Institute of Theoretical Physics
University of Cologne

Diploma Thesis

Analysis and Simulation of Scientific Networks

Felix Pütsch*

July 14, 2003

supervised by Prof. D. Stauffer

Hereby, I confirm (according to the Prüfungsordnung of July 12, 1996, §20(5)) having composed this diploma thesis alone, using no other than the mentioned sources and tools. Citations have been marked.

Hiermit versichere ich gemäss §20(5) der Prüfungsordnung vom 12. Juli 1996, dass ich diese Diplomarbeit alleine erstellt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Zitate wurden kenntlich gemacht.

* fxp@thp.uni-koeln.de
printed July 16, 2003

Though the mountains divide
And the oceans are wide
It's a small world after all

R. M. and R. B. Sherman

Contents

1. Introduction

1.1. Networks

In recent times, hearing the word *network* immediately arouses the idea of physically wired networks as those formed by telephone lines or computer links. Though, *network* is a concept a good deal more general than only this.

Mathematically spoken, networks are *graphs*, i.e. a set of *nodes* (of whatever kind) connected by *edges* (links, connections) between certain pairs.

This abstraction has been known for a long time. Probably the first paper of graph theory was written by Euler [1], the so-called "bridge problem of Königsberg". Euler discusses whether or not it is possible to make a round walk, passing of each of Königsberg's nine bridges exactly once (figure **??**).

The concept of networks can be applied to lots of theoretical or experimental subjects [2–4], nodes being people [5], Internet servers [6], scientists [7, 8] or others, the range of links comprises e-mails [9], friendships [5], citations [10, 11] and more.

Thus, there are numerable different kinds of networks, physical ones (e.g. hard wired) as well as logical (e.g. dependencies) or social ones (e.g. contacts, friendships), stretching out to topics far from wired networks [9, 12]. The area is under vigorous research. Good reviews can be found in [2–4, 13, 14].

1.2. Six Degrees of Separation

Out of personal experience, nearly everybody has been confronted with what we call *small world effect*. There are numerous examples:

At a party, we find out to know some stranger we just started talking to by only a few middle-persons (or technically spoken we are only separated from him by a low degree). E.g., he could be our street neighbors' colleague's son. We hear "My god, how world is small".

Rumors are another example. We are astonished to experience the pace at which they spread. After a few hours and thus only a few possibilities of telling rumors to others, whole city seems to know.

Milgram [15] made an experiment on this. He instructed a set of people to try to send a letter to some stranger, only by using personal contacts. He found out that an astonishing short chain of social links is needed for this task, which entered in everyday's language as *Six Degrees of Separation*. Recently, this has been reviewed on a more popular basis by a German weekly newspaper [5].

1.3. Small World Effect

Six Degrees of Separation is only one manifestation of a more general principle: the *small world effect* [16–18].

Observations of many real-world net-

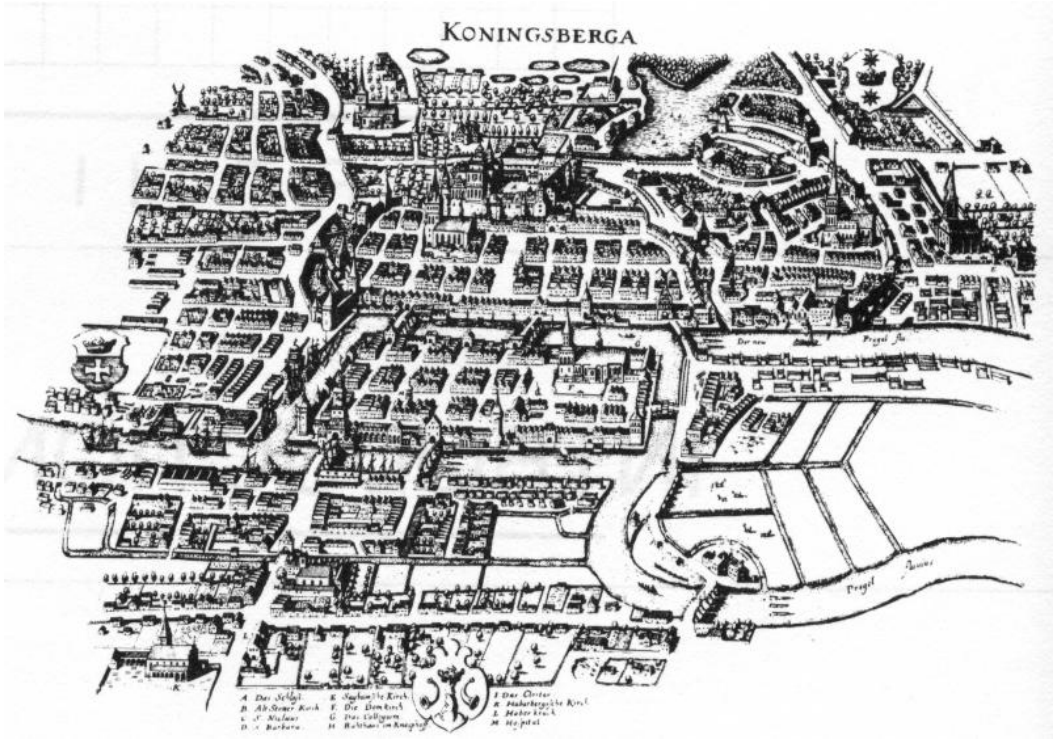

Figure 1.1.: Königsberg bridge problem: Is it possible to make a round trip, passing each bridge exactly once? [1]

works in computer science, biology, chemistry, linguistic, sociology, etc. have revealed a crucial difference from regular lattices.

Regarding average (or sometimes maximum) path lengths on such networks, we would expect to see an increase linearly with the number of nodes. Instead, we examine distances growing logarithmically with system size.

This behavior is not only an amusing effect but has far-spreading consequences [19]. Prominent examples are Internet's stability against attacks [20], disease spreading [21, 22] or path finding strategies [23, 24].

1.4. Science Collaboration Networks

In context of science, the network between scientists as nodes of the graph is of particular interest. This network belongs to the group of social ones, with humans as nodes. Unlike many other forms of social relationships, that are mostly quite difficult to capture objectively, the field of published papers is very widespread covered by the *Science Citation Index* [25] and so easily available to research.

2. Network Models

There are many types of networks competing to describe observations made in socio-physics. After discussing which measurements describe a given networks structure, we will give a short overview about what we think to be the most important ones and discuss advantages and possible disadvantages.

2.1. Measurements

2.1.1. Small World Effect

As illustrated in the introduction, we are interested in the correlation of network size and average (maximum) path lengths. We investigate if there is linear, logarithmic or other behavior. In case of a logarithmic one, the net is said to show the *Small World Effect*.

2.1.2. Clustering

In friendship network, we find friends of one person often to be friends themselves. This is true for most social networks and even other ones. Links are not spread randomly but arranged in clusters.

To describe thing mathematically, we introduce a *clustering coefficient* C_i of a node i describing the portion of m established links between all k_i next neighbors compared to the maximum possible number of $M = \binom{m}{2}$, i.e.

$$C_n = \frac{m}{M}.$$

This value is averaged over all vertices to give a clustering coefficient C for the whole network.

Typical values experienced are far above results expected for random networks [2, p. 50].

2.1.3. Scale-Free Behavior

A third observation regarding social networks is its distribution of degrees. Regarding frequency of vertices of given coordination numbers, we do not find an exponential but a power law [26].

In all, we have three possible means to classify networks. Many social graphs show small path lengths, high clustering and scale-free behavior.

2.2. Regular Lattices

The simplest form of a lattice is a symmetrical formation of nodes connected by edges between all pairs (or all pairs of adjacent) nodes as shown in figure 2.1**a,b**. Reasons to choose this linking are e.g. to simulate neighborship in a town etc.

Such network show a high degree of clustering, as wished. The average path lengths are very long, though and scale with system size. So, small world behavior cannot be found which makes the model inappropriate for our needs.

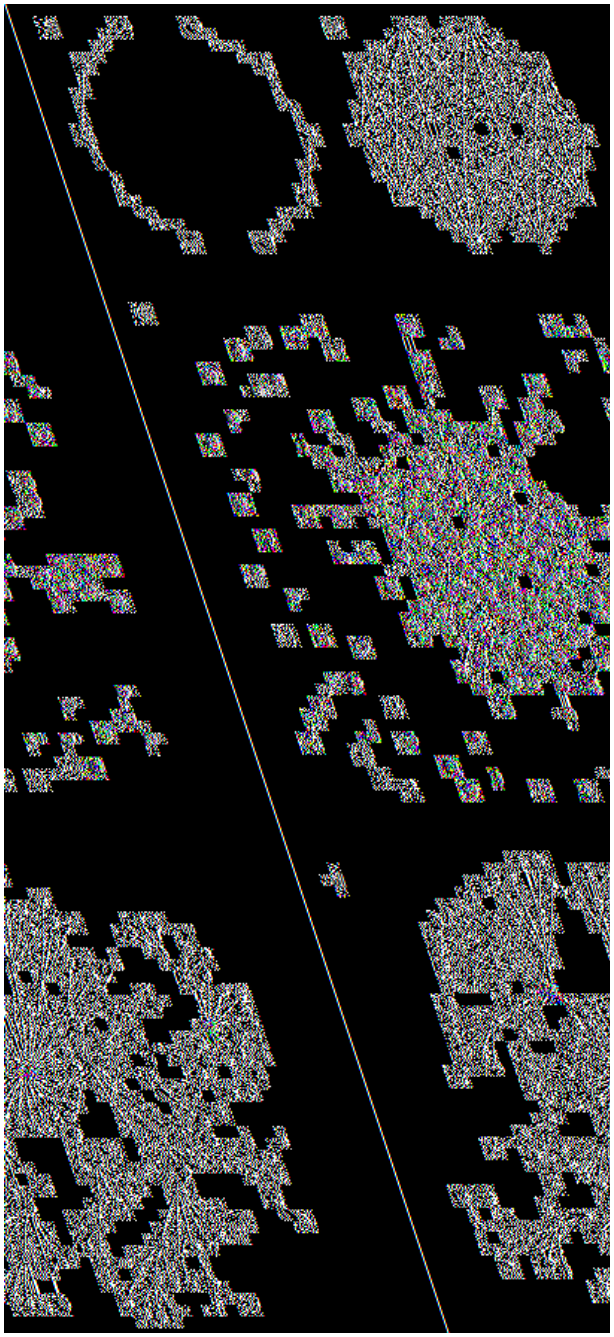

Figure 2.1.: network types: **a, b** regular lattices, **c** random network, **c** scale-free graph [14]

2.3. Erdős-Rényi Random Networks

Random networks are the extremum on the other side of the spectrum. A number of nodes is wired by pure chance, i.e. we throw the dices to select two nodes and place an edge between them.

