UMEÅ UNIVERSITY
Institute of Technology

MASTER'S THESIS
20 p (30 ECTS credits)
2005-08-31

Network modeling and multi-step degree correlations

Sebastian Bernhardsson

Master's Thesis in Engineering Physics preformed at
the Nordic institute for Theoretical Physics (NORDITA) in Copenhagen, Denmark.

Network modeling
and
multi-step degree correlations

By

Sebastian Bernhardsson

Thesis submitted to the Department of Physics at the University of Umeå in partial fulfillment of the requirement for the degree of Master of Science in Engineering Physics.

Supervisors: Petter Minnhagen and Kim Sneppen.

Examinator: Mattias Marklund.

Abstract

Our lives are effected daily by the features and properties of complex networks that surrounds us everywhere. Internet, the World Wide Web, friendship networks and protein networks in our cells are some of them. Research in the field has shown that many of these real world networks have a very broad degree distribution, close to power-laws. This means that participants with a small number of connections dominate the networks while there are just a few with a very large number of connections. These highly connected agents influences a big part of the network, which makes them interesting and probably very important for the network function. In part I, a self-organizing model of merging and regeneration is presented to create directed networks with power-law degree distributions similar to several real world networks. The model could for example describe the dynamics of companies buying up each other at the same time as new ones are started. In part II the structure of real world networks are investigated in the sense of degree sequences when stepping out in the network, a multi-step degree correlation measurement is presented. This measurement is also normalized with the one step degree correlation profile. The positioning of highly connected nodes relative to each other is one of the structural properties that are analyzed and most of the networks are shown to demonstrate hub separation.

Contents

Chapter 1

Introduction

Have you ever been amazed over how fast everyone you know knows about the extremely embarrassing thing you did just the day before? Or how, almost evertime you're on a vacation abroad and you meet a fellow countrymen you seem to have a mutual friend? The answers lie in the field of complex networks which, like the friendship networks mentioned above effects our lives daily. Internet, the world wide Web and protein networks in our cells are other examples. Studying complex networks originates from graph theory which was born as early as in the 18th century to study problems like, how to take trips which visits certain sites exactly once [5]. The field took a big leap when fast computers with a high computational capacity became available since the computers gave the scientists the opportunity to preform 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 of the networks from completely different parts of our world, like those mentioned above, have a common feature. They consists of a very large number of nodes with a low number of connections and a few number of nodes with a very high number of connections. The distribution of connections among the nodes follows close to a power-law [4]. This makes everything even more interesting. For instance, the World Wide Webb is constructed by millions of people creating sites every day which they link to other sites of their particular interest. Considering the vast divergence in human interests, why isn't this network completely random [3]? How the structure is connected to the function is also a very interesting question. Even though a lot of these networks have similar distributions of connections they are constructed to do different things and thus should have different structures! Or?

This thesis is about both the process of evolving directed networks through a merging and regeneration model and mapping the structure of six real world networks using a multi-step degree correlation measurement.

Chapter 2

Complex networks

2.1 What is a complex network?

The statement that something is a complex network but something else isn't has to be supported by some concepts and definitions. Below follows some clues.

2.1.1 What is a network?

A *network* (sometimes also called a *graph*) typically describes a system where the basic parts (agents) are interconnected, often via some sort of information flow. One can think of cars on a road, gossip between friends, electricity between powerstations or money between companies. A network can have any size, i.e number of interconnected agents, or shape.

2.1.2 What does complex mean?

It is often difficult to separate the meaning of the words *complicated* and *complex*. What does the latter word mean in present context? There are probably almost as many definitions of a complex system as there are scientists working in the field, but here are two examples: *A complex system is one that by design or function or both is difficult to understand and verify* [18] and *A complex system is one in which there are multiple interactions between many different components* [15]. Another common explanation is that for a complex system, the whole is greater than the sum of its parts. This can be understood better by looking at a soccer team. It's impossible to say for sure which team are going to win a match just by looking at the individual players. The way they play together is a crucial factor. The word *complicated* in this context has a somewhat different connotation, an example of a complicated system could instead be a machine with a large number of parts. If one sum up the contribution of each part one would get the action the machine was built to perform. A common example of a complicated system like this is an airplane.

2.2 Terminology

Nodes

The participants of a network (people, powerstations, companies etc.) are called *nodes*. They build up the network by different kinds of communication via links between them. The size of a network is the total number of nodes in it, N.

Links

The connections between different nodes are called *links*. These can be directed or undirected (fig. 2.1) depending on the system in question. For example a friendship network is often undirected (if A is a friend of B, B is most likely a friend of A) while the World Wide Web is a directed network since the links between pages goes one way (if A is linked to B, B doesn't have to be linked to A).

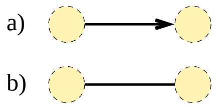

Figure 2.1: a) directed- b) undirected link

Degree

The number of links that are connected to a node will here be called the degree (k) of that node (sometimes also called the connectivity of a node). So k_i is the degree of node i. In the case of a directed network a node has both an in- and an out-degree, denoted $k_{i,in}$ and $k_{i,out}$ respectively. That is, the number of links that goes *from* a certain node and the number of links that goes *to* that node. The total number of links in an undirected system is $k_{tot} = \frac{1}{2}\sum_i^N k_i$ and the average degree is $k_{av} = 2k_{tot}/N$, where the two comes, in both cases, from the fact that each link has two ends. In the directed system the total number of links is $k_{tot} = \sum_i^N k_{i,in} = \sum_i^N k_{i,out}$ and the average degree is $k_{av} = k_{av,out} = k_{av,in} = k_{tot}/N$. Nodes that have a degree much larger than the average degree are called *hubs*.

Degree distribution

One well defined characteristic of a network is its degree distribution which is the distribution of nodes with a certain degree. If one then normalize it with the total number of nodes in the system one get a probability distribution function, $P(k)$. That is, the probability that an arbitrary node has the degree k. Consequently for a directed network one has both an in- and an out-degree distribution describing the system. Another common way of presenting the degree distribution is as a cumulative degree distribution, $P(>k)$. This distribution is giving the probability that an arbitrary node has a degree *larger* than k. The cumulative degree distribution is defined as

$$P(>k) = \int_k^\infty P(\tilde{k})d\tilde{k}. \tag{2.1}$$

Shortest path

A path in a network is a sequence of nodes that one has to go through to get from one node to another one, moving along the links between them. Since the number of paths between two nodes (and the maximum path length) diverges with increasing system size, one often use the *shortest path*. The *shortest path* is the path that has the smallest sequence of nodes between a pair. It can also be measured in steps (or length) as the smallest number of steps between two nodes, which is equal to the number of links one has to pass.

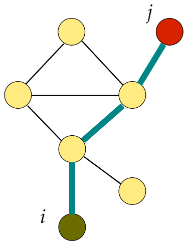

Figure 2.2: Shortest path of length 3 between node i and j.

Diameter of a network

The diameter of a network is a measurement of the size of a network in terms of distances, it is often defined as the longest shortest path or the average shortest path between two nodes in the network. Both definitions have their strengths and weaknesses but gives a hint about what kind of distances the network is dealing with. Is it really dense (small diameter) or is it smeared out (large diameter). In this thesis the diameter will be used only in general terms and comparisons between networks where both definitions can be applied.

For further reading, see reference [6].

2.3 Different types of networks

2.3.1 Structure

The structure of a network depends on how the links are distributed in the system. An increase in the total number of links often reduces the structural properties of the system. In the limit of $k_{tot} = N(N-1)$ all nodes will be linked to all other nodes and the system isn't really a complex system any more. Another structureless but complex system is the *classical random*

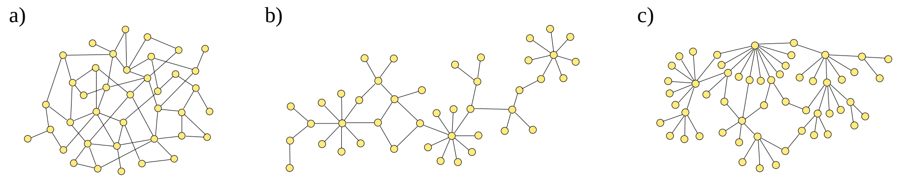

Figure 2.3: a) Network without visible structure. b) Network with a string-like structure. c) Network with a tree-like structure.

graph (also called the Erdős-Rényi model). This network is constructed by having a fixed number of nodes and then adding links one by one completely at random. This gives a homogeneous distribution of links in the system. If one instead grow a network, where at each step a node is added and a link (or several links) is put in between two random nodes, one get a random network where the links are not distributed homogeneously in the system. The oldest nodes will get more links since they have had more chances of getting links. Both these random networks has a structure similar to figure 2.3 a. These models are elaborated more on page 8 in reference [6]. Networks that are found in the real world can have a more or less non-random structure. Two examples of different structures are shown in figure 2.3 b, and c.

2.3.2 Degree distributions

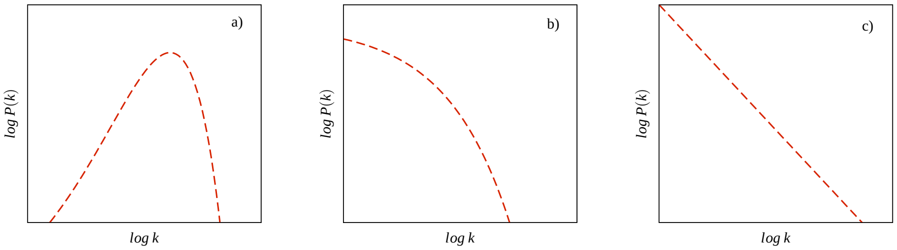

Figure 2.4: a) Poisson distribution with a characteristic scale of the average degree. b) Exponential distribution with a characteristic scale of the average degree. c) Power-law distribution without a characteristic scale.

Poisson distribution

The *classical random graph* has, in the limit of infinite size, a Poisson degree distribution,

$$P(k) = \frac{e^{-\bar{k}}\bar{k}^k}{k!} \tag{2.2}$$

where $\bar{k}$ is the average degree of the system. The characteristic scale of distribution is the average degree.

Exponential distribution

The growing random graph has a exponential degree distribution,

$$P(k) \propto e^{-k/\bar{k}} \tag{2.3}$$

where $\bar{k}$ is the average degree of the system. This distribution also has the average degree as the characteristic scale.

