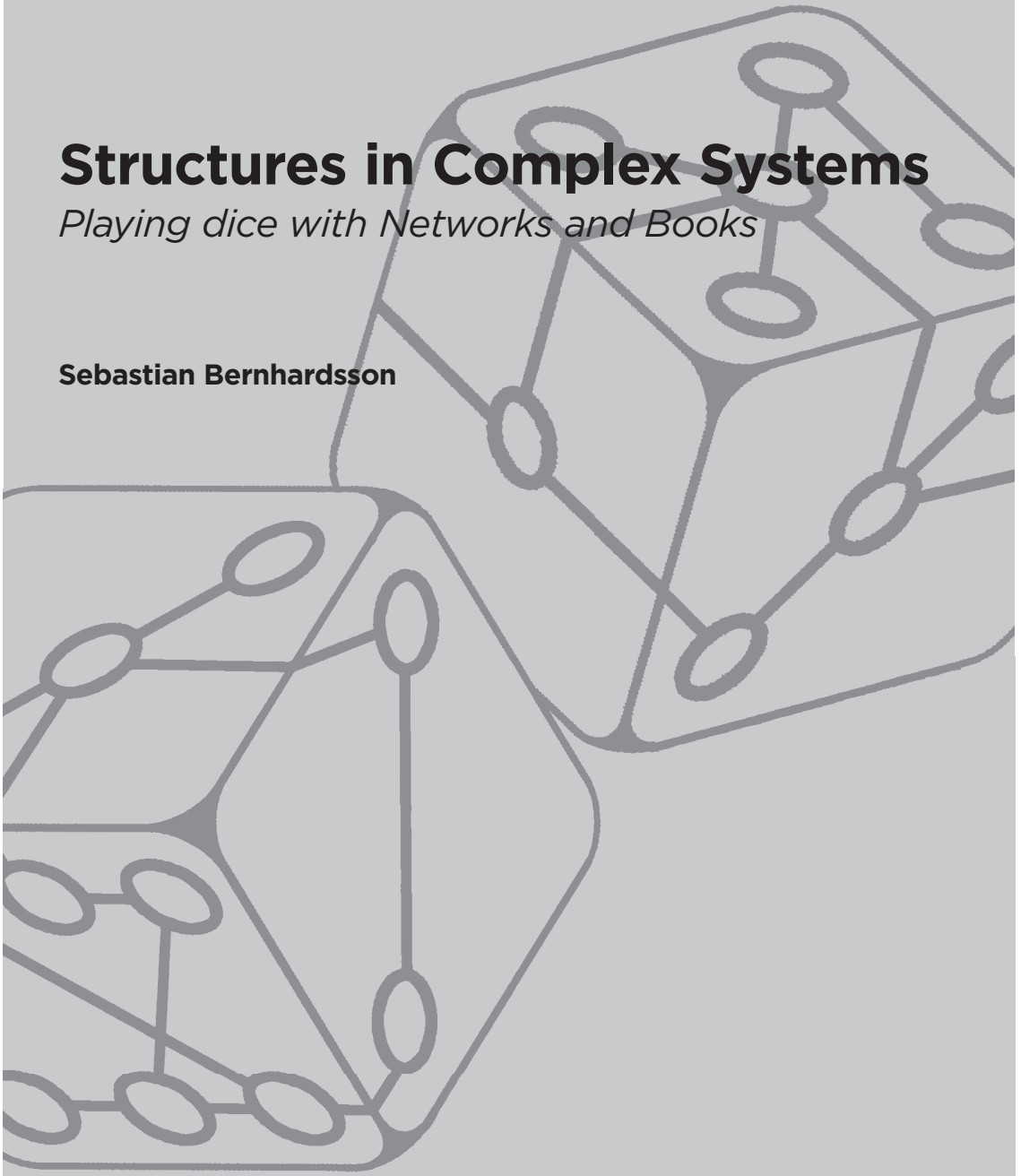
Structures in Complex Systems

Playing dice with Networks and Books

Sebastian Bernhardsson



Department of physics
Doctoral Thesis 2009

Structures in Complex Systems

Playing dice with Networks and Books

Sebastian Bernhardsson

Department of physics
Doctoral Thesis 2009

Department of Physics
Umeå University
SE-90187 Umeå, Sweden

ISBN: 978-91-7264-910-1

Printed by Print & Media, Umeå 2009

Abstract

Complex systems are neither perfectly regular nor completely random. They consist of a multitude of players who, in many cases, play together in a way that makes their combined strength greater than the sum of their individual achievements. It is often very effective to represent these systems as networks where the actual connections between the players take on a crucial role. Networks exist all around us and are an important part of our world, from the protein machinery inside our cells to social interactions and man-made communication systems. Many of these systems have developed over a long period of time and are constantly undergoing changes driven by complicated microscopic events. These events are often too complicated for us to accurately resolve, making the world seem random and unpredictable. There are however ways of using this unpredictability in our favor by replacing the true events by much simpler stochastic rules giving effectively the same outcome. This allows us to capture the macroscopic behavior of the system, to extract important information about the dynamics of the system and learn about the reason for what we observe. Statistical mechanics gives the tools to deal with such large systems driven by underlying random processes under various external constraints, much like how intracellular networks are driven by random mutations under the constraint of natural selection. This similarity makes it interesting to combine the two and to apply some of the tools provided by statistical mechanics on biological systems. In this thesis, several null models are presented, with this view point in mind, to capture and explain different types of structural properties of real biological networks.

The most recent major transition in evolution is the development of language, both spoken and written. This thesis also brings up the subject of quantitative linguistics from the eyes of a physicist, here called linguaphysics. Also in this case the data is analyzed with an assumption of an underlying randomness. It is shown that some statistical properties of books, previously thought to be universal, turn out to exhibit author specific size dependencies. A meta book theory is put forward which explains this dependency by describing the writing of a text as pulling a section out of a huge, individual, abstract mother book.

Sammanfattning

Komplexa system är varken perfekt ordnade eller helt slumpmässiga. De består av en mängd aktörer, som i många fall agerar tillsammans på ett sådant sätt att deras kombinerade styrka är större än deras individuella prestationer. Det är ofta effektivt att representera dessa system som nätverk där de faktiska kopplingarna mellan aktörerna spelar en avgörande roll. Nätverk finns överallt omkring oss och är en viktig del av vår värld , från proteinmaskineriet inne i våra celler till sociala samspel och människotillverkade kommunikationssystem. Många av dessa system har utvecklats under lång tid och genomgår hela tiden förändringar som drivs på av komplicerade småskaliga händelser. Dessa händelser är ofta för komplicerade för oss att noggrant kunna analysera, vilket får vår värld att verka slumpmässig och oförutsägbar. Det finns dock sätt att använda denna oförutsägbarhet till vår fördel genom att byta ut de verkliga händelserna mot mycket enklare regler baserade på sannolikheter, som ger effektivt sett samma utfall. Detta tillåter oss att fånga systemets övergripande uppförande, att utvinna viktig information om systemets dynamik och att få kunskap om anledningen till vad vi observerar. Statistisk mekanik hanterar stora system pådrivna av sådana underliggande slumpmässiga processer under olika restriktioner, på liknande sätt som nätverk inne i celler drivs av slumpmässiga mutationer under restriktionerna från naturligt urval. Denna likhet gör det intressant att kombinera de två och att applicera de verktyg som ges av statistisk mekanik på biologiska system. I denna avhandling presenteras flera nollmodeller som, baserat på detta synsätt, fångar och förklarar olika typer av strukturella egenskaper hos verkliga biologiska nätverk.

Den senaste stora evolutionära övergången är utvecklandet av språk, både talat och skrivet. Denna avhandling tar också upp ämnet om kvantitativ linguistik genom en fysikers ögon, här kallat linguafysik. Även i detta fall så analyseras data med ett antagande om en underliggande slumpmässighet. Det demonstreras att vissa statistiska egenskaper av böcker, som man tidigare trott vara universella, egentligen beror på bokens längd och på författaren. En metabokteori ställs fram vilken förklarar detta beroende genom att beskriva författandet av en text som att rycka ut en sektion ur en stor, individuell, abstrakt moderbok.

Publications

The thesis is based on the following publications (reprinted with the kind permission of the publishers):

I S. Bernhardsson and P. Minnhagen. *Models and average properties of scale-free directed networks.* Physical Review E **74** (2006), 026104.

II J.B. Axelsen, S. Bernhardsson, M. Rosvall, K. Sneppen and A. Trusina. *Degree landscapes in scale-free networks.* Physical Review E **74** (2006), 036119.

III J.B. Axelsen, S. Bernhardsson and K. Sneppen. *One hub-one process: A tool based view on regulatory network topology.* BMC Systems Biology **2** (2008), 25.

IV P. Minnhagen, S. Bernhardsson and B.J. Kim. *Scale-freeness for networks as a degenerate ground state: A hamiltonian formulation.* Europhysics Letters **78** (2007), 28004.

V P. Minnhagen and S. Bernhardsson. *Optimization and scale-freeness for complex networks.* Chaos **17** (2007), 2.

VI P. Minnhagen and S. Bernhardsson. *The blind watchmaker network: Scale-freeness and evolution.* PLoS ONE **3** (2008), (2):e1690.

VII S. Bernhardsson and P. Minnhagen. *Selective pressure on the metabolic network structures as measured from the random blind watchmaker network.* Manuscript (2009).

VIII S. Bernhardsson, L.E da Rocha Correa, and P. Minnhagen. *Size dependent word frequencies and translational invariance of books.* Physica A **389** (2010), 330–341.

IX S. Bernhardsson, L.E da Rocha Correa, and P. Minnhagen. *The meta book and size-dependent properties of written language.* New Journal of Physics (2009), accepted.

Other publications by the author not included in the thesis are:

- P. Minnhagen, B.J. Kim, S. Bernhardsson and Gerardo Cristofano. *Phase diagram of generalized fully frustrated xy model in two dimensions*. Physical Review B **7** (2007), 224403.

- P. Minnhagen, B.J. Kim, S. Bernhardsson and Gerardo Cristofano. *Symmetry-allowed phase transitions realized by the two-dimensional fully frustrated xy class*. Physical Review B **78** (2008), 184432.

- B. J. Kim, P. Minnhagen and S. Bernhardsson, *Phase Transitions in Generalized XY Model at $f = 1/2$ (Proceeding of APPC10)*. Journal of the Korean Physical Society **53** (2008), 1269.

- S.K. Baek, P. Minnhagen, S. Bernhardsson, K. Choi K and B.J. Kim. *Flow improvement caused by agents who ignore traffic rules*. Physical Review E **80** (2009), 016111.

- P. Minnhagen, S. Bernhardsson and B.J. Kim. *The groundstates and phases of the two-dimensional fully frustrated xy model*. International Journal of Modern Physics B **23** (2009), 3939-3950.

- S.K. Baek and S. Bernhardsson. *Comment to "Comments on 'Reverse auction: The lowest unique positive integer game' "*. Submitted (2009).

Preface

Our world can at times seem random, or unpredictable, without any real underlying purpose. Chains of seemingly unrelated events often lead to unexpected, circumstantial incidents, and we call it chance. After 20 years of making a random walk in life I started my random walk in physics. Later I also came into contact with problems and questions from outside the borders of physics, and I have to say that our world is really an amazing place. It does not matter if we are talking about the birth of a star, black holes, how life came to be, how languages have developed or how the stock market works. There are interesting questions everywhere. Our society has, however, divided science into fields, or disciplines, in order to make it easier for us to handle. But as a consequence we have created a void in between these disciplines, or fuzzy borders where they overlap, and few people have felt the urge to go there in the past. This is however about to change. There has been an increasing activity in interdisciplinary sciences during the recent years and scientists from different fields are coming together and collaborate in growing numbers. Attacking problems from different angles and with different view points is very healthy for the progress of science. The challenge is to find a common ground and learn to decode each other's vocabulary.

I think it is safe to say that this thesis is quite interdisciplinary. It is built on different research projects with questions and data imported from various fields other than physics. The common theme here is symbolized by the dices on the cover. These dices represent the underlying randomness of a system which we try to explore and exploit in order to give possible explanations to observed non-trivial properties.

The dices also signify what type of random events we are talking about. The outcome of a dice is random because the process is chaotic. The outcome of a throw depends on the initial velocity and rotation which, depending on the starting height, determine how it will collide with the table. The consecutive bounces, in there turn, depend on how the dice hits the table, and on the properties of the surface it bounces on before it comes to a rest. The point is that all these steps can be exactly calculated and repeated if we knew, and could re-create, all the above mentioned conditions of the throw precisely. However, if the initial condition of the throw is changed just a tiny bit, the outcome will change dramatically. So, if we do

not know everything about the throw with very good precision, we can just as well guess the outcome. Which is what we do. We simplify the process by saying that it is random and assign probabilities to the different outcomes. The same is true in a classical view of a system of particles in a box exchanging energy and momentum via collisions, or the mutation of a specific base pair in the DNA due to radiation. The point is that *randomness in this sense reflects nothing more than a lack of adequate knowledge*.

In order to make up for this lack of knowledge we zoom out and exchange the microscopic events with much simpler stochastic rules so that they represent the effective outcome of the system. This allows us to make predictions and draw conclusions about the macroscopic properties of the system. For example, by assigning equal probability to each side of a dice we can make predictions about how often certain numbers will come up. We can, in the same way, conclude to what extent a dice is biased, or affected by constraints, by the way the outcome deviates from the expected result. With this approach in mind statistical mechanics provides the means and tools to deal with large chaotic systems under the influence of external constraints, or forces like gravity or magnetism. In a similar way Darwinian evolution can be described as a random process, driven by mutations, under the constraints of natural selection. The mutations constitute the engine and natural selection is controlling the steering wheel.

By using the tools from statistical mechanics on biological systems we zoom out and try to find the simplest representation of the underlying process in an attempt to describe the dynamics of the system and with the hope to learn about how the constraints of natural selection has affected its structure.

This thesis is divided into three parts where the first chapter is an introduction to network science including a guided tour through the terminology and some of the main issues, concepts and models that have awoken peoples' interest in the field.

My goal of the second chapter is to give an overview of statistical mechanics and hopefully giving an understandable description of how randomness comes up, and is dealt with, in physics. And ultimately how it can be applied to networks.

Finally the third chapter is about word frequencies and quantitative linguistics which I here call *linguaphysics* (in conformity with the term 'econophysics'). We move to this field because the statistical approach and modeling of these systems are very similar to those used in network science. Also, this system is free from the complexity related to patterns of connections and there is a huge amount of data available, making it a natural step to take when studying randomness in complex systems.

As a final remark, before I leave you to unravel the mysteries of the randomness in your world, I quote the words of Eric Hoffer: "Creativity is the ability to introduce order into the randomness of nature".

Acknowledgment

There are many people involved in the making of this thesis and to whom I would like to show my deepest gratitude. I especially would like to thank:
Petter Minnhagen for taking me in and letting me fully explore and experience the world of science. And with his never ending optimism and infinite well of ideas making it fun and interesting to go to work every day.

The other group members in Umeå: Martin Rosvall, Seung Ki Baek, Andreas Grönlund, Petter Holme, Ala Trusina, Luis E.C. da Rocha and Beum Jun Kim, for collaborations, feedback and friendship.

The people in the C-mol group in Copenhagen: Kim Sneppen, Sandeep Krishna, Mogens Høgh Jensen, Jacob Bock Axelsen, Mille Micheelsen, Namiko Mitarai, Ludvig Lizana, Philip Gerlee and all the rest for making every stay in Copenhagen fun, interesting and productive.

All the employees of the department of physics for making the work place a social environment with a lot of fun and interesting discussions during the coffee break. Especially the administrative staff, both in Umeå and Copenhagen, and Jörgen Eriksson for always lending a helping hand when ever needed.

And last, but not least, my family and friends for all the support. A lot of extra credit is in order for my wonderful fiancee, Frida Hägglund, for her tremendous patience and understanding during the time I was completely lost in my thesis cocoon.

Contents

Chapter 1

Complex Networks

Have you ever been amazed by the speed at which some news reaches the people around you? Or by the fact that when you meet a complete stranger you often seem to have a mutual acquaintance? The explanations to many such everyday phenomenas can be found in the field of complex networks, which studies interconnected systems where the patterns of interactions between the constituents play an important role. These networks affect our lives daily, like the Internet, the world wide Web and the protein networks in our cells.

The field of complex networks originates from graph theory which was born as early as in the 18th century from studying problems like how to visit all the cities in a country without crossing ones own path [16]. The field took a big leap when fast computers with a high computational capacity became available since it gave scientists the opportunity to perform fast simulations on large systems. During the recent years the field has been dominated by measuring real world networks, trying to find connections between the structure and the function of a network and to understand the process of evolving networks. It was found that many networks from completely different parts of our world, like those mentioned above, have some common features. Many real world networks, for example, display a small world effect meaning that all entities of the network are separated by only a small number of steps. Another common property is that most entities of a network have only a low number of connections while a few entities are very well connected [11]. This is usually referred to as a broad distribution of connections, also called scale-free [10]. The questions that arose as a consequence of these findings were regarding the universality of such properties and the functional abilities that come with them [4, 20]. What kind of processes are behind the assembly of these networks, creating the observed structures?

There are a number of books available on this topic both with a popular scientific approach (e.g. *Six Degrees: The Science of a Connected Age* by D.J. Watts)[86, 9] and those giving a more technical description of network science [19, 27].

1.1 Definition of nodes, links and complex networks

Most people have a fairly good idea of what is meant by the word "network". However, to rid us from the risk of misunderstandings we need a clear definition. A network, or graph, is a web of connections, or links, which represents some sort of information flow between the constituents that make up the network. These constituents, usually referred to as nodes or vertices, can take the form of people, animals, computers, web pages, proteins, metabolites etc. Furthermore, the information flow can represent gossip, the flow of nutrients up the food chain, electrical impulses, switching a gene on or off, and much more.

The condition on the links to represent some sort of "information flow" is used to highlight the fact that even though we could, with enough imagination, construct an almost endless number of networks, many of them would not pass as "real" networks in this sense. If two stocks seems to go up and down together in a correlated fashion, we could be tempted to put a link between them and make a network of companies. But the fact that they are correlated do not have to mean, and it usually doesn't, that the increase in the stock prize of one company is the direct cause of the increase in the stock prize of the other. The conventional use of the term "network" requires a direct cause and effect relationship.