Such graphs have been first proposed by Solomonoff and Rapoport [27] and have been extensively studied by Erdős and Rényi [28]. Actual results have been reviewed in [29]. A typical result can be seen in figure 2.1**c**.

We find that average path lengths behave logarithmically with network size. While this is as desired for small world simulation, obviously there is no clustering.

2.4. Watts-Strogatz Small-World Networks

The idea is plausible to try combining both presented models to sum up their corresponding advantages. A big step towards this goal was done by Watts and Strogatz [16].

Their model starts with a circular graph that is regularly wired (figure 2.2). Step by step, edges are chosen by chance and rewired to an arbitrary destination node. Thus, a small fraction of links are long-range ones. To illustrate, this could be habitants of a street of neighbors having relationships with far-away relatives.

At first, the model seems to fulfill our desires. It shows small average path lengths as well as high clustering. Looking at the degree distribution, i.e. the frequency of nodes of a certain degree, we find strong differences from real-world data as there is no scale-free behavior.

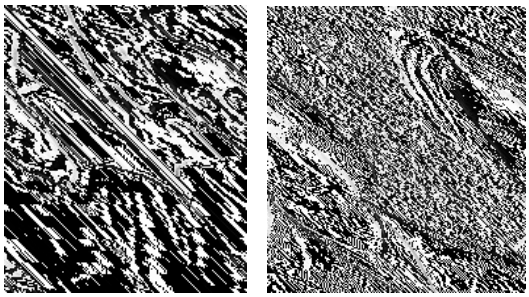

Figure 2.3.: Barabási and Albert [30]

2.5. Barabási-Albert Network Model

Barabási and Albert [30] started a new idea. Their model consists of two ingredients: *growth* and *preferential attachment*.

We start with a graph of $m_0 = 3$ vertices, each one connected to each other. Now, in each time step, we add a node that is connected to others by $m = 3$ links. The new node being one side of the links, the other one is chosen at random from the existing network. The probability of a vertex being selected is proportional to the number of links already attached to it.

To stay in the image: If you already have lots of friends, you are more likely to get new ones. "The rich get richer."[1]

These rules result in a network (figure 2.1**d**) that is capable of reproducing small-world behavior, as well as being scale-free. Research has found good collapse with

[1] "Whoever has will be given more, and he will have an abundance." [31].

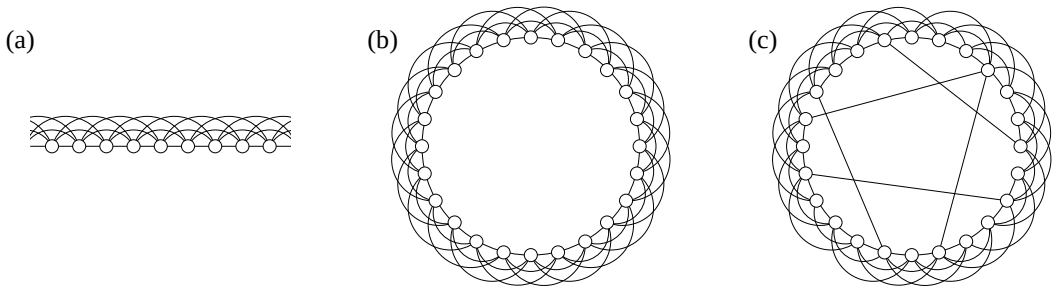

Figure 2.2.: Watts-Strogatz network model: We start with a regular lattice (a) formed to a ring (b) and re-wire a small fraction of links to random destinations (c) [18]

empirical networks, included e.g. the world wide web [6]. Good introductions can be found in [32, 33].

Clustering is present to a certain degree, but still much too small regarding experimental values.

2.6. New Approaches

Recently, new network models have been developed to cope with inconveniences encountered with present ones.

Ravasz and Barabási [35] examined networks of a self-similar structure imitating the idea of hierarchical organization in sociology. Combining high clustering and scale-free behavior, their model does not show short path lengths, though.

Klemm and Eguíluz [34][2] developed an auspicious model joining all three demands in one network. The authors present a generalization of the Barabási-Albert model, adding aging of nodes and some random behavior.

There will be further research to be done on this model to verify if it copes with reality.

An overview of all models can be found in figure 2.4.

[2] cf. also [36]

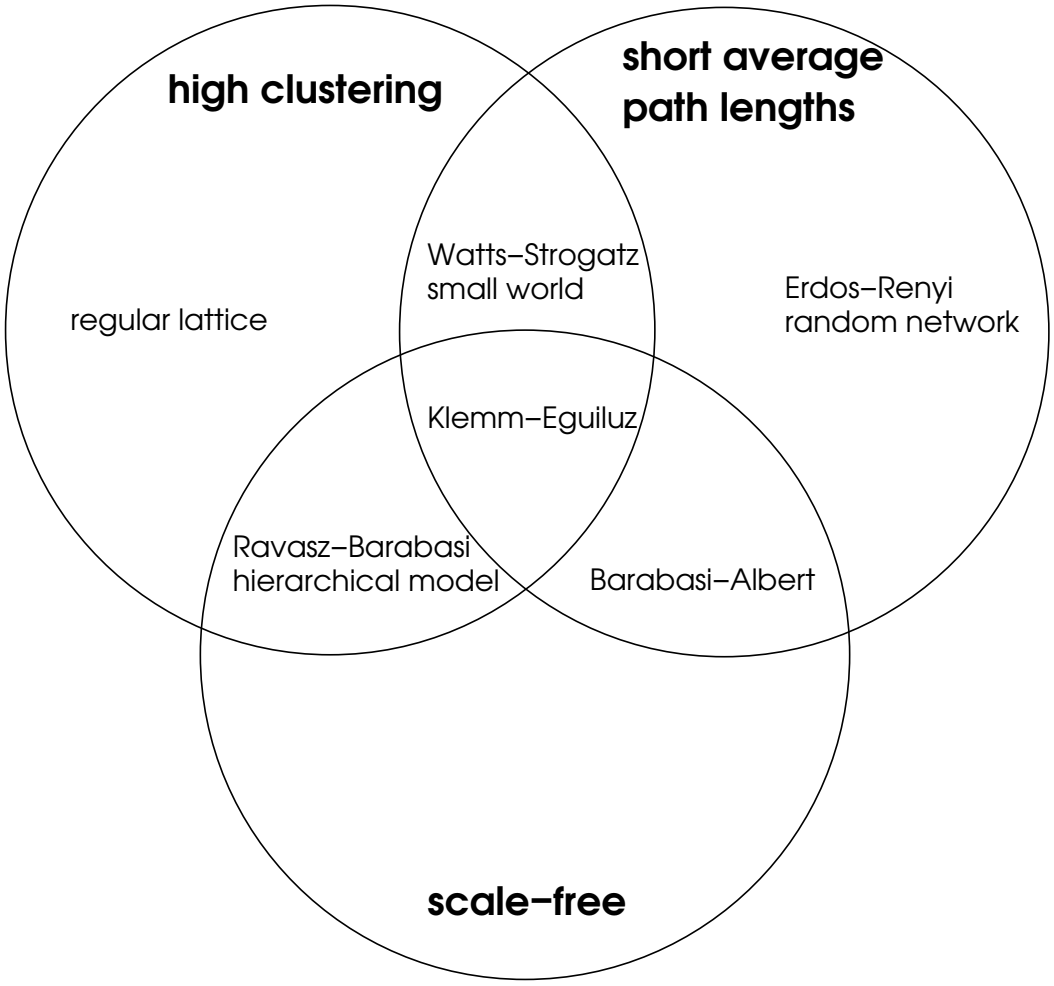

Figure 2.4.: overview over recent network models (Erdős and Rényi [28], Watts and Strogatz [16], Barabási and Albert [30], Klemm and Eguíluz [34], Ravasz and Barabási [35])

3. Empirical Collaboration Network

In context of science, the network between scientists as nodes of graph is of particular interest.

3.1. Typology

First, we want to deal with the definition of a collaboration graph. As to the nodes, we have the choice to identify each vertex either with an author or with a paper.

The second possibility is also area of research [37], but we think studying the relationship of scientists as the paper's authors offers more insight in how research works. So we will make each scientist a node of our network.

As what concerns the edges, there are basically two possible choices—both covered equally by the database used [25].

3.1.1. Citation Graph

We might chose to consider citations from one author to another as links [10], thus resulting in a directed graph.

Starting at a given paper, we can enlarge our network by following links recursively up to a certain depth, e.g. by *depth-first* or *width-first* algorithms. Each new work will cite several to many still un-included. Roughly spoken, the number of publications to include will raise exponentially with the maximum depth chosen.

Quickly, we arrive at huge amounts of data. Additionally, there is no canonical end of the hunt for new links. In the extreme case we could be caught in a giant cluster containing all or nearly all of the papers ever published. We see no possibility to narrow this down in a reasonable way without fear of introducing arbitrary boundary conditions.

3.1.2. Collaboration Graph

Second possibility to define edges of a graph is creating links by co-authorship in one or several papers [7, 8]. If n scientists publish a paper together, they are connected to each other by $\binom{n}{2}$ edges.

As an additional advantage, we have the choice to start with an arbitrary set of authors, establishing links between them by looking at all papers they are involved. This will result in a graph of limited size.

Of course, we should think carefully about reasonable selection, to avoid edge effects. We will discuss this in the next section.

3.2. Building the net

3.2.1. Proceeding

As solution, we choose the following proceeding: We start with one paper. As one part of our work will be the comparison of real world data to Barabási-Albert networks, we take the corresponding paper [30] as center of investigation.

In order to determine the set of authors we want to deal with, we select all 185 papers that cite this paper.[1] Secondly, we construct a list of unique authors from all these papers. A first approach delivers 559 scientists, whereof some turned out to be identical but appearing in particular papers with typos. We finish with a set of 555 authors to whom we attribute consecutive numbers.

The last step of the network creation process consists in establishing links between all these authors. This is done by selecting one paper after the other and introducing connections between each possible pair of this paper's authors (i.e. $\binom{n}{2}$ links for n authors).

Eventually, this gives us a graph of 555 nodes representing scientific collaboration in the area of Barabási-Albert networks.

The network size is relatively small compared to all data in the *Science Citation Index* [25] (approx. 10^7 papers). Studying properties of this subnet, we hope getting an insight to what leads to the structure observed. Verification with bigger networks remains a task for the future.