Power-law distribution

The power-law distribution looks like

$$P(k) \propto k^{-\gamma}, \tag{2.4}$$

where γ is the exponent of the distribution. Plotted in a log-log scale the slope of the curve is equal to the exponent (fig. 2.4 c). This distribution doesn't have a characteristic scale and is therefore often called *scale-free*. All real networks has of course a finite size, which gives a size dependent cut-off where the distribution ends. A cumulative plot of a power-law distribution gives

$$P(>k) \propto k^{-\gamma+1}, \tag{2.5}$$

since it's an integration of $P(k)$ over k (see section 2.2 *degree distribution*).

To get more insights in the subject see page 12 of reference [6].

2.3.3 Real world networks

Real networks can be found almost everywhere and in all contexts. In this thesis a couple of real world networks spanning from human transport networks, to modern information networks to biological networks will be examined. Table 2.1 presents a description of them and figure 2.5 shows their degree distributions. "E-c" stands for *Escherichia Coli* (usually called E-coli), "C-e" for *Caenorhabditis elegans*, "prot." for proteins, "metab." for metabolic and "p-l" for power-law. Networks that are directed has an out- and an in-degree distribution.

The world wide web network is a small piece extracted from a larger piece ($3.25 \cdot 10^5$ nodes) which was downloaded from the home page of A.L. Barabasi [20]. Here, the smaller piece of the WWW is used because the calculations would take too much time otherwise. In order to get a good representation of the larger network, the small piece was extracted using a breath-first algorithm with a probability condition inversely proportional to the degree of the selected node. That is, at each step, a node, i, is selected with the probability $P(select) = 1/k_i$. When the extraction get stuck, is starts all over from the top and tries again to select the nodes that previously was denied. This makes sure that the small network roughly has the same degree distribution as the bigger one but with a reasonable size dependent cut-off.

Network	Nodes	Links	Size	k_{av}	degree distr.	Ref.
Stockholm	Roads	Intersections	3325	1.5	Expon. tail	[16]
US Airports	Airports	Flight routes	332	6.4	Expon. tail	[19]
WWW	Web pages	Hyperlinks (URLs)	9999	2.2	Out:p-l ($\gamma = 2.6$) In: p-l ($\gamma = 2.3$)	[20]
Internet	Routers	connections	6474	1.9	p-l ($\gamma = 2.2$)	[21]
YPD (yeast)	Proteins	Prot.-prot. interactions	848	2.1	Out: Expon. tail In: Expon. tail	[8],[9]
E-c prot.	Proteins	Prot.-prot. interactions	1522	2.7	Out:p-l ($\gamma = 2$) In: Expon. tail	[11],[12]
E-c metab.	Chemicals	Chemical reactions	851	4.6	Out:p-l ($\gamma = 2.2$) In:p-l ($\gamma = 2.2$)	[20]
C-e metab.	Chemicals	Chemical reactions	503	4.3	Out:p-l ($\gamma = 2.2$) In:p-l ($\gamma = 2.2$)	[20]

Table 2.1: Descriptions of eight real networks ranging from different man made systems to different biological systems. "E-c" stands for *Escherichia Coli* (usually called E-coli), "C-e" for *Caenorhabditis elegans*, "prot." for proteins, "metab." for metabolic and "p-l" for power-law. Networks that are directed has an out- and an in-degree distribution.

2.4 Tools for working with networks

2.4.1 Randomization

3 It is often instructive to compare the structure of the network with a random version of the same network. This is to show the structural differences from a random network. Often it is also important to keep certain features when randomizing in order to exclude them as the reason for the differences. Since the degree distribution is the most striking feature of a network, the randomization used in this thesis will keep the degree distribution fixed [14]. The rewiring is done in the following way (fig. 2.6 a):

1. Randomly pick two links in the system, each with the probability $1/k_{tot}$. The first connected to nodes i and j and the second to nodes l and m.

2. Swap the links so that i connects to m and l to j. If these new links already exists the swap is canceled and the procedure starts over at step 1 again.

This randomization will give a structureless network and keep the degree of each node and thus also keep the degree distribution of the whole network constant. In the case of a directed network the links are swapped so that both the in- and the out-degree of each node is kept constant (fig. 2.6 b).

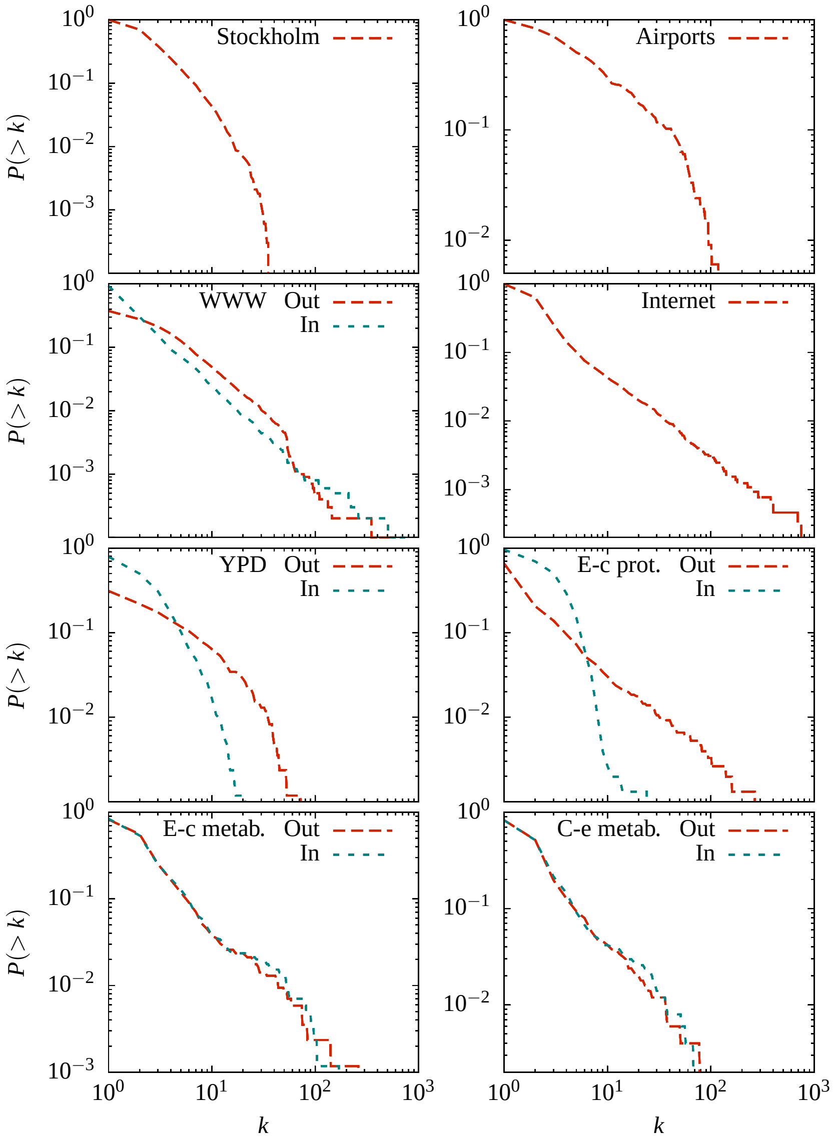

Figure 2.5: Cumulative plot of the degree distribution for eight different real networks. In the case of a directed network the plot is showing both the in- and the out-degree distribution.

2.4.2 Degree correlation profile

Networks with the same degree distribution can have very different structures. One measurement to capture some of these differences is the *degree correlation profile* [14]. The degree correlation profile of a network shows if there are any non-random patterns of connections between specific sizes of nodes. that is, if low degree nodes are more connected to each other than in the random version of the same network or if they are more connected

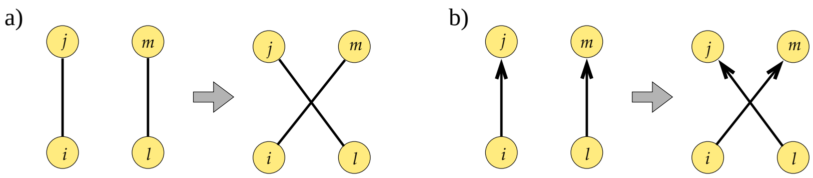

Figure 2.6: a) Undirected randomization that keeps the degree of each node. b) Directed randomization that keeps both the in- and the out-degree of each node.

to high degree nodes instead. Here, the degree correlation profile will be defined as

$$R(k_i, k_j) = \frac{n(k_i, k_j)}{< n_{random}(k_i, k_j) >}, \tag{2.6}$$

where $n(k_i, k_j)$ and $< n_{random}(k_i, k_j) >$ is the number of links that are connected between a node of degree k_i and a node of degree k_j for the real and the randomized version respectively. The $<>$ means an average over many randomization. Since the degree distribution isn't continuous over all sizes between 1 and N and that there are very few links between nodes of exact sizes, groups of nodes (bins) are used instead. R will then be a measurement of the number of links that goes between nodes in certain bins. The fewer bins one uses the better statistic one gets (more nodes to average over), but the amount of information it gives about the network decreases. Three bins will be used in this thesis, *low*, *medium* and *high*. The boundaries are

$$\begin{aligned} 0 \leqslant Low &\leqslant k_{max}^{1/6} \\ k_{max}^{1/6} < Medium &\leqslant k_{max}^{1/2} \\ k_{max}^{1/2} < High &\leqslant k_{max} \end{aligned}$$

where k_{max} is the degree of the largest hub. The R values makes up a 3×3 matrix (low-low, low-medium, medium-low etc.) and is plotted as a 2D surface with a color scheme representing the R value.

2.4.3 Z-score

Measurements like the degree correlation profile are based on a comparison between a number for the network that is being studied and an average number for the randomized versions of the same network. The difference in the numbers also has to be confirmed by how likely it is that these numbers comes from the same distribution assuming that they are normally distributed. The *Z-score* in this case is the difference between a number and a mean, normalized with the standard deviation of the parent distribution of the mean [14]. The Z-score gives the difference in units of standard deviations and thus a significance value. The Z-score is given by

$$Z = \frac{x_i - \bar{x}}{\sigma}, \tag{2.7}$$

where x_i is the number to be tested, $\bar{x}$ and σ is the mean and the standard deviation respectively of the distribution of numbers from the randomized networks. This means that if the Z-score is equal to two, the probability of drawing a number $x \geqslant x_i$ from the distribution of numbers from the randomized networks is 2.3 procent and for $Z = 3$ the probability is 0.1 procent. In this thesis the latter significance level will be used which means that all Z-scores that are bigger than three or smaller than minus three will be considered as significant.