So, what do we mean by a *complex* network? The word "complex" is another term with as many meanings as there are scientists. This term is sometimes separated from the word "complicated" by the notion of solubility. A complicated problem is difficult but straight forward to solve, while a complex system includes some higher order structure which is unreachable to us. The phrase "the whole is more than the sum of its parts" says it pretty well. The performance of a complex network is not only dependent on the nodes but also on the interactions. Or, as an example, the success of a soccer team is not determined only by the names of the players but also on how they play together. Another definition of a complex system is a system which is neither perfectly regular nor completely random. That is, a complex system has nontrivial structures which are indeed difficult to deal with analytically.

The number of nodes in a network will here be denoted as N, and the total number of connections as M. Since a link has two connections (ends), the number of links is $M/2$. The degree, or connectivity, refers to the number of links attached to a certain node, and will be denoted as k. Thus, the average degree in a network is $\langle k \rangle = M/N$. The links can also be directed or undirected, depending on the actual meaning of the link, and if information can flow both ways. The Internet, and most friendship networks are undirected while, for example, the World Wide Web and food webs are directed. A reindeer will not suddenly eat a wolf. Directed links are usually illustrated by arrows with an outgoing- and an ingoing end. This means that a node has both an out- and an in-degree (for a directed network the total number of in-connections, out-connections and links are the same).

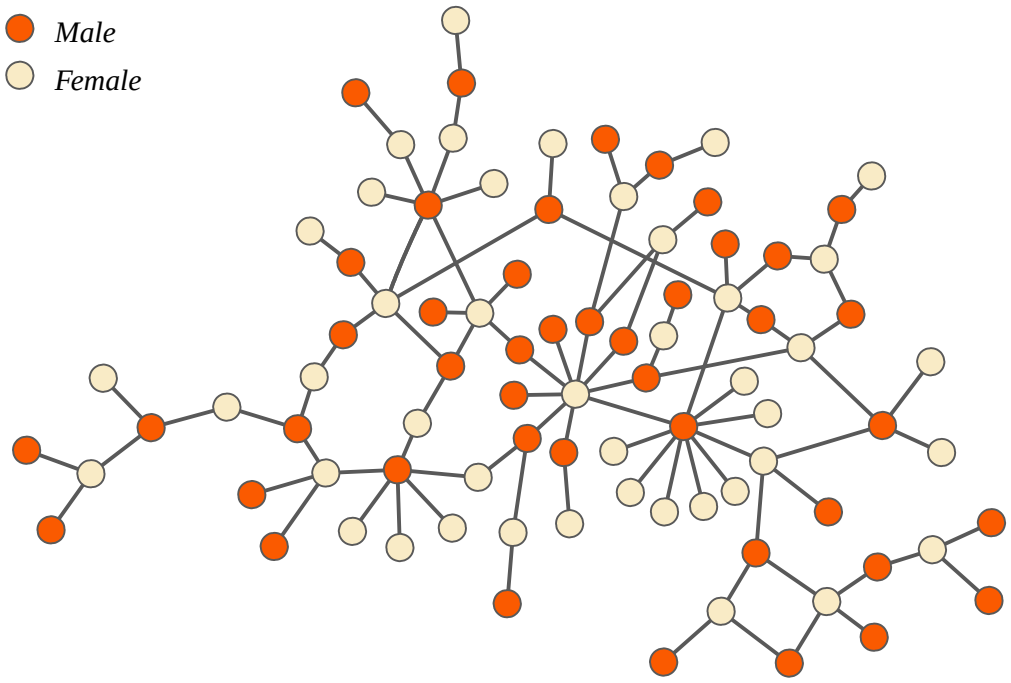


Figure 1.1: *Dating network for celebrities in USA [35]. The two well connected persons in the middle are Gwyneth Paltrow and Leonardo DiCaprio*

1.2 Real networks

The field of complex networks strongly rely on the existence of good data. That is, real-world networks mapped down to a set of nodes and links. Luckily, our world is full of them.

1.2.1 Social networks

Social scientists have been collecting data of human social interactions for a long time and there are many data sets available on this topic [76]. Examples of such networks are friendship networks of school children [32] dating networks (e.g. from on-line dating services or for celebrities as shown in Fig. 1.1 [35]), co-authorships [62, 12], business relationships [37], sexual contacts [53] and many more. Social networks are usually highly clustered (see section 1.3.4) meaning that they are locally tightly connected.

1.2.2 Infrastructural networks

Infrastructural networks are man made constructions with the purpose to ease our daily life and enhance the communication in our society. These kinds of networks

often have geometrical constraints imposed on them since we are confined to live on a 2D surface. Examples of such networks are the Internet, car roads, flight roots between airports, and the World Wide Web (which does not have any geometrical constraints).

Internet

The Internet is a huge, fast growing, network of computers communicating with each other by sending digital packages of information. Usually a zoomed-out version, to the autonomous-system level, is used to decrease the number of nodes of the system [31]. An autonomous system is simply put a collection of connected computers with a common IP prefix. These computers communicate with others through the Internet using a common routing policy. The Internet has a hierarchical structure [85] where the biggest hubs are connected to each other and to medium sized nodes. These, in turn, are connected to smaller nodes, and so on.

Roads

The roads we use when driving to work, visiting our friends and family, or when going shopping, make up a very important infrastructural network for the functioning of a society. Goods and people are being transported, and we all want to reach our destination as fast and easy as possible. There are two common ways of representing a system of roads as a network. The first one is to use the intersections as nodes and the streets as links [22]. This makes sense in the way that cars are flowing on the links between intersections. However, for many purposes a node should be the start-, and the end point when traveling through a network. When driving to visit a friend, your home address (a road) is the starting point, and your friends address (another road) is the endpoint. So, in this sense it might be better to make a representation where the roads are nodes and the links represent the crossing of two roads, symbolizing the fact that it is possible to make a turn on one road to end up on the neighboring road [75].

Airports

Every day, thousands of airplanes fly between airports all over the world, making up a network of flight routes [23]. The common practice is to use only links representing regular flights with some lower bound frequency of departure. The links can be weighted according to flight frequency, number of passengers or amount of cargo being transported, depending on the interest of the study [24]. Airports are also highly hierarchical (as described for the Internet).

World Wide Web

The World Wide Web (WWW) is a network of hyperlinks, connecting web pages [3, 1]. This network is by definition directed since web page administrators can only create links from their own web page to other pages, and not the other way around. But, the other pages can, of course, link back, creating a double link, which together works in practice like an undirected link. The WWW is a virtual network and thus has no geometrical constraints. It is also a huge network which has been growing extremely fast during the last 15 years. An interesting feature is a peculiar bow tie like structure [29]. It turns out that about a forth of all the web pages are a part of a strongly connected giant component (SCGC), where all pages can reach each other. Another forth is a part of a giant connected "in-component", where everyone can reach pages down stream of themselves leading to the SCGC. A similar giant connected "out-component", leading away from the SCGC, also constitute approximately a forth of the pages. The final forth consists of tendrils leading out from the in-component and in to the out-component plus pages isolated from the main bulk.

1.2.3 Intracellular networks

Important and interesting networks can be found also in living cells. For example, the proteins that preform all the daily tasks needed to keep us alive are working together in elaborate webs of interactions. There exist several types of protein networks, dealing with different types of interactions. Two examples are protein-protein networks and regulatory networks. The metabolism is another type of network where food is transformed into more usable molecules.

Protein-protein networks

Protein-protein interactions are physical interactions which are extremely important and used in almost all cellular processes. In protein-protein networks the interactions represent the ability for a pair of proteins to bind and form a complex. Since virtually all proteins can bind under the right conditions a threshold on the probability of binding is needed to weed out "unimportant links" and to avoid a fully connected network. Such a threshold brings in a subjective element in the analysis of the system [45].

Regulatory networks

The DNA is the blueprint of life. And not only does it encode for all the proteins needed for sustaining life, it also encodes for the mechanism of controlling *when* they are needed. The DNA is regulating itself by giving some proteins control over the production of other proteins. Thus, a protein can turn another protein *on* or

off by blocking (repressing) or activating (promoting) the read off (transcription) of the gene in question. These proteins and their regulatory interactions create a regulatory network where the links are directed and with the properties of turning its neighbors *on* or *off* [30, 42, 60].

Metabolic networks

Food need to be digested in order to extract the key molecules used as energy sources in molecular processes. Once the raw food has been taken in by a cell, it is transformed in chains of reactions, catalyzed by enzymes, until the desired products are produced. Each reaction has substrates as input and products as output and the metabolic network can be represented in three different ways: As a reaction network where different reactions are connected if the output of one reaction is the input of another. As a substance network where the substances are linked together if one substance is needed in the making of another. And, finally, as a bipartite network where there are two kinds of nodes, reactions and substances, connected in an alternating fashion. A substrate is linked to a reaction for which it is an input, and the reaction is, in its turn, linked to the substances it produces [56, 55, 48].

1.3 Network structures and properties

The structure, or topology, of a network is about what kinds of patterns of connections exist in the network. In order to investigate what organizational principles and evolutionary rules there are governing the structure of real world networks, the structure needs to be quantified and measured. This is also necessary when trying to classify different types of networks. The structure of a network is presumably also important for its function. It affects a networks resilience against random failures [4] and breakdowns [43], as well as the speed at which signals (e.g. deceases) can spread through a network [66].

Many measures have been developed over the years to capture various properties of networks. They range from large scale properties like modularity (subgraphs with more internal links than links to the outside) [70] to motifs of different shapes and sizes [78] and down to point properties of nodes. This section is devoted to some of the simpler, and much studied, properties of networks.

1.3.1 Degree distribution

The degree is in many cases a property which reflects the importance, or the role, of a node in a network [4]. An important feature of the whole network is then the distribution of degrees. That is, the number of nodes, $N(k)$, or the fraction of nodes, $P(k) = N(k)/N$, with a certain degree k. The system size is related to the degree distribution in the following way

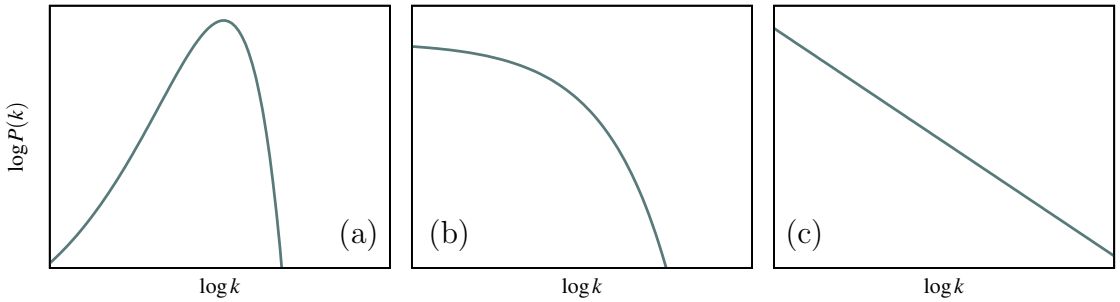


Figure 1.2: *Discrete probability distributions in log-log scale: (a) Poisson, (b) exponential and (c) power law.*

$$\sum_k N(k) = N \quad \Rightarrow \quad \sum_k P(k) = 1 \tag{1.1}$$

$$\sum_k kN(k) = M \;\Rightarrow\; \sum_k kP(k) = \langle k \rangle \tag{1.2}$$

Poisson

An important bench mark in network science has been the "random" Poisson distribution. It has been widely used as a null model representing a random network (see section 1.4.2). The Poisson distribution is described by the expression

$$P(k) \propto \frac{\langle k \rangle^k}{k!}, \tag{1.3}$$

which is peaked at the average value and has a very fast decaying tail on both sides (Fig. 1.2a). The distribution coincides with the Gaussian distribution at high values of $\langle k \rangle$.

Exponential

Another common distribution in nature is the exponential distribution (Fig. 1.2b). This distribution, given by Eq. 1.4, is monotonically decreasing, but a characteristic scale, proportional to the average value, determines the rate of decrease.

$$P(k) \propto \exp(-k/k_0) \tag{1.4}$$

where $k_0 \propto \langle k \rangle$.

Power law

The field of complex networks exploded in the late nineties when it was discovered that many real networks exhibit the peculiar property of having a degree distribution described by a power law. A power-law distribution is broad (or heavy tailed) in the sense that, even though most nodes will have small degrees, there exist a few nodes with a huge amount of connections (Fig. 1.2c). The power-law distribution is given by

$$P(k) \propto k^{-\gamma}, \tag{1.5}$$

where the exponent γ is a positive number, usually between 1 and 3 for real systems [11]. This distribution is often referred to as "scale free" since it is scale invariant according to the relation $P(ak) = A(a)P(k)$.

Presentation

A convenient way of plotting the degree distribution is to use logarithmic scale. This gives a more detailed view of the behavior for large k and a quick hint to what extent the distribution follows a power law since there is a linear relation between $\log P(k)$ and $\log k$.

There are also two conventional ways of increasing the range, and reducing the fluctuations, of a distribution generated by a stochastic process. One is to plot the cumulative distribution given by

$$F(k) = \sum_{k'=k}^{\infty} P(k'), \tag{1.6}$$

which is the fraction of nodes that have a degree larger than, or equal to, k. For a power-law distribution the exponent is simply decreased by one since the primitive function of $k^{-\gamma}$ is $k^{-(\gamma-1)}$.

Another way is to bin the data with an increasing bin size. That is, take the average value of the points in each bin and display them at the center of the bin. The size of bin i could, for example, follow the expression $S_i = 2^{i-1}$ (the first bin contains $k = 1$, the second $k = 2$ and 3, the third $k = 4, 5, 6$ and 7, and so on), which makes the bins equally separated in log scale. This method works well for monotonically decreasing functions, like the exponential or the power law, where there is a falloff in statistics with increasing k.

1.3.2 Shortest path

A measure of distance in a network is the number of steps needed to go from one node to another[1]. Since there might exist a very large number of possible paths connecting two nodes, the shortest-path length is usually used. This can be motivated by a simple flow analogy: If information flows down all links with equal speed, and all links have the same length, then a receiving node will first get the information through the shortest path.

The size of a network is naturally determined by the number of nodes and links, but, a simple network representation does not have a spatial extension. In order to get a feeling for the "volume" of a network, a common practice is to measure a diameter. Several definitions have been proposed, but the most popular one is probably the average shortest-path length [3] as described by

$$D = \langle d \rangle = \frac{1}{N(N-1)} \sum_{i}^{N} \sum_{j>i}^{N} d_{ij}, \tag{1.7}$$

where N is the number of nodes and d_{ij} is the shortest path between node i and j.

Another definition is to use the longest shortest path between any two nodes [69].

1.3.3 Centrality

There are situations when it might be crucial for the problem at hand, or simply just fun, to find the most important nodes in a network. It could be persons that have a high risk of being infected by a disease or computers that are vital for the transmission of digital messages. There exists several measures designed to capture these nodes, all with the common goal of quantifying some sort of *central* role in the network. Two commonly used definitions are *betweeness centrality* and *closeness centrality* [36].

Betweeness centrality

One way of defining an important node is through the number of shortest paths that it is a part of. If one passes a certain node very often when moving between random pairs, then this node can be considered as a very central node. It also means that if this node is removed then many shortest paths are made longer, or it might even break up the network into disconnected pieces.

[1] If the links are weighted according to a distance related quantity then the distance might be measured as the sums of the weights along a certain path

Closeness centrality

To be a good broadcaster in a network a node should be as close to all other nodes as possible for the messages to quickly reach its destinations. This leads to a centrality measure where the most important node is the one with the smallest average shortest path to all other nodes. Nodes with high closeness centrality (small average shortest path) often have a high degree since each link constitutes a short cut in the network.

1.3.4 Clustering coefficient

The clustering coefficient (CC) is a measure of the number of triangles existing in a network, normalized by the possible number of triangles that could exist. A triangle in a social network means that if A and B are friends, and A and C are friends, then B and C are also friends. A subgraph with a high density of triangles implies a tightly connected module.

A local clustering coefficient, introduced by Watts and Strogatz (1998) [87], counts the number of triangles involving a certain node, divided by the total number of possible triangles that could be formed in the neighborhood of that node. The local CC for node i is then

$$C_i = \frac{2N_\triangle}{k_i(k_i-1)}, \tag{1.8}$$

where $N_\triangle$ is the number of triangles (three nodes where everyone is connected to everyone) and k_i is the degree of node i. A total average CC can then be calculated as

$$C = \frac{1}{N_{k>1}} \sum_{i,k_i>1} C_i, \tag{1.9}$$

were $N_{k>1}$ is the number of nodes with a degree larger than one.

Another definition was introduced by Barrat and Weigt (2000) [13] as a global clustering coefficient defined as

$$C = 3\frac{N_\triangle}{N_\wedge}, \tag{1.10}$$

where $N_\triangle$ is the total number of triangles and $N_\wedge$ is the total number of triplets (three nodes where at least one node is connected to the two other nodes) in the network.

The two definitions are the same when calculating the CC of a single node.

1.3.5 Degree correlations

Degree correlations is a "one step" local measure in the sense that it addresses the question *who is connected to whom*, with identities represented by degrees. That is,

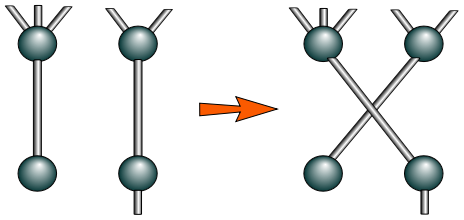


Figure 1.3: *Randomization scheme keeping the degree of each node fixed.*

do low degree nodes tend to connect to high degree nodes or do they prefer other nodes with low degree?

The degree-correlation profile was introduced by Maslov and Sneppen (2002) [59, 61] with the aim to measure the quantity $P(k_i, k_j)$, which is the probability for a node of degree k_i to be connected to a node of degree k_j. The correlation matrix, $R(k_i, k_j)$, is then calculated by taking the ratio of the number of links connecting nodes of certain degrees in the observed network and the average result for a random null model.

$$R(k_i, k_j) = \frac{P_{obs}(k_i, k_j)}{P_{rand}(k_i, k_j)}. \tag{1.11}$$

The null model is furthermore designed to keep the degree of every node fixed since degree correlations depend strongly on the degree distribution. The randomization is done by picking two random links and exchanging the connections of two of the nodes, as illustrated in Fig. 1.3. To assure good statistics even for large k-values (for which there exist only a few nodes) the range is binned with an increasing bin size for increasing k (e.g. bin 1 contains $k = 1, 2, 3$, bin two $k = 4...10$, bin three $k = 11...30$ etc.).

