On The Observational Robustness of Genetic Networks

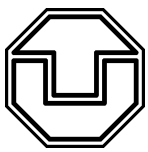
Diplomarbeit
zur Erlangung des akademischen Grades
Diplom-Physiker

vorgelegt von

Mathias Kuhnt

geboren in Zerbst, am 11. Juni 1977

Institut für Theoretische Physik
Fachrichtung Physik
Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden



Februar 2006

1. Gutachter:	Prof. Dr. Sigismund Kobe (Technische Universität Dresden)
2. Gutachter:	Prof. Dr. Martin Greiner (Justus Liebig Universität, Gießen und Siemens AG, München)

Die Arbeit wurde am 28. Februar 2006 eingereicht.

Abstract

Several models are proposed in the literature to simulate the evolution of the network structure of protein interaction networks. This work aims to evaluate these models by looking beyond the degree distribution. Therfore, the gene-duplication and mutation models are compared with available protein interaction data of *Saccharomyces cerevisiae* (Bakers yeast) by taking into account that the observed structure of protein interaction networks is corrupted by many false positive and false negative links. This observational incompleteness is abstracted and modeled by random link removal, addition and exchange as well as random subnetwork sampling and a specific, experimentally motivated (spoke) link rearrangement. The impact of these error algorithms on the structural properties of gene-duplication and mutation network models is studied. Whereas the network properties appear to be robust against the first four types of random perturbations, the spoke error algorithm changes the degree distribution, degree correlation, clustering coefficient and motif structure of the gene-duplication and mutation models largely and brings them closer to the yeast observations.

Zusammenfassung

Verschiedene Modelle zur Simulation der strukturellen Eigenschaften von Proteininteraktionsnetzwerken wurden in der Literatur vorgeschlagen. Diese Arbeit hat zum Ziel, diese Modelle zu bewerten, wobei über die bloße Wahrscheinlichkeitsverteilung von Knoten mit bestimmter Anzahl von Nachbarn hinausgegangen wird. Dafür werden die Genduplizerungs- und Mutationsmodelle mit verfügbaren Proteininteraktionsdaten der *Saccharomyces cerevisiae* (Bäckerhefe) verglichen, wobei beachtet wird, dass die beobachtete Struktur von Proteininteraktionsnetzwerken durch eine Vielzahl von fälschlich angenommenen oder nicht angenommenen Links verschlechtert ist. Diese Unvollständigkeit der beobachteten Daten wird sowohl durch zufälliges Löschen, Hinzufügen und Austauschen von Links als auch durch eine zufällige Auswahl von Telnetzwerken und durch ein spezifisches, biologisch motiviertes (Spoke) Neuordnen von Links abstrahiert und modelliert. Der Einfluss dieser Fehleralgorithmen auf die strukturellen Eigenschaften von Genduplizierungs- und Mutationsmodellen wird untersucht. Bezüglich der ersten vier Fehleralgorithmen erscheinen die Netzwerkeigenschaften robust gegen zufällige Störungen. Im Gegensatz dazu ändert der Spoke Fehleralgorithmus die Wahrscheinlichkeitsverteilung von Knoten mit bestimmter Anzahl von Nachbarn, die Korrelation der Nachbaranzahl an benachbarten Knoten, den Clustercoeffizienten und die Motifstrukturen dieser Genduplizierungs- und Mutationsmodelle weitgehend und bringt sie in bessere Übereinstimmung mit dem beobachteten Hefenetzwerk.

Contents

1 Introduction

The network science became important in the 1930's with the investigation of social systems [1, 2, 3]. It bases on the simplification of complex systems towards a picture of reality where only nodes and links exist to describe actors like people, computers or proteins and their interactions. This is a very helpful way of information reduction and opens a gate towards a different look at a world of complex systems.

Even if network research appears to be not as young anymore, many networks and many applications of network research are still waiting to be investigated. Research has started with friendship networks [1, 2], and today a large amount of networks is identified like the structure of the world wide web as well as metabolic-, citation-, railway-, wireless communication- and of course protein interaction networks, to name a few examples from sociology, engineering and biology [2]. All these examples have paved the road for a relatively young branch of physics, the Statistical Physics of complex networks [2, 4, 5, 6].

For the study of the functions of complex networks, it is essential to study their topological features [2, 4, 5, 7, 8]: Is the network hierarchical or egalitarian? Is there redundancy? Are the friends of my friends also friends? These topologies are important if questions are asked about the dynamical properties of a network: How fast is information spread in the network? How robust is the network against perturbations? That is, where statistical physics comes in.