3.2.2. Visualization

To get an idea of what we are dealing about, we visualize the graph using a spring model [38, 39]. In order to give manageable results we remove a paper on the *Human Genome Project* [40] with 274 authors. Brief examination yields that this is no harm, as scientists participating in this work did not cooperate with others in our graph, and form a big cluster on their own. The result is shown in figure 3.1.

[1] We have to be careful not to mix citation data from different dates as new papers are continuously added to the database. Base of our investigation is October 21st, 2002.

3.3. Analysis

3.3.1. Authors per paper

authors	frequency
1	37
2	69
3	47
4	21
5	6
6	4
7	1

Table 3.1.: Authors per paper

First thing we are interested in is the frequency distribution of papers per author. We expect to see many papers with few authors and *vice versa* (table 3.1).

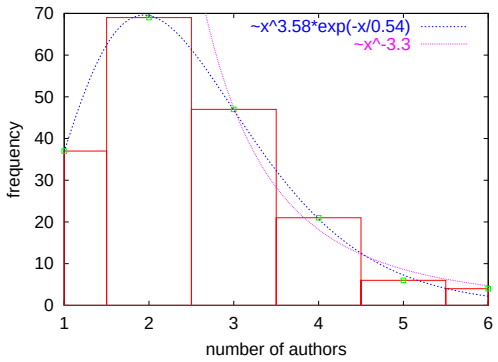

Figure 3.2.: frequency distribution of the number of authors per paper

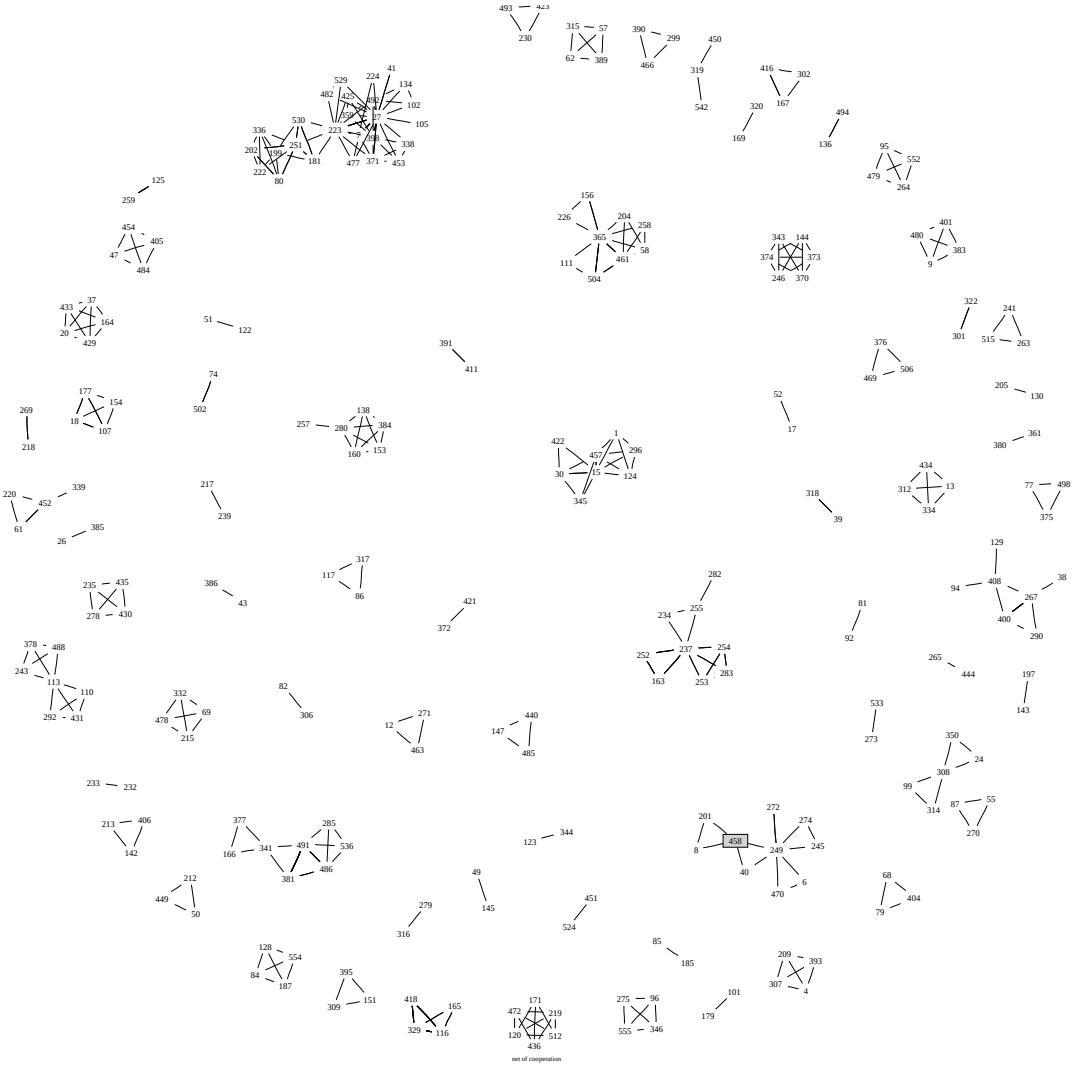

Figure 3.1.: collaboration network with 555 nodes (plotted using *GraphViz* package [38, 39])

Indeed, the considered graph (figure 3.2) shows this behavior with one remarkable exception. There are much too few papers written by only one author. This could be due to the fact that collaboration helps in science, but the more (scientific!) partners you have, the slower gets your scientific output as communication overhead increases.

In other words: establishing scientific relationships with other authors is not easy. You have to agree on the field of research, coordinate your efforts etc. Postulating that cooperation with more scientists is always favorable, we can explain the statistics by difficulty of finding new partners. This even increases corresponding to the number of co-workers you already have, as additional coordination is needed. The risk of research overlap raises, too.

The power law predicted by Lotka [11] with an exponent of -2 cannot be confirmed. This could be due to insufficient statistics for this test. Other recent studies of collaboration networks found an exponent of 2.1 or 2.4 [41] which is another indication for statistical errors predominating our results of study.

3.3.2. Connections per author

Next, we study the number of connections per author, which is the number of other scientists an author ever published papers with. This number is weighted by the number of papers, i.e. a coauthor with whom a scientist published n papers contributes n connections (table 3.2).

Again, we expect to see a frequency decrease with increasing number of connections. The experimental data (figure 3.3) shows smaller frequencies for "isolated" au-

links	weighted	unique
0	16	16
1	51	67
2	75	69
3	61	67
4	24	27
5	21	22
6	10	2
7	5	6
8	4	3
9	2	
10	3	
11	1	
13	2	1
14	1	
15	1	
17	1	1
20	2	
29	1	
273	274	274

Table 3.2.: Connections per author

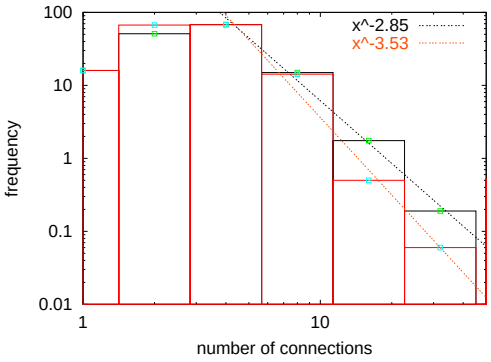

Figure 3.3.: frequency distribution of the number of connections per author (black:weighted—grey:unique)

thors that never publish with others as well as for authors with only one coauthor. This is comparable to the effect observed in the last graph. The most productive seem to be authors with two or three colleagues they are working with.

Statistical data in the area of highly connected authors shows a truncated power law. The exponent of approximately -2.85 falls well in the region bounded by analysis of other scale-free networks (`www`: around 2.3 [6, 26, 42]). The sharp or exponential cutoff at very high connection numbers has been reported for other networks, too [7, 43]. Mossa et al. [44] offer an explanation using a model with limited (local) information on the network. Surely, no scientist knows all others, so this could lead to the observed effect.

3.3.3. Double vs. unique links

We are interested how things change when we cease weighting connections by number of papers published together, i.e. we only take into account how many other *unique* scientists a researcher published papers with. Results can be found in table 3.2.

We see that despite the different numbers, results are qualitatively the same. Scientists working together with two other authors are the most productive.

Depending on whether your glass is half full or half empty there are two contrary explanations:

1. It is common practice that you name persons as authors of your work that did not contribute to it, out of a feeling of debt, may it be sponsors or others.
2. Science lives from cooperation. Working together on one subject increases scientifical output whilst reducing errors.

The author of this paper will not judge.

3.3.4. Cluster sizes

size	frequency
1	16
2	28
3	15
4	13
5	2
6	3
7	2
8	2
9	2
10	1
26	1
274	1

Table 3.3.: Cluster sizes

Our last focus is on subnets of science that exist in our net of collaboration. Authors group into several clusters by connections established between them. We investigate the frequency of clusters of a given size. Our expectation is getting a frequency increase for growing cluster sizes up to a peak, and then a decay as clusters grow very big, comparable to the statistics we saw already.

The experimental data (table 3.3, figure 3.5) shows this behavior, but with one surprise: although the most frequent cluster size is 2 due to a big number of publications

