. The function $z \mapsto I(z)$ is called the *rate function*. It plays an important role in large deviation theory and has some typical properties that we will discuss in the next subsection.
2. The theorem holds with the same rate function for $\mathbb{P}(\frac{1}{n} S_n \leq a)$ aif $a < \mathbb{E}[X_1]$, as can be seen from the substitution $X_1 \rightarrow -X_1$.
3. Equation (2) in Theorem 7.1 can be written as

$$\mathbb{P}\left(\frac{1}{n} S_n \geq a\right) = e^{-I(a)n + O(n)}.$$

4. The technique of exponentially tilting the probability measure through the Cramér transform combined with the CLT is a type of argument common in large deviation theory. By finding $\tau = \tau(a)$ with a as in Cramér's Theorem we get a new tilted probability distribution with mean a , so indeed the large deviation event $\{S_n \geq a\}$ is typical for this new distribution.

We will now give two examples in which we calculate the rate function.

Example 7.5. Suppose that X_1 is normally distributed with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$. We can find its moment-generating function in standard textbooks on probability theory:

$$\varphi(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}, \quad t \in \mathbb{R},$$

so that

$$I(z) = \sup_{t \in \mathbb{R}} \left[(z - \mu)t - \frac{1}{2}\sigma^2 t^2 \right], \quad z \in \mathbb{R}.$$

The function $t \mapsto (z - \mu)t - \frac{1}{2}\sigma^2 t^2$ is continuous, strictly concave and bounded from above. Setting the derivative equal to zero we find that the maximum is obtained at $t = \tau$ given by

$$\tau = \frac{z - \mu}{\sigma^2}.$$

Therefore the rate function is given by

$$I(z) = \frac{(z - \mu)^2}{2\sigma^2}, \quad z \in \mathbb{R}.$$

Example 7.6. Consider a sequence of coin tosses:

$$\mathbb{P}(X_1 = 0) = \frac{1}{2}, \quad \mathbb{P}(X_1 = 1) = \frac{1}{2}.$$

Then

$$\varphi(t) = \frac{1}{2}(1 + e^t), \quad t \in \mathbb{R},$$

so that

$$I(z) = \sup_{t \in \mathbb{R}} [zt - \log(1 + e^t)] + \log 2, \quad z \in \mathbb{R}.$$

Suppose that $z \notin [0, 1]$. Then $t \mapsto zt - \log(1 + e^t)$ is unbounded from above and therefore $I(z) = \infty$. If $z \in [0, 1]$, then $t \mapsto zt - \log(1 + e^t)$ is continuous, strictly concave and bounded from above. Setting the derivative equal to zero we find that the maximum is obtained at $t = \tau$ given by the relation

$$e^{-\tau} + 1 = \frac{1}{z}.$$

Eliminating τ in favor of z , we find $I(z) = \log 2 + z \log z + (1 - z) \log(1 - z)$, the relative entropy of $(z, 1 - z)$ with respect to $(\frac{1}{2}, \frac{1}{2})$. We conclude

$$I(z) = \begin{cases} \log 2 + z \log z + (1 - z) \log(1 - z), & \text{if } z \in [0, 1], \\ \infty, & \text{else.} \end{cases} \quad (3)$$

Figure 7.1 shows the graph of this rate function.

7.3 The rate function

Let us look more closely at the rate function. Two key properties that all rate functions obtained by Cramér's Theorem have in common are:

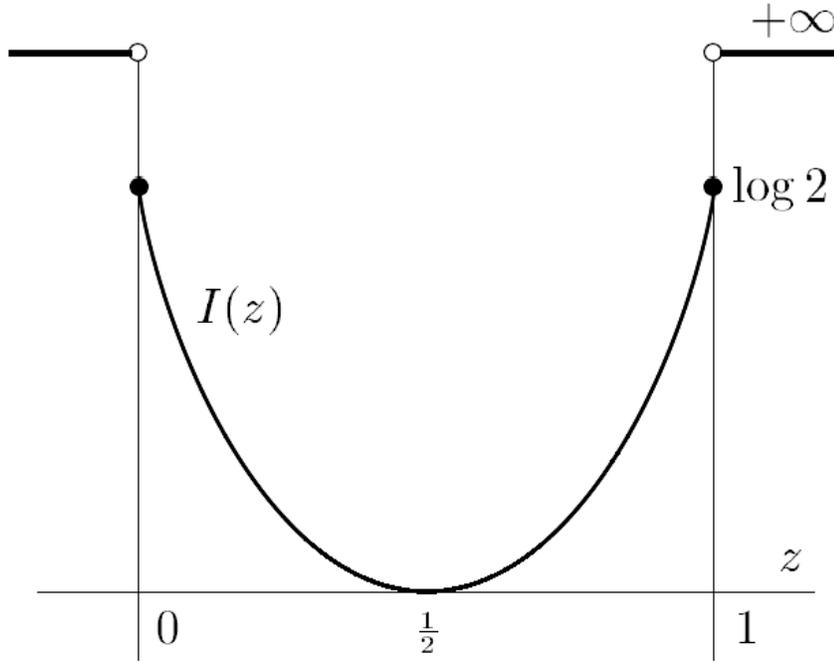


Figure 7.1: The rate function in example 7.6.

1. I is continuous and strictly convex on $\text{int}(\mathcal{D}_1)$, where $\mathcal{D}_1 = \{z \in \mathbb{R} : I(z) < \infty\}$.
2. $I(z) \geq 0$ with equality if and only if $z = \mu$.

Together they tell us that the further z is from the mean μ , the faster the probability $\mathbb{P}(\frac{1}{n}S_n \geq a)$ decays with n , and also that Figure 7.1 is qualitatively typical for rate functions obtained by Cramér's Theorem.

The zero in property 2 corresponds to the SLLN, since it implies that $I(\mu \pm \epsilon) > 0$ for all $\epsilon > 0$. Because $\mathbb{P}(\frac{1}{n}S_n \geq \mu \pm \epsilon) = e^{-nI(\mu \pm \epsilon) + O(n)}$, it follows that

$$\sum_{n \in \mathbb{N}} \mathbb{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) < \infty \quad \forall \epsilon > 0.$$

Applying the Borel-Cantelli lemma we get the SLLN.

We end this subsection with presenting a key principle in large deviation theory. If $a > \mu$, then $I(z) \geq I(a)$ for all $z \geq a$ because of properties 1 and 2. Therefore Cramér's Theorem can be reformulated as saying that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}S_n \in A\right) = - \inf_{z \in A} I(z) \quad \text{with } A = [a, \infty).$$

This formula can be proven to remain true when A is a sufficiently “nice” subset of \mathbb{R} . The reformulation can be interpreted as saying that *any large deviation is done in the least unlikely of all the unlikely ways*, i.e., the event with the smallest rate function dominates because all other events cost exponentially more. This is also why Cramér’s Theorem remains true with the same rate function for $\mathbb{P}(\frac{1}{n}S_n \in da)$.

7.4 Sanov’s Theorem for the empirical measure

Let us consider the case where the X_i take values in a *finite* set:

$$X_i \in \Gamma = \{1, \dots, r\} \subset \mathbb{N}$$

X_1, X_2, \dots are i.i.d. with marginal law $\rho = (\rho_s)_{s \in \Gamma}$ satisfying $\rho_s > 0 \forall s \in \Gamma$.

Let δ_x denote the point-mass at $x \in \mathbb{R}$. The *empirical measure*

$$L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

is a random probability measure on Γ that records the values that are observed in the sequence X_1, \dots, X_n . Consider the probability simplex in \mathbb{R}^r :

$$\mathcal{M}_1(\Gamma) = \left\{ \nu = (\nu_1, \dots, \nu_r) \in [0, 1]^r : \sum_{s=1}^r \nu_s = 1 \right\},$$

which corresponds to the set of probability measures on Γ . For $\mu, \nu \in \mathcal{M}_1$, define the total variation distance

$$d(\mu, \nu) = \frac{1}{2} \sum_{s=1}^r |\mu_s - \nu_s|,$$

which turns \mathcal{M}_1 into a Polish space, i.e., a complete separable metric space. Because of the SLLN, we know that

$$d(L_n, \rho) \xrightarrow{a.s.} \mu \text{ as } n \rightarrow \infty.$$

Sanov’s theorem for the empirical measure tells us what the large deviations of L_n away from ρ are.