Newman (2002) [67] suggested another measure, called the *assortativity*, based on the Pearson correlation coefficient which ranges between the values -1 and 1. The Pearson correlation coefficient measures the linear dependence between two random variables and can be written as

$$r = \frac{\langle jk \rangle - \langle j \rangle \langle k \rangle}{\sigma_j \sigma_k}, \tag{1.12}$$

where $\langle ... \rangle$ stands for an ensemble average, j and k are the outcome of the two random variables and σ is the standard deviation. For the assortativity in a network $\langle ... \rangle$ means an average over all links, and j and k are the degrees of the nodes on either side of a link. The variables j and k cannot be separated in an undirected network (there is no "left" or "right" on a link). To get around this problem the term $\langle j \rangle \langle k \rangle$ is replaced by $\langle \bar{k} \rangle^2$ where $\bar{k} = \frac{1}{2}(j + k)$ is the average degree of the two connected nodes. The resulting formula then becomes

$$r = \frac{4\langle jk\rangle - \langle j+k\rangle^2}{2\langle j^2+k^2\rangle - \langle j+k\rangle^2}. \tag{1.13}$$

The value still ranges between -1 and 1, where -1 means perfect disassortative mixing (connected nodes have very different degrees) and 1 means perfect assortative mixing (connected nodes have the same degree). The assortativity measure can also be designed to capture different types of node correlations other than the degree, depending on what kinds of node characteristics exist in the data (e.g. language, race, age etc) [68].

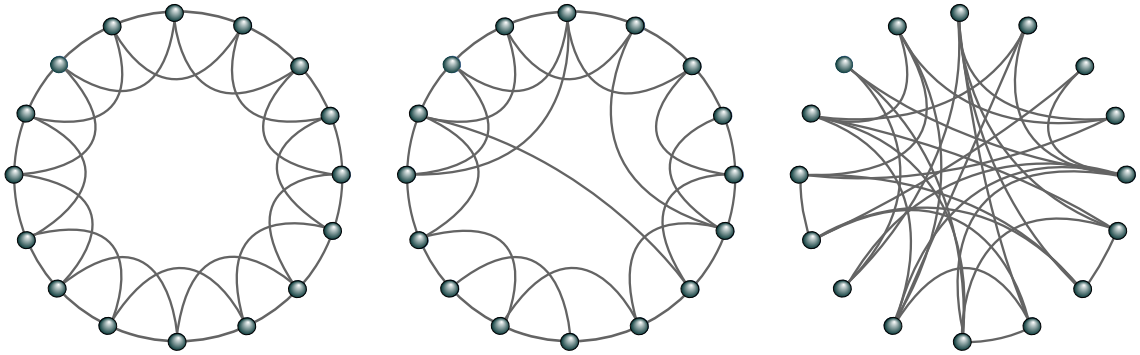
1.4 Network models

Models are used to give hints to the origin of some observed property and to teach us something about how the system works. If a model accurately contains every possible action that takes place in the system we have not really gained any new knowledge. So, a good model should therefore be able to reproduce the desired property by only a few simple rules, suggesting that these rules are possibly the most important reasons for the appearance of a particular property. In an attempt to reproduce structures of real networks many authors have developed models of the evolution, or assembly, of networks.

1.4.1 Small world

In 1967, Milgram preformed a famous experiment where he sent out letters to randomly selected persons in USA, asking them to forward the letter to a predetermined target person [47]. The only catch was that the letter was only allowed to be sent to a friend on a first name basis. The task was thus to choose a friend believed to be closer (geometrically, professionally etc.) to the target person. This friend in his/her turn had to forward the letter again, and so on, until the final destination was reached. When collecting the letters that made it all the way (about 25%), Milgram found that the average number of steps taken to reach the target person was six. And thus the expression "six degrees of separation". This is also called a *small-world phenomenon* which implies that people on our planet are much closer connected than we might imagine at first.

In 1998, Watts and Strogatz [87] developed a network model (WS) describing the situation of having short distances between nodes and high clustering at the same time, which they appropriately gave the name 'small-world' networks. The model contains a continuous transition from a perfectly regular network (Fig. 1.4a), a lattice, to a completely irregular network (Fig. 1.4c), a random graph. The lattice has high clustering but a very large diameter, while a random network has a very low clustering but a small diameter. The region between these two extremes was



(a) $p = 0$ (b) $p = 0,05$ (c) $p = 1$

Figure 1.4: *The WS model where the parameter p gives the probability for a link to be randomly rewired: (a) A perfectly regular network $p = 0$. (b) For a small p short cuts are created shrinking the diameter of the network. (c) A perfectly irregular network is created when $p = 1$ with a small diameter and low clustering.*

explored by introducing a probability, p, for a link to be randomly rewired, and thus to create a short cut in the system. Consequently, $p = 0$ corresponds to the lattice and $p = 1$ to the random network as shown in Fig. 1.4. It turned out that it takes just a few rewirings ($p \sim 0.01$, around 1% of the links are rewired) to get a small diameter that scales as $D \propto \ln N$, instead of the linear dependence on N which is the case for the lattice. On the other hand, the clustering coefficient do not reach the small values similar to those of random networks before a large fraction ($p > 0.5$) of the links has been rewired. Thus, there exists a 'small-world' region for small $p > 0$. These results have many practical implications where one is the spreading of diseases on social networks. It can be shown that the spreading of diseases on a network is much faster if it exhibits small-world properties since the short cuts connects otherwise distant parts of the network. But it is, at the same time, difficult for individuals to be aware of these short cuts since the local structure (e.g. clustering) is very weakly affected by a few random rewirings.

1.4.2 ER model

An a priori assumption, or approximation, when considering a real network could be that it is random in the sense that there is no preference for anyone to be connected to anyone else in particular. Erdős and Rényi developed a random graph model in 1959, usually referred to as the ER-network [72], in which every pair of nodes have the same probability, p, to be connected. The algorithm is very simple:

Start with N disconnected nodes.

(i) Pick a pair of nodes that have not been picked before.

(*ii*) Put a link between them with probability p.

These steps are then repeated until all pairs have been picked. The ER-model is suitable for analytic calculations due to the lack of structure in the network, like degree correlations (all nodes, regardless of their degree, have the same probability to be connected to any other node).

The expectancy value of the average degree is

$$\langle k \rangle = (N-1)p \approx Np, \tag{1.14}$$

and the degree distribution can be found by realizing that the probability for a certain node to have degree k follows the binomial distribution

$$P(k) = p^k (1-p)^{N-1-k} \binom{N-1}{k}. \tag{1.15}$$

That is, the probability to get k links, times the probability to *not* get $N-1-k$ links, times all the combinations in which this happens.

The binomial distribution coincides with the Poisson distribution in the large N limit, which leads to a degree distribution given by

$$P(k) = \frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!}. \tag{1.16}$$

Even the clustering coefficient, as defined by Eq. 1.10, is easy to calculate analytically since given an open triplet, the probability for the remaining two nodes to be connected is p. This means that the clustering coefficient can be calculated by

$$C_{ER} = \frac{pN_\wedge}{N_\wedge} = p. \tag{1.17}$$

The parameter p regulates the number of links in the network and thus to what extent the network is connected. When increasing p the network gets more dense and the chance of having a path between every node in the network increases. There exists a percolation threshold, for large N, at $M = N$ ($p \approx 1/N$). When the number of links is smaller than this threshold the network consists of small, isolated, components with sizes of order $\mathcal{O}(\log N)$. When the system is above the threshold the network becomes almost completely connected and a single giant component is formed with a size of order $\mathcal{O}(N)$ [69]. The ER-network also exhibits small-world properties in the sense that all nodes can be reached through a small number of steps. It has been found that above the percolation threshold, the average shortest-path follow the expression [17]

$$\langle d \rangle_{ER} \sim \ln N / \ln \langle k \rangle. \tag{1.18}$$

1.4.3 BA model

The first model to reproduce the power-law behavior of real networks was presented by Barabási and Albert (1999) [10] and has been the inspiration of much work in the field since then. The model is a special case of the Simon model (see section 3.4.1), as pointed out by Bornholdt et al. (2001) [18], and is based on growth and preferential attachment of links. The latter element is motivated by the rich-get-richer phenomenon [25] which addresses the notion that it is much easier to make money if you have a lot of money already. Or, it is easier to make new friends if you have many friends to start with, and are well known in the community. The algorithm of the Barabási and Albert (BA) model is:

Start with a small set of nodes and links.

(i) Add a new node with m links.

(ii) Attach each of the new links to an existing node, i, with a probability proportional to the degree of that node, $p_i \propto k_i$.

These steps are then repeated until the network consists of the desired number of nodes, N.

The networks produced by this algorithm will have, in the large N limit, an average degree of

$$\langle k \rangle \approx 2m, \tag{1.19}$$

and the degree distribution will follow a power law with the exponent $\gamma = 3$, independent of the parameter m. Worth noting is that the power-law behavior cannot be obtained by simple preferential attachment, without growth, or by using only growth with uniform attachment. They are needed together. Also, the preferential element has to be linear.

Since it was first introduced by Barabási and Albert, many proposed extensions and modifications of the model have seen the light of day. Most of them are concerning the preferential element of the model but there are also versions including rewiring of links and removing of nodes [2].

1.4.4 Merging model

The *merging-and-regeneration* model was first introduced in the field of networks by Kim et al. (2005) [50], and has been used to model, for example, the size distribution of solar flares (sun spots) [64, 82][2]. The merging element has also been used in non-network models to reproduce the size distribution of ice crystals and the length distribution of α-helices in proteins [33]. The model was constructed for undirected

[2]The articles [64] and [82] has a publication date of 2004 but they are both citing the preprint of Kim et al.

networks and based on the notion that systems should continuously try to optimize their function. Since the main function of many systems is to transfer information, the idea was to develop a dynamical process where the signaling capability was increased. The two lines of action used was shortening of signaling paths, and growth of signaling hubs. As progress goes on, several smaller routers in the Internet can be exchanged by a larger, and faster, router, making it possible to send information in a more efficient way. At the same time, new computers, or routers, are added as the network extends, and more people get connected to the Internet. This results in the following algorithm:

Start with N nodes and M links, connected in an arbitrary way.

(i) Pick a node, i, and one of its neighbors, j, at random.

(ii) Merge the two nodes by letting node i absorb node j and all of its links, except those they previously shared. The resulting node will thus get the degree $k_i + k_j - u$, where u is the number of links that were discarded in order to avoid double links and self loops.

(iii) To keep the number of nodes fixed, add a new node with degree r and connect it to r random nodes.

When repeated many times, a steady state situation is reached when $u = r$. That is, the number of lost links in the merging step equals the number of added links in the regeneration step, on average. The model creates scale-free networks with a power-law exponent between -3 and -2 with increasing r. A slight modification of the model where both nodes to merge are picked randomly generates an exponent around 1.5.

1.5 Summary of papers

1.5.1 Paper I

In the first paper entitled *Models and average properties of scale-free directed networks* we extend the merging model, described in section 1.4.4, to directed networks and investigate the emerging scale-free networks. Two versions of the model, *friendly-* and *hostile merging*, are described and it is shown that they represent two distinctly different types of directed networks, generated by local update rules. Also, two minimalistic model networks, *model A* and *B*, are introduced as prototypes of these two kinds of networks. Furthermore, it is shown that the distinctive features of the two network types show up also in real networks from the realm of biology, namely *metabolic-* and *transcriptional networks*.

The measures used to classify these directed networks is the in- and out-degree distribution, the average in-degree of a node as a function of its out-degree, the

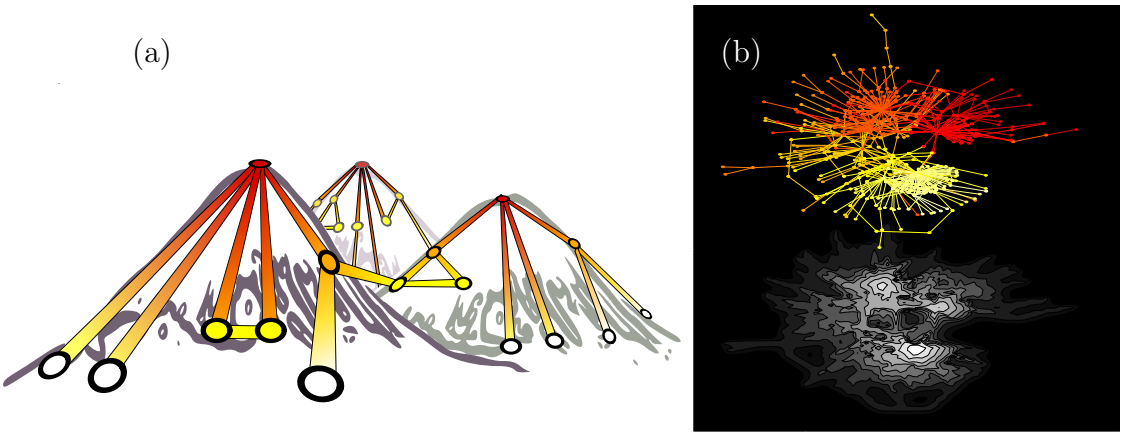


Figure 1.5: *Landscape analogue: (a) Landscape analogue where high degree nodes have a high altitude. The color coding represent a node property proportional to the degree of the node (red high, white low). (b) Network with separated hubs and ridged landscape generated by the algorithm described in the text. The color coding of the network represent a node property other than the degree (here a random number), and for the landscape (contour map) it represents the degree (white high, black low).*

spread of the in-degrees of nodes with a certain out-degree, and finally the portion of nodes with only in-, only out- or with both in- and out-links.

It turns out that metabolic networks belong to type A of directed networks (model A and friendly merging) where the in- and out-degree distributions are identical, and there is a linear dependence between the average in-degree and the out-degree of a node, $\langle k_{in} \rangle_{k_{out}} \approx k_{out}$. This is a non-trivial property which can be analytically shown to hold for degree distributions following a power law but not for e.g. Poisson. Furthermore, the spread follows a power law with a slope close to $-1/2$ and there are about the same portion of nodes with only in-links as there are nodes with only out-links.

The regulatory network of yeast, on the other hand, belong to type B (model B and hostile merging) with a broad out- and a narrow in-degree distribution. The in-degree of a node, as well as the spread of in-degrees, are in this case independent of the out-degree. Also, the fraction of nodes with only in-links is far greater than those with only out-links.

1.5.2 Paper II

In the paper *Degree landscapes in scale-free networks* we generalize the degree-organizational view of real-world networks with broad degree distributions [73, 61, 59, 67, 85]. We present a landscape analogue where the altitude of a site is proportional to the degree of a node (Fig. 1.5a) and measure the widths of, and the

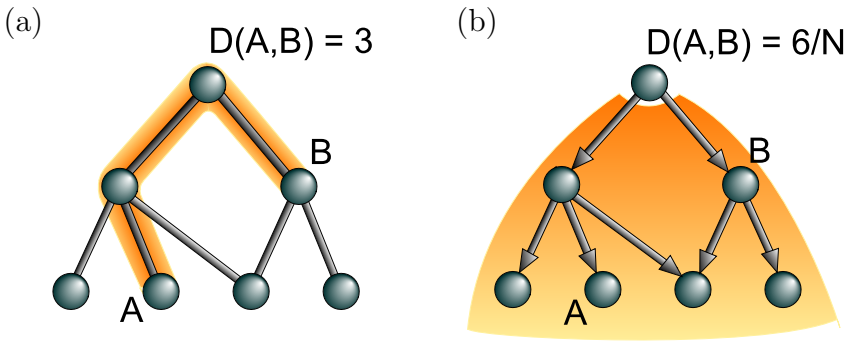


Figure 1.6: *Two definitions of GO distance: (a) A direct distance as the shortest path between two nodes, A and B. (b) A hierarchical distance as the fraction of nodes downstream of the closest "ancestor" of two nodes, A and B.*

distances between, mountain peeks in a network. It is found that the Internet and the street network of Manhattan display a smooth, one mountain, landscape while the protein network of yeast has a rough landscape with several separated mountains. It is speculated that these structures reflects the type of node property (e.g. degree, functional ability etc.) that is crucial to the organizational principle of the network. With this in mind, we suggest a method for generating ridged landscapes where a random rank is assigned to every node, symbolizing the constraints imposed by the space the network is embedded in. The constraint can be associated to spatial or in molecular networks to functional localization. The network is then randomized keeping each individual degree (see section 1.3.5) but where nodes of similar ranks are connected. When introducing a small error rate, the algorithm creates small-world networks with ridged landscapes (Fig. 1.5b) similar to those seen in many biological networks. Also, the rank gradient is still preserved which was supposedly the original organizational goal.

1.5.3 Paper III

In the paper *One hub-one process: A tool based view on regulatory network topology* we extend the work done in paper II by studying the similarity of node properties as a function of distance in the regulatory network of yeast. In other words, we try to find a real version of the gradient displayed in Fig. 1.5b. Using the Gene Ontology (GO) Consortium annotations [54] we show that locality in the regulatory network is associated to locality in biological processes, and only weakly related to the functional ability of a protein.

The GO database is in the form of an acyclic directed graph (similar to a tree with connected branches) which organize proteins according to a predefined categorization. Lower ranking proteins in a GO-graph share large scale properties with higher ranking proteins, but are more specialized. There are three different catego-

rizations: (P) Which type of *biological process* a protein takes part in (cell division, metabolism, stress response etc.). (F) What kind of *molecular function* a protein performs (transcription factor, DNA binding etc.). (C) In what *cellular component* a protein is acting (nucleus, membrane etc.).

The similarity in the node property of two nodes is measured as a GO-distance, D, for the three categorizations respectively. We define two different distance measures capturing two separate definitions of closeness. The first measure is a direct distance. That is, the shortest path length between the two nodes in the GO-graph (Fig. 1.6a). The other measure is a hierarchical distance which gives a large distance between two nodes that are close to the root, but on different branches. The hierarchical distance is defined as the fraction of nodes downstream of the lowest common "ancestor" of the two nodes (Fig. 1.6b).

By using the method introduced in paper II we rewire the network with a bias towards closeness in the GO-graph. The results indicate that nodes downstream of a hub, in the real network, has been brought together with maximum bias on process closeness.

Overall we suggest that the topology of the yeast network is governed by processes located on hubs, each consisting of a number of tools in the form of proteins with quite different functional abilities. Our findings also suggest that the rewiring of links play a bigger role than gene duplication [84] during the network evolution.