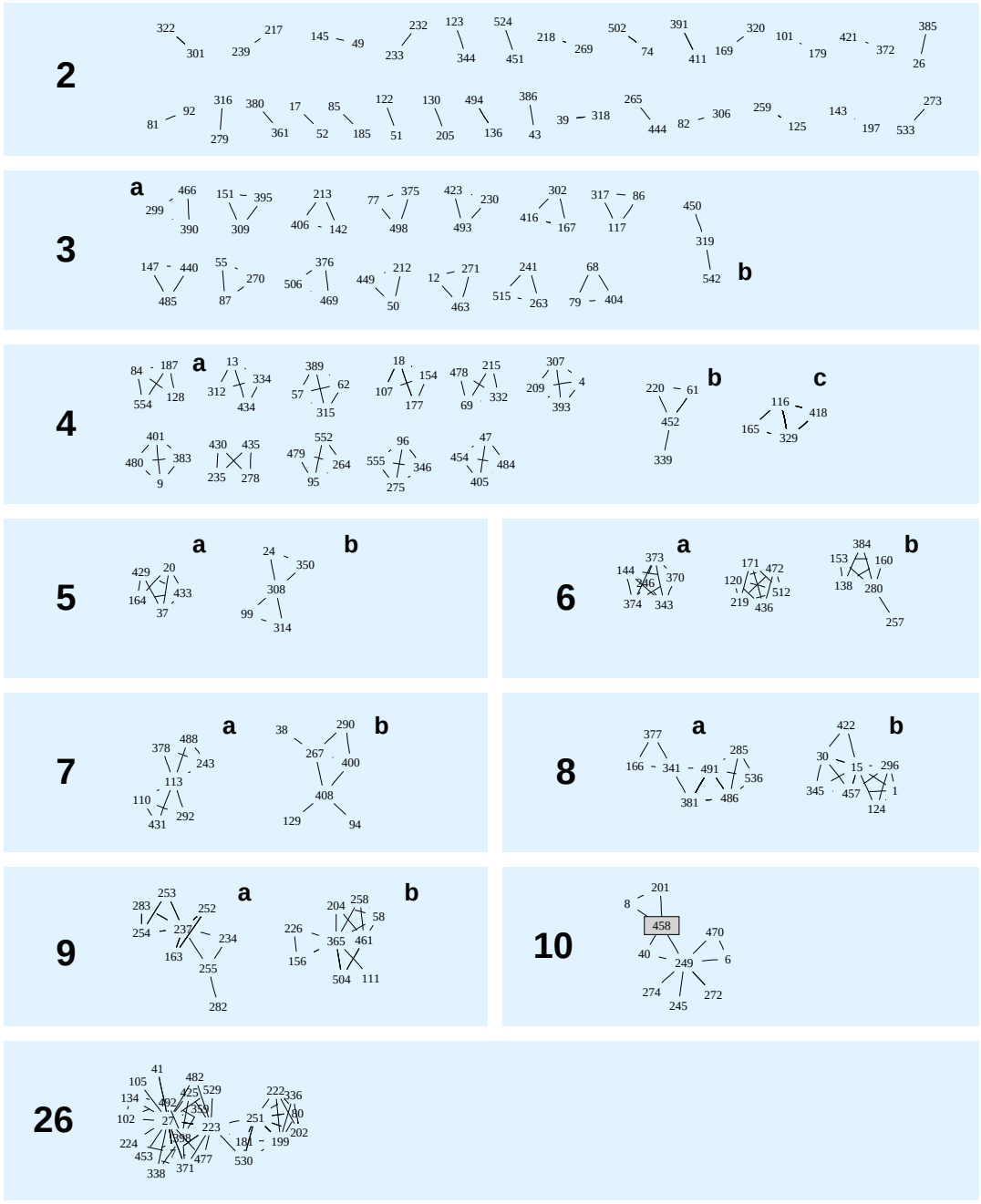

Figure 3.4.: Collaboration clusters, ordered by size of participants. The shaded box no. 458 represents my professor, D. Stauffer.

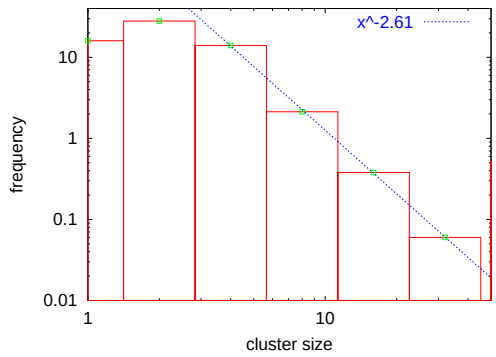

Figure 3.5.: frequency distribution of the cluster size

with two authors, most scientists maintain collaboration with three others.

The can only be due to scientists being member of research groups involved in different themes, thus connecting different clusters formed by single papers. This can directly be verified by the graphical representation (figure figure 3.4) of the clusters, ordered by size.

3.4. Comparison

We have collected some statistical figures to express the structure of the network in concern. Now we want to find out whether classical or current network models bear similar results.

3.4.1. Erdős-Rényi random graphs

Connections per author In a random graph the distribution is a binominal one, i.e. the probability of a node with k connections in a net with N nodes is

$$P(k) \propto \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

In the limit of large N this approaches a Poisson distribution around the expectation value $\langle k \rangle = pN$

$$P(k) \propto \mathrm{e}^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}.$$

This is contrary to the power-law statistic observed in the collaboration network.

Cluster sizes For random graphs, percolation theory predicts that the cluster size distribution shows an exponential decay for big cluster sizes [2, 45].

Again, the considered network show rather a power-law decay than an exponential one.

Unsurprisingly, the structure of scientific collaboration differs basically from that of a random network.

3.4.2. Watts-Strogatz small-world networks

Connections per author The degree distribution of Watts-Strogatz small-world networks is similar to that of a random graph [2]. It has a peak and decays exponentially for large connection numbers, contrary to the collaboration network.

Cluster sizes The usual case in Watts-Strogatz networks is re-wiring of only a small portion of links. Thus, the network stays well connected, mostly forming one giant cluster.

This model does not describe scientific collaboration as well.

3.4.3. Barabási-Albert networks

Connections per author Barabási-Albert networks show a vertex degree distribution as $P(k) \propto k^{-3}$ [30, 46].

In our science collaboration network we found out exponents of 2.85 resp. 3.53 which is only a slight deviation.

Cluster sizes In Barabási-Albert networks, new sites are added with links to already existing nodes. Consequently only one giant cluster forms. Obviously this differs crucially from the net of co-authorship.

4. Spin models

4.1. Leadership effect

4.1.1. Ising model

In 1925, Ising [47] published a paper on a model of spin interaction that later became very famous. The idea to this had been given to him by his teacher Lenz [48], so it is sometimes referenced as Lenz-Ising model.[1]

The idea is to consider spins (e.g. on a square lattice) and an interaction Hamiltonian

$$H = -\sum_{i \neq j} J_{i,j} S_i S_j$$

where S_i are the spins and $J_{i,j}$ is a matrix describing the interaction forces. Usually we consider the case

$$S_{i,j} = \begin{cases} J: & i,j \text{ nearest neighbors} \\ 0: & \text{else,} \end{cases}$$

i.e. only allow equal interaction between nearest neighbors ($J > 0$ for ferromagnetic behavior).

In the following chapter, we investigate how such a model behaves on our constructed collaboration network.

[1] A generalization of the Ising model is the Potts model [49, 50]. Instead of Ising spins with two possible states $+1$ and -1, Potts allows $k \geq 2$ different spin values. The Hamiltonian is

$$H = -J \sum_{i \neq j} \delta_{i,k},$$

i, j being nearest neighbors.

Applying it to the scientific network, I find results very similar to those of Ising's model.

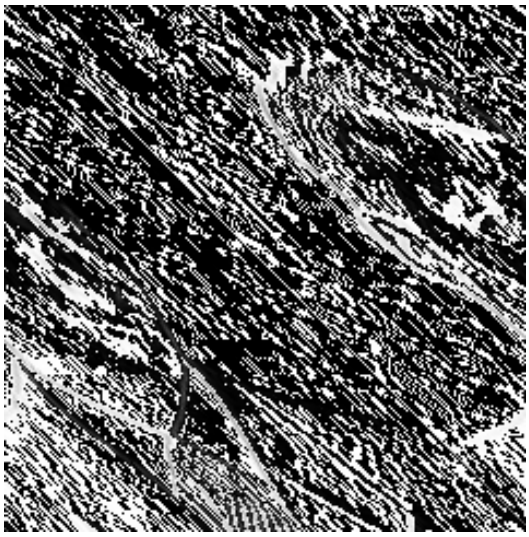

Figure 4.1.: Ising [47]

We use a *Metropolis* [51] Ising model, i.e. probabilities for a single spin flip of $p \propto \mathrm{e}^{-\Delta E/k_\mathrm{B}T}$ if $\Delta E > 0$ and 1 otherwise. To determine ΔE we sum up spins of all vertices connected to a given node. In principle, we have the choice between two proceedings:

- consider only unique links between two nodes
- count a connection several times according to the number of links, i. e. the number of papers the corresponding authors published together.

Both possibilities have been examined.[2]

[2] change the switch `NODOUBLE` in line 17 of source

4.1.2. Phase transition

The results of both experiments show qualitative similarity. We observe a rounded phase transition at about $k_B T/J = 0.8$ opposed to a value of about 2.3 on the regular square lattice.

A closer look reveals that decay of magnetization with raising temperature is exponential. This result corresponds to research on Ising models on Barabási-Albert networks by Aleksiejuk, Holyst, and Stauffer [52], who also found an exponential law.

Anyhow, the critical temperatures found by me and by Aleksiejuk et al. [52] differ by more than one order of magnitude. This can easily be explained by different coordination numbers in both networks. The collaboration graph holds a maximum of 14 neighbors of a single vertex, the graph of Aleksiejuk et al. [52] exceeds this by several orders of magnitude. This makes it far more "difficult" to break the ferromagnetic bonds, resulting in a higher critical temperature.

4.1.3. Degree distribution

Dorogovtsev et al. [53] studied random graphs with given degree distributions $P(k)$ of a vertex of degree k. They deduced an estimate for the critical temperature of an Ising model on such networks as

$$\frac{J}{k_B T_c} = \frac{1}{2} \ln \left(\frac{\langle k^2 \rangle}{\langle k^2 \rangle - 2\langle k \rangle} \right).$$

Considering the collaboration network as a random graph with given degree distribution (table 4.1), we use their formula and get $T_c/J = 2.91$, counting only unique links between scientists. Using all links, we find $T_c/J = 5.46$.

degree	several	unique
0	16	16
1	51	67
2	75	69
3	61	67
4	24	27
5	21	22
6	10	2
7	5	6
8	4	3
9	2	
10	3	
11	1	
13	2	1
14	1	
15	1	
17	1	1
20	2	
29	1	

Table 4.1.: Degree distribution

Both values are far from critical temperatures observed in simulation. This is a strong clue towards the statement that collaboration networks are crucially different from random networks, even with the same degree distribution.

4.1.4. Spin flip model

Following a suggestion of Holyst[3], we can determine the importance of most connected authors of our collaboration network by successive flipping of most connected

code in section B.1

[3] personal correspondence, cf. [52]

spins and pinning them in their new position.

In other words: After some time of equilibration, we chose the author who has most connections to others, and change his/her spin permanently to a value of -1, opposite to all others (at $T = 0$, or nearly all others else). Subsequently, we allow the system to relax some time, after which we permanently flip the second most connected spin, and so on.

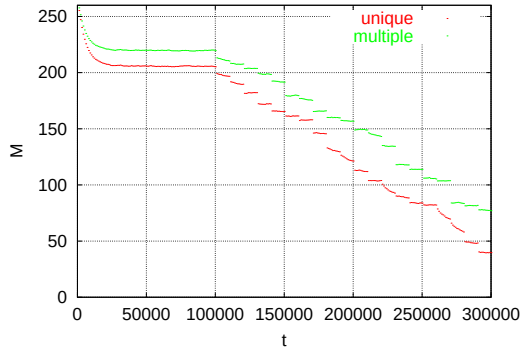

Figure 4.2.: Ising model with successive spin flips. After 10^5 steps of equilibration, we flip the most connected spin and stick it to its new value. After some relaxation of 10^4 steps, this step is repeated. Network with multiple and network with unique links used. Averaged over 1000 runs. $T = 0.2$.