Sometimes one wants to compare to numbers where both are an average over a large number of samples. In this case the question is how likely it is that these two means are the same. The Central limit theorem states that a sum of independent, identically distributed random variables approaches a normal distribution as the sample size approaches infinity (page 284 of reference [7]). This means that the average of a large sample is normally distributed with the mean equal to the average value and the standard deviation equal to the standard deviation of the individual observations divided by the square root of the sample size. The hypothesis that two averages ($\bar{x}_1$ and $\bar{x}_2$), made up of independent normally distributed random variables, are the same gives the Z-score

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{2.8}$$

(page 422 of reference [7]). The difference between them will be given in units of standard deviations and the significance level used will be the same as in the previous case above.

Part I

Chapter 3

Merging and regeneration

3.1 Background

The field of studying networks is very young but there has been a lot of studies made on real networks and especially their degree distribution. These studies show that a surprisingly large number of networks have a broad (scale-free) degree distribution. Biological networks (proteins, reactions etc.), social network s (e.g. sexual relationships, research collaborations) and modern information networks (internet, the world wide web etc.) all have this common feature [4]. So, an obvious question became, why this is the case, is there a universal rule in nature that governs the build up of these networks? R. Albert et al. presented a proposal that said that a scale-free network is roboust against random attacks (since a random attack most likely will hit a low degree node) but this also means that such a network is weak against deliberate attacks (hitting a high degree node will influence the network a lot). The idea is then that nature perhaps would like to protect itself against random mutations and thus develops this scale-free feature [1]. Several models have been presented to explain this phenomenon (the most famous one is probably *preferential attachment* [2]) and they all most likely give a good clue of whats really going on. Beom Jun Kim et al. developed a model in 2004 which give self organized scale-free networks based on merging and regeneration [13]. The model turned out to be very robust and gave a scale free degree distribution. Even though a lot has been published on this subject not much has been done with directed networks. This, in spite of the fact that many of the real networks are directed! Many biological networks are directed not to mention the WWW, and it's easy to imagine others, like for example companies investing money in other companies. The question is now, what happens if we treat the network in the merging process as directed instead? A directed network has one dimension more of complexity since each node now have an in- and an out-degree and thus also an in- and an out-degree distribution. Since a node is connected with its neighbors through in- and out-links there might be an asymmetry due to how one treat the different links in the model. Will this give different results?

Due to the directedness of the network this can be done in several ways. Here, two types will be considered, *Friendly merging* and *Hostile merging.*

3.2 The model

The model is based on merging and regeneration of nodes. That is, two neighboring nodes merge to one larger node and a new small node is put into the system to keep the system size constant. This model can be seen as the evolution of companies that invest money in other companies and by that have some control over them. So, if company A has invested in company B there is a link from A to B. This means that the big companies have a lot of in- and out-links, since they have a big turnover of money, and small companies will have few links. The model is then to let the companies buy up each other and restart from scratch until a steady state has been reached and the average size of a company is constant.

3.2.1 Friendly merging

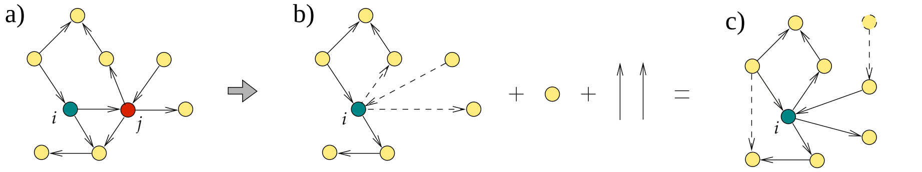

Figure 3.1: a) Node i is randomly picked to merge with one of its out-links neighbor, j. b) Node i gets all the links that are connected to j except the ones they have in common and the ones pointing to each other ($N_{common,in}$ and $N_{common,out}$), so one node and two links are taken away from the system. Node i thus get the in- and out-degree, $k_{i,in} = k_{i,in} + k_{j,in} - N_{common,in}$ and $k_{i,out} = k_{i,out} + k_{j,out} - N_{common,out}$. c) One node and two links are put in at random to keep the system size constant.

If company A makes a friendly takeover and buys up company B, all the assets of B will belong to A. The companies that had money invested in B will now have investments in A instead and the ones with investments in both of them will now only have investments in A. The same thing happens for those that both A and B have investments in. So, *Total merging* means that a node gets all the links from the neighbor that it merges with (except the ones it already has). In network language, the following will be done at each step (fig. 3.1):

1. Randomly pick a node, i, with in- and out-degree $k_{i,in}$ and $k_{i,out}$.
2. Randomly pick one of its neighbors, j, with in- and out-degree $k_{j,in}$ and $k_{j,out}$, through one of the out-links of i.

3. Move all the links (in- and out-links) connected to j, so that they connect to i instead. Node i will now have the in- and out-degree $k_{i,in} = k_{i,in} + k_{j,in} - N_{common,in}$ and $k_{i,out} = k_{i,out} + k_{j,out} - N_{common,out}$. The term $k_{common,in/out}$ is the number of links that will disappear because i and j have links to the same nodes or to each other, since it is not allowed to have several links to the same node or links pointing to itself.

4. Put in a new node, l, with degree zero, $k_{l,in} + k_{l,out} = 0$, and r number of links that connects randomly in the system.

The number r is a parameter that decides how many links there are in the system. When two nodes merge, several links will be removed ($k_{d,in/out}$) from the system. If r is bigger than that, the total number of links in the system will increase until $< k_{d,in} + k_{d,out} >= r$ and steady state is reached. And of course the opposite will occur if r is smaller than the number of links removed at each step.

3.2.2 Hostile merging

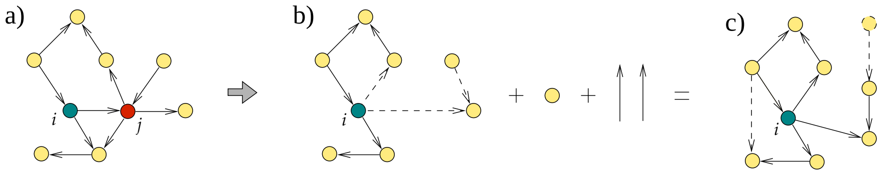

Figure 3.2: a) Node i is randomly picked to merge with one of its out-links neighbor, j. b) Node i gets all the out-links that are connected to j except the ones they have in common and the ones pointing to each other ($N_{common,in}$ and $N_{common,out}$), so one node and two links are taken away from the system. Node i thus get the in- and out-degree, $k_{i,in} = k_{i,in} - N_{common,in}$ and $k_{i,out} = k_{i,out} + k_{j,out} - N_{common,out}$ and the end of the links pointing to j is moved to a random node. c) One node and two links are put in the system at random to keep the system size constant.

If company A instead makes a hostile takeover and "steal" all the assets of B, the companies that had money invested in B will not be allowed to have control over A. So in this case these companies will be forced to sell their parts in company B and invest the money somewhere else. And again, in the case of both A and B having money invested in the same company, this company will, after the takeover, only have A as an investor. *Hostile merging* thus means that a node gets all the out-links from the neighbor that it merges with (except the ones it allready has). This can then be translated to the following update rule:

1. Randomly pick a node, i, with in- and out-degree $k_{i,in}$ and $k_{i,out}$.

2. Randomly pick one of its neighbors, j, with in- and out-degree $k_{j,in}$ and $k_{j,out}$, through one of the out-links of i.

3. Move all the out-links starting at j, so that they start at i instead, and move all the links coming in to B so that they point to a random node. Node i will now have the in- and out-degree $k_{i,in} - N_{common,in}$ and $k_{i,out} + k_{j,out} - N_{common,out}$ respectively. All the nodes that had links to B will have the same degrees as before. The term $N_{common,out}$ is the number of links that will disappear because i and j have links to the same nodes or to each other.

4. Put in a new node, l, with degree zero, $k_{l,in} + k_{l,out} = 0$, and r number of links that connects randomly in the system.

The number r is the same parameter as for *Total merging.* Notice also that in this case $N_{common,in}$ can only be one or zero depending on if company B also have investments in A $(i \rightleftharpoons j)$.

3.3 Results

3.3.1 Friendly merging

Figure 3.3 a) is showing the cumulative out-degree distribution for three different system sizes of networks constructed with the friendly merging model. The out-degree distribution is a power-law $(P(k) \propto k^{-\gamma})$ with an exponent of $\gamma = 2.2$ for $r = 8$ and reaches, for large systems, over about 2 order of magnitudes. The slope of the degree distribution is not dependent on the system size but it is dependent on the parameter r. In figure 3.3 b) one can see that the degree distribution is shifted to the right when r is increased (more links are added to the system) and the slope gets a little bit flatter. Eventually the power-law will break down because k_tot is approaching $N(N-1)$. One could expect that there would be a difference between the in- and the out-degree since there is a preferential merging to nodes with a high in-degree (high in-degree gives more nodes that have the potential to merge with you). But the in- and the out-degree distribution is exactly the same (fig. 3.3 c), which indicates that the friendly merging is completely symmetric in how the in- and the out-links are treated. As figure 3.3 d) shows, this seems to originate from the fact that each node has, on average, the same in- and out-degree. Figure 3.3 e) and f) is showing the dynamics of how the the degree of the largest node evolves during the merging process for a system of size $N = 5 \cdot 10^3$.