The next challenge is to model specific networks, to ask the question whether complex systems can be reproduced at least in their basic properties with the actual knowledge. In addition, it points out, where the limits of our knowledge are. However, when it comes to evaluate the quality of the description of real networks by these models, it has to be taken into account that real networks also hide their structures. Not only when investigating sexual relationships, network scientists deal with a huge uncertainty of their data. Furthermore, actual examined networks are huge. This makes the first problem even more severe. Hence, it is not possible anymore just to have a look at a network and to decide on its features. But this makes network research even more fascinating, and with their larger size and hence their better statistical basis, network properties can be compared between completely different fields with larger confidence.

In this work, the focus will be on biological, namely protein interaction networks in the *Saccharomyces cerevisiae* (Bakers yeast) cell. The cell consists, amongst others, of thousands of different types of proteins. These proteins carry out most biological functions by binding to each other. This binding or interaction occurs between proteins in very different strengths and for very different periods of time. By considering all

proteins as nodes and their interactions as links, a large network arises, see Fig. 1.1.

Protein interaction networks include all of the mentioned challenges. They are huge, consisting of about 5 000 nodes and 30 000 links between them, and they manage to escape the biologists exploration rather effectively. Although estimates say that the total number of links in these networks is about 30 000, only 15 000 are known today. Additionally one half of the known links is wrong [9, 10, 11]. Another challenge is that only little is known about the processes that take place within the cell through the interaction of all these proteins. Even though few functions of special proteins are known, protein interaction network models are used to get an insight on the function of the cell from the network topology of its constituents. Furthermore, the identification of substructures in the network topology bridges the knowledge of local functions and the investigation of the entire structure.

In recent studies, several ways are proposed to model protein interaction networks [12, 13, 14, 15, 16, 17]. They simulate their evolution, starting from only a few proteins up to the about 5 000 known today. In the models discussed here [12, 13, 15], only two basic mechanisms are included: The duplication and mutation of genetic information within an organism. Nevertheless, these models can describe basic topological features very well.

But it must be asked, in how far any topological properties, found in real protein interaction networks, can be trusted if they are compared with the proposed models. When it came to evaluate protein interaction models, model networks were compared to available interaction data. As depicted in Fig. 1.2 this disregards the influence of false links in protein interaction data. This work aims to find a model counterpart of interaction data by the application of error algorithms to gene-duplication and mutation models.

Besides that, the changes in the network properties after the application of certain error algorithms help to decide how robust found properties of real protein interaction networks are against perturbations that result from a large uncertainty of biological mapping methods. The influence of errors on network topologies has been studied in several publications. The robustness of the network functionality has been discussed in a general context by analyzing the overall connectivity of the network [19, 20, 21] after the random or directed removal of links. In [22], the degree distribution and further observables have been investigated also after link exchange and addition. Furthermore, several very general sampling models have been proposed [23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33] that mainly base on random selection of links, random walk and spanning trees. For this work, random link removal, addition and exchange are applied. Moreover, random walk and spanning tree algorithms are introduced that incorporate some aspects of experimental biases. Finally, a dedicated algorithm is presented that has its focus on very specific errors that are made during the mapping of protein interactions.

The next chapter is dedicated to a brief introduction of the Statistical Physics of

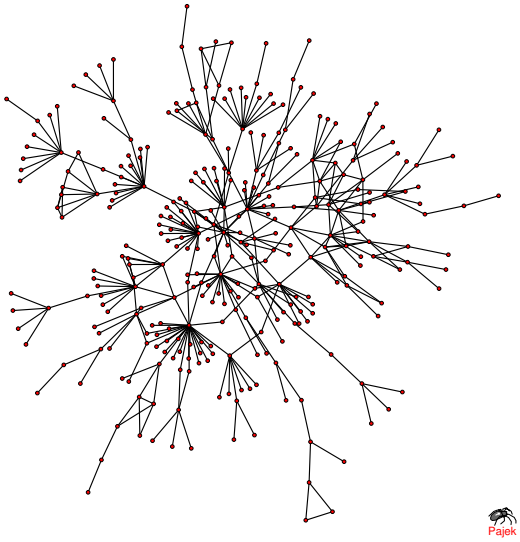


Figure 1.1: Part of the protein interaction network of yeast. The picture contains 320 proteins (nodes) and 390 interactions (links) that are part of the largest connected cluster of all interactions mapped by the yeast-two-hybrid method that are listed in the MIPS database which is included in the GRID database [18]. The spatial positions of the nodes are arbitrary and only determined by the Pajek program which was used to illustrate the network.