Results are shown in figure 4.2. We observe two things:

1. Even after switching 20 most connected spins, the system does not flip in the opposite state with *all* spins pointing down. In simulations of Aleksiejuk et al. [52], less than 6 spins were enough to flip a whole network of $30,000$ nodes.

   This is quite obviously due to the fact that we don't have a contiguous graph, but one consisting of different clusters. A spin flip in one cluster is not able to affect spins in others.

   In pictures of spins representing opinions (yes/no, etc. [54]), this means a few authors with view differing from the broad mass of scientists are hardly capable of changing the global opinion, may they even be the most connected (known) ones.

2. Allowing multiple links in our net, we expect the magnetization to break down much faster, as a flip of a spin is able to influence others in a stronger way.

   Yet, the simulation shows contrary results. The graph containing only unique links shows a much steeper decay of magnetization (figure 4.2).

   Possible explanation is the fact, that choosing most connected spins in a network only having unique links picks authors with connections to many other authors, whereas in a network allowing multiple links, there are as well spins connected in a strong manner to only few others.

   It seems that in order to spread new opinions, it is more advantageous to have small influence on many other people than a big impact on only few ones.

4.2. Cluster limited Ising models

The network we are looking at consists of many distinct clusters of different sizes (fig-

ure 3.4). We may ask if they differ regarding their properties, or if they behave alike.

4.2.1. Proceeding

We split up the network into sub-nets, each containing one cluster, numbered sequentially, *2, 3a, 3b,... 26* (numbers and letters as in figure 3.4).

On each net, we run an Ising model for temperatures from 0.1 to 6.9 in 0.2 steps. Each of these simulations runs for 10^6 steps, with magnetization measured every 100 steps, thus resulting in 10^4 measurements per run, to give good statistics.

4.2.2. Results

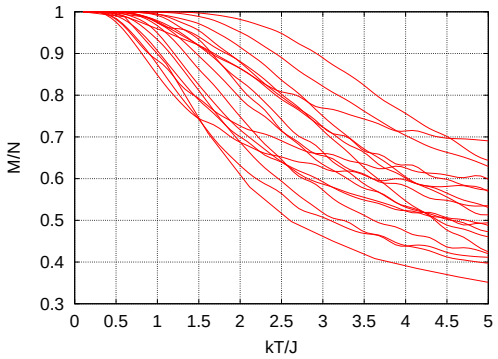

Figure 4.3.: Ising model on the different clusters of a collaboration net, averaged over 10^4 measurements per temperature and net.

Examining the results (figure 4.3), we see different curves that all decay with raising temperature, but show no apparent similarities. We wonder why the curves seemingly do not converge to zero but to finite values.

Obviously, the network is so small that macroscopic magnetization flips occur frequently, even at moderate temperatures. That means, expectation value of magnetization[4] at high temperatures is not zero, but something around one!

4.2.3. Bias adjustification

To validate this hypothesis, we simulate the networks at very high temperature ($k_\mathrm{B}T/J = 50$), in order to determine $M_\infty = M(T = \infty)$ (table 4.2).

net	M_∞
2	1.03
3a	1.71
3b	1.51
4a	1.59
4b	1.54
4c	1.55
5a	2.02
5b	1.96
6a	2.00
6b	1.94
7a	2.27
7b	2.23
8a	2.26
8b	2.29
9a	2.55
9b	2.55
10	2.54
26	4.25

Table 4.2.: Bias

[4] all over this publication we consider the (unsigned) value $|M|$ as magnetization, not M! Doing the latter leads to false results. E.g. at low temperatures, averaging M over very long times would give zero, as flips of the whole system occur (though at very low probabilities).

Using these values, we rescale our simulation results from figure 4.3 by use of the scaling function[5]

$$\frac{M}{N} \longrightarrow \frac{M - M_\infty}{N - M_\infty},$$

where N is the total number of authors in the cluster (figure 4.4).

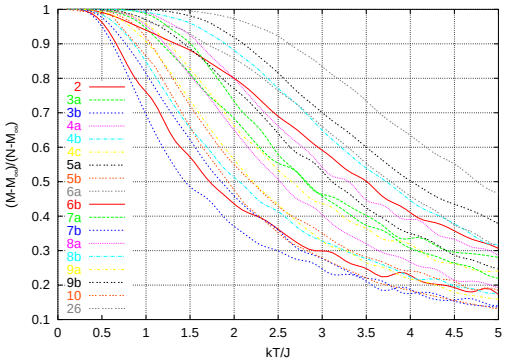

Figure 4.4.: Same data as in figure 4.3 but adjusted to a common scale from 0 to 1 by eliminating bias from random fluctuations.

We find a much cleaner image. Apart from one exception (*6b*) all curves are parallel up to the $M = 0.5$-line, and even beyond there are very few crossings.

Each different cluster can now be characterized by the temperature ϑ at which it achieves $M = 0.5$. This leads to a sort of "melting temperature" (table 4.3).

To get an idea of this temperature's meaning, we sort the cluster's graphical representations by ϑ (figure 4.5).

Is seems plausible that ϑ is a measure of coherence or connectiveness of a cluster. Single bonds lead to lower melting temperatures, fully connected subsets to higher ones.

[5] This function is a linear approximation that gives 1 for $M \to N$ and zero for $M \to M_\infty$.

net	ϑ
2	1.7
3a	2.8
3b	1.5
4a	3.3
4b	2.1
4c	2.6
5a	4.0
5b	2.2
6a	4.8
6b	3.5
7a	2.8
7b	1.8
8a	2.5
8b	3.7
9a	2.2
9b	3.1
10	1.9
26	3.8

Table 4.3.: Melting temperatures

This offers a possible explanation why *6b* shows a different behavior from all others in figure 4.4: This cluster consists of two parts. One of them is a completely connected set of five nodes, the other one a single vertex. Both are linked by one single edge. Probably, this "conflict of interest" leads to the observed anomality.

Additionally, connectedness described by ϑ seems to be crucially different from the classification given by standard clustering coefficient. E.g. the net consisting of three completely connected vertices *3a* yields a clustering coefficient of 1 but a low melting

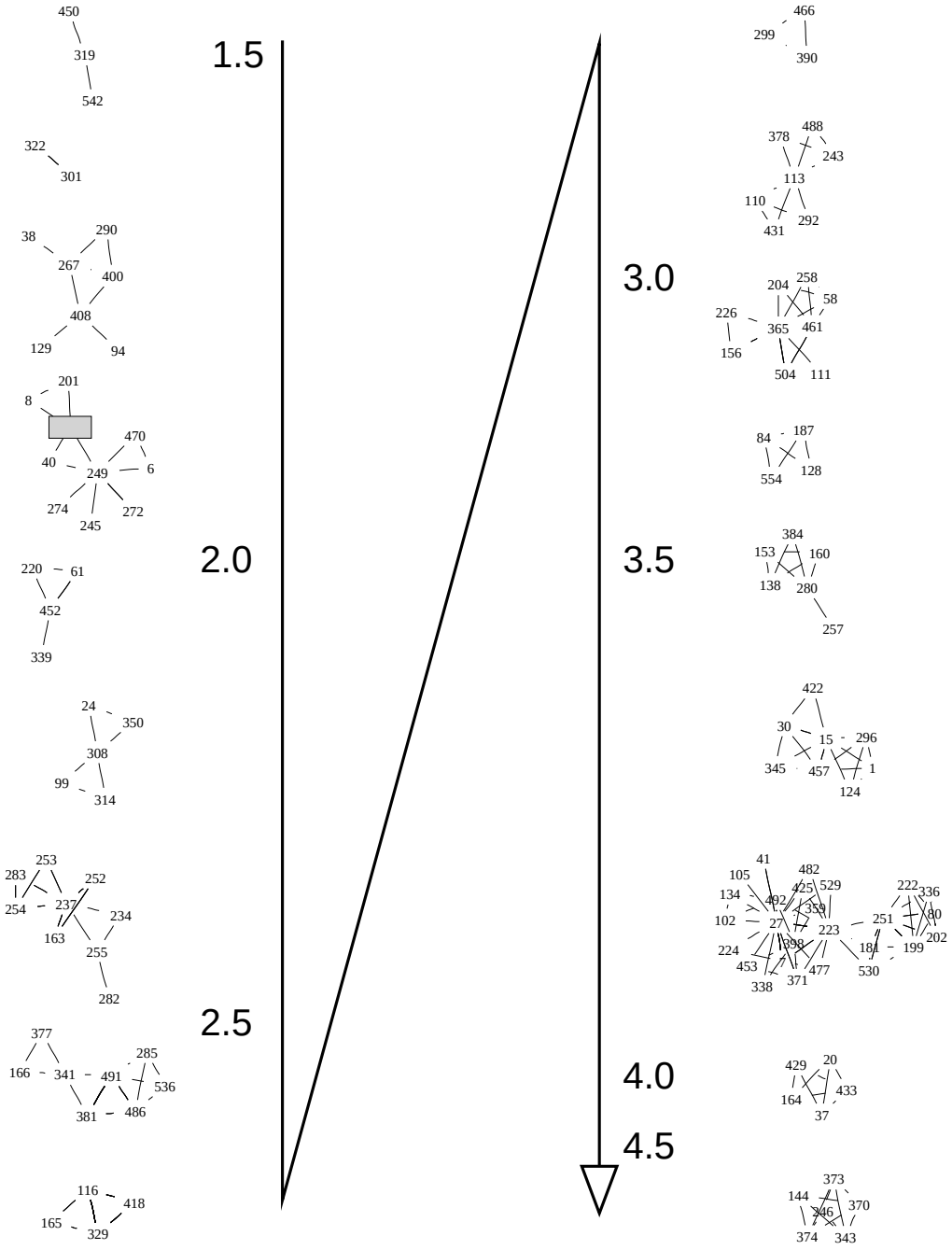

Figure 4.5.: collaboration clusters from figure 3.4, ordered by ϑ

temperature.

4.2.4. Linear Relationship

Surprisingly, we find a linear relationship between N and E/ϑ (figure 4.6). Thus, we postulate

$$\frac{E}{\vartheta} = aN - b$$

and conclude a formula for ϑ:

$$\vartheta_{\text{calc}}(E, N) = \frac{E}{aN - b}.$$

Fitting parameters to our measurements (yielding $a = 0.72, b = 0.89$), this gives good prediction of melting temperatures. Results can be seen in figure 4.7, as well as a diagram showing errors being inferior to 10% in most cases.

$$\vartheta_{\text{calc}} = \frac{E}{aN - b} \overset{N\to\infty}{\longrightarrow} \frac{\langle k \rangle}{2a}.$$

We see that, in the limit of high N, the melting temperature ϑ is proportional to the average number of edges per site $\langle k \rangle = 2E/N$, a result known from *mean field theory*.