One can see that it stabilizes already at a number of mergings around 2 times the number of nodes in the system. These figures are also showing the symmetry between the in- and the out-degrees. To investigate further why this is the case one can write down the rate equations for the degree of

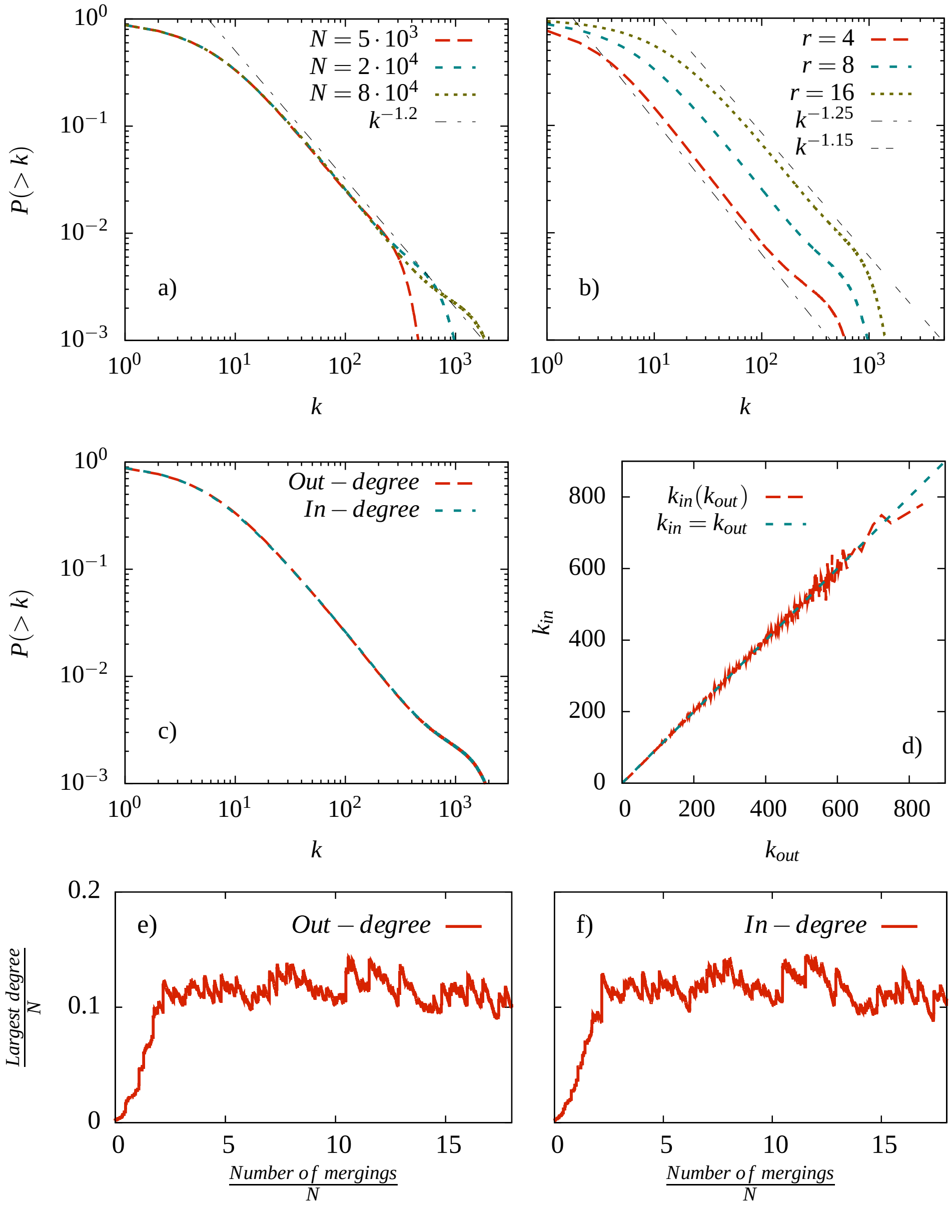

Figure 3.3: Friendly merging: a) Cumulative plot of the out-degree distribution for system sizes $N = 5 \cdot 10^3$, $2 \cdot 10^4$, $8 \cdot 10^4$ ($r = 8$). The fitted line is a power-law with an exponent $\gamma = 2.2$. b) Cumulative plot of the out-degree distribution for $r = 4$, 8, 16 ($N = 2 \cdot 10^4$). The two fitted lines are power-laws with exponents $\gamma = 2.25$ and $\gamma = 2.15$. c) Cumulative plot of the in- and the out-degree distribution ($r = 8$ and $N = 8 \cdot 10^4$). d) Average in-degree as function of out-degree for a node ($N = 5 \cdot 10^3$). d), f) Largest out- and in-degree respectively as function of merging steps ($N = 5 \cdot 10^3$).

a node. Since the merging is symmetric when i merges with j in the sense that it doesn't matter which node is removed and which one is kept, the equations can be written as

$$\begin{aligned} N\frac{\partial k_{i,out}}{\partial t} &= \frac{k_{i,in}}{k_{l,out}}P(k_{i,out},k_{l,out})(k_{l,out}-N_{common,il,out})+ \\ &P(k_{i,out},k_{j,out})(k_{j,out}-N_{common,ij,out})- \\ &P(k_{i,out},k_{m,out})\frac{C_{out}(k_{i,out})\cdot k_{i,out}}{k_{m,out}}+r \end{aligned} \tag{3.1}$$

$$\begin{aligned} N\frac{\partial k_{i,in}}{\partial t} &= \frac{k_{i,in}}{k_{l,out}}P(k_{i,in},k_{l,in})(k_{l,in}-N_{common,il,in})+ \\ &P(k_{i,in},k_{j,in})(k_{j,in}-N_{common,ij,in})- \\ &P(k_{i,in},k_{m,in})\frac{C_{in}(k_{i,in})\cdot k_{i,in}}{k_{m,out}}+r \end{aligned} \tag{3.2}$$

The first term on the right hand side (in equation 3.1 and 3.2) is the number of links that node i gets at each time step (merging step) when node l merges with node i. The second term is when node i merges with node j, the third term is the number of links that node i looses when two of its neighbors merges with each other, and the forth term is the number of links that node i gets from the random links that are put in the system. $P(k_{x,in/out},k_{y,in/out})$ is the probability that a node of degree $k_{x,in/out}$ is linked to a node with degree $k_{y,in/out}$ and $C_{in/out}(k_{i,in/out})$ is the fraction of possible mergings that the neighbors of node i can do. Both $C_{in/out}(k_{i,in/out})$ and $P(k_{x,in/out},k_{y,in/out})$ are impossible to write out exact due to the complexity of the network. $P(k_{x,in/out},k_{y,in/out})$ depends on a non-trivial degree correlation, but in this case the out- to out-degree correlation is the same as the in- to in-degree correlation in the steady-state (see fig. 3.4 a and b). This makes it possible to take out $P(k_{x,in/out},k_{y,in/out})$ from equation 3.1 minus equation 3.2. So, in the steady-state we have the following difference in rates between the in- and the out-degree of a random node:

$$\begin{aligned} \left(\frac{\partial k_{i,out}}{\partial t}-\frac{\partial k_{i,in}}{\partial t}\right) &\propto \frac{k_{i,in}}{k_{l,out}}(k_{l,out}-k_{l,in})+(k_{j,out}-k_{j,in}) \\ &-\frac{k_{i,in}}{k_{l,out}}(N_{common,il,out}-N_{common,il,in}) \\ &-(N_{common,ij,out}-N_{common,ij,in}) \\ &-(C_{out}(k_{i,out})\frac{k_{i,out}}{k_{m,out}}-C_{in}(k_{i,in})\frac{k_{i,in}}{k_{m,out}}). \end{aligned} \tag{3.3}$$

The two first terms on the right hand side does'nt say very much since they represents the degree of a neighbor. The third and the fourth terms on the other hand says that if node i has more common out-links than in-links with a neighbor, the growth if the out-degree will be slowed down compared to the growth of the in-degree. Simply because the number of common links

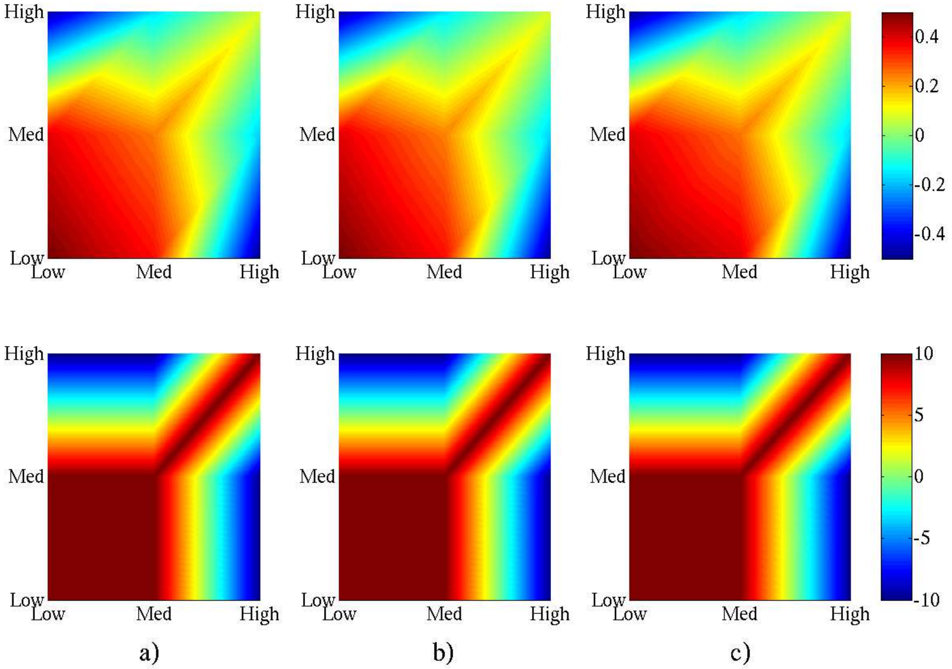

Figure 3.4: Degree correlations in the friendly merging network. a) out-out correlation, b) in-in correlation and c) out-in correlation. The upper row shows a color representation of $log(R)$ (R from eq. 2.6) and the lower row of the Z-score, Since the R-value from both the merging and the randomized version is an average over many networks, the Z-score comes from equation 2.8. Values close to zero in the upper row means a very small difference from the random version and in the lower row it means a difference that isn't significant.

for i and j is proportional to $k_i k_j$. The same thing will happen to the last two terms: more neighbors gives a higher probability that some of these neighbors are connected to each other and thus have the chance to merge. So, this gives a hint about why this model is so symmetric. Simply put: the more friends you have, the more friends you risk to lose.