Chapter 2

Statistical Mechanics and Networks

Einstein once said "God doesn't play dice" when referring to the, at the time, new ideas of quantum mechanics. However, he only objected to the lack of determinism of the fundamental laws of quantum mechanics. Many aspects of physics are well described by statistical mechanics and in this case the whole concept is based on probabilities and dice throwing.

In fact, Einstein himself[1] used statistical mechanics to solve the problem of Brownian motion (first observed as a pollen particle in water, moving around in an irregular fashion), confirming the existence of atoms and molecules [28]. It is the collective motion of all the surrounding molecules that is responsible for this irregular movement by pushing on the pollen particle in random directions. The randomness is caused by the simple fact that the particle is sometimes hit from the left and sometimes from the right, and chance plays a role just like when throwing a dice. Statistical mechanics gives the tools to describe and predict many properties of large ensembles under the influence of such random events. In this chapter the concept of entropy and maximization of the entropy will be described, both in general and for networks. A broader and more thorough, but easy going, introduction to statistical mechanics and thermodynamics can be found in the first chapters of the book *Molecular Driving Forces: Statistical thermodynamics in Chemistry and Biology* by Dill, Bromberg and Stigter [26].

2.1 The concept of entropy

Entropy is one of the key concepts in thermodynamics and statistical mechanics. In thermodynamics, entropy is defined as a macroscopic quantity measuring the level of order of a systems. A highly ordered system has a low entropy while a very disordered system has a high entropy. The second law of thermodynamics then states that the entropy of a system can spontaneously only increase. That is, an

[1] and Marian Smoluchowski independently

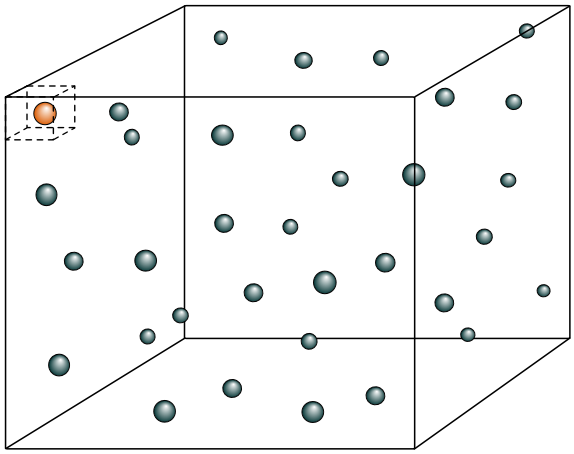


Figure 2.1: *Particles in a box. The small dashed cuboid represents one possible location for the orange (light gray) particle. The total number of possible locations is thus the number of cuboids that can be fitted inside the box.*

isolated system will tend to increase its disorder. To make this more intuitively clear, imagine a system of gas molecules confined in a box. We then implement a constraint on the system by manually forcing all the gas molecules to be closely packed together in one of the corners of the box. We have now forced the system into an ordered state with lower entropy[2]. However, when relaxing the constraint we expect the gas to spread out in the box, due to the thermal motion of the molecules, leading to a spontaneous increase in the disorder. We would be very surprised if the opposite happened.

In statistical mechanics, however, entropy is defined as a microscopic quantity measuring the multiplicity of a system. Here multiplicity means the number of microscopic states (microstates) that a system could be in, given a certain macroscopic state (macrostate). A macrostate is an observed state of the whole system, and in the gas case having all the molecules spread out over the whole box can be considered as one macrostate and having them confined in a smaller volume (e.g. in one of the corners) another. But, in both cases each molecule could be located in many different places, and every configuration of the molecules location (inside the confined volume) is a microstate. One can think of this as a big Rubik's cube where every little cuboid is a possible location for a molecule (see Fig. 2.1). It then follows that the entropy will be larger if the molecules are spread out over a larger volume (larger number of "cuboids"), which is consistent with the thermodynamical definition based on increased disorder.

[2]Usually the state of a gas particle includes both its position and velocity. In this example the velocity is excluded for the sake of simplicity.

These two definitions are connected by the Boltzmann expression

$$S = k_B \ln \Omega, \tag{2.1}$$

where S is the entropy, Ω is the number of microstates and k_B is the Boltzmann constant (in thermodynamics $k_B = 1.380662 \cdot 10^{-23}\ JK^{-1}$). The entropy is extensive which means that the total entropy of two systems equals the sum of the two entropies. That is,

$$S_{A+B} = k_B \ln(\Omega_A \Omega_B) = k_B \ln \Omega_A + k_B \ln \Omega_B = S_A + S_B. \tag{2.2}$$

2.1.1 The maximum entropy principle

In statistical mechanics a macrostate is described by a frequency distribution of outcomes, $n(i)$, giving the number of constituents with outcome i. In the case of the gas example this distribution describes how often a certain location (cuboid) is occupied by a molecule, when taking a time average. To give another example one can think of a coin. A coin can give two different outcomes: head (H) or tail (T). Making a sequence of coin flips (a microstate) will then generate a distribution function (a macrostate) for the number of heads and tails. The entropy of the system is thus a measure of the number of microstates that are enclosed in this distribution function. For example, a certain sequence of coin flips can look like

$$HHTHTH \tag{2.3}$$

giving the macrostate $n(H) = 4,\ n(T) = 2$. But the sequences

$$HTHHTH,\ THHHTH,\ THHTHH, etc.$$

are also giving the same macrostate of four heads and two tails. The number of possible microstates, giving the same macrostate, can be calculated as follows: For the first element in the sequence we have N constituents (e.g. number of coin flips) to choose from, giving N combinations. For the second element we have one less constituent available to choose from giving us $N - 1$ possible combinations. Continuing this reasoning down to the last element gives us the expression $N(N-1)(N-2)...1 = N!$. This is the total number of configurations that can be constructed out of N distinguishable outcomes. However, in most cases constituents with the same outcome are not distinguishable. For example, exchanging two coins, both with the outcome H, do not give a new microstate. We cannot tell them apart. So, there are, for each configuration, $n(i)!$ permutations for outcome i giving the same microstate. Taking this degeneracy into account gives the final expression

$$\Omega = \frac{N!}{\prod_i^s n(i)!}, \tag{2.4}$$

where s is the number of possible outcomes (e.g. two for a coin and six for an ordinary dice). Note that describing a sequence of heads and tails as flipping one coin N times, is equivalent to describing it as N coins being flipped only once[3].

Using Sterling's approximation ($x! \approx (x/e)^x$) simplifies Eq. 2.4 into

$$\begin{aligned} \Omega &\approx \frac{(N/e)^N}{\prod_i^s (n(i)/e)^{n(i)}} \\ &= \frac{N^N}{\prod_i^s n(i)^{n(i)}}. \end{aligned} \tag{2.5}$$

Finally, taking the logarithm of both sides and defining the normalized frequency distribution $p(i) = n(i)/N$,[4] gives the formula for the entropy per constituents

$$\tilde{S} = \frac{1}{N} \ln \Omega = -\sum_i^s p(i) \ln p(i), \tag{2.6}$$

where $\tilde{S} = S/k_B N$. The macrostate, $p(i)$, that maximizes Eq. 2.6, and thus the entropy, is the uniform distribution $p(i) = 1/s$. This seems intuitively reasonable since it means that if we flip a coin many times we should get, on average, equally many heads and tails. Or, leaving the gas alone would give a situation where each location in the box has the same chance of being occupied by a molecule. Note, however, that this is only true if the coin is unbiased and if there is nothing from the outside influencing the positions of the molecules in the box. Phrased in the language of statistical mechanics it means that there must be an equal probability for the system to be in any microstate in order to spontaneously reach the macrostate with the maximum entropy. So, from all the possible macrostates that can be created from flipping a coin many times, or from leaving a gas alone in a box for a long time, there is one which largely dominates all the others in the number of microstates. When picking a microstate randomly, and uniformly (which is what one does when making a series of coin flips), it will basically always belong to the dominating macrostate. The principle of maximum entropy thus states that the macrostate with the highest entropy is the most likely one to be observed. It is like drawing lottery were the contestants have different number of lottery tickets, and the person with the most tickets have the highest chance of winning.

2.1.2 The Boltzmann distribution law

The maximum entropy solution presented in the previous section (given by a uniform distribution function), is obtained by assuming that there are no constraints on the

[3] Assuming that all coins are statistically equivalent.

[4] It is important to note that even though $p(i)$ can be regarded as a probability function, it do not refer to the real underlying probability of an outcome, but to the probability that can be inferred from an observation.

system. However, many systems are operating under various constraints. It could be gravity pulling the gas molecules in the above example, or indeed a false dice. In this case the maximum entropy solution is the macrostate with the largest number of microstates, but which is at the same time satisfying the constraints. A problem of maximization (or minimization) under various constraints can be solved by using variational calculus. Let us assume the constraint is some property of the system, E, which should be kept constant. This constraint can be written as

$$\sum_i^s E(i)n(i) = E \;\Rightarrow\; \sum_i^s E(i)p(i) = \langle E \rangle, \tag{2.7}$$

where $\langle E \rangle = E/N$. We also have to make sure that the function $p(i)$ is normalized, giving the second constraint

$$\sum_i^s n(i) = N \;\Rightarrow\; \sum_i^s p(i) = 1. \tag{2.8}$$

The next step is to maximize the entropy, $\tilde{S}$ (Eq. 2.6), given these constraints, and thus maximize the auxiliary function

$$\Phi[p(i)] = -\sum_i^s \Big(p(i)\ln p(i) + \alpha\Big(p(i) - 1\Big) + \beta\Big(E(i)p(i) - \langle E \rangle\Big)\Big), \tag{2.9}$$

where α and β are Lagrange multipliers. The maximum is found by fixing the derivative with respect to $p(i)$ to be zero for all i, which gives

$$\begin{aligned} \frac{\partial}{\partial p(i)}\Phi[p(i)] = 0 \;&\Rightarrow \\ \ln p(i) = -1 - \alpha - \beta E(i). \end{aligned} \tag{2.10}$$

Solving Eq. 2.10 then gives the maximum entropy solution

$$\begin{aligned} p(i) &= \exp(-1 - \alpha - \beta E(i)) \\ &= A\exp(-\beta E(i)). \end{aligned} \tag{2.11}$$

The actual values of the Lagrange multipliers α and β can be found by simultaneously solving Eq. 2.7 and 2.8 after substituting Eq. 2.11. An example of the result for a dice with an average outcome of 3 (instead of 3.5 for an unbiased dice) is shown in Fig. 2.2. Also, the physical meaning of these multipliers can be interpreted by examining the rule [49]

$$\alpha = \frac{\partial \tilde{S}^*(N,E)}{\partial N} \tag{2.12}$$

$$\beta = \frac{\partial \tilde{S}^*(N,E)}{\partial E}, \tag{2.13}$$

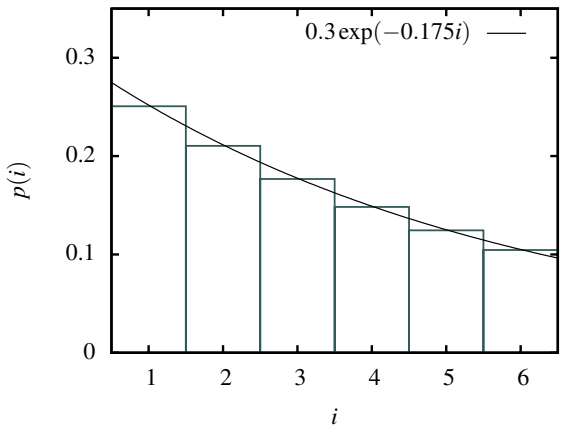


Figure 2.2: *The maximum entropy solution for the outcome of a six-sided dice with* $E(i) = i$ *under the constraint* $\langle i \rangle = 3$.

where $\tilde{S}^*$ is the maximum entropy. The meaning of α and β is thus the rate of increase of the maximum entropy with the number of constituents and with the quantity E, respectively.

Dividing both sides of Eq. 2.11 by Eq. 2.8 leaves the expression unchanged and we get

$$\begin{aligned} p(i) &= \frac{p(i)}{\sum_i^s p(i)} = \frac{\exp(-1-\alpha)\exp(-\beta E(i))}{\sum_i^s \exp(-1-\alpha)\exp(-\beta E(i))} \\ &= \frac{\exp(-\beta E(i))}{\sum_i^s \exp(-\beta E(i))}. \end{aligned} \tag{2.14}$$

Equation 2.14 is called the Boltzmann distribution law and the quantity in the denominator is a normalization factor called the partition function. In statistical physics and thermodynamics the quantity $E(i)$ is usually an energy controlling the system. For a gas in a box under influence of gravity it involves the potential energy of the molecules. The Boltzmann distribution law says that the probability for a constituent (e.g. a molecule or a coin flip) to have a certain outcome with energy $E(i)$ is proportional to the quantity $\exp(-E(i))$. The higher the energy, the lower the probability (for a given β).

2.1.3 The Boltzmann factor and the Metropolis algorithm

In thermodynamics, E is the internal energy of the system and it can be shown that the Lagrange multiplier in the previous section must be $\beta = 1/k_B T$ (in accordance with Eq. 2.13), where T is the temperature in units of Kelvin. This means that the probability to observe an outcome of a certain energy increases with increasing

temperature. Or, in other words, the Boltzmann distribution function flattens out when the temperature is increased so that it becomes more likely to get an outcome of higher energy. It is an interplay between energy minimization and entropy maximization. All the fundamental forces in nature struggle to relax everything into its lowest energy level (the ground state). But at the same time their worst enemy, the second law of thermodynamics, is working against them. The spontaneous increase in entropy is pushing the system, by the means of thermal noise, towards higher energies. An oxygen molecule is pulled down towards the earth's surface by gravity, but the oxygen molecule is also moving in random directions, and is constantly colliding with other molecules, which keeps it from falling all the way to the ground. The source for the power of the entropy is the temperature, and the higher the temperature, the stronger it pushes. The quantity $\exp(-E(i)/k_BT)$ is usually referred to as the Boltzmann factor. Even though this quantity is not strictly a probability (since it is not normalized) it gives the relative probability for a certain outcome. In fact, using the Boltzmann factor one can get the ratio of probabilities for two outcomes as

$$\frac{p(i)}{p(j)} = \frac{\exp(-E(i)/k_BT)}{\exp(-E(j)/k_BT)} = \exp(-\Delta E_{ij}/k_BT), \tag{2.15}$$

where $\Delta E_{ij} = E(i) - E(j)$. That is, if $\Delta E_{ij} = k_BT$, then the probability to observe an outcome of energy $E(j)$ is around 2.7 times higher than to observe an outcome of energy $E(i)$.

This simple expression (Eq. 2.15) turns out to be a very powerful tool when simulating stochastic processes. In 1953, Nicholas Metropolis et al. suggested a Markov chain Monte Carlo algorithm for generating random samples from a probability distribution that is difficult to sample from directly [63][40]. This algorithm has been frequently used in numerical statistical physics together with Eq. 2.15. The algorithm works in the following way:

Start with a system of N elements (e.g. molecules, coins, spins etc.). Define an energy function, E, which depends on the outcome of all the constituents. Make a random swap of the outcome of one of the constituents[5] (e.g. a head is exchanged for a tail) and calculate the energy of the new microstate. This change should then be accepted with a probability equal to Eq. 2.15. By drawing a random number from a uniform distribution between zero and one, $U(0,1)$, a decision can be made to accept a swap if

$$U(0,1) < \exp(-\Delta E/\tilde{T}), \tag{2.16}$$

where $\tilde{T} = k_BT$ and $\Delta E = E_{new} - E_{old}$. If the above condition is not fulfilled then the old microstate is recovered. When repeated over and over again, this scheme pulls the system towards lower energies since every swap giving a microstate with lower energy is accepted (ΔE is negative). But, at the same time, the entropy increase is pushing in the other direction since some random swaps, giving higher

[5] satisfying the condition that every microstate has the same probability to occur

energies, are also accepted. The balance is again determined by the temperature $\tilde{T}$. At infinite temperature, all swaps are accepted and the entropy dominates, while at zero temperature no swaps giving higher energies are carried out. The latter case results in a system at its ground state[6]. Fixing the temperature and letting the system reach its equilibrium then enable us to measure the system variables as a function of the temperature.

The beauty of this method is that it can be used for any system, even systems where there is no real temperature or energy present. We can simply define a temperature and an energy function relevant to the system and the problem in hand. Of course, the actual value of the energy and the temperature we are measuring at has no real physical meaning, but we can reach a ground state in an unbiased random fashion and we can study both the actual ground state and the approach to this ground state.

2.2 Master equation and detailed balance

In a real, physical, system the different microstates are realized by a process where the constituents move, or jump, from one outcome to another. Particles in a box are, for example, moving around, changing their direction and speed due to collisions. It could also be coins that are constantly being flipped. Making an observation of the outcomes at a certain time catches the system in a certain microstate.

A master equation gives a general description of the time evolution of the probability, $p(i,t)$, for a constituent (e.g. particle, coin) to have outcome i at time t, given the nature of the process. The process it self is represented by transition rates describing the chance for a constituent to jump to another outcome. The master equation for a discrete set of outcomes can be written as

$$\frac{\partial p(i,t)}{\partial t} = \sum_j \big[T_{ij}p(j,t) - T_{ji}p(i,t)\big], \tag{2.17}$$

where T_{ij} is the transition rate for jumping from outcome j to outcome i (Fig. 2.3). In other words, the change of the probability to have outcome i equals the sum of the out- and in flows to- and from other outcomes. When the sum in Eq. 2.17 equals zero the system is in steady state. A more strict equilibrium situation is when the terms in the sum equals zero separately. This is called *detailed balance* and is defined as

$$T_{ji}p(i) = T_{ij}p(j), \tag{2.18}$$

for all i and j.

If all the transition rates are the same, then detailed balance is obtained when $p(i) =$ constant for all i and we obtain the unconstrained maximum entropy solution.

[6]Usually a simulated annealing routine, where the temperature is decreased slowly, has to be

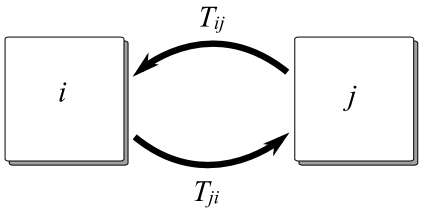


Figure 2.3: *Two outcomes i and j with the transition rates T_{ij} and T_{ji} of jumping from one to the other.*

2.3 Entropy of networks

In section 2.1 the entropy was presented as a measure of the number of possible microstates of a system, like gas molecules or even coins. This concept can also be applied to networks, but it is a little bit more complicated than a sequence of heads and tails when defining an outcome, and thus a microstate. We follow the ideas of E.T. Jaynes (1957) [46] by making the connection between states and different distinguishable ways of distributing objects, like links between nodes. But what is distinguishable and what is indistinguishable for a network? What are the simplest combinatorial entities to which we can assign equal probability? The implicit assumption is then that if the resulting macrostate matches that of a real system, then these entities effectively gives a good representation of the true behavior of the system. These questions will be addressed in this section by mapping a network onto a set of balls and boxes.