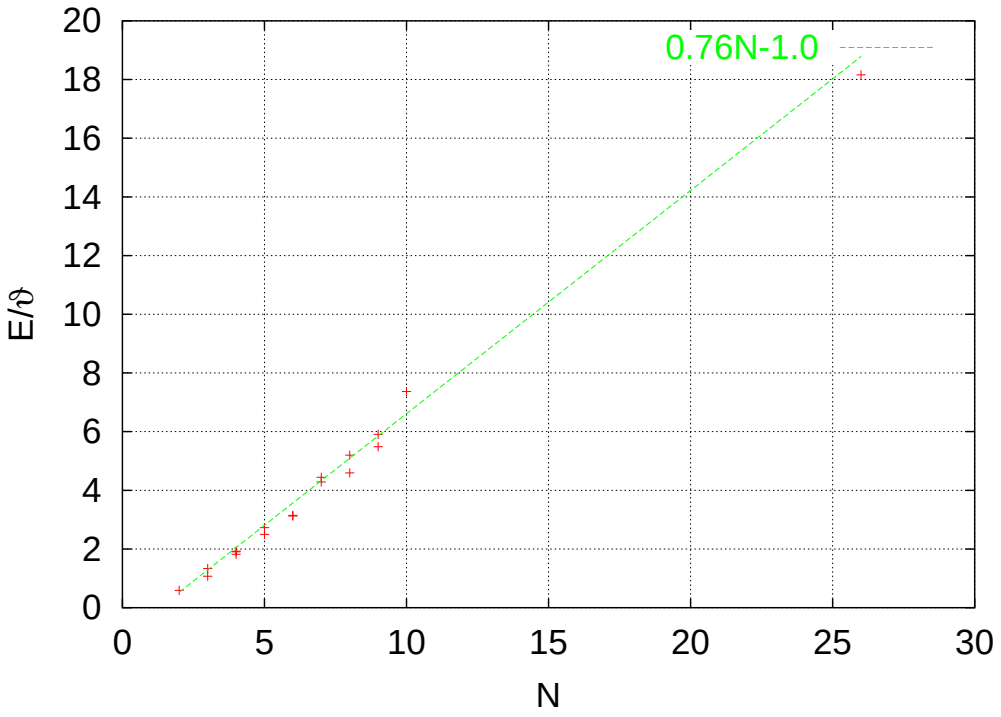

Figure 4.6.: E/ϑ vs. N shows a surprisingly linear correlation.

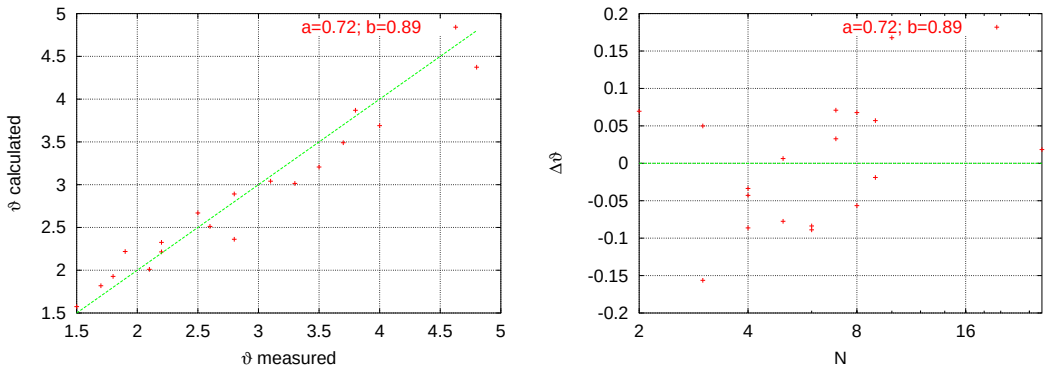

Figure 4.7.: ϑ determined by $\vartheta_{\mathrm{calc}}(E,N) = \frac{E}{aN-b}$ vs. measurements. E is the cluster's total number of edges.

5. Barabási-Albert network models

5.1. Modified Barabási-Albert model

Network model of Barabási and Albert [30] was introduced in section 2.5. We pointed out that it shows rather good fits with empirical networks, but lacks support for disjointed ones, as the algorithm only delivers one giant cluster.

Thus, to cope with networks consisting of several components, we must modify the model. We chose a very simple approach: In each step of adding nodes, we start a new cluster of $m_0 = 3$ nodes with probability p.

Vertices added in consecutive time steps can connect to any node in any component respecting the same probability rule as in the standard model.

- In the case $m = 1$, components can only grow (*isolated clusters*), whereas
- in the case $m > 1$, new nodes are able to connect two or more existing components of the network (*merging clusters*).

5.2. Simulation

In order to compare simulation results with real-world data from a scientific collaboration network [55], we let the network grow up to the size of 555 nodes. For proper statistics, this is repeated 10^4 times.

5.2.1. Isolated clusters

scale-free behavior In case of $m = 1$, i.e. considering *isolated clusters*, we can be sure to get scale-free behavior within the distinct clusters, as the probabilities for attachment of a new node to an existing one are the same as in a single Barabási-Albert network (modulo a constant factor due to a new node having the "choice" between different clusters to connect to).

However, complete network is *a priori* not necessarily scale-free, as total statistics is made up by the sum of all scale-free subnetworks or clusters. So, we have to focus later on the question, if scale-free behavior prevails.

cluster size distribution Next, we examine the number of clusters of different sizes (figure 5.1). Obviously, we find that high probability of starting a new net leads to many smaller networks, whereas low values privilege bigger networks. Yet, we make an interesting observation: low probabilities lead to a cluster-size distribution that is not monotonic any more, but favors big networks.

Looking at figure 5.1 which shows the number of points in clusters of a given size instead of the sheer cluster count, makes this more plausible.

- For $p = 0$, we will see a graph $\propto \delta(555)$, as there is only one giant cluster,

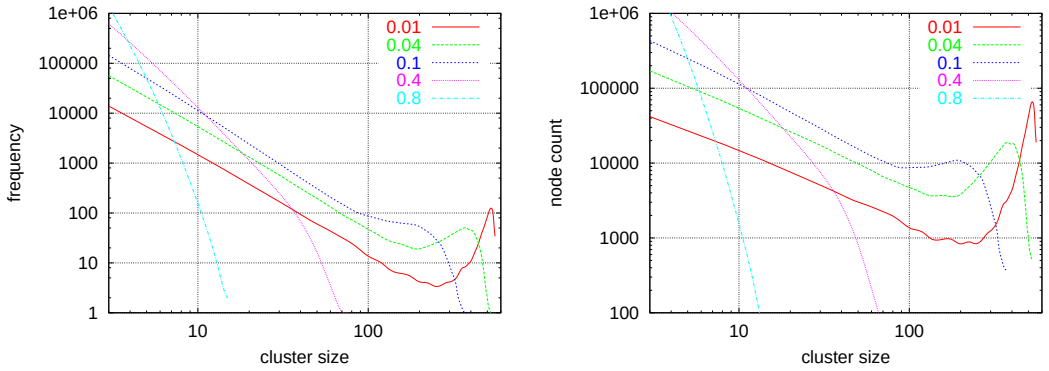

Figure 5.1.: Frequency of clusters (left) resp. number of nodes in clusters of a given size (right) vs. cluster size at different probabilities for a new net. Simulation was run 10^4 times with a network growing up to 555 nodes. The curve for $p = 0.01$ is the one with the rightmost peak; to the left follow the other p-values in descending order.

- for $p = 1$ a graph $\propto \delta(m_0 = 3)$, because there are only embryonic subnets.

- What we observe for $0 < p < 1$ is the transition between both extremes.

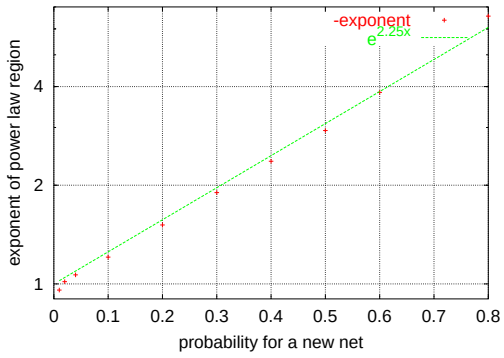

Figure 5.2.: Negative exponent of the power law part of the curves in figure 5.1 vs. probability p for a new net. The line corresponds to $\text{exponent} = -\text{e}^{2.25p}$.

power law region For all p, we start with a power law region, regarding the distribution of small and medium cluster sizes. The exponent varies with network-birth probability p. In the semi-logarithmic plot in figure 5.2 it is shown, that the exponential relation $-\text{e}^{2.25p}$ describes our data rather well. Of course, this formula cannot be true for general p as for $p \to 1$ we expect $m \to -\infty$!

Regarding empirical data from sections 3.2's collaboration graph, we find good overlap with the model (figure 5.3). Interestingly, the model is even able to explain facts formerly regarded as statistical anomalies, as the observation of a giant cluster of a size exceeding largely all others in the network (section 3.2.2).

5.2.2. Merging clusters

Now, we modify the model by examining $m > 1$. In this case, newly added vertices develop several links to existing nodes (and thus existing clusters), being able to con-

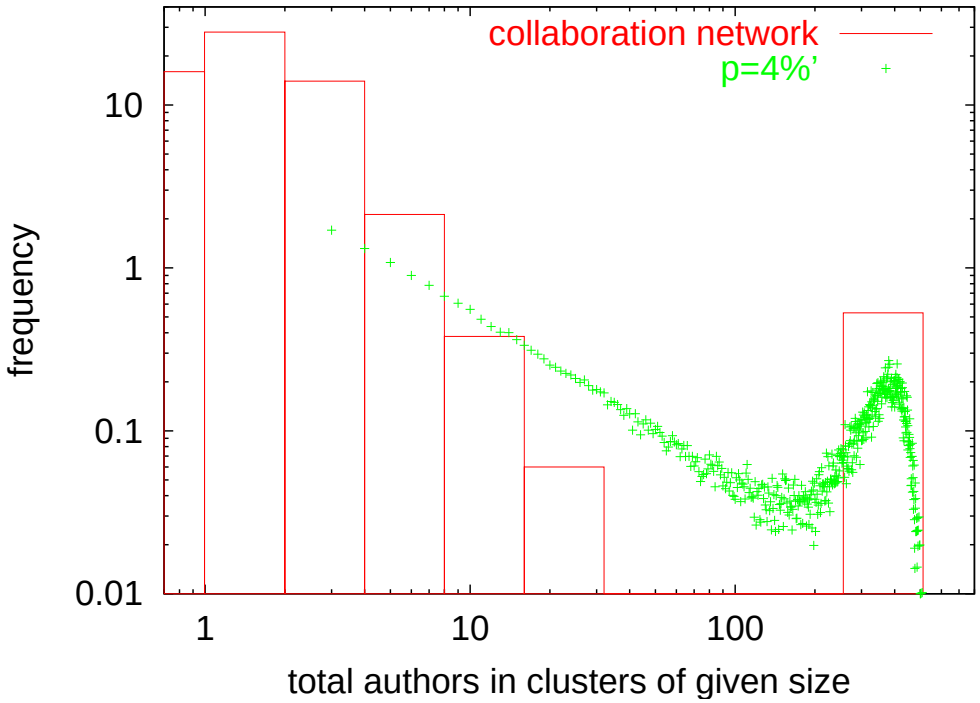

Figure 5.3.: Semi-logarithmic plot comparing the simulation with $p = 0.04$ using the *isolated clusters* model of figure 5.1 and statistical data from a science collaboration network (section 3.2).

nect hitherto separated networks. In this paper, we limit our considerations on the standard Barabási-Albert case $m = m_0 = 3$.