The friendly merging model creates a network with a very non-random degree correlation. Figure 3.4 is showing three different degree correlation matrises (as described in section 2.4.2) where figure a is the out-out correlation, b is the in-in correlation and c is the out-in correlation. All of them look exactly the same (as suggested in eq. 3.3) which also shows the complete symmetry between the in- and the out-links. The upper row in the figure is showing $log(R)$, where the R value comes from equation 2.6 and the lower row is showing the Z-score, equation 2.8. The plots show that low-low, low-medium, medium-medium and medium-low degree nodes has a significantly higher number of connections than in the random case. On the other hand, high-low, high-medium, low-high and medium-high degree nodes has a significantly smaller number of connections. The high-high degree nodes

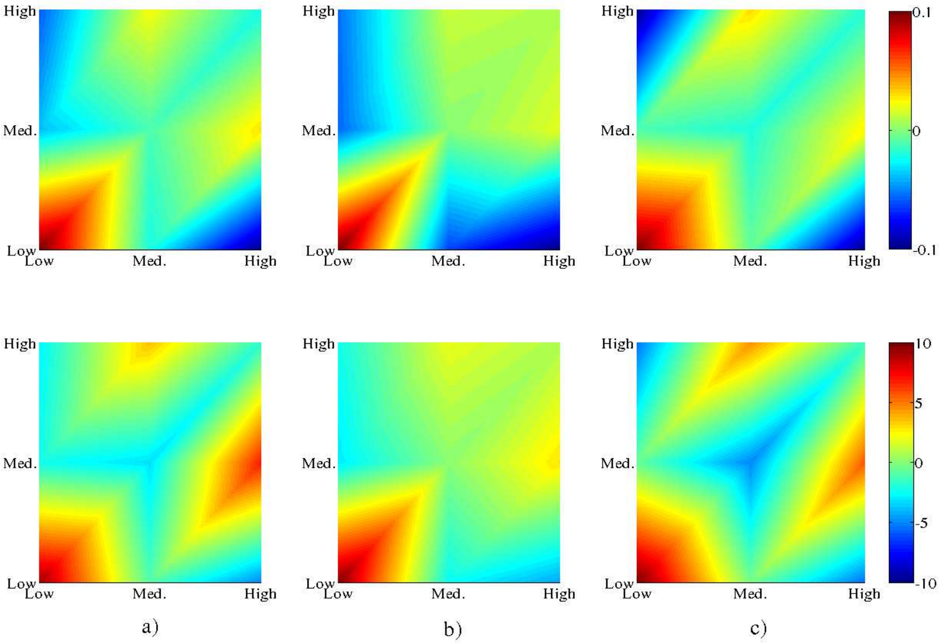

Figure 3.5: Degree correlations in the E-coli metabolic network. a) out-out correlation, b) in-in correlation and c) out-in correlation. The upper row shows a color representation of $log(R)$ (R from eq. 2.6) and the lower row of the Z-score, Since the R-value from the randomized version is an average over many networks, the Z-score comes from equation 2.7. Values close to zero in the upper row means a very small difference from the random version and in the lower row it means a difference that isn't significant.

doesn't show a significant difference at all. Since hubs tend to connect to each other in a random network this implies that high connected nodes are connected to each other also in the friendly merging network.

As a final observation one can see that the metabolic networks in section 2.3.3 have close to the same in- and out-degree distribution and with an exponent of $\gamma = 2.2$. They too actually approximately follows the pattern shown in figure 3.3 d, with, on average, the same in- and out-degree of each node. One can see in figure 3.5 that the E-coli metabolic network has, like the friendly merging, more links between low degree nodes and fewer links between low to high and high to low than random. The three degree correlation profiles (out-out, in-in and out-in) also looks pretty much the same. Is there a connection between the metabolic networks and the friendly merging model?

3.3.2 Hostile merging

The hostile merging also gives a power-law but with an exponent around $\gamma = 1.55$, with $r = 8$, for the out-degrees distribution. The in-degree distribution in this case has an exponential tail (fig. 3.6 a and c). This model is also size independent and the parameter r has the same effect as for the friendly merging with a decreasing exponent for an increasing r (fig. 3.6 a and b) even though the difference is smaller in this case ($\gamma = 1.55$ for $r = 4$ and $\gamma = 1.6$ for $r = 16$). Figure 3.6 is showing the average in-degree of a node as a function of its out-degree and they are not proportional to each other. From figure 3.6 e) and f) one can see that it takes more merging steps to reach a steady state , around ten times the number of nodes in the system, then for the friendly merging. And even though the size of the largest in- and the out-degree is completely different they reach the steady state more or less at the same time. The largest in-degree also fluctuates much less then the largest out-degree which makes sense since the distribution of degrees it can increase with is much more narrow.

The degree correlation of the hostile merging doesn't show as much structure as for the friendly merging. Figure 3.7 *a* shows the out-out, *b* the in-in and *c* the out-in degree correlation profile. One can see in figure *a* that for the out-degrees there are significantly more links between low-low and low-medium degree nodes and significantly less links between high-low and high-medium degree nodes, than in the random version. The other bins shows a somewhat significant, but very small, difference. The fact that there are so little structure for the out-out could indicate that the narrow in-degree distribution could make it difficult to get out-out correlations. In figure *b* and *c* there are a significantly larger number of links between the low and medium bins than in the random case. The high bin seems to have roughly the same number of links to all bins as in the random version for *b* and only a small difference for *c*.

The comparison with degree distributions for this model would be the *YPD* and *E-c prot.* in figure 2.5. They both have a much broader out-degree than their in-degree.

3.4 Conclusions

The friendly merging gives a robust power-law degree distribution for directed networks. The in- and the out-degree distributions are also exactly the same which means that there, surprisingly, isn't any asymmetry between the in- and the out-degrees even though there is a preferential merging for nodes with high in-degree.

The hostile merging also gives robust power-law degree distribution but only for the out-degree. This model is remarkably close to the random merging considered in the undirected case [13]. In that case random nodes where merged with each other without any regard of the under laying network. But this is a very network-based model and still get pretty much the same result! This can be explained by the randomness of the in-degree. The in-degree distribution is narrow (has an exponential tail) and with a more random

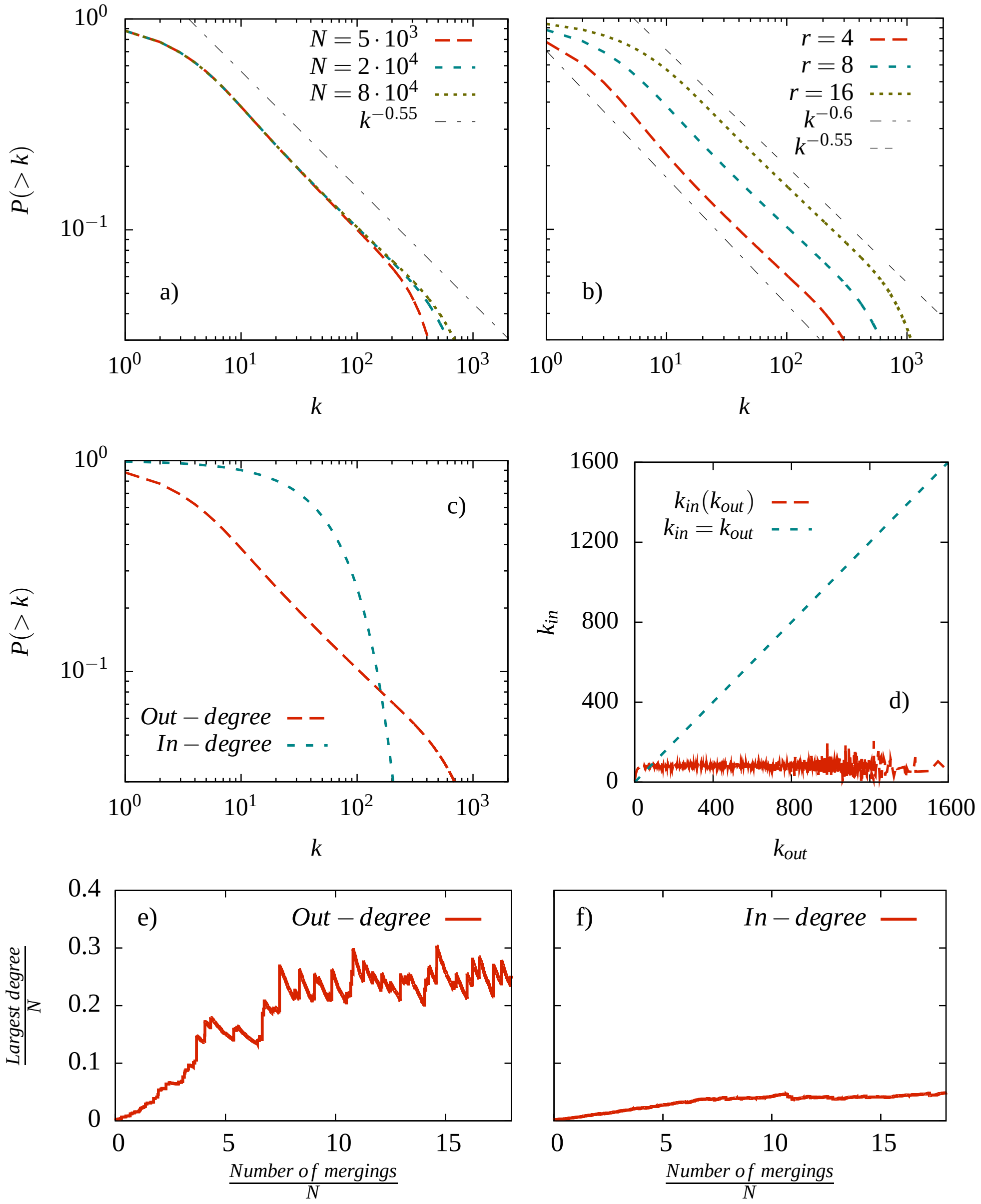

Figure 3.6: Hostile merging: a) Cumulative plot of the out-degree distribution for system sizes $N = 5\cdot10^3,\ 2\cdot10^4,\ 8\cdot10^4$ $(r = 8)$. The fitted line is a power-law with an exponent $\gamma = 1.55$. b) Cumulative plot of the out-degree distribution for $r = 4,\ 8,\ 16$ $(N = 2\cdot10^4)$. The two fitted lines are power-laws with exponents $\gamma = 1.6$ and $\gamma = 1.55$. c) Cumulative plot of the in- and the out-degree distribution ($r = 8$ and $N = 8\cdot10^4$). d) Average in-degree as function of out-degree for a node $(N = 5\cdot10^3)$. d), f) Largest out- and in-degree respectively as function of merging steps $(N = 5\cdot10^3)$.

structure than in the friendly merging. This makes a hostile merging over a link very close to a random merging.