2.3.1 Definition of a microstate

A network can be mapped onto a set of balls and boxes where each node corresponds to a box and each link end corresponds to a ball (Fig. 2.4). If the link ends are numbered in a way that preserves the information of which boxes are interconnected, then this mapping is exact. In Fig. 2.4 this is done by using a system where ball 1 is connected to ball 2, 3 is connected to 4, and so on.

Nodes are traditionally the important entities in network science. They are in some sense real, unique individuals and thus distinguishable (people, webpages, airports etc.) while the links are just means to represent their connections. The approach presented here takes on a view point where a node is defined by its links. That is, who you are is defined by what you do, or whom you associate with. Changing a link changes who you are. For example, a protein is defined by its sequence of amino acids. But this sequence will also determine what kind of interactions it will have with other molecules in the cell. So, in some sense a protein is defined by its

implemented in order to avoid getting stuck in local minimas.

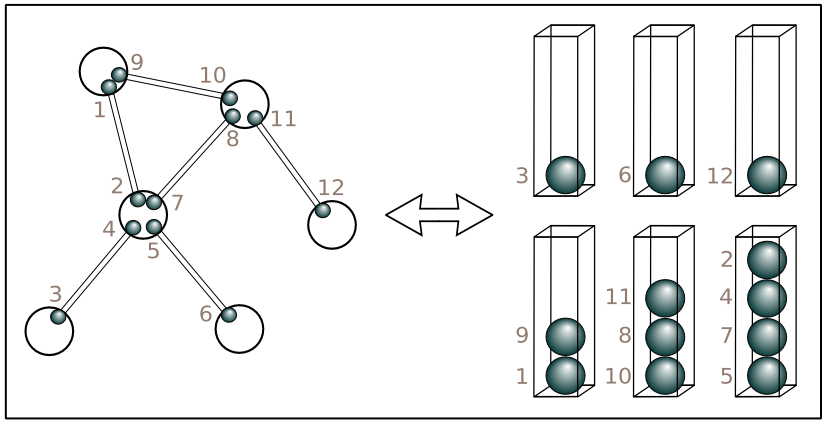


Figure 2.4: *Mapping of a network onto a set of balls and boxes. Each box on the right corresponds to a node to the left. And, each ball to the right corresponds to a link end to the left. For example, the top left box with one ball maps to the bottom left node with one link.*

interactions, its links.

The balls-and-boxes model is effective and relatively simple to handle and can be used for many different types of problems [14, 21, 15]. However, it is fairly difficult in this case to deal with network constraints. These constraints can include disallowing nodes to link to itself, disallowing multiple links between nodes, and keeping a network connected. It would then be necessary to always keep track of which balls are in which box and if they are connected in an "illegal" manner. These constraints will however be neglected here and dealt with in the next section. Furthermore, the fraction of loops and double links in a network has been shown to vanish in the infinite size limit [27].

Equal outcomes were treated as indistinguishable in the previous sections. In this case, however, we can define different types of microstates where different parts of an outcome are indistinguishable. The system now consists of N boxes being the system components making up the microstates, and M balls representing their outcome. The outcome of a box is which balls, and the number of balls (size), k, it has, where the balls and their internal order can be either distinguishable or indistinguishable in different ways. And, a macrostate is the distribution of sizes, $N(k)$. So, distributing the M balls over the N boxes then generates a microstate with a corresponding macrostate. The question is: Which size distribution has the most number of microstates? Below, three different definitions of a microstate, and the resulting maximum entropy solutions, are presented.

Distinguishable balls without internal order

We will start with the case of distinguishable balls without internal order which means that we know which balls are in which boxes, but we do not know (or care about) their order. So, all the permutation's of the balls between boxes gives new microstates but reshuffling the balls inside a box makes no difference (in the same way as exchanging two heads in previous example of coins made no difference). The total number of configurations that can be created using M balls is $M!$. But, as before, exchanging the position of two boxes of the same outcome (size) gives the same microstate so we must divide the number of configurations by the term $N(k)!$ for each size. We also have a $k!$ degeneracy of microstates for each box due to the lack of an internal order. The total number of microstates is then given by the formula

$$\begin{aligned} \Omega &= \frac{M!}{\prod_k N(k)!k!^{N(k)}} && (2.19) \\ \Rightarrow \ln\Omega &\approx M\ln M - M - \sum N(k)[\ln N(k) - 1] - \sum N(k)\ln k! && (2.20) \end{aligned}$$

Since the number of boxes and balls are fixed we have the two constraints

$$\sum_k N(k) = N \quad \text{and} \quad \sum_k kN(k) = M. \tag{2.21}$$

We can again maximize the entropy $\ln\Omega$ under these constraints by using variational calculus and introducing the two Lagrange multipliers a and b, respectively. The solution is given by

$$\begin{aligned} &-\ln N(k) + 1 - 1 - a - bk = 0 \\ \Rightarrow \quad & N(k) = A\frac{\exp(-bk)}{k!}, \end{aligned} \tag{2.22}$$

where $A = \exp(-a)$ is the normalization constant. The expression given by Eq. 2.22 can also be written as

$$N(k) = A\frac{B^k}{k!}, \tag{2.23}$$

where $B = \exp(-b)$ is a constant. This solution can be recognized as the Poisson distribution which is also, as we saw in chapter 1, the degree distribution for the ER network with $B = M/N = \langle k\rangle$.

The meaning of the enumeration in this case can be thought of as the time ordering of when the ball entered the system. This is equivalent to linking up nodes one by one and then asking how many different networks one can make, given a

certain distribution of degrees, $N(k)$. So, we start with a given set of boxes and then put ball 1 in an arbitrary box, and then ball 2, ball 3, and so on. In this way there is only one way in which the same balls can be distributed inside one box, and that is with the lowest number first and the highest last. This means that the Poisson distribution gives the most number of possible different networks.

Distinguishable balls with internal order

Next we move to the case of distinguishable balls with internal order. In this case we know exactly which balls are in which box and the order in which they arrived to the box. For example, in Fig. 2.4, the balls in the largest box has the order: first 5, then 7, next 4 and last 2. This means that every permutation of the balls, in a box as well as between boxes, give different microstates. That is, we have no degeneracy except the $N(k)!$ for each size, and the total number of microstates is

$$\Omega = \frac{M!}{\prod_k N(k)!} \tag{2.24}$$

$$\tag{2.25}$$

The maximum entropy solution is found in the same way as before and is given by

$$N(k) = A\exp(-bk), \tag{2.26}$$

where $A = \exp(-a)$. Thus, this definition of a microstate gives the Boltzmann distribution as the maximum entropy solution.

The meaning of the enumeration of the balls in this case is quite different from the previous one. Here, the number of a ball signifies to which box it connects to and the fact that each link on a node connects to another node means that they are distinguishable. So, in this case we ask how many different networks we can make, including all the different time orders in which it can happen. A more local question is how many different other nodes a node, with a certain degree, can connect to and in how many time orders those connections can be made. This is equivalent to enumerating the balls by which box it connects to and then distribute them over the boxes in all possible ways (which includes the possibility to connecting to oneself or another node several times). This means that the distinguishability of the balls can be confined to within a box, and that enumerating them from 1 to M is an analogous way of simplifying the problem, giving the same result. We still have $M!$ ways of distributing the balls.

There is actually one more way of getting the same result. The balls in this case are completely indistinguishable, inside as well as between boxes. The only thing we can distinguish between are the boxes and their sizes. So, now we have $N!$ ways of distributing box sizes among the N distinguishable boxes. For example, box number 1 got k balls, box 2 got k' balls, and so on. However, it does not matter if box 1 got

k' balls and box 2 got k balls instead if $k = k'$, so we again have a $N(k)!$ degeneracy per box size. The number of states is then

$$\Omega = \frac{N!}{\prod_k N(k)!} \tag{2.27}$$

with the maximum entropy solution $P(k) = A\exp(-bk)$.

Distinguishable balls with cyclic internal order

The third case is again treating the balls as distinguishable, but with a cyclic internal order. That is, each cyclic permutation of the balls in a box gives the same microstate leading to a k degeneracy of microstates for each box. The expression for the number of microstates is for this case given by

$$\Omega = \frac{M!}{\prod_k N(k)!k^{N(k)}} \tag{2.28}$$

with the maximum entropy solution

$$N(k) = A\frac{\exp(-bk)}{k}. \tag{2.29}$$

Here the enumeration has the same meaning as in the previous case but including a different kind of time ordering. Here we ask the question of how many different networks we can make, including the different ways it can be done in the sense of a relative order between the links. A cyclic degeneracy is a more abstract concept in this context, but it roughly means that a ball only cares about if a certain other ball entered the box right before it, and not the absolute time it got there. So for example, all the cases where ball A is the first one or where ball B is right before A, are the same.

General case

One can of course think of many more variations by simply defining a degeneracy function, $f(k)$, so that

$$\Omega = \frac{M!}{\prod_k N(k)!f(k)^{N(k)}} \tag{2.30}$$

and

$$N(k) = A\frac{\exp(-bk)}{f(k)}. \tag{2.31}$$

The three cases described above, however, have a combinatorial meaning.

It should be mentioned that other types of entropy measures have been developed like for example applying Eq. 2.6 to the betweeness distribution (number of paths through a certain node, described in section 1.3.3) [83].

2.3.2 Variational calculus using a random process

As mentioned before, it is difficult to handle network constraints in analytic calculations. This is mainly due to the difficulties one faces when trying to formulate expressions governing the structure of the network. That is, who is connected to whom, and who do a nodes neighbors connect to, etc. One way of dodging this problem is to construct a random process which is sampling the microstate in an unbiased way, while dealing with the constraints by simple rejection and acceptation. The unbiased way of doing this would be to jump between microstates with uniform probability. That is, move the balls around in such a way that every microstates has the same probability to occur. However, we are only concerned about the macrostate of the system and what we can observe is the box sizes. Thus, since many microstates correspond to the same macrostate, an efficient way of solving the problem is to map the jumping between microstates to directly making changes in the box sizes. Moving a ball from one box to another changes their size and thence the macrostate of the system. So, to be fair, this move should be weighted proportional to the amount of microstates, Q, that could be reached by making this exchange. The implementation of the weights could be done in the same way as for the metropolis algorithm where a uniform random number is drawn and checked against the normalized weights. A more efficient way would be to instead pick the boxes with the correct probability directly, if possible. This would remove the need to reject many moves, which saves computational time. A way to do this could be to create an array of box labels where the number of times each label appears is proportional to the weight of that box. Boxes are then chosen by picking elements of the array with uniform probability. Once the weights are found and the algorithm for picking boxes is constructed, any constraint can be implemented by a simple check and rejection scheme.

The algorithm for finding the maximum-entropy solution through a random process goes like this: Start with M balls distributed over N boxes in an arbitrary way. Pick two boxes, A and B, with a probability proportional to their weights, q_A and q_B, and move a ball from box A to box B. Performing these swaps many times leads to a steady state with the correct maximum-entropy solution. The actual weights, and how to find them, for the previously described definitions of a microstate are presented below.

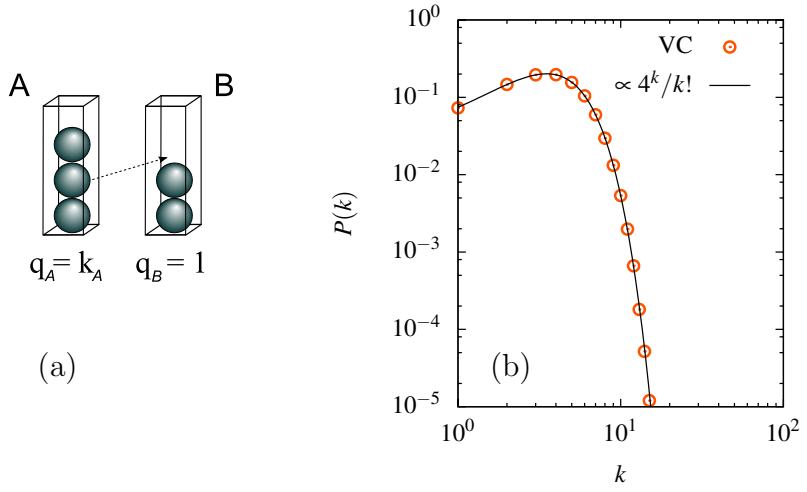


Figure 2.5: *Distinguishable balls without internal order: (a) Process for unbiased sampling of states. (b) The result of repeating the algorithm of choosing a box with a probability $p \propto q_A$ and moving one ball to another box chosen with a probability $p \propto q_B$, for $M/N = 4$.*

Distinguishable balls without internal order

The number of microstates that can be reached by moving a ball from one box to another is illustrated in Fig. 2.5a. Since the balls are distinguishable there are $q_A = k_A$ possibilities when picking a ball from box A. But, we only have $q_B = 1$ place to move the ball to, because the internal order does not matter and all we know is that the ball has been moved to box B. Thus, when constructing the algorithm for sampling the microstates, we can pick boxes uniformly and reject or accept them according to the total weight $Q = q_A q_B = k_A$, much like the metropolis algorithm. A more computationally efficient way is to pick the first box proportional to $q_A = k_A$ (i.e. its size) and the second box proportional to $q_B = 1$, without the need of rejection, as described above.

If repeated enough times, the algorithm will equilibrate at the maximum entropy solution for the given definition of microstate. as shown in Fig. 2.5b. In network science this algorithm is usually thought of as the most basic randomization scheme creating an Erdős-Rényi network (see chapter 1) from any starting network. The arguments presented here explains why this is so.

Distinguishable balls with internal order

As illustrated in Fig. 2.6a, we here have $q_A = k_A$ balls to choose from in box A since the balls are distinguishable. When placing the same ball in box B there are $q_B = k_A + 1$ possible locations to choose from, due to the presence of an internal

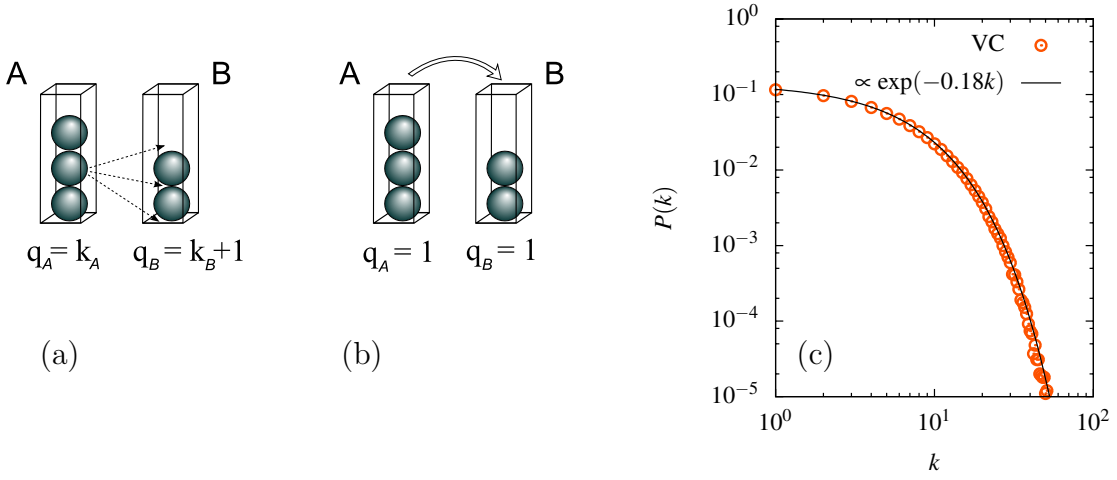


Figure 2.6: *Process for unbiased sampling of states for (a) distinguishable balls with internal order and (b) indistinguishable balls. (c) The result of repeating the algorithm of choosing a box with a probability $p \propto q_A$ and moving one ball to another box chosen with a probability $p \propto q_B$, for $M/N = 4$.*

order. So, the total weight for moving a ball from one box to another is in this case $Q = q_A q_B = k_A(k_B + 1)$. The more efficient implementation is to simply pick box A proportional to its size and box B proportional to its size plus one.

For the other case, giving the same solution, we instead have only one choice per box since it does not matter which ball we pick in the box. Also, there is only one place to put in the second box since there can not exist an internal order for indistinguishable balls (Fig. 2.6b). Thus, the algorithm in this case is to choose two boxes with uniform probability and move one ball from one to the other.

Both algorithms gives the result presented in Fig. 2.6c.

Distinguishable balls with cyclic internal order

For the case of a cyclic internal order we have, as shown in Fig. 2.7, $q_A = k_A$ balls to choose from in box A since the balls again are distinguishable. And, we have $q_B = k_B$ locations to place the ball in box B because the bottom and top position gives the same microstate due to the cyclic degeneracy. The total weight is then $Q = q_A q_B = k_A k_B$, or the boxes should be picked with a probability proportional to their sizes. This case amounts to the same algorithm as for the preferential urn model presented in Ref. [71], though on a different motivative ground.

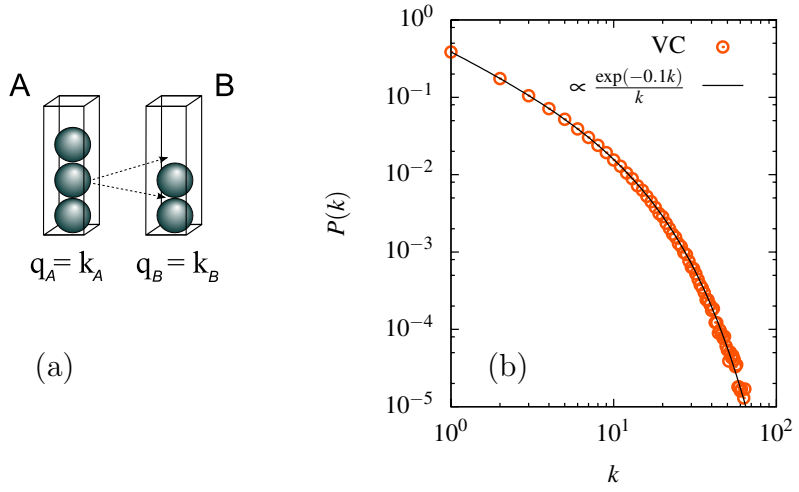


Figure 2.7: *Distinguishable balls with cyclic internal order: (a) Process for unbiased sampling of states. (b) The result of repeating the algorithm of choosing a box with a probability $p \propto q_A$ and moving one ball to another box chosen with a probability $p \propto q_B$, for $M/N = 4$.*

General case

In general, any system undergoing a stochastic process in equilibrium has in fact maximized the entropy for some definition of a microstate. In this case the microstates are defined by the process itself. Sometimes this definition can be very hard to describe combinatorially. For example, an ordinary dice has the outcomes 1 to 6, and if it is rolled in an unbiased way they all have the same chance of appearing. But, if the player decides to re-role every time 6 comes up, the process has changed and is no longer unbiased with respect to the outcomes 1 to 6. It is now unbiased with respect to 1 to 5 instead.