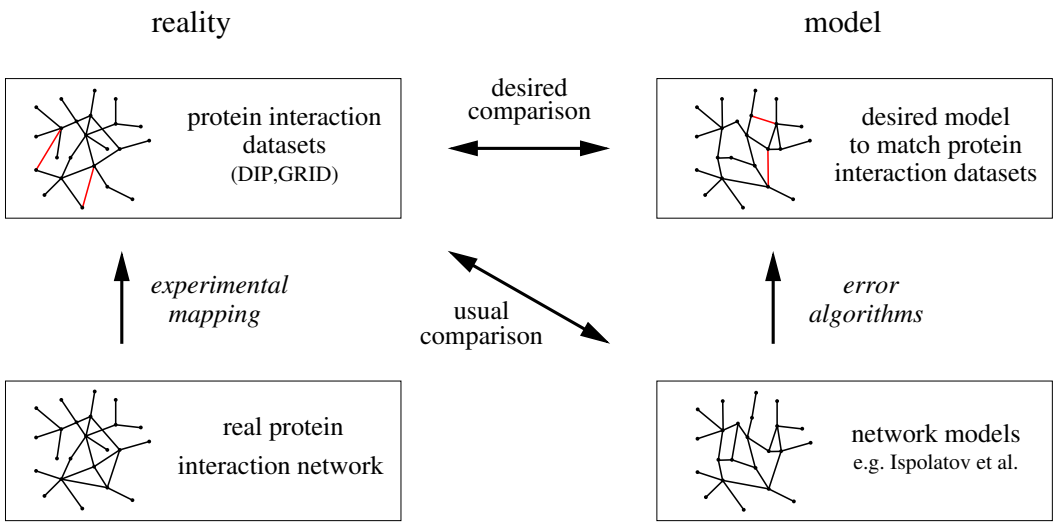


Figure 1.2: Model and Reality: on the left, the true protein interaction network (bottom) and its corrupted representation in actual datasets (top). Hence, the modeling must consist out of two steps: The model that corresponds to the real protein interaction network and a model with a superimposed error algorithm which corresponds with protein interaction datasets.

Complex Networks, which provides the instruments used in this work. Chapter three gives an overview over the biological background of proteins and their interactions combined with a discussion of the applied methods, these interactions are found with. This is followed by the introduction of gene-duplication and mutation models and their discussion. Chapter four focuses on observational incompleteness of yeast data. The biases and drawbacks of the three most important mapping methods will be discussed, and it will be simulated in how far the properties of an underlying network are changed by several error algorithms. It is then asked how well the biases found for mapping methods are reflected in error algorithms. Chapter five closes this thesis with a summary and a short outlook.

2 Statistical physics of complex networks

A network, or in mathematical terms a graph $\mathcal{G}\{\mathcal{N},\mathcal{L}\}$, is represented by its *nodes* $i \in \mathcal{N}$ and a set of *links* ($l_{ij} \in \mathcal{L}$) between nodes i and j. These networks can be represented by their *adjacency matrix* with the elements

$$a_{ij} = \begin{cases} 1 & \text{if a link between node } i \text{ and node } j \text{ exists,} \\ 0 & \text{else.} \end{cases} \tag{2.1}$$

In principle, all network properties follow from the adjacency matrix. In the case of an undirected network, the matrix becomes symmetric with the elements $a_{ij} = a_{ji}$. Generalizations towards directed links and weighted networks are straightforward but not necessary for the focus of this thesis on protein interaction networks. Mutual interactions are generally undirected. Also the largely differing strength of binding is disregarded in the protein interaction network models studied here.

In this chapter an overview over network properties will be given, followed by some general network types and models.

2.1 Network properties

In graph theory and in applied network research, several topological measures have been developed like degree distribution, degree correlation, clustering coefficient and motif structure, to name some of the most important.

2.1.1 Degree distribution

The *degree* k_i counts the number of neighbors attached to the node i by links [2, 4, 5]:

$$k_i = \sum_{j \in \mathcal{N}} a_{ij}. \tag{2.2}$$

To achieve an overall measure for the network, this degree is averaged over all nodes leading to an *average degree* $\langle k \rangle$

$$\langle k \rangle = \sum_k k p(k). \tag{2.3}$$

This average degree is then related to the total number of nodes N and links L in the network as

$$\langle k \rangle = \frac{2L}{N}. \tag{2.4}$$

The factor 2 results from the fact that every link has two node ends. A more detailed way to characterize the entire network is to order the nodes with their degrees. After normalization over the total number of nodes and in the limit of large numbers, this *degree distribution* gives the probability $p(k)$ for a node to have degree k_i:

$$p(k) = \frac{1}{N} \sum_{i \in \mathcal{N}} \delta_{k_i k}. \tag{2.5}$$

The degree distribution is used to generally classify network types. For example a sloping degree distribution indicates a hierarchical network, while a Poisson distribution indicates the absence of such a hierarchy.