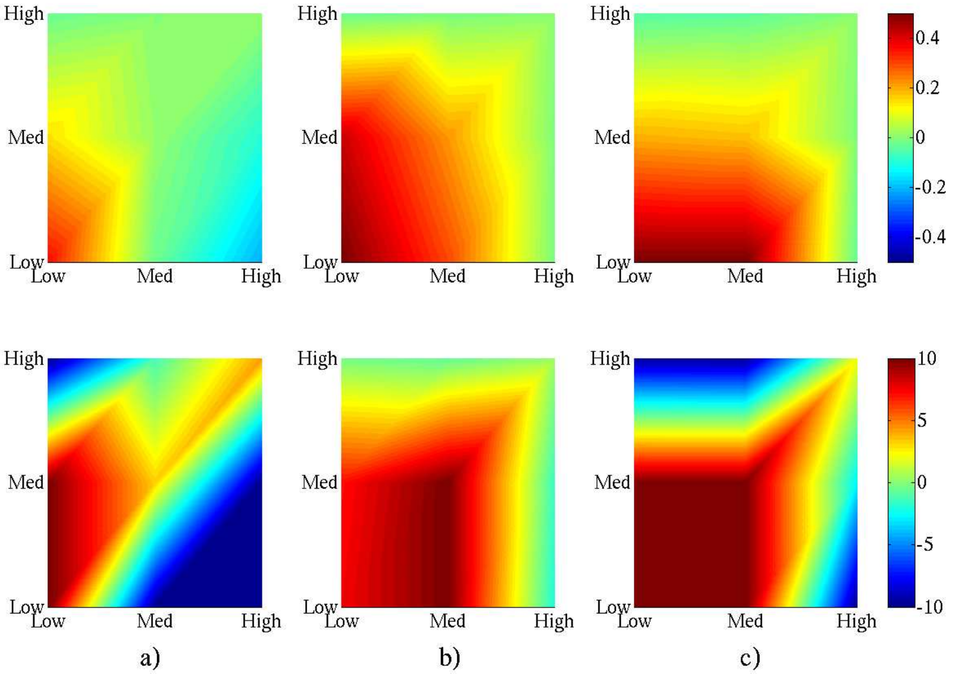

Figure 3.7: Degree correlations in the hostile merging network. a) out-out correlation, b) in-in correlation and c) out-in correlation. The upper row shows a color representation of $log(R)$ (R from eq. 2.6) and the lower row of the Z-score. Since the value from both the merging and the randomized version is an average over many networks, the Z-score comes from equation 2.8. Values close to zero in the upper row means a very small difference from the random version and in the lower row it means a difference that isn't significant.

Even though merging may not be the underlying process of how these biological networks evolved, it shows that these characteristic features of some of the biological networks aren't unique and that they are fairly easy to reproduce. It gives a hint about what's important in the understanding of biological networks and maybe there are things beyond the degree distribution that are more important. Either the degree distribution has little to do with the function of the network, which means that the degree distribution comes from something else, or these types of broad degree distributions are universal for many different functions.

Part II

Chapter 4

Multi-step degree correlation

4.1 Background

A very important question in network science is how the structure of a network is linked to its function. If networks with different functions all have a random structure it would indicate a separation between the structure and the function. Thus a necessary signature for coupling between function and structure is a non-random structure. As mentioned in part I, many of the real networks have a very broad degree distribution. These networks stand out from networks with narrow degree distribution because of the highly connected nodes, hubs, that influences a very big part of the network. If the hubs play a crucial role for the network function, their position in the network should not be random. In a random network the hubs tend to link to each other ($P(k_i, k_j) \propto k_i \cdot k_j$) and create a highly connected core. This core will enhance the small world features (a small number of steps is necessary to reach all nodes in the system) [17] observed already in networks with narrow degree distributions. In this part, several real world networks will be examined to answer the questions: are there any non-random degree correlations at longer steps than one and if so, is it independent on the one step degree correlation? Can the multi-step degree correlation also give a hint about the relative position of hubs!?

4.2 Randomization keeping the degree correlation constant

The algorithm for the randomization keeping the degree correlation constant is a Metropolis algorithm with simulated annealing. The degree correlation is measured in the same bins as in section 2.4.2 and thus makes up a 3×3 matrix. Each element in the matrix contain the number of links that goes between nodes of the size that the element in question represents. In each step a random swap is made in the same way as in section 2.4.1, but with the probability of the Boltzmann factor

$$P(swap) = e^{-\Delta E/T}, \tag{4.1}$$

where ΔE is the energy difference that the swap would give and T is a scaling factor, a sort of "temperature". The energy function is defined as

$$E = \sqrt{\sum_{i,j=1}^{i,j=3} (\mathbf{A}_{ij} - \mathbf{T}_{ij})^2}, \tag{4.2}$$

where $\mathbf{A}_{ij}$ is the number of links in the system that goes from a node in bin i to a node in bin j and $\mathbf{T}_{ij}$ is the number of links that the final system should have between bin i and j, $\mathbf{T}$ is here called the *target matrix.* The energy function is thus a scalar distance between the two matrises and E is equal to zero if the two matrises are the same. So, a swap is automatically made if it moves the degree correlation matrix ($\mathbf{A}$) closer to the target matrix ($\mathbf{T}$) which gives a negative ΔE. But, if it moves it away from the target matrix (positive ΔE) the swap is only made with the probability in equation 4.1. In this case however, we start from the target matrix and want the degree correlation matrix to move away from the it in order to span over as much of the configuration space as possible (get a network that is as random as possible) and then get back to the target matrix. This is done by starting at a high T which makes it easy to accept forbidden swaps and then slowly lower the temperature so that the energy function settles in a global minimum ($E = 0$). Thus, the resulting network will have exactly the same degree distribution and degree correlation matrix as before the randomization.

If one wants to keep the complete degree correlation exact, each link has to go between nodes of exactly the same size before and after the randomization. This however puts a huge constraint to the number of possible networks one can create. The outcome of this will be a very large overlap (fraction of links that goes between exactly the same nodes as before the randomization) between the real network and the randomized version of it. By keeping only three bins, one makes sure that the overlap is small and that most of the original structure of the real network is destroyed in the same time as the degree correlation is roughly constant. How one chooses the boundary's of the bins is very important, especially in the lower range, and shouldn't be chosen arbitrarily. The degree correlation matrix is symmetric for a undirected network but not for a directed network.

4.3 Multi-step degree correlation

The multi-step degree correlation is a measure of the average degree as a function of steps taken out in the network. That is, starting from a certain node, what is the average degree one encounters at each distance, stepping out in the network along the shortest paths? This is then averaged, per step, over all nodes in the system that can reach that many steps. The equations looks like

$$< K_{l,i} >= \frac{1}{N_{l,i}} \sum_{j=1}^{N_{l,i}} k_{l,i,j}, \tag{4.3}$$

$$< K_l >= \frac{1}{N_l} \sum_{i=1}^{N_l} < K_{l,i} > . \tag{4.4}$$

$< K_{l,i} >$ is the average degree l steps out in the network from node i, $N_{l,i}$ is the number of nodes that node i can reach in l steps and $k_{l,i,j}$ is the degree of node j that node i can reach in l steps. $< K_l >$ is the total average degree that an arbitrary node encounters l steps out in the network and N_l is the number of nodes that has at least one shortest path to another node of length l. In order to get a separation between nodes of different sizes, $< K_l >$ is calculated from three bins defined in the same way as in section 2.4.2. That is, what does the average degree look like stepping out from a node in a certain bin. To normalize it $< K_l >$ for a real world network is divided by the same quantity averaged over many randomized versions of the same network, $< K_l >_{random}$.

4.4 Results

The global degree correlation (eq. 4.4) is calculated for six real world networks, *Stockholm*, *US Airports*, *WWW* (World Wide Web), *Internet*, *YPD* (yeast protein-protein) and *E-coli* (protein-protein). The data for these networks is presented in section 2.3.3. The WWW, YPD and E-coli is in principle directed networks with large asymmetries between in- and out-degrees. These asymmetries are maintained in the randomization by keeping both the in- and the out-degrees of every node (section 2.4.1). It is also the directed degree correlation profile, out-in, that is kept when randomizing keeping the degree correlation profile (section 4.2). In order to facilitate comparisons with other networks the directionality are ignored in all the analysis.

Figure 4.2 is showing the global degree correlation for the six networks. The continuous line is showing the average result starting from all nodes. The dashed lines are showing the result for starting at nodes with a degree in bin *Low*, *Med.* and *High*. The figure clearly shows that most of these networks has a very non-random structure in the sense of degree sequences along shortest paths.

Due to the small world effect from hubs, all the shortest paths longer than a couple of steps will go through a hub. Often already at step one or two. Since it's an average over all nodes the total average degree as a function of steps will decrease almost exponentially for both the randomized version and its parent network (see figure 4.4): most nodes reaches a hub in the first step, a smaller number of nodes reaches a hub in two steps and so on. So, it's really the differences in the slopes of the curves that is informative and this information is given in the $< k > / < k_r >$ curve.

Figure 4.1 shows a very simple example of the differences between stepping out in a network with hubs separated from each other (to the left in the figure) and a randomized version of the same type of network (to the right in the figure). In a network with a power-law degree distribution the low degree nodes will dominate the total average of the global degree correlation (since they are so many) so they are considered to give total average in the example. The basic idea is that a network with hubs separated by low and medium degree nodes and otherwise only connected to low degree nodes gives a sick-sack pattern while the randomized version is more uniform. In the network to the left in figure 4.1, the low degree nodes will hit a hub in the first step and in the fourth step. In the randomized case the low degree nodes hits both hubs and low or medium nodes in the first step which averages to medium. They will also pass the hubs in a short number of steps and hit only low and medium degree nodes in their fourth step. This gives rise to the curve in the bottom of the figure. Notice that the diameter of the network shrinks a lot when randomized and many more nodes can be reached in the same number of steps than in the parent network. It is of course not this simple in real world networks but the basic shape of the $< k > / < k_r >$ curve gives a clue about the structure of the network. A dip in step two is a strong indication of hub separation which effectively means less links between hubs than in the randomized case. and if it takes several steps to "recover" from the dip could mean that it takes many steps to pass the hubs in the randomized case.

The results for a network that has hubs connected to each other and to medium, medium connected to high, medium and to low and finally low connected to medium and to low degree nodes will be almost the opposite of figure 4.1. Now the low degree nodes will hit other low and medium degree nodes and the hubs will hit other hubs and medium degree nodes while in the randomized case all nodes will hit all types of other nodes. This means that the $< k > / < k_r >$ curve will start below one for the total average and above one for the high bin. For the total average the curve will start to increase when stepping out in to the network until it hits the core (all the connected hubs). This will be the maximum and the curve will go down again along with the path hitting medium and low degree nodes until the end of the network is reached.

All the data points presented in figure 4.2 and 4.3 have a significant difference between k and k_r due to the large sample size. Here the z-score presented in equation 2.8 has been used.