2.4 Summary of papers

2.4.1 Paper IV

In paper IV, *Scale-freeness for networks as a degenerate groundstate: A Hamiltonian formulation*, we give a statistical mechanical formulation of networks, addressed through the analogous non-network model of balls and boxes. An equivalent description would be that a node corresponds to a town and its degree to the number of inhabitants. We start from a basic rewiring scheme where a random person meets another random person and moves to his/her town with a certain probability. By

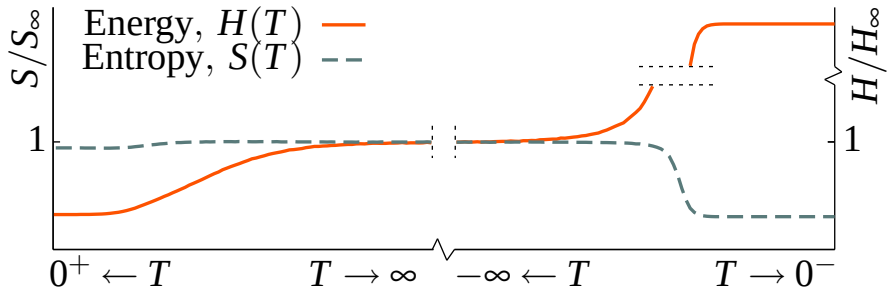


Figure 2.8: *The energy and the entropy as a function of temperature, divided by their respective values at infinity.*

writing down the master equation for this process, and finding the solution to the detailed balance conditions, we derive a Hamiltonian for the system. The energy, the entropy (given by Eq. 2.6), as well as the degree distribution and its fluctuations are investigated at various temperatures. Figure 2.8 shows the energy and the entropy as a function of temperature, ranging from zero (from the positive temperature side) to the positive infinite temperature and from minus infinity to zero (from the negative side). The groundstate, at zero(+) temperature, of the Hamiltonian is the scale-free power-law distribution, $P(k) = Ak^{-\gamma_0}$ which turns out to have a small energy but a reasonably large entropy. That is, the groundstate is highly degenerate with many possible microstates. At infinite temperature the solution is $P(k) = A\exp(-bk)/k$ and for temperatures in the range $0^+ < T < +\infty$ the solution is $P(k) = A\exp(-bk)/k^\gamma$ where $1 < \gamma < \gamma_0$.

The result for a negative infinite temperature is the same as for the positive, but when increasing the temperature towards 0^- the result is quite different. The negative groundstate is when all boxes has the same number of balls (or at least only two different sizes if the ratio between the number of balls and boxes is not an integer). This solution has a very high energy and a low entropy. Also, for an intermediate temperature $-\infty < T < 0^-$ a distribution emerges which is very similar to the Poisson.

2.4.2 Paper V

In paper V, entitled *Optimization and scale-freeness for complex networks*, we generalize the concept of entropy for networks. By mapping a network onto a set of boxes (nodes) and distinguishable balls (link-ends) we consider different definitions of a microstate and their corresponding maximum entropy solutions. We argue that the maximum entropy solution $P(k) = A\exp(-bk)/k$ is a random degree distribution with respect to a set of network states. Furthermore, the question of what kind of bias is needed to impose on the system in order to change the distribution to a gen-

eral power law is addressed, and a type of box information measure is suggested. It is shown that the maximum entropy solution then becomes $P(k) = A\exp(-bk)/k^c$ and that the power-law distribution, with $b = 0, c = c_0$, maximizes also the box information.

So far we have only been dealing with the average distribution $P(k)$. However, since the maximum entropy principle depends on an underlying stochastic process, the distribution must fluctuate. These fluctuations are studied by calculating the distribution of system state probabilities, p_i (the probability to find the system in state i with the distribution $n_i(k)$ close to $P(k)$), a Hamiltonian, H, can be derived so that $p_i \propto \exp(-H/T)$. This Hamiltonian turns out to be the same as the one derived in paper IV. Again, the pure power-law distribution is the groundstate which corresponds to small fluctuations (the distribution p_i is dominated by a single state i).

Finally, the consequences for directed networks are addressed by randomly giving undirected links a direction. By using the measures introduced in paper I it is shown that the maximum entropy arguments of a random network are consistent with model A and metabolic networks.

2.4.3 Paper VI

In paper VI, *The blind watchmaker network: Scale-freeness and evolution* the discussion of possible maximum entropy solutions for networks is taken one step further. We show that there exist another, process driven, definition of a microstate giving the distribution $P(k) = A\exp(-bk)/k^2$ as the maximum entropy solution. Moreover, when the system constraints are implemented the resulting distribution overlaps very well with those of metabolic networks (see Fig. 2.9). This means that, in contrary to the suggestions from paper IV and V, this power-law distribution can arise without any bias, and thus is obtained at infinite "temperature" with maximum fluctuations. These results implies that natural selection has exerted no or very little pressure on the network degree distribution and that its broad behavior is simply a side effect of the stochastic element in Darwinian evolution.

Here we consider the previously described case of distinguishable balls with a cyclic degeneracy but with the additional constraint that the only way to pick a box is to pick a ball. That random mutations target the links makes sense for metabolic networks since the enzymes that catalyze the reactions are encoded in the DNA. A mutation could then change the enzyme to catalyze a different reaction and thus effectively move around the links in the system. As a consequence an extra weight of k is assigned to each box giving the weights $q_A = k_A^2$ and $q_B = k_B^2$ for picking the boxes A and B.

When mapping back to a network one more step must be added to the algorithm. Before moving a link-end from node A to node B a check has to be made so that no constraints (e.g. multiple links and loops) are violated. If the move is rejected

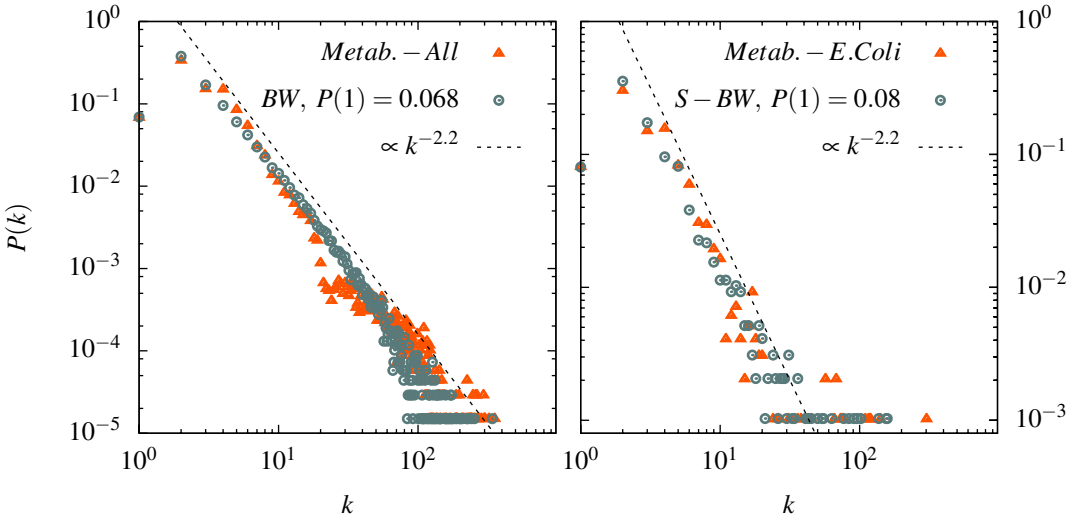


Figure 2.9: *Comparison between the Blind watchmaker (BW) model with a fixed $P(1)$, and metabolic networks: (a) Average over 107 metabolic- and BW networks. (b) A snap shot of the BW networks and the metabolic network of E. Coli.*

another link-end on the same node is chosen and tested against the constraints. If no links on node A is allowed to be moved two new nodes are selected according to the corresponding weights. The resulting steady-state network is entitled *the blind watchmaker network*.

An additional constraint needs to be implemented when comparing to metabolic networks. This constraint restricts the number of one-degree nodes to about $5-8\%$ of the nodes in the network. This means that there are very few substances that are involved in only one reaction and that this reaction has only one substrate or only one product, which could be a chemical constraint.

2.4.4 Paper VII

In Paper VII, *Selective pressure on the metabolic network structure as measured from the random blind watchmaker network*, we investigate the network structure phase space of the Blind watchmaker (BW) network, represented by a clustering-assortativity space [44] (see Fig. 2.10). We compare the behavior of the BW null model to the same real data as used in paper VI, metabolic networks, and show that their network structure, quantified by these two measures, fits inside the BW phase space. However, the real data display a non-randomness in their structure by deviating from the random expectation value of the null model. This finding implies a selective pressure on metabolic networks towards lower assortativity and higher clustering coefficient. It is also shown that when selecting BW networks with low assortativity (of the same magnitude as for the metabolic networks) the

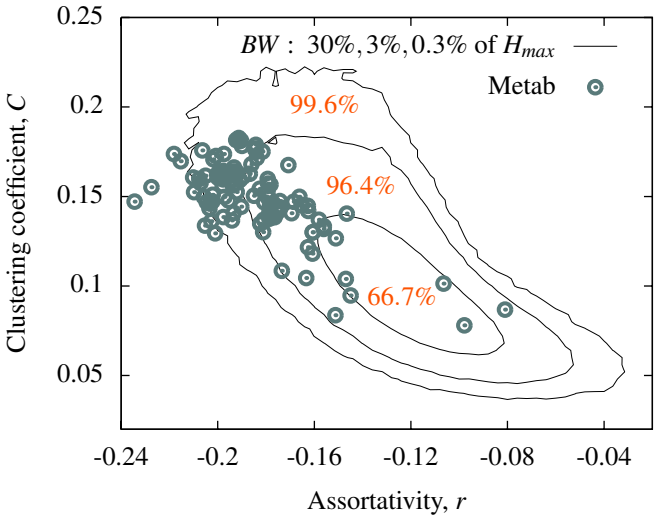


Figure 2.10: *The clustering-assortativity space for the BW network (contour lines) and the 107 metabolic networks (circles). The contour lines encompasses* 67.2%*,* 96.4% *and* 99.6% *of the generated BW networks.*

degree distribution is affected only slightly. However, this small change seem to increase the similarity between the distributions. When randomizing the metabolic networks, keeping their degree distributions fixed, their structural properties are quite unaffected. This suggests that much of these structures are encoded in the degree distribution itself. When using the BW random process as null model the degree distribution is allowed to fluctuate, so when selecting for a particular value of the assortativity it will take the easiest route towards this goal. And this involves changing the degree distribution. The BW and the metabolic networks seem to behave in a similar way compared to their respective randomized counterparts with a fixed degree sequence.

A feature of the metabolic network ensemble is that they have quite different sizes (number of nodes and links), which is not taken into account in Fig. 2.10. When studying the size dependency on the C-r measurements it is found that the two systems behave alike when fixing N and varying $\langle k \rangle$. On the other hand, when fixing $\langle k \rangle$ and varying N they behave differently. That is, when increasing N the assortativity for metabolic networks stays roughly constant while it increases for the BW networks. This result suggests that the selection towards lower assortativity is stronger for larger networks.

Chapter 3

Linguaphysics

The development of language is one of the major transitions in evolution [81] where we acquired the ability to communicate and transfer information between individuals and even between generations. To reach the level of intelligence needed for self-awareness, to understand action and consequences, and later obtaining the gift of speech made it possible for us to warn others of danger, communicate about our needs and discuss plans for the future. With the development of the written language information could be spread to more individuals and saved for future generations. This is also a big advantage in the evolutionary sense. Using the analogy by Bergstrom et al, one can imagine evolution as the process of sending information from a parent to a child. The child wakes up in an unknown world with only a letter from her parents explaining how to survive. This is her DNA. Without exploring the world the child now has to write new letters to her forthcoming offspring. She can either simply copy the old letter or make some small changes (mutations), hoping, without knowing, that these changes will make her child more fit for this world. This is a slow process full of mistakes leading to children dying from misleading letters. On the other hand, the children that got better letters have evolved and with a higher chance of survival, get the opportunity to write many more letters. With the invention of language, however, the parents can complement the letter with accurate information about how to survive, after having explored the world themselves. This capability ensures a higher rate of survival in the offspring and could make humans acquire skills that are very difficult to encode in the DNA, such as growing crops or developing medicines. It might thus be "easy" to understand, in the context of evolution, why we developed the skill of language, but it is much harder to explain the reasons for, or even the properties of, its structure.

This chapter is about the large scale structures in texts, about existing models, and about how it all relates to the underlying random entities in written language.

3.1 Physics - A jack of all trades

Physicists are infamous for the habit of bashlessly taking on problems outside the "physics borders". Social and cognitive sciences has certainly not been spared and the quantitative linguist Gabriel Altmann writes in despair [5]

> *Today there are many physicists counting letters, hoping to find physical laws behind them. And once in a decade they discover that letters behave like mesons and create a wonderful theory. It is not valid, but it is wonderful.*

However, the cooperation between the two disciplines has in many cases been fruitful and he admits that

> *...the engagement of physicists in linguistics was almost always associated with progress. It is not so much the mathematical apparatus they bring in, but rather the way of thinking which, due to their education, is quite "natural" to them but fully foreign to linguists who have a very modest mathematical and "non-linguistic" knowledge.*

There are, of course, also dangers associated with applying lines of thinking and established theories from one field, to problems in other fields since the basic entities might just be different. Altmann finishes with the observation

> *In any case, linguists know that "to have a body" does not mean "to be physicist" but physicists mostly do not know that "to speak a language" does not mean "to be linguist".*

The work presented in this thesis on word frequencies, however, is not dealing with traditional questions in linguistics like the structure of language in the sense of grammar nor does it concern semantics. We are not studying morphology, syntax or phonology. We are interested in text as data. We look for universal properties of the data and ways to mathematically describe its behavior.

Our strategy, which might indeed be deemed as a typical physicists approach, is to make collections of books for single authors and in addition to use periodic boundary conditions in order to optimize the statistics.

3.2 Definition of letters, words and texts

In this context a text is a series of words, where each word is a sequence of letters separated by any other symbol, e.g. blanks, punctuation marks etc. A letter is an alphabetic symbol, *a-z* (*a-ö* in Swedish), with no distinction between lower and upper case letters. The apostrophe is also included to accommodate the grammatical rules for showing *possession* and marking *omission* of letters. As a result of these definitions the word *"it's"* is counted as one word (different from both "it" and "is") and *"its"* as another.

The number of different (unique) words in a text will be denoted as N and the total number of words (the length of the text) as M. Consequently, the average usage of a word is M/N. The number of words with occurrence, or frequency, k is denoted $N(k)$ and the normalized frequency distribution is given by $P(k) = N(k)/N$. That is, the fraction of words with frequency k.

3.3 Empirical laws

An empirical law is a mathematical expression describing a property observed in data obtained through experiments or observations. It is thus not derived from some basic principles which explains the observed behavior. Two such laws are Heaps'- and Zipf's law where the latter has had a big impact also outside the field of quantitative linguistics [79][88][52].

3.3.1 Heaps' law

In the nineteen seventies Harold Stanley Heaps proposed an empirical law describing the relation between the number of unique words, N, and the total number of words M in a text [41]. That is, for every word that is written, M is increased by one. But, N is only increased when a new word is written, that has never been used before in the text. The law states that

$$N(M) = \kappa M^{\alpha}, \tag{3.1}$$

where κ and α are positive constants. κ usually lies in the range 10 to 100 for a typical book and $\alpha < 1$ since N cannot grow faster than M by definition. For real texts, however, $N(M = 1) = 1$. That is, the first word is always unique. This means that the proportionality constant κ in Heaps' law should be equal to one, and $N(M) = M^{\alpha}$. It is often the case that even the first couple of words are different giving close to a linear relation between N and M for small M. For example, in the sentence "Once upon a time there was a boy" the first repetition occurs at the seventh word giving $N(6) = N(7) = 6$ and $N(8) = 7$. However, as we continue to write we tend to repeat previous words more and the rate of adding new words

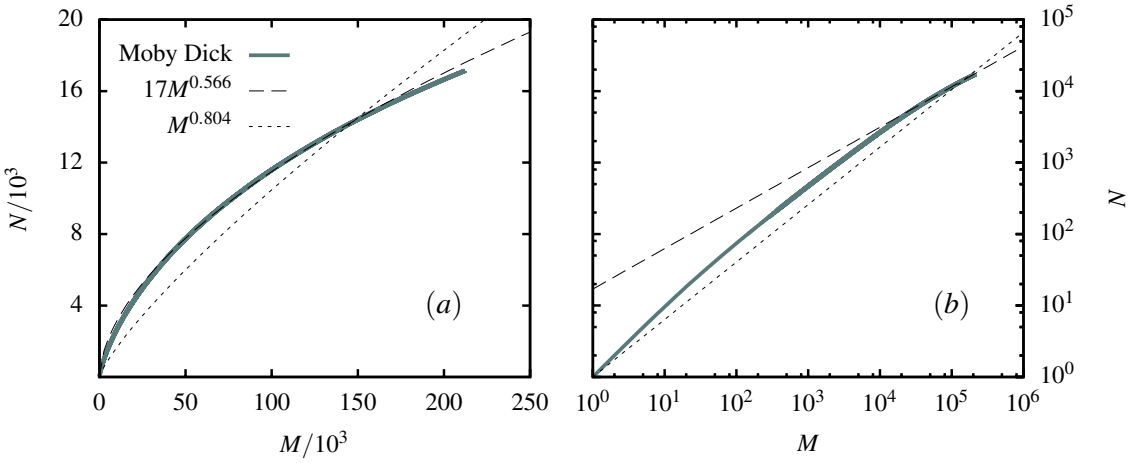


Figure 3.1: *Heaps law: Fits of Heaps law to real data from the novel* Moby Dick *by Herman Melville in linear (a) and logarithmic scale (b). The long dashed curve corresponds to κ as a free parameter and the dotted curve to $\kappa = 1$ in Eq. 3.1.*

decreases. This constitutes a problem when trying to fit the functional form of Heaps' law to real data: We cannot both have a linear relation for small M, and a decreasing rate of adding new words as M gets larger. This is illustrated in Fig. 3.1 which shows a fit of Heaps' law to the $N(M)$-curve of the book *Moby Dick* by Herman Melville, both for a free κ and $\kappa = 1$ in linear (a) and logarithmic scale (b).