Theorem 7.7. *Let (X_i) be i.i.d. random variables satisfying*

$$X_i \in \Gamma = \{1, \dots, r\} \subset \mathbb{N}$$

X_1, X_2, \dots are i.i.d. with marginal law $\rho = (\rho_s)_{s \in \Gamma}$ satisfying $\rho_s > 0 \forall s \in \Gamma$,

and let $L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$. Then, for all $a > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a^c(\rho)) = - \inf_{\nu \in B_a^c(\rho)} I_\rho(\nu),$$

where $B_a(\rho) = \{\nu \in \mathcal{M}_1(\Gamma) : d(\nu, \rho) \leq a\}$, $B_a^c(\rho) = \mathcal{M}_1(\Gamma) \setminus B_a(\rho)$ and

$$I_\rho(\nu) = \sum_{s=1}^r \nu_s \log \left(\frac{\nu_s}{\rho_s} \right)$$

(with $\inf_\emptyset I_\rho = \infty$).

Proof. Let

$$K_n = \left\{ k = (k_1, \dots, k_r) \in \mathbb{N}_0^r : \sum_{s=1}^r k_s = n \right\},$$

so that $\frac{1}{n}K_n \subset \mathcal{M}_1(\Gamma)$ for all $n \in \mathbb{N}$. The distribution of L_n is multinomial:

$$\mathbb{P} \left(L_n(s) = \frac{k_s}{n} \quad \forall s \right) = n! \prod_{s=1}^r \frac{\rho_s^{k_s}}{k_s!}, \quad k \in K_n.$$

For $k \in K_n$, let $v_n(k) = \frac{1}{n}k \in \mathcal{M}_1(\Gamma)$. Also let

$$Q_n(a) = \max_{k \in K_n : v_n(k) \in B_a^c(\rho)} \left(n! \prod_{s=1}^r \frac{\rho_s^{k_s}}{k_s!} \right),$$

to get the estimates

$$Q_n(a) \leq \mathbb{P}(L_n \in B_a^c(\rho)) \leq |K_n| Q_n(a),$$

where $|K_n| = \binom{n+r}{r-1} = O(n^{r-1})$. Applying Stirling's formula and noting that $\sum_{s=1}^r k_s = n$, we get

$$\frac{1}{n} \log \left(n! \prod_{s=1}^r \frac{\rho_s^{k_s}}{k_s!} \right) = \sum_{s=1}^r \frac{k_s}{n} \left(\log \rho_s - \log \frac{k_s}{n} \right) + O \left(\frac{\log n}{n} \right) \quad \text{uniformly on } K_n.$$

Here the sum in the RHS equals $-I_\rho(\nu(k))$, so that

$$\begin{aligned} \frac{1}{n} \mathbb{P}(L_n \in B_a^c(\rho)) &= O \left(\frac{\log n}{n} \right) + \frac{1}{n} \log Q_n(a) \\ &= O \left(\frac{\log n}{n} \right) - \min_{k \in K_n : v_n(k) \in B_a^c(\rho)} I_\rho(\nu_n(k)). \end{aligned}$$

For the final part of the proof we claim that

1. $\bigcup_{n \in \mathbb{N}} \{v_n(k) : k \in K_n\}$ is dense in $\mathcal{M}_1(\Gamma)$.
2. $\nu \mapsto I_\rho(\nu)$ is continuous on $\mathcal{M}_1(\Gamma)$.

both of which are straightforward to check. Together they imply that for every $\nu \in \mathcal{M}_1(\Gamma)$ there exists a sequence $(k_n)_{n \in \mathbb{N}}$, with $k_n \in K_n$ for all n , such that

$$\lim_{n \rightarrow \infty} d(\nu_n(k_n), \nu) = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} I_\rho(\nu_n(k_n)) = I_\rho(\nu).$$