A quick analysis of figure 4.2 tells us the following about these real world networks:

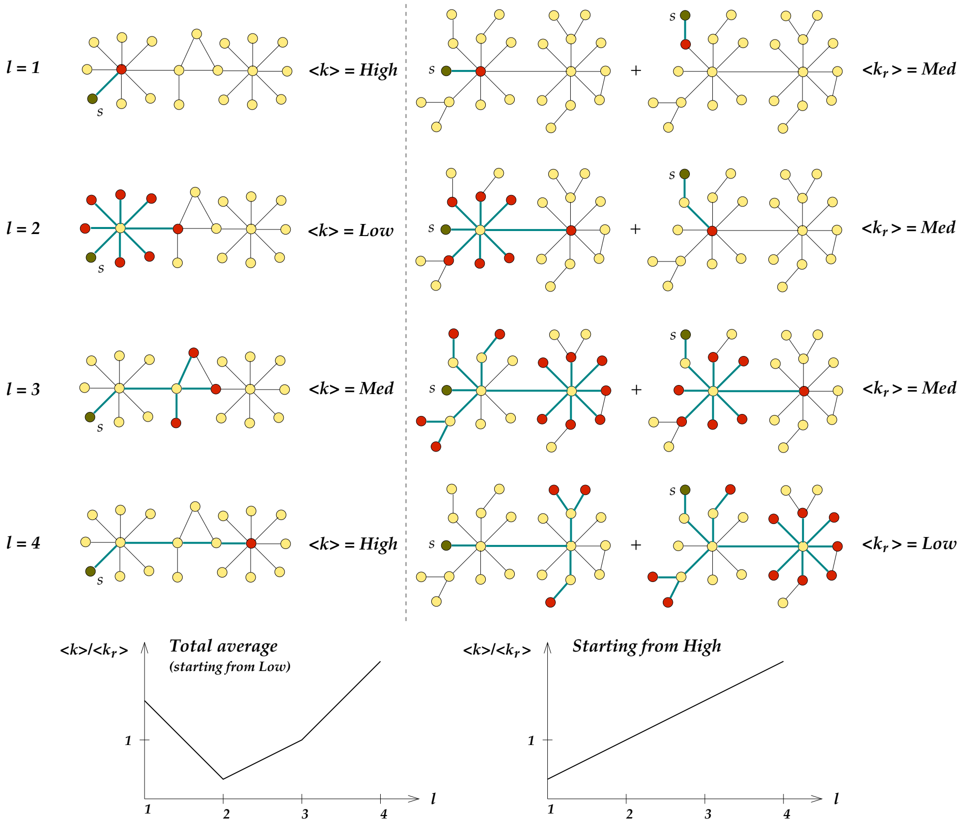

Figure 4.1: The figure is showing a very simplified example of the curve $< k > / < k_r >$ as a function of l for an extreme case and an explanation for the shape of the curve. The upper part is showing average degree sequences when stepping out from a typical node (s) in an example network (left) and a randomized version of the same type of network (right). l is the number of steps out in the network and $< k >$ and $< k_r >$ is the average degree for the example network and the randomized version respectively. The example network has the properties of hubs separated by low and medium nodes and the hubs are connected to a lot of low degree nodes. In the randomized version the hubs are connected to each other and also to both medium and low degree nodes. In a scale-free network the low degree nodes will dominate the total average because of their large number. This will give the curve in bottom to the left. Starting from a hub will instead give the curve to the right.

Stockholm

The curves for Stockholm has a very small dip for several steps before they cross over the line $< k > / < k_r >= 1$. The dip is showing that there is a small hub separation but since the degree distribution is quite narrow, the fluctuation in the degrees in the network is not very large which explains the small amplitude of the curves.

Airports

This network is quite close to random at short steps. The hubs are linked to each other a little bit more than in the randomized version but they are also linked to medium and low degree nodes. This makes sense since the flight routes are constructed in a way to make the small world effect as large as possible, you don't want to change plane to many times flying from one side of the country to the other side. The large cut-off in the degree distribution (fig. 2.5) tells us that there are many hubs of approximately the same size in the network which gives a large average degree and a very connected core. This also gives a very short diameter of the network which is why the curves in the figure stops at four steps.

WWW

This piece of the WWW is showing a clear dip for the total average very close to the one in figure 4.1 and it tells us that the hubs are separated more than in the randomized case. This makes sense since really big sites doesn't link to other big sites of the same type because of mutual clientele. The fact that the curve goes down again in the end means that the hubs have been passed even in the real world network with hubs separated. The curve for starting from a high degree node is also showing similarities with hub separation. High degree nodes are connected to low and vice verse.

Internet

Internet also has a dip in the total average, i.e. hub separation, but the recovery is slower and $< k >$ never seems to be larger then $< k_r >$ except at step one. This could mean that the hubs are not really separated by several steps but only in the sense of fewer links among them than random. Another interesting feature is that the High-curve and the Low-curve are almost reflections of each other.

YPD

This network is also showing a dip but has, like Stockholm, a more narrow degree distribution than for example Internet and WWW. This again means that the fluctuations in degrees are smaller which gives a smaller dip. It is very clear though that the curve goes up in the end which points to a conclusion that some hubs are actually separated by at least two steps.

E-coli

The E-coli curve looks a little bit funny. It goes up at step two, down at step three and up again at step five. The curve also starts at one which indicates that the number of connections between hubs are approximately the same

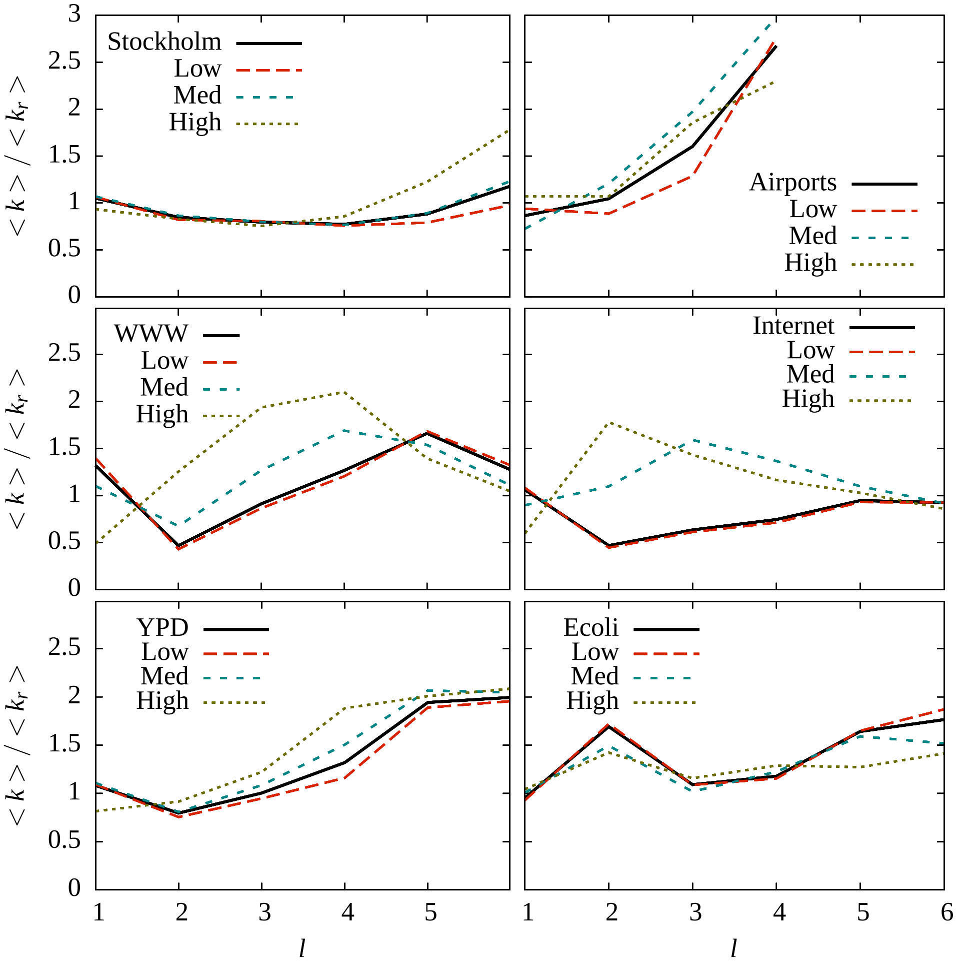

Figure 4.2: The figure is showing, for six real world networks, the average degree as function of the distance from a typical node divided by the same quantity averaged over many randomized versions of the same network (eq. 4.4). The randomization used is the one described in section 2.4.1 which keeps the degree distribution constant. The curve *Low* is an average over starting at a node with a degree in the low bin, *Med* starting in the medium bin and *High* starting in the high bin (see section 2.4.2 for bin boundaries).

as in the random case. The shape of this curve could come from a structure where a few large hubs are connected to several other hubs while in the random case these links are more homogeneously distributed among the hubs. This would mean that one would almost always hit one of these larger hubs at step two but not in the random case and at step three they have been past more often than in the random case. The fact that the curve goes up in the end indicates that there are several hubs that are separated by several steps. This network has a very connected hub, a protein called *RNAP70*, which is connected to about half of the nodes in the network including several other hubs.