3.3.2 Zipf's law

When studying the occurrences of words in the written language, George Kingsley Zipf (1935), made an empirical observation that the most common word appears about twice as many times as the second most common word, and about three times more often than the third most common word [89][90][91]. When sorting, or ranking, all the words in a text according to their frequency, a rank distribution, known as Zipf's law, can be written as

$$k(r) \propto 1/r, \tag{3.2}$$

where $k(r)$ is the frequency of the word with rank r (the most frequent word has $r = 1$). It also means that the *number* of words with frequency k equals the number of discrete rank values giving a frequency in the range $[k, k + 1)$. Thus, by calculating the absolute value of the slope of the inverted rank distribution as $|\Delta r/\Delta k| \approx |dr/dk|$, we can obtain the frequency distribution

$$P(k) \propto 1/k^2. \tag{3.3}$$

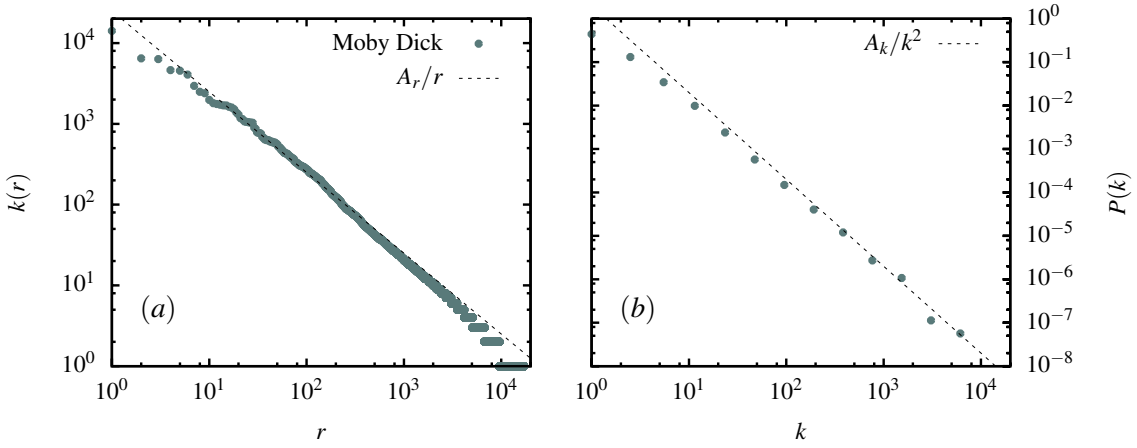


Figure 3.2: *Zipf's law: The rank- (a) and the frequency distribution (b) for the novel* Moby Dick *by Herman Melville (circles) together with the theoretical Zipf's law (dashed lines), where A_r and A_k are constants. The frequency distribution has been binned as described in section 1.3.1.*

Figure 3.2 is illustrating Zipf's law by showing the rank- (a) and the frequency distribution (b) for the novel *Moby Dick*, together with the corresponding theoretical curves for comparison.

Zipf's law is sometimes written in a more general form as $k(r) \propto r^{-\beta}$, where β is usually close to one. Several modifications and extensions to Zipf's law have been presented over the years, both for the actual functional form [57] and its content [65][39]. It has for example been suggested that real data display two scaling regimes with different exponents β [34].

3.4 Models

Such a seemingly universal and interesting empirical law as that of Zipf of course inspires the development of many models with the aim to explain the observed behavior [38][74], although very little attention has been devoted to the empirical observation described by Heaps' law. The two most famous models which reproduce Zipf's law are probably the *stochastic* model by Simon and the *optimization* model by Mandelbrot. These two models will be presented in this section together with an alternative version of Mandelbrots model, called *random typing*.

3.4.1 The Simon model

In the article *On a Class of Skew Distribution Functions* from 1955 [80], Herbert A Simon proposed a stochastic model which generates frequency distributions similar

to those observed in many real systems, e.g. word frequencies, city sizes and income. The model for writing a text is built on two assumptions:

1 The probability to repeat a word that has already appeared exactly k times in the text is proportional to $kN(k)$.

2 There is a constant probability, ε, to write a new word that has not been written before.

These assumptions lead to an algorithm where, at every time step, a new word is written with probability ε or an old word is repeated with probability $1-\varepsilon$. Moreover, the repeated word is chosen uniformly from the words already existing in the text. If this algorithm is run for T time steps the number of total words will be $M = T$. It then follows that the number of different words will be $N(M) = \varepsilon M$ since every $\frac{1}{\varepsilon}$th words is new, and the average frequency is given by $\langle k \rangle = M/N = 1/\varepsilon$.

It is also fairly easy to write down a rate equation for the repetition of a word with a certain frequency k. When calculating the chance of repeating this word the number of favorable choices to make is k, and the possible number of choices is the same as the total number of words, T. The rate at which the frequency of a word is increasing is thus given by

$$\frac{\partial k}{\partial T} = \frac{(1-\varepsilon)k}{T}, \tag{3.4}$$

where the term $(1-\varepsilon)$ is the probability to actually repeat an old word instead of writing a new one. The solution to Eq. 3.4 is

$$k(T) = Ce^{(1-\varepsilon)\ln T}, \tag{3.5}$$

where C is the constant of integration. The time t at which the word was first introduced in the text gives the boundary condition $k(t) = 1$ leading to the final expression

$$k(T,t) = e^{(1-\varepsilon)\ln T/t} = \left(\frac{T}{t}\right)^{1-\varepsilon}. \tag{3.6}$$

$k(T,t)$ is thus the average frequency of a word introduced at time t in a text of length T. We can from this obtain the frequency distribution in a non-rigorous, heuristic way, similar to the derivation for Zipf's law. By inverting Eq. 3.6 we get

$$t(k) = k^{-\rho}\, T, \tag{3.7}$$

where $\rho = \frac{1}{1-\varepsilon}$. Since we are only dealing with discrete values, the number of words with frequency k equals the number of introduction times resulting in a frequency in the range $[k, k+1)$. This number is given by the absolute value of the

slope of Eq. 3.7 as $|\Delta t/\Delta k|$, where $\Delta k = 1$. Approximating the slope as a derivative gives the expression for the frequency distribution

$$P(k) \propto k^{-(\rho+1)}. \tag{3.8}$$

According to Eq. 3.8 the power-law exponent ranges between the extreme points $\gamma = 2$ ($\varepsilon = 0$, only repetition) and infinity ($\varepsilon = 1$, only unique words).

ε is equal to 0.5 for the BA-model described in chapter 1 since every time a link is added to a new node the degree of an old node is also increased, resulting in the exponent $\gamma = 1 + \frac{1}{1-0.5} = 3$.

3.4.2 Optimization

A popular approach to explain Zipf's law has been the notion that this feature is the result of an optimization [6][8]. Benoit Mandelbrot suggested in 1953 that natural languages has minimized the ratio of cost to information content, leading to the behavior observed by Zipf [57]. His view was that the information content, $H(r)$, of a word with rank r is related to its normalized frequency, $f(r)$ $(= k(r)/M)$ through $H(r) = -\log_2 f(r)$. This expression states that rare words (low f) contains more information than common words, although the precise form is not clearly motivated. The average information content per word is consequently

$$H = -\sum_r f(r) \log_2 f(r), \tag{3.9}$$

which can be recognized as the Shannon entropy of the distribution $f(r)$ [77].

Furthermore, the cost of using a particular word is $C(r)$ resulting in an average cost per word given by

$$C = -\sum_r f(r)C(r). \tag{3.10}$$

The task is then to find the rank distribution, $f(r)$, which minimizes the ratio C/H under the constraint $\sum_r f(r) = 1$. Variational calculus (see section 2.1.2) gives the solution

$$f(r) \propto 2^{-HC(r)/C}. \tag{3.11}$$

Mandelbrot further argued that the cost function should be proportional to the length (number of letters) of a word, L. If one assumes that the same is true also for the rank (short words has low rank since low cost gives high frequency) and that the number of unique words of a certain length grows exponentially with L, then

$$r(L) = \sum_{L'=1}^{L-1} K^{L'} \approx K^{L-1} \propto K^L$$
$$\Rightarrow \quad L(r) \propto \log_2 r, \tag{3.12}$$

where K is the number of letters in the alphabet. The expression for the cost function then becomes $C(r) \propto L(r) = C_0 \log_2 r$.[1] This specific form of the cost function also happens to be the only one giving a power-law rank-distribution. Inserting the cost function in Eq. 3.11 gives

$$f(r) \propto r^{-B}, \tag{3.13}$$

where $B = HC_0/C$. When evaluating the values of H and C by inserting Eq. 3.13 back into Eq. 3.10 and 3.9 it turns out that the exponent B must be infinite to get a self-consistent solution. This was of course bad news, but to get around the problem Mandelbrot suggested an extended form of the cost function, namely $C(r) = C_0 \log_2(r + r_0)$. The rank distribution then becomes

$$f(r) \propto (r + r_0)^{-B}, \tag{3.14}$$

which is commonly referred to as the Zipf-Mandelbrot law. Evaluating C and H this time gives the Zipfian exponent $B = 1$ when $r_0 \to \infty$.

The optimization model by Mandelbrot might be phenomenologically appealing but it suffer from several weak points. First of all, even though real data display a plateau like behavior for small r (as described by the parameter r_0), the function given by Eq. 3.14 is hard to fit to any real data since r_0 needs to be very large in order to get the correct exponent. Despite this fact, the Zipf-Mandelbrot law is often used as a parametrization to independently fit B and r_0 to real data.

It has also been shown empirically that the number of different words of the same length do not grow exponentially in natural languages (which was the requirement for getting a power law). In fact, it is a non-monotonic function with a maximum at around 7 or 8 letters[2] [58].

3.4.3 Random typing

Mandelbrot also made the connection between his model and a model of randomly typing letters, since a maximization of the entropy implies a highly disordered,

[1] Manin [58] instead tried to explain this form as the amount of information needed to retrieve a word of rank r from a memory. That is, he suggests that $\log_2 r$ is the number of bits needed to specify the address of the rth object in an array. This is however incorrect since the address of all objects will have the same length $\log_2 R$, where R is the number of objects.

[2] Checked for English and Russian.

random, system. Also, if words are written by randomly (and uniformly) picking among $K+1$ letters (where the extra letter constitutes a blank) then the number of possible different words of length L is $N(L) = K^L$, precisely as required by Mandelbrot's cost function. As shown by Wentian Li in 1992, random texts produced in this manner display a rank distribution similar to that of Zipf's law [51]. When typing random letters, the chance to repeat a word of length L is proportional to K^{-L}. This means that the frequency of words of length L also follows the same expression

$$f(L) \propto K^{-L}. \tag{3.15}$$

As a consequence, short words are more likely to be repeat than long words, and will thus have smaller ranks. The same arguments that led to Eq. 3.12 also holds in this case and consequently $L(r) \propto \log_2 r$. Substituting $L(r)$ into Eq. 3.15 gives the rank distribution

$$f(r) \propto 1/r. \tag{3.16}$$

The scheme presented here creates a step like rank distribution since all words of the same length will have the same frequency. However, Li showed that the curve can be made smoother by implementing different probabilities for different letters. The weak point of the model is again the fact that real words do not follow an exponential relationship between the number of words and the length.

3.5 Summary of papers

3.5.1 Paper VIII

In the paper *Size dependent word frequencies and translational invariance of books* we show that the word frequency distribution (wfd) can to a good approximation be described by the functional form $P(k) = A\exp(-bk)/k^\gamma$, and that the power-law slope seem to change with the size of the text, in contrary to Zipf's law and the predictions made by previously presented models. Moreover, it is found that the size dependence on the wfd for different books is very similar to the outcome of sectioning down a long text (pulling out a section of size M from a text of size M'). This sectioning can be mathematically described by a Random Book Transformation based on binomial coefficients generalized for any partitioning (dividing a book into n pieces) and is given by

$$P_M(k) = C\sum_{k'=k}^{\infty} P_{M'}(k')(n-1)^{k'-k}\frac{1}{n^{k'}}\binom{k'}{k} \tag{3.17}$$

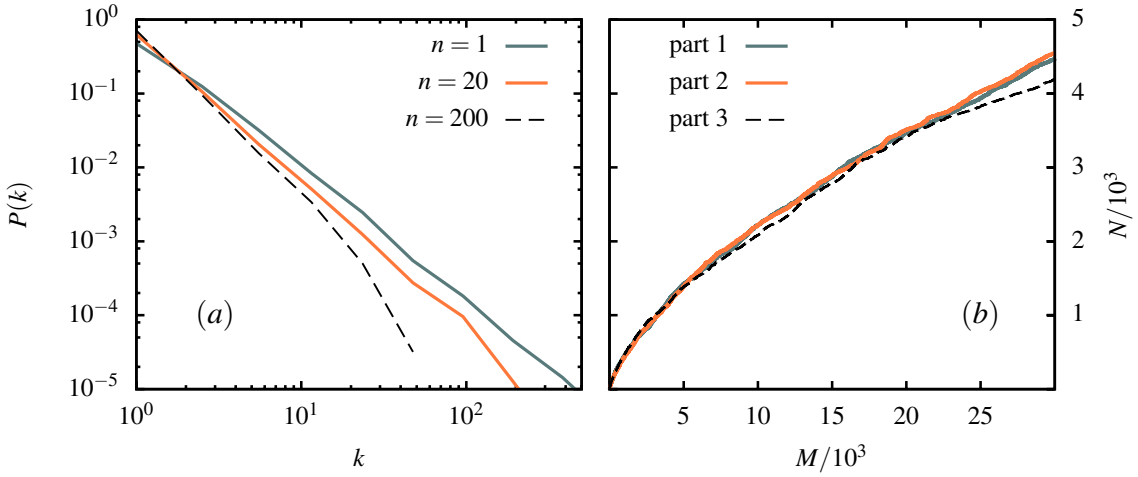


Figure 3.3: *Statistical properties of the novel* Howards End *by E.M. Forster: (a) Size dependency of the wfd shown for different section sizes (M'/n) of the novel. (b) The $N(M)$-curve for three different starting positions in the novel.*

where $n = M'/M$ and C is the normalization constant. The word 'Random' stands for the fact that this transformation only gives the correct result if the words in a book are uniformly distributed.

We also compare statistical properties of real books to the Simon model and a random null model. It is shown that real texts and the null model, where the words have been reshuffled preserving the individual frequencies, show a similar behavior, which is not picked up by the Simon model. One such property is the $N(M)$-curve which for real books deviate very little from the null model. It is also shown that, statistically speaking, there is no such thing as a beginning nor an end of a book. Real books are translational invariant. By this we mean that there are no visible trends for $N(M)$-curves constructed from different starting positions in a book. This is not the case for a text written by the Simon model. The stochasticity of the model will always create books that have more unique words in the end, than in the beginning. In an attempt to quantify how uniformly the words are distributed in a real book, we measure how many words of a certain frequency have been encountered when half of the book has been read. It is found that a big majority of the words in the novel *Howards End* by E.M. Forster are within two standard deviations of what is expected for the null model. On the other hand, real books tell a story and there will always be context related deviations. That is, some rare words which are very important for a small part of the book can give a strong signal of non-randomness.

Furthermore, we present a procedure, based on the RBT, for fitting a function to stochastically generated data.

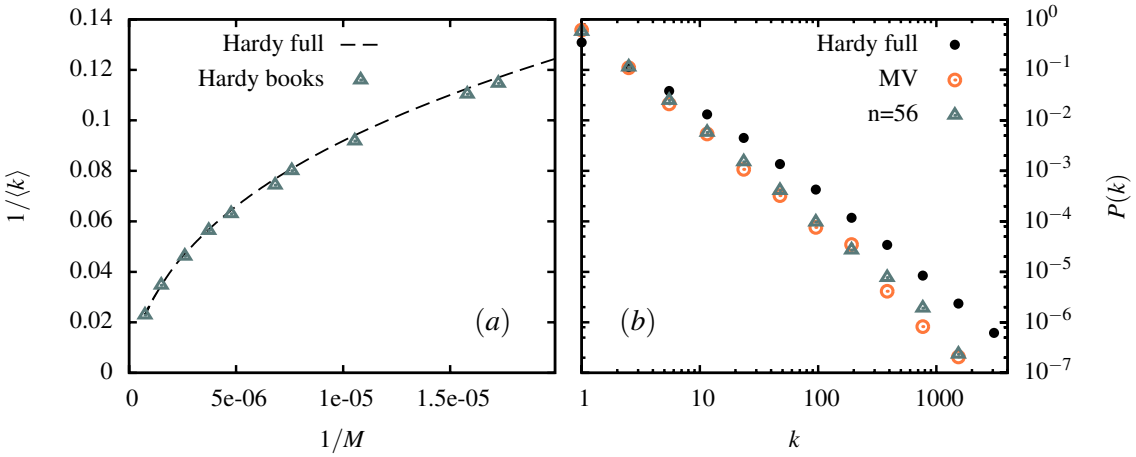


Figure 3.4: *Evidence in favor of the meta book concept from books by Tomas Hardy: (a) The average usage of words for different books (triangles) and for sections pulled out of the full collection of books (dashed line), as a function of the size, M. (b) The wfd for a short story (squares) and for a section of the same size as MV (triangles) pulled out of the full collection of books (circles).*

3.5.2 Paper IX

In the second paper about word frequencies, entitled *The meta book and size dependent properties of written language*, we present the meta book concept as a way to describe the size dependency on some of the statistical properties observed in books. The idea is that the writing of a text can be described by a process where the author pulls a piece of text out of a large mother book (the meta book) and puts it down on paper. This meta book is an imaginary infinite book which gives a representation of the word frequency characteristics of everything that an author could ever think of writing. This has nothing to do with semantics and the actual meaning of what is written, but rather to the extent of the vocabulary, the level and type of education and the personal preferences of an author.