Using different p, we quickly recognize that low and medium probabilities make the simulations nearly always end up with a single giant cluster containing all vertices. Points of interest are higher p in the region of 60–90%.

cluster size distribution Again, we plot the total number of nodes contained in clusters of a given size (figure 5.4). For small cluster sizes, we observe an non-uniform behavior of the graph. The explanation is as follows: newly born clusters have a size of $m_0 = 3$ and thus appear very often. Also, cluster of sizes 4 or 7 are very probable, whereas a cluster of size 5 is very rare, because it can only be formed by a new cluster to which two new ones have connected without gluing it to a second cluster.

In a semi-logarithmic plot (figure 5.4), we find a parabolic dependence for high cluster sizes (i.e. a Gaussian distribution around a mean depending on p). Appearently, the *merging clusters* cannot cope with reality.

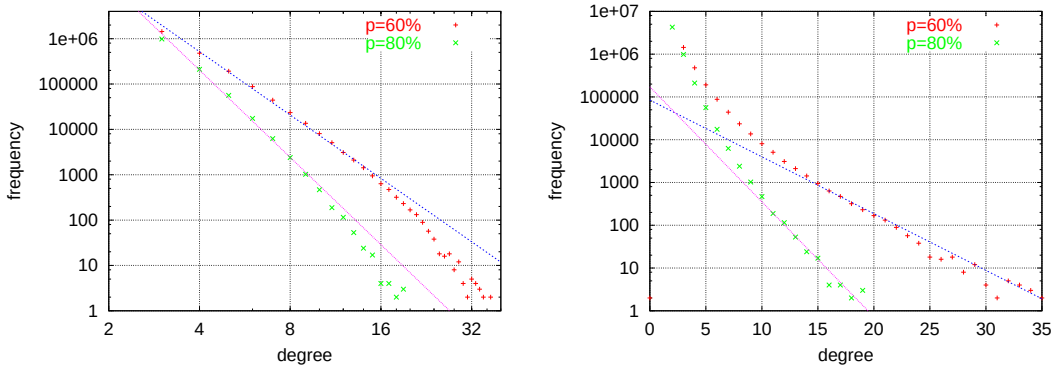

Figure 5.5.: Frequency of nodes with a certain degree. Simulation was run 10^4 times with networks growing up to 555 nodes. Left plot is linear, right plot semi-logarithmic. $M = 1$ (isolated clusters).

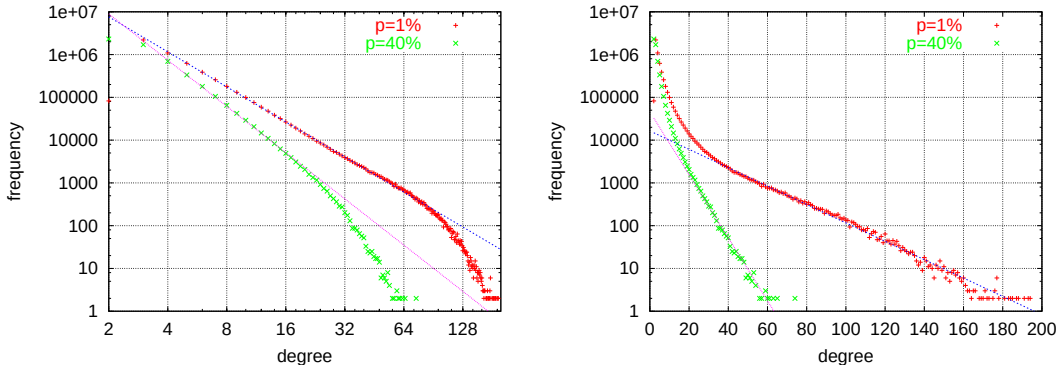

Figure 5.6.: Same plots as figure 5.5 using $M = 3$ (merging clusters).

5.2.3. scale-free behavior

In figure 5.5 we can see that there is no pure scale-free behavior. There seems to be power-law behavior for small degrees and an exponential cutoff (figure 5.5) at higher values. Similar results have been observed by Newman [7] for collaboration networks.

One could argue that this effect is due to the fact that we do not plot the degree distribution for single clusters, but for the whole set of them. This demur only counts at first sight, though. At $p = 80\%$ we have several small clusters but virtually only one giant cluster dominating the degree distribution for high degrees. So, the fact of averaging of many different sized clusters should manifest mainly in the area of small degrees opposite to our observations.

Mossa et al. [44] offer a possible explanation for the exponential cutoff encountered. They use a model which attributes to each node only a restricted knowledge on the network, i.e. the vertex is not able

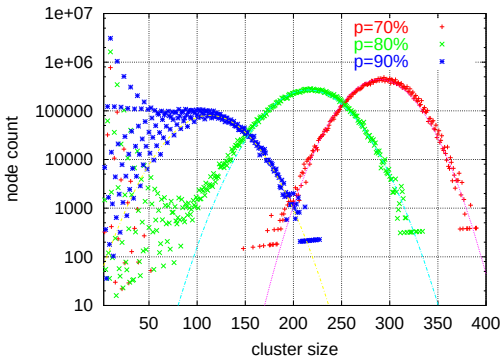

Figure 5.4.: Number of nodes in clusters of a given size vs. cluster size at different probabilities for a new net. Simulation was run 10^5 times with a network growing up to 555 nodes.

to consider the whole graph's structure, but only a subset according to its limited view.

M	p	$\ln(N_0)$	k	σ.
1	1%	16.8	2.27	60.
1	40%	17.1	2.45	9.2
3	60%	18.1	3.1	5.1
3	80%	20.1	4.8	3.5

Table 5.1.: Coefficients for figure 5.7

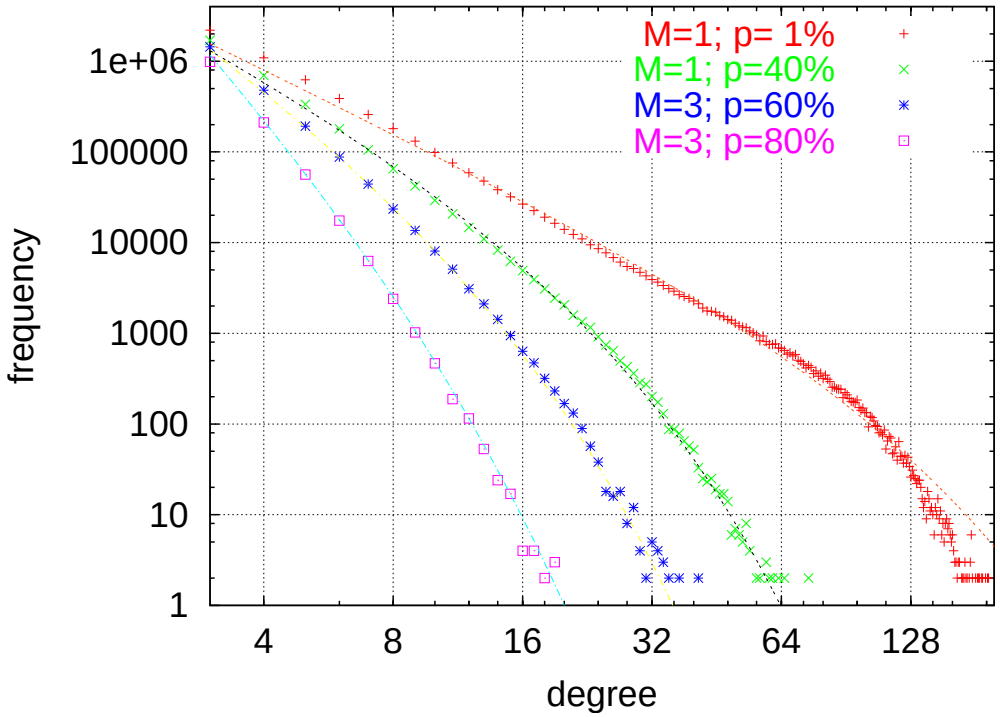

Figure 5.7.: Plots of figure 5.5 and figure 5.6 show good fit to exponentially truncated power laws $y = N_0 x^{-k} \mathrm{e}^{-x/\sigma}$.

6. Conclusion

Dealing with real-world networks, scientists found three properties predominating:

- short average path lengths (*Small World Effect*),
- scale-free behavior,
- high clustering.

Different models were developed to cope with this challenge, each having different advantages and disadvantages. The model of Barabási and Albert [30] is a promising one, but lacks support for discontiguous networks.

We constructed a network of co-authorship with 555 authors. Only scientists were chosen that cite a specific paper [30]. The resulting net shows scale-free characteristics but differs substantially from accepted computer models' results.

Simulating Ising models on the network reveals strong robustness against disturbances (spin flip experiment/leadership effect) and shows coherence with *mean field theory*: We find the critical temperature of subnets of our graph being proportional to the average number of edges per site, in the limit of a high node count.

In order to overcome the mentioned disadvantages of the Barabási-Albert model, we developed a modified version, allowing formation of multiple clusters. We saw a strong dependence of a node's edges count on the network structure, separating two cases: *isolated clusters* and *merging clusters*.

Only the first case leads to results fitting reality. Comparison with statistics from our collaboration net shows similar behavior and is even able to explain facts at first regarded as statistical anomalies as the observation of a giant cluster of a size exceeding largely all others in the network. Even exponential cutoff of nodes with high degrees, as encountered empirically, is reproduced.