Since $B_a^c(\rho)$ is an open set, we get

$$\limsup_{n \rightarrow \infty} \min_{k \in K_n : v_n(k) \in B_a^c(\rho)} I_\rho(\nu_n(k)) \leq I_\rho(\nu) \quad \forall \nu \in B_a^c(\rho).$$

We conclude that

$$\limsup_{n \rightarrow \infty} \min_{k \in K_n : v_n(k) \in B_a^c(\rho)} I_\rho(\nu_n(k)) \leq \inf_{\nu \in B_a^c(\rho)} I_\rho(\nu).$$

Since the reverse inequality is immediate, this finishes the proof. \square

Sanov's theorem says that $I_\rho(\nu) = H(\nu|\rho)$, the *relative entropy of ν with respect to ρ* . This rate function is always finite and satisfies the same two properties as discussed for rate functions given by Cramer's Theorem. The theorem remains true when $B_a^c(\rho)$ is replaced by some other sufficiently "nice" subset of $\mathcal{M}_1(\Gamma)$, conforming to the key principle that any large deviation is done in the least unlikely of all the unlikely ways. It is possible to generalize the theorem to *N -word empirical measures* where N successive values are recorded at each instant of time. The definition of the distance can be varied as well.

References

- [1] W.Th.F. den Hollander, *Large Deviations*, Fields Institute Monographs 14, American Mathematical Society, Providence RI, 2000, x + 143 pp., ISBN 0-8218-1989-5.
- [2] T. Suartini, and D. Garlaschelli, *Analytical maximum-likelihood method to detect patterns in real networks*, New Journal of Physics 13.8, 083001 (2001).
- [3] J.W. Gibbs, *Elementary Principles in Statistical Mechanics*, Yale University Press, Yale, C.T. (1902).
- [4] L. Boltzmann, *On the relationship between the second law of the mechanical theory of heat and the probability calculus*, Wiener Berichte 2 (76), 373 (1877).
- [5] R.S. Ellis, K. Haven, and B. Turkington, *Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles*, J.Stat.Phys. 101, 999 (2000).
- [6] R.S. Ellis, K. Haven, and B. Turkington, *Nonequivalent statistical equilibrium ensembles and refined stability theorems for most probable flows*, Nonlinearity 15, 239 (2002).
- [7] M. Blume, V.J. Emery, and R.B. Griffiths, *Ising model for the λ transition and phase separation in He^3 - He^4 mixtures*, Physical Review A 4.3, 1071 (1971).
- [8] J. Barré, D. Mukamel, and S. Ruffo, *Inequivalence of ensembles in a system with long-Range interactions*, Phys. Rev. Lett. 87, 030601 (2001).
- [9] R.S. Ellis, H. Touchette, and B. Turkington, *Thermodynamic versus statistical nonequivalence of ensembles for the mean-field Blume-Emery-Griffiths model*, Physica A 335, 518 (2004).
- [10] D. Lynden-Bell, *Negative specific heat in astronomy, physics and chemistry*, Physica A 263, 293 (1999).
- [11] P.H. Chavanis, *Gravitational instability of isothermal and polytropic spheres*, Astron. & Astrophys. 401, 15 (2003).
- [12] M.D. Agostino, *Negative heat capacity in the critical region of nuclear fragmentation: an experimental evidence of the liquid-gas phase transition*, Phys. Lett. B473, 219 (2000).
- [13] H. Touchette, *The large deviation approach to statistical mechanics*, Physics Reports 478.1, 1 (2009).
- [14] H. Touchette, R.S. Ellis, and B. Turkington, *An introduction to the thermodynamic and macrostate levels of nonequivalent ensembles*, Physica A: Statistical Mechanics and its Applications 340.1, 138 (2004).
- [15] H. Touchette, *Equivalence and nonequivalence of ensembles: Thermodynamic, macrostate, and measure levels*, Preprint, arXiv:1403.6608 [cond-mat.stat-mech] (2014).

- [16] E.A. Bender, *The asymptotic number of labeled graphs with given degree sequences*, Discrete Math., 10 (1974), pp. 217223.
- [17] B.D. McKay, and N.C. Wormald *Asymptotic enumeration by degree sequence of graphs with degrees $o(\sqrt{n})$* , Combinatorica 11 (1991), 369382.