One evidences in favor of the meta book concept is the average usage of a word, $\langle k \rangle$, as a function of the length of the text. As seen from Fig. 3.4a this quantity for a real book (points) is very similar to what we get when pulling small sections out of a much larger book (line).

Another evidence is the word frequency distribution (wfd). As first pointed out in paper VII, and here shown in Fig. 3.4b, the wfd for a short story is also very similar to the result of pulling a small section out of a large book.

An analytic relationship between the wfd and an extended version of Heaps' law ($N(M) = M^{\alpha(M)}$) is derived giving the resulting size dependent wfd

$$P_M(k) = A\frac{e^{-b_0 k/M}}{k^{1+\alpha(M)}}. \tag{3.18}$$

A, in Eq. 3.18, is the normalization constant and b_0 is an author-specific constant. The extended Heaps' law is given by $N(M) = M^{\alpha(M)}$, with a monotonically decreasing α. The real data can be well described by the parametrization $\alpha(M) = 1/(u \ln M + v)$ which has the asymptotic value $\alpha(M \to \infty) = 0$. The resulting wfd for the infinite meta book is thus

$$P_\infty(k) = \frac{A}{k}, \tag{3.19}$$

in contrary to A/k^2 given by Zipf's law. In practice, though, b_0/M and $\alpha(M)$ will never be exactly zero.

It is also shown that the behavior of the $N(M)$-curve (Heaps' law) can be reproduced by the Random Book Transformation which is in fact a mathematical formulation of the meta book concept. It is thus shown that the size dependence in the power-law slope is linked to the behavior of the $N(M)$-curve, which in its turn can be described by the meta book concept.

An interesting note is that the relation between the exponents $\gamma = 1+\alpha$ has also been shown to hold for the frequency distribution of family names. It was shown that the growth in the number of family names in Korea corresponds to $\alpha = 0$ and that the frequency distribution follows the expression $P(k) \sim 1/k$ [50][7].

Chapter 4

Summary and Discussion

This thesis has been about organizational principles of complex systems based on an underlying randomness. Due to the complicated and chaotic nature of the microscopic events that has shaped biological networks we needed to zoom out and simplify. We investigated, following the lines of statistical mechanics, the simple stochastic rules that can be assigned to a system and the different kinds of results that emerge as a consequence of those rules. The processes that come with these rules have also been used to compare the null models to real data in order to learn about the possible constraints these systems could be working under.

We found that the degree seems to be the important characteristics in the organizational principles of *who should be connected to whom* for engineered communication networks. Examples of such networks are the Internet and street networks where signals often needs to be sent from one end of the network to the other. On the other hand, biological systems like protein-protein networks, or transcriptional networks, display a structure which, in the latter case suggest a shift in focus towards what type of biological process the node (protein) are involved in. It also seems reasonable to speculate that a similar type of gradient as the one displayed in Fig. 1.5b also exists in many other types of systems. It could for example represent political views in a social network.

We also mapped a network onto a set of balls and boxes which opens up a whole new regime of combinatorial possibilities. Different definitions of a microstate gave different results as measured by the degree distribution. It was shown that one of these definitions together with a set of process based rules gave a result very similar to that of metabolic networks. This would imply that the constraint of natural selection has had very little influence on the degree distribution of these networks.

The question of whether or not links in a specific network are in fact distinguishable is a difficult nut to crack. However, It seems reasonable that links on a certain node are distinguishable since they connect to different nodes. Another meaning of the labels on the balls could be possible weights on the links. These weights could represent the strength of binding or the rate of a reaction. It also makes sense that

random mutations target the links in metabolic networks since a mutated enzyme might catalyze another reaction and thus changes the links between the substances in the network. Even though it is very hard to map the actual way links are rewired to a definition of a microstates, a good agreement between a maximum entropy solution and real data can, in general, give a hint to the actual underlying rules of the system. For example, Max Planck was forced to introduce discontinuous energy levels with a separation proportional to a constant (Planck's constant) in order for the entropy of a black body to match experimental data. Later, it was shown that this assumption in fact represented the correct behavior of the system.

Also the meaning of a time order on the links is difficult concept to wrap once head around. One analogy could be the assembly of a bookshelf from IKEA. Even though the function of the final product is independent of how it was constructed, some moves must be made before others (side walls must be mounted before the shelf's can) to even make it possible to reach its final state. So, the time order of the assembly do matter and we can distinguish them by the fact that some lead to a dead end in the construction.

In the third chapter we moved to a system which is not very different from that of networks but one which lacks the complexity of the connections, namely texts and books. Also here we set up to determine to what extent words are randomly distributed over a text and how this can be used to extract information about the system. Books tell stories and of course a text is not completely random. It is shown that some context related words gives very strong signals of non-randomness. However, filling words (e.g. "the", "of", "and" etc.), which are the most frequent words in a text, are to a large extent homogeneously distributed over the text. Actually, only the 7 most frequent words in Moby Dick (out of around 17 000) constitute more than 20% of the whole book, making these words a dominate factor in the statistics (95 words can be accounted for about half of the book). Also rare words gives good signals of randomness when grouped together in frequency classes. During this work we found a very interesting size dependency in the word-frequency distribution of real books. This led us to formulate the meta book concept which describes the process of writing as pulling sections out of a big abstract mother book, a meta book. The data also discriminates authors, which implies that the meta book represents our personal way of writing, and not only a statistical result of the structure of the language itself. Nevertheless, the theoretical infinite limit for all authors is the same and this work shifts the problem from trying to recreate Zipf's law to trying to explain the structure of the meta book. Also, the fact that there is a difference between authors means that each meta book is unique and can be seen as a linguistic fingerprint.

The meta book concept is probably also quite general and seems to be applicable to, for example, family name distributions.

Bibliography

[1] H. Jeong A.-L. Barabási, R. Albert and G. Bianconi, *Power-law distribution of the world-wide web*, Science **287** (2000), 2115.

[2] R. Albert and A.-L. Barabási, *Statisitcal mechanics of complex networks*, Rev. Mod. Phys. **74** (2002), 47–97.

[3] R. Albert, H. Jeong, and A.-L. Barabási, *Diameter of the world-wide web*, Nature **401** (1999), 130–131.

[4] R. Albert, H. Jeong, and A.-L Barabási, *Error and attack tolerence of complex networks*, Nature **406** (2000), 378–382.

[5] G. Altmann, *On the symbiosis of physicists and linguists*, Romanian Reports in Physics **60** (2008), no. 3, 417–422.

[6] M.V. Arapov and Y.A. Shrejder, *Zakon cipfa i princip dissimmetrii sistem*, Semiotics and Informatics **10** (1978), 74–95.

[7] S.K. Baek, H.A.T. Kiet, and B.J. Kim, *Family name distributions: Master equation approach*, Phys. Rev. E **76** (2007), 046113.

[8] V.K. Balasubrahmanyan and S. Naranan, *Algorithmic information, complexity and zipf's law*, Glottometrics **4** (2002), 1–26.

[9] A.-L. Barabási, *Linked: How everything is connected to everything else and what it means*, Plume, 2003.

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

[11] A.-L. Barabási and E. Bonabeau, *Scale-free networks*, Scientific American **288** (2003), 60–69.

[12] A.-L. Barabási, H. Jeong, E. Ravasz, Z. N'eda, A. Schuberts, and T. Vicsek, *Evolution of the social network of scientific collaborations*, Physica A **311** (2002), 590–614.

[13] A. Barrat and M. Weigt, *On the properties of small-world network models*, Eur. Phys. J. B **13** (2000), 547–560.

[14] P. Bialas, Z. Burda, and D. Johnston, *Phase diagram of the mean field model of simplicial gravity*, Nucl. Phys. B **542** (1999), 413–424.

[15] G. Bianconi, *Degree distribution of complex networks from statistical mechanics principles*, cond-mat/0606365 (2006).

[16] N.L. Biggs, E.K. Lloyd, and R.J. Wilson, *Graph theory 1736-1936*, Clarendon Press, Oxford, 1976.

[17] B. Bollobás, *The diameter of random graphs*, Trans. Amer. Soc. **267** (1981), 41–52.

[18] S. Bornholdt and H. Ebel, *World wide web scaling exponent from simon's 1955 model*, Phys. Rev. E **64** (2001), 035104.

[19] S. Bornholdt and H.G. Schuster, *Handbook of graphs and networks: From the genome to the internet*, Wiley-VCH, 2003.

[20] S. Bornholdt and K. Sneppen, *Robustness as an evolutionary principle*, Proc. Roy. Soc. Lond. B **267** (2000), 2281–2286.

[21] Z. Burda, J. D. Correia, , and A. Krzywicki, *Statistical ensemble of scale-free random graphs*, Phys. Rev. E **64** (2001), 046118.

[22] A. Cardillo, S. Scellato, V. Latora, and S. Porta, *Structural properties of planar graphs of urban street patterns*, Phys. Rev. E **73** (2006), 066107.

[23] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani, *The role of the airline transportation network in the prediction and predictability of global epidemics*, PNAS **103** (2004), no. 7, 2015–2020.

[24] L.E.C. da Rocha, *Structural evolution of the brazilian airport network*, J. Stat. Mech **P04020** (2009).

[25] D.J. de S. Price, *A general theory of bibliometric and other cumulative advantage processes*, J. AMer. Soc. Inform. Sci. **27** (1976), 292–306.

[26] K.A Dill, S. Bromberg, and D. Stigter, *Molecular driving forces: Statistical thermodynamics in chemistry and biology*, Taylor & Francis, Inc., 2003.

[27] S.N. Dorogovtsev and J.F.F Mendes, *Evolution of networks: From biological nets to the internet and www*, Oxford University Press, 2003.

[28] A. Einstein, *über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen*, Annalen der Physik **17** (1905), 549–560.

[29] A. Broder et al, *Graph structure in the web*, Computer Networks **33** (2000), no. 1-6, 309–320.

[30] M.C Costanzo et al., *Ypd™, pombepd™ and wormpd™: model organism volumes of the bioknowledge™ library, an integrated resource for protein information*, Nucleic. Acids. Res. **29** (2001), 75–79.

[31] M. Faloutsos, P. Faloutsos, and C. Faloutsos, *On power-law relationships of the internet topology*, Computer Communications Review **29** (1999), 251–262.

[32] T.J. Fararo and M. Sunshine, *A study of a biased friendship network*, Syracuse University Press, 1964.

[33] J. Ferkingoff-Borg, M.H. Jensen, P. Olsen, and J. Mathiesen, *Diffusion, fragmentation and merging processes in ice crystals, alpha helices and other systems*, Dynamics of Complex Interconnected Systems: Networks and Bioprocesses, vol. 232, 2006, pp. 61–70.

[34] R. Ferrer and R.V. Sole, *Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited*, Journal of Quantitative Linguistics **8** (2001), 165–174.

[35] K. Fjällborg, *Alla vägar bär till gwyneth*, Aftonbladet (2003).

[36] L.C. Freeman, *Centrality in social networks conceptual clarification*, Social Networks **1** (1979), 215–239.

[37] J. Galaskiewicz and P.V. Marsde, *Interorganizational resource networks: Formal patterns of overlap*, Social Science Research **7** (1978), 89–107.

[38] P. Guiraud, *The semic matrices of meaning*, Social Science Information **7** (1968), no. 2, 131–139.

[39] L.Q. Ha, E.I. Sicilia-garcia, J. Ming, and F.J. Smith, *Extension of zipf's law to words and phrases*, Proceedings of the 19th International Conference on Computational Linguistics (COLING), 2002, pp. 315–320.

[40] W.K. Hastings, *Monte carlo sampling methods using markov chains and their applications*, Biometrika **57** (1970), no. 1, 97–109.

[41] H.S. Heaps, *Information retrieval: Computational and theoretical aspects*, Academic Press, 1978.

[42] P.E. Hodges, W.E. Payne, and J.I. Garrels, *The yeast proteome database (ypd):a curated proteome database for saccharomyces cerevisiae*, Nucleic Acids Res. **1** (1998), 68–72.

[43] P. Holme, *Edge overload breakdown in evolving networks*, Phys. Rev. E **66** (2002), 036119.

[44] P. Holme and J. Zhao, *Exploring the assortativity-clustering space of a network's degree sequence*, Phys. Rev. E **75** (2007), 046111.

[45] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, *A comprehensive two-hybrid analysis to explore the yeast protein interactome*, Proc. Natl. Acad. Sci. USA **98** (200), 4569.

[46] E.T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. **106** (1957), no. 4, 620–630.

[47] T. Jeffrey and S. Milgram, *An experimental study of the small world problem*, Sociometry **32** (1979), no. 4, 425–443.

[48] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabási, *The large-scale organization of metabolic networks*, Nature **407** (2000), 651–654.

[49] H. Karabulut, *The physical meaning of lagrange multipliers*, Eur. J. Phys. **27** (2006), 709–718.

[50] B. J. Kim, A. Trusina, P. Minnhagen, and K. Sneppen, *Self-organized scale-free networks from merging and regeneration*, Eur. Phys. J. B **43** (2005), 669–672.

[51] W. Li, *Random texts exhibit zipf's law-like word-frequency distribution*, IEEE Transactions on Information Theory **38** (1992), no. 6, 1842–1845.

[52] W. Li and Y. Yang, *Zipf's law in importance of genes for cancer classification using microarray data*, J. Theoretical Biology **219** (2002), no. 4, 539–551.

[53] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, and Y. Åberg, *The web of human sexual contacts*, Nature **411** (2001), 907–908.

[54] et al. M. Ashburner, *Gene ontology: tool for the unification of biology. the gene ontology consortium*, Nat. Genet. **25** (2000), 25.

[55] H. Ma and A.-P. Zeng, *The connectivity structure, giant strong component and centrality of metabolic networks*, Bioinformatics **19** (2003), 1423–1430.

[56] ———, *Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms*, Bioinformatics **19** (2003), 270–277.

[57] B. Mandelbrot, *An informational theory of the statistical structure of languages*, Butterworth, Woburn, 1953.

[58] D. Yu. Manin, *Mandelbrot's model for zipf's law: Can mandelbrot's model explain zipf's law for language*, Journal of Quantitative Linguistics **16** (2009), no. 3, 274–285.

[59] S. Maslov and K. Sneppen, *Specificity and stability in topology of protein networks*, Science **296** (2002), 910.

[60] ———, *Computational architecture of the yeast regulatory network*, Phys. Biol. **2** (2005), 94.

[61] S. Maslov, K. Sneppen, and A. Zaliznyak, *Detection of topological patterns in complex networks: correlation profile of the internet*, Physica A **333** (2004), 529.

[62] G. Melin and O. Persson, *Studying research collaboration using co-authorships*, Scientometric **36** (1998), 363–377.

[63] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, Journal of Chemical Physics **21** (1953), no. 6, 1087–1092.

[64] P. Minnhagen, M. Rosvall, K. Sneppen, and A. Trusina, *Self-organization of structures and networks from merging and small-scale fluctuations*, Physica A **340** (2004), 725–732.

[65] M.A. Montemurro, *Beyond the zipf-mandelbrot law in quantitative linguistics*, Physica A **300** (2001), 567–578.

[66] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, *Epidemic outbreaks in complex heterogeneous networks*, EPJ B **26** (2002), no. 4, 521–529.

[67] M.E.J. Newman, *Assortative mixing in networks*, Phys. Rev. Lett. **89** (2002), 208701.

[68] ———, *Mixing patterns in networks*, Phys. Rev. E **67** (2003), 026126.

[69] ———, *The structure and function of complex networks*, SIAM Review **45** (2003), 167–256.

[70] ———, *Modularity and community structure in networks*, PNAS **103** (2006), 8577–8582.

[71] J. Ohkubo, M. Yasuda, and K. Tanaka, *Preferential urn model and nongrowing complex networks*, Phys. Rev. E **72** (2005), 065104.

[72] P. Erdős and A. Rényi, *On random graphs*, Publicationes Mathematicae **6** (1959), 290–297.

[73] R. Pastor-Satorras, A. Vazquez, and A. Vespignani, *Dynamical and correlation properties of the internet*, Phys. Rev. Lett. **87** (2001), 258701.

[74] L. Pietronero, E Tosatti, and A Vespignani, *Explaining the uneven distribution of numbers in nature: the laws of benford and zipf*, Physica A **293** (2001), 297–304.

[75] M. Rosvall, A. Trusina, P. Minnhagen, and K. Sneppen, *Network and cities: An information perspective*, Phys. Rev. Lett. **94** (2005), 028701.

[76] J. Scott, *Social network analysis: A handbook*, 2nd ed., Sage Publications, 2000.

[77] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal **27** (1948), no. 3, 379–423.

[78] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, *Network motifs in the transcriptional regulation network of escherichia coli*, Nature Genetics **31** (2002), 64–68.

[79] Z. K. Silagadze, *Citations and the zipf-mandelbrot law*, Complex Systems **11** (1997), no. 6, 487–499.

[80] H.A. Simon, *On a class of skew distribution functions*, Biometrika **42** (1955), 425–440.

[81] J.M Smith and E. Szathmáry, *The major transitions in evolution*, Oxford University Press, New York, 1995.

[82] K. Sneppen, M. Rosvall, A. Trusina, and P. Minnhagen, *A simple model for self-organization of bipartite networks*, Europhys. Lett. **67** (2004), no. 3, 349–354.

[83] K. Sneppen, A. Trusina, and M. Rosvall, *Hide-and-seek on complex networks*, Europhys. Lett. **69** (2005), no. 5, 853.

[84] R.V. Sole, R. Pastor-Satorras, E. Smith, and T.B. Kepler, *A model of large-scale proteome evolution*, Adv. Complex Syst. **5** (2002), 43.

[85] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen, *Hierarchy measures in complex networks*, Phys. Rev. Lett. **92** (2004), 178702.

[86] D.J. Watts, *Six degrees: The science of a connected age*, W. W. Norton & Company, 2003.

[87] D.J. Watts and S. Strogatz, *Collective dynamics of 'small-world' networks*, Nature **393** (1998), 440–442.

[88] Y. Yonezawa and H. Motohasi, *Zipf-scaling description in the dna sequence*, 1999, 10th Workshop on Genome Informatics, Japan.

[89] G. Zipf, *Selective studies and the principle of relative frequency in language*, Harvard University Press, Cambridge, 1932.

[90] ———, *The psycho-biology of language: An introduction to dynamic philology*, Mifflin Company, Boston, 1935.

[91] ———, *Human behavior and the principle of least effort*, Addison-Wesley, Reading, 1949.