Re-evaluating the model with a higher number of authors would lead to better statistics and greater reliability.

A. Acknowledgements

I would like to thank D. Stauffer[1] for giving me the idea of the subject and supporting me by comments and discussions during my research.

Thanks to M. Abd-Elmeguid[2] for co-judging this work.

My thanks for writing excellent software goes to the authors of

- LaTeX, BibTeX, pdfLaTeX, dvips
- LaTeXpackages: KOMAScript, natbib, custom-bib, graphicx, listings, units, hyperref, hypernat, colortbl, color
- SciTE
- gnuplot
- Ruby, C++, Perl, bash
- dotty, neato[39]

Thanks to A. Sindermann for supporting my work by addressing computer network problems.

Special thanks to K. Godthardt and my family for supporting me.

[1] Institute of Theoretical Physics, University of Cologne

[2] II. Institute of Experimental Physics, University of Cologne

B. Source code

B.1. Ising model

This C++ program simulates an Ising model on a given graph. In section 4.1.1 it was used on our collaboration network.

(left out for technical reasons; can be found in postscript version of this document)

B.2. Spin flip model

This C++ program simulates an Ising model on a given graph. In regular time intervals, the most connected spins (hard coded) are pinned to an *up* position. In section 4.1.4 this was used on our collaboration network.

(left out for technical reasons; can be found in postscript version of this document)

B.3. Modified Barabási-Albert model

This Ruby program creates modified Barabási-Albert models with a given final size. In section 5 this was used on our collaboration network.

```
#!/home/fxp/bin/ruby -w

$P = 0.8
$RUNS = 1E0.to_i
$M = 1

$cluster_initial_size = 3

$MAXINT = 2147483647*2+1

class Random
  @@ibm = 1
  def rnd(max=nil)
    @@ibm *= 65539 # 16807 # 65539
    @@ibm &= $MAXINT
    max ? @@ibm*max/$MAXINT : @@ibm
  end
end

class Vertex
  attr_reader :connections
  attr_accessor :nr
  @@total_vertices = 0
  def initialize
    @connections = []
    @@total_vertices += 1
    @nr = @@total_vertices
  end
  def inspect
    "I'm node nr. #{@nr} connected to #{@connections.collect{|c| c.nr}.
        join(",")}."
  end
  def add_link(partner)
    @connections <<= partner
  end
  def connect(partner)
    add_link(partner)
    partner.add_link(self)
  end
end

class BA_net
  def initialize(prob_new=0)
    @p_new = (prob_new * $MAXINT).to_i
    @nodes = []
    @kertesz = []
    @r = Random.new
    100.times { @r.rnd }
    start_new_net
```

```
  end
  def size
    @nodes.length
  end
  def start_new_net
    max = @nodes.length
    $cluster_initial_size.times { @nodes << Vertex.new }
    $cluster_initial_size.times { |i|
      orig = i + max
      dest = ((i+1) % $cluster_initial_size) + max
      @nodes[orig].connect(@nodes[dest])
      @kertesz << orig << dest
    }
  end
  def inspect
    @nodes.collect{|node| node.inspect}.join("\n") + "\n#{@kertesz.
        inspect}"
  end
  def add_node
    if @r.rnd < @p_new
      start_new_net
    else
      @nodes << new_node = Vertex.new
      l = @kertesz.length
      $M.times {
        dest = @kertesz[@r.rnd(l-1)]
        new_node.connect(@nodes[dest])
        @kertesz << @nodes.length-1 << dest
      }
    end
  end
  def check_subtree(node, subtree)
    return if !node.nr
    subtree << node.nr
    node.nr = nil
    node.connections.each { |n| check_subtree(n, subtree) }
  end
  def analysis
    @nodes.each { |node|
      next if !node.nr
      tree = []
      check_subtree(node, tree)
      # p tree.sort
      # [1] [2,3] [4,5,6,7] [8...]...
      bucket = tree.length
      if !$statistik[bucket]
        $statistik[bucket] = 1
      else
        $statistik[bucket] += 1
      end
    }
  end
end
```

```
$statistik = {}
$RUNS.times { |i|
  $stderr.print "#{i}..." if i%100==0
  mynet = BA_net.new($P)
  begin
    mynet.add_node
  end while mynet.size < 555
  # p mynet
  mynet.analysis
}
$statistik.keys.sort.each { |l| printf "%3i %6.4f\n",l,$statistik[l
    ]} # .to_f/$RUNS }
```

C. Figures

The figures found in this publication were made by myself, except the following:

- figure 1.1 was found on http://-www.math.colostate.edu/~betten/-courses/M501/combi.html.
- figure 2.1 was found in [14].
- figure 2.3 was found on http://-www.science.nd.edu/physics/-Faculty/barabasi.html.
- figure 2.3 was found on http://-www.phys.psu.edu/~ralbert/.
- figure 2.2 was found in [18].
- figure 4.1 was found on http://-www.physik.tu-dresden.de/itp/-members/kobe/isingphbl/.

Bibliography

[1] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commetarii Academiae Scientiarum Imperialis Petropolitanae* **8** (1736).

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

[3] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Adv. Phys.* **51**, 1079–1187 (2002).

[4] A.-L. Barabási. *Linked. The New Science of Networks* (Perseus, Cambridge, Massachusetts, 2002).

[5] K. Rieger. Sind Sie mit Marlon Brando befreundet? *Die Zeit* **44**, BL21 (1999).

[6] 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).

[7] M. E. J. Newman. Scientific collaboration networks. Network construction and fundamental results. *Phys. Rev. E* **64**, 01631 (2001).

[8] M. E. J. Newman. Scientific collaboration networks. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 01632 (2001).

[9] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E* **66**, 035103 (2002).

[10] S. Redner. How popular is your paper? an empirical study of the citation distribution. *Europ. Phys. J. B* **4**, 131–134 (1998).

[11] A. J. Lotka. The frequency distribution of scientific productivity. *J. Wash. Acad. Sc.* **16**, 317–323 (1926).

[12] F. Liljeros, C. R. Edling, et al. The web of human sexual contacts. *Nature* **411**, 907 (2001).

[13] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[14] S. H. Strogatz. Exploring complex networks. *Nature* **410**, 268–276 (2001).

[15] S. Milgram. The small world problem. *Psych. Today* **61** (1967).

[16] D. J. Watts and S. H. Strogatz. Small world. *Nature* **393**, 440–442 (1998).

[17] D. Watts. *Small worlds: the dynamics of networks between order and randomness* (Princeton University Press, Princeton, New Jersey, 1999).

[18] M. E. J. Newman. Models of the small world—a review (2000). ArXiv:cond-mat/0001118.

[19] M. Gladwell. *The Tipping Point. How little things can make a big difference.* (Little Brown and Company, Boston, Massachusetts, 2000).

[20] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.* **85**, 21, 4626–4628 (2000).

[21] A. L. Lloyd and R. M. May. How viruses spread among computers and people. *Science* **292**, 1316–1317 (2001).

[22] M. E. J. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E* **66**, 035101 (2002).

[23] B. J. Kim, C. N. Yoon, S. K. Han, and H. Jeong. Path finding strategies in scale-free networks. *Phys. Rev. E* **65**, 027103 (2002).

[24] J. M. Kleinberg. Navigation in a small world. *Nature* **406**, 845 (2000).

[25] Isi web of science (2002). http://www.isinet.com/-isi/products/citation/wos/index.html.

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

[27] R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bull. Math. Biophys.* **13**, 107–227 (1951).

[28] P. Erdős and A. Rényi. On random graphs. *Pub. Mathem.* **6**, 290–297 (1959).

[29] M. E. J. Newman. Random graphs as models of networks (2002). ArXiv:cond-mat/0202208.

[30] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[31] St. Matthew. New Testament. In *The Bible*, chap. 12, 13 (God, ca. 10AD).

[32] D. Cohen. All the world's a net. *New Scientist* **174**, 24–29 (2002).

[33] A.-L. Barabási and E. Bonabeau. Scale-free networks. *Sci. Amer.* **3**, 50–59 (2003).

[34] K. Klemm and V. M. Eguíluz. Growing scale-free networks with small-world behavior. *Phys. Rev. E* **65**, 057102 (2002).

[35] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E* **67**, 026112 (2003).

[36] K. Klemm and V. M. Eguíluz. Highly clustered scale-free networks. *Phys. Rev. E* **65**, 036123 (2002).

[37] A. F. J. van Raan. Fractal dimension of cocitations. *Nature* **347**, 626 (1990).

[38] AT&T. Graphviz—open source graph drawing software (2002). http://www.research.att.com/-sw/tools/graphviz/.

[39] S. C. North. *Drawing graphs with neato*. AT&T (2002).

[40] J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

[41] A.-L. Barabási, H. Jeong, Z. Nèda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A* **311**, 590–614 (2002).

[42] A.-L. Barabási. The physics of the web (2001). http://www.physicsweb.org/article/-world/14/7/09.

[43] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proc. Nat. Acad. Sci. USA* **97**, 21, 509–512 (2000).

[44] S. Mossa, M. Barthélemy, H. E. Stanley, and L. A. N. Amaral. Truncation of power law behaviour in "scale-free" network models due to information filtering. *Phys. Rev. Lett.* **89**, 208701 (2002).

[45] D. Stauffer and A. Aharony. *Introduction to Percolation Theory* (Taylor and Francis, London, 1994), second ed.

[46] A.-L. Barabási, R. Albert, and H. Jeong. Mean field theory for scale-free random networks. *Physica A* **272**, 173–187 (1999).

[47] E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschr. f. Phys.* **31**, 253–258 (1925).

[48] W. Lenz. Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. *Phys. Zeitschr.* **21**, 613–615 (1920).

[49] R. B. Potts. Some generalized order-disorder transformations. *Proc. Cambridge Phil. Soc.* **48**, 106 (1952).

[50] F. Y. Wu. The Potts model. *Rev. Mod. Phys.* **54**, 235–268 (1982).

[51] N. Metropolis, A. W. Rosenbluth, and A. H. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 6, 1087–1092 (1953).

[52] A. Aleksiejuk, J. A. Hołyst, and D. Stauffer. Ferromagnetic phase transition in Barabási-Albert networks. *Physica A* **310**, 260–266 (2002).

[53] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Ising model on networks with an arbitrary distribution of connections. *Phys. Rev. E* **66**, 016104 (2002).

[54] K. Kacperski and J. A. Hołyst. Phase transitions as a persistent feature of groups with leaders in models of opinion formation. *Physica A* **287**, 631–643 (2000).

[55] F. Pütsch. Analysis and modeling of science collaboration networks. *Adv. Compl. Sys.* (2003). Submitted on May 6th.