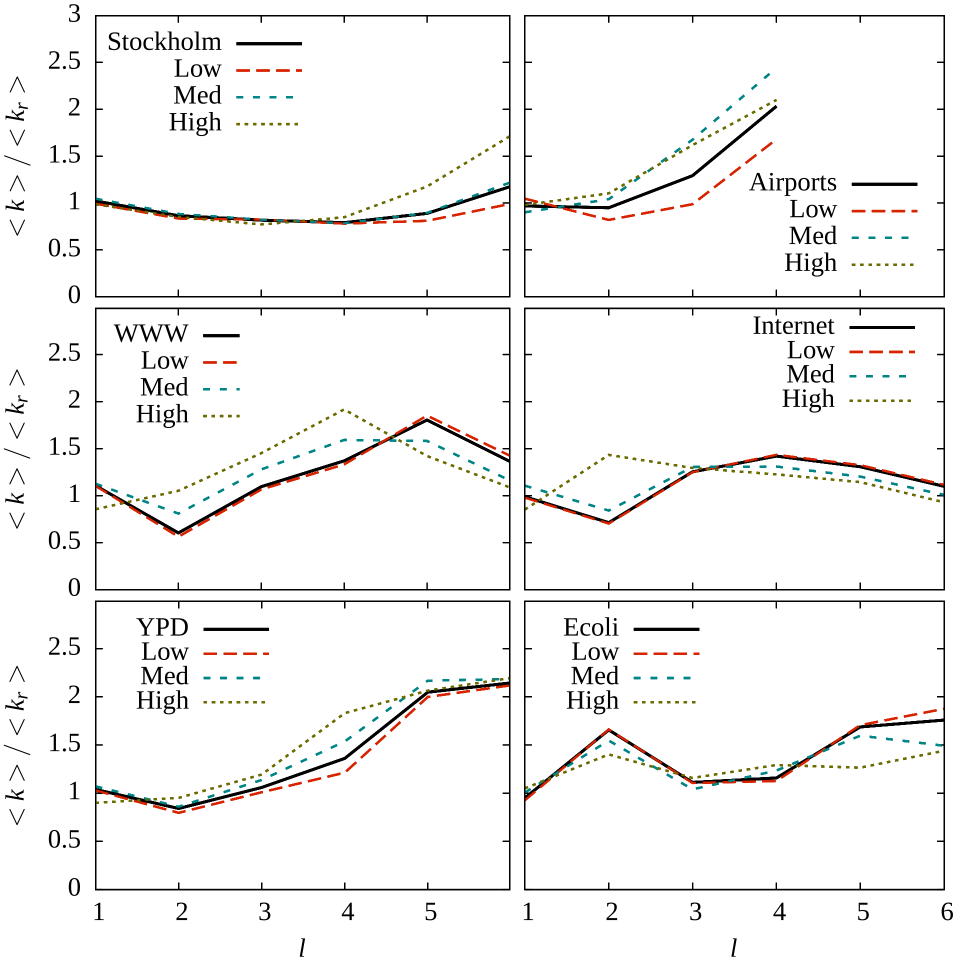

Figure 4.3: The figure is showing, for six real world networks, the average degree as function of the distance from a typical node divided by the same quantity averaged over many randomized versions of the same network (eq. 4.4). The randomization used is described in section 4.2 which keeps both the degree distribution and the degree correlation profile. The curve *Low* is an average over starting at a node with a degree in the low bin, *Med* starting in the medium bin and *High* starting in the high bin (see section 2.4.2 for bin boundaries).

One thing that should be tested is if these non-random patterns are a direct result of the one step degree correlation or if there really is something beyond the first step. Figure 4.3 is showing the same plots as figure 4.2 but the randomization has kept the three bin one step degree correlation profile (section 4.3). The amplitudes of the deviation from the $< k > / < k_r >= 1$ line seems to decrease but the over all shape of the curves are still there. The curves would have started at the value one if the degree correlations had been kept exact but in this case it can at the best move much closer.

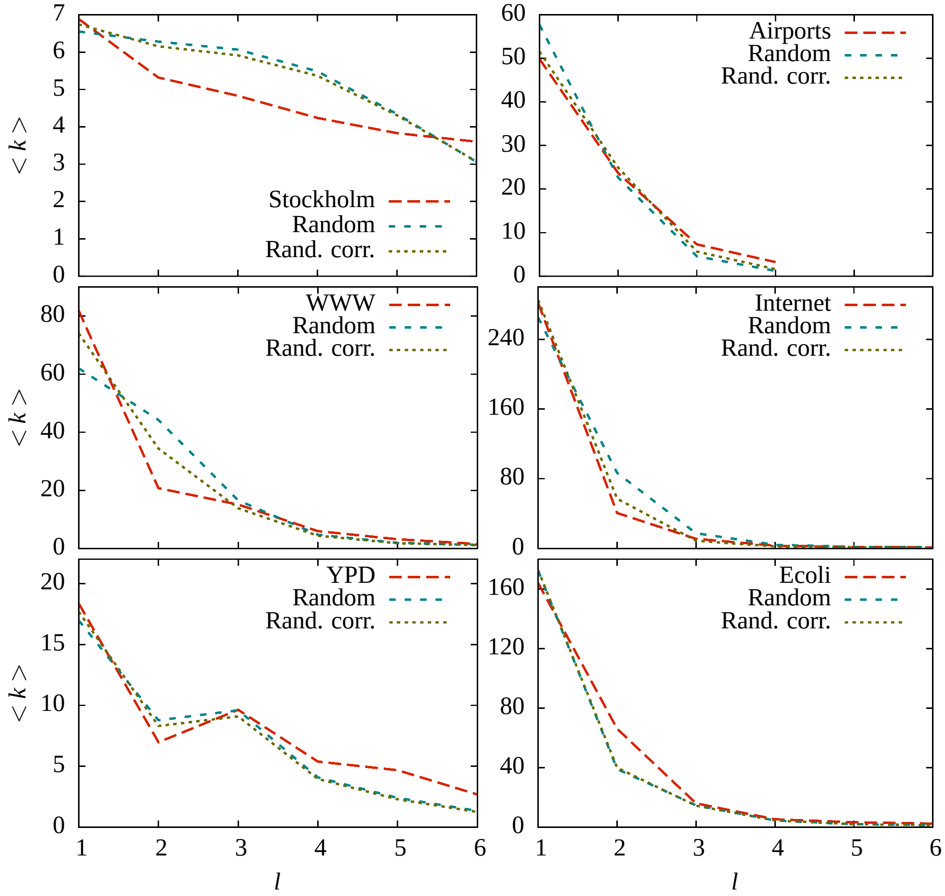

Figure 4.4: The figure is showing, for six real world networks, the average degree as function of distance from a typical node. The three curves in each plot is showing the result for the real world network, the randomized version from figure 4.2 (*Random*) and the randomized version from figure 4.3 where the degree correlation profile is kept constant (*Rand.corr.*). The error bars are smaller then the thickness of the curves.

4.5 Conclusions

The real world networks examined here clearly possesses non-random structures in the sense of degree sequences along shortest paths in the network. The structure can be traced for several steps and gives more insight about the network than the one step degree correlation profile. It is also clear that the multi-step degree correlation isn't effected very much by keeping the one step degree correlation constant when randomizing. The amplitude is without doubts effected but not the over all shape of the curves.

The positioning of the hubs are very important in this measurement and the results are showing that the hubs are not positioned randomly but, in most cases, separated from each other.

Future projects

The work behind this thesis has opened for a lot of questions and future projects, here are some ideas:

- **Connecting the Merging model to real world systems:** What kind of real world systems could be described by the merging model, what similarities do they have and what are the differences? Is it reasonable?

- **Improve the method to extract smaller networks:** How can the method to extract smaller networks from a larger one (in section 2.3.3) be improved. What kind of features is it capturing and what kinds is it not? The same question for a simpler method without the probability condition. Is that the method most often used, and in that case, what kind of errors could be expected when using it? And finally, if a method is developed that keeps the most important features of a network, could it be used to find out how for example different measurements are scaling with the size of the network?

- **Use the randomization keeping the degree correlation profile to create other profiles:** It can sometimes be handy to have a network with a certain degree correlation profile. The method described in section 4.2 could be used to create wanted profiles. How would for example a completely democratic network (all nodes have the same probability to be connected to nodes of all sizes) look like?

- **Develope randomizations which keeps other features of a network:** In the quest for connecting the structure of a network with its function, questions like "what are the key features that explains a certain behavior of a network?", is very important. A randomization like the one in section 4.2 but with another energy function could be a big help for finding the answers.

Acknowledgements

I would like to deeply thank Petter Minnhagen and Kim Sneppen for giving me this great opportunity and guiding me through all the work it took to complete this thesis. Also for sending me to the Geilo school in Norway and to Krogerup where I got the chance to listen to good speakers and talk to interesting people. It has been really fun and stimulating at NORDITA and I have learned a lot. Tord Oscarsson and Patrik Norqvist too deserves a thank you for introducing me to Petter Minnhagen. I would also like to thank all the members of the complex group at NORDITA, the Niels Bohr Institute, and Umeå university for all the help, advice and non the least, the friendship.

Bibliography

[1] R. Albert, H. Jeong, and A.-L. Barabási, *Error and attack tolerance in complex networks.* **Nature 406**, 378 (2000)

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

[3] A.-L. Barabási, R. Albert and H. Jeong, *Scale-free characteristics of random networks: The topology of the World Wide Web.* **Physica A 281**, 69-77 (2000)

[4] A.-L. Barabási and E. Bonabeau, *Scale-Free Networks.* **Scientific American 288**, 60-69 (2003)

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

[6] S.N. Dorogovtsev and J.F.F. Mendes, *Evolution of Networks.* Oxford university press, 2003

[7] H. Frank and S. Althoen, *Statistics, concepts and applications.* Cambridge university press, 1994

[8] P.E. Hodges, A.H. McKee, B.P. Davis, W.E. Payne and J.I. Garrels, *The Yeast Proteome Database (YPD):a model for the organizition and presentation of genome-wide functional data.* **Nucleic Acids Res.**, Jan 1;27(1):69-73 (1999)

[9] 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.**, Jan 1;26(1):68-72 (1998)

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

[11] P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, J. Collado-Vides, S.M. Paley, A. Pellegrini-Tool, C. Bonavides and S. Gamma-Castro, *The EcoCyc Database.* **Nucleic Acids Res.**, Jan 1;30(1):56-8 (2002)

[12] I.M. Keseler, J. Collado-Vides, S. Gamma-Castro, J. Ingraham J, S. Paley, I.T. Paulsen, M. Peralta-Gil and P.D. Karp, *EcoCyc:a comprehensive database resource for Escherichia coli.* **Nucleic Acids Res.**, Jan 1;33(Database issue):D334-7 (2005)

[13] B.J. Kim, A. Trusina, P. Minnhagen and K. Sneppen, *Self-organized scale-free networks from merging and regeneration.* **Eur.Phys.J.B 43.** (2004)

[14] S. Maslov and K. Sneppen, *Specificity and Stability in Topology of Protein Networks.* **Science 296**, 910 (2002)

[15] D. Rind, *Complexity and Climate.* **Science 284**, No 5411 (1999)

[16] M. Rosvall, A. Trusina, P. Minnhagen and K. Sneppen, *Networks and Cities: An Information Perspective.* **Phys.Rev.Lett. 94**:2, 028701 (2005)

[17] D. Watts and S. Strogatz, *Collective dynamics of 'small-world' networks.* **Nature 393**, 400 (1998)

[18] G. Weng, U. Bhalla and R. Lyengar, *Complexity in Biological Signaling Systems.* **Science 284**, No 5411 (1999)

[19] From Pajek dataset at: http://vlado.fmf.uni-lj.si/pub/networks/data/

[20] Homepage of A.-L. Barabasi: http://www.nd.edu/∼networks/resources.htm

[21] Website maintained by the NLANR Measurement and Network Analysis Group at: *http : //moat.nlanr.net/*