2.1.2 Network connectivity

Crucial for the function of complex networks is the overall connectivity of a network. Also in protein interaction networks, the connectivity of every protein to any other over a finite number of intermediary proteins appears to be essential for cell regulation [34]. It is supposed that only this enables the control of the production of proteins. As most cellular functions of proteins are carried out by their interactions, proteins become dysfunctional when they lose their ability to interact. This would not prevent the formation of multiple communities, which have no connection between each other. However, the data of protein interaction networks shows that only very few proteins emerge that are not part of the main community. In the current yeast protein interaction datasets consisting of $\approx 4\,800$ nodes, only 0.2% are found in communities of three nodes and 1.6% of nodes are found in pairs. Following this argumentation, only the main communities are discussed in this work.

The main community is called *giant component* $\mathcal{G}_{\mathrm{gc}}$ if it consists of the large majority of all nodes [2, 4]. In random networks, the appearance of such a giant component is determined by a construction parameter. If this parameter passes a threshold, a transition of the network towards the emergence of a giant component occurs. This construction parameter can be e.g. the total number of links L for a constant number of nodes N.

The *average path length* $\langle d \rangle$ characterizes the giant component further. It is the average over all shortest distances d_{ij} between any node i and j:

$$\langle d \rangle = \frac{\sum_{ij} d_{ij}}{N(N-1)}. \tag{2.6}$$

The path length d_{ij} is defined to be finite within the giant component and infinite in all other cases. A related measure is the *network diameter*, as the maximum of all shortest paths:

$$d_{\max} = \max_{i,j}(d_{ij}). \tag{2.7}$$

An algorithm to find the shortest paths and some examples are given in [35].

2.1.3 Betweenness centrality and community structure

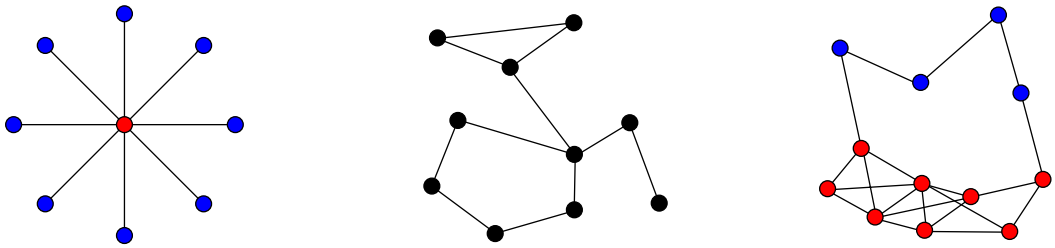
With the knowledge of the set of all shortest paths d_{ij}, a further measure counts the number of shortest paths passing over every node or link. This measure is called *node- or link betweenness centrality*.

The betweenness centrality is an important measure to gain a deeper insight beyond the pure connectivity. If the betweenness centrality fluctuates strongly and if nodes or links exist that participate in a large number of shortest paths, this is an evidence that the network is not homogeneously connected but parts (communities) exist that have only few connections between each other. In contrast, if the betweenness centrality is rather homogeneous for all nodes or links, no communities can be distinguished. To be a little bit more practical: To cut a regular lattice in several parts, many links have to be removed, while for the branch of a tree only one cut is necessary. Newman and Girvan [36] proposed an algorithm for the so called *community structure* based on betweenness centrality:

Calculate betweenness centrality for all links
repeat
  Find link with highest betweenness centrality
  (If more than one are found, choose one at random)
  Mark link as removed
  Calculate In-betweenness for remaining links
until no link remains unmarked.

With this algorithm, links with the highest betweenness centrality are removed step by step. By recalculating the betweenness centrality for the remaining network, it is assured that changes due to former link removal are taken into account. This is crucial to measure a correct community structure but results in a very time-consuming algorithm.

While links are removed, the network breaks apart step by step into an increasing number of isolated communities, and such a community is assigned to every node. Based on the original network, the *quality factor Q* is calculated as a normalized measure of how many links exist within a community and how many links connect the identified communities. The fraction of inter- and intra-community links is a measure for the strength by which communities are separated from each other. The quality factor Q is defined as follows:



a) b) c)

Figure 2.1: Three examples for different degree correlations: a) a hierarchical network with disassortative behavior, where a highly connected node (red) is attached to lowly connected nodes (blue), b) a random network with a constant degree correlation where no preference is visible and c) a clustered network, where highly connected nodes are more likely connected to other highly connected nodes and lowly connected to other lowly connected nodes.

$$Q = \sum_k \left\{ e_{kk} - \left(\sum_l e_{kl} \right)^2 \right\}, \tag{2.8}$$

where indices indicate communities and e_{kl} is the number of links between a community k and l.

While the network is broken into more and more communities, the quality factor is recalculated. The maximum of the quality factor indicates in how many communities the network is separated for a maximized proportion of inter- and intra community links. The value of Q at its maximum gives the separation strength of communities within the network.

2.1.4 Degree correlation

The *degree correlation* is another crucial observable for the function of signal transmission and functional control in protein interaction networks [2, 4, 37, 38]. It is a network-hierarchy measure to determine whether sparsely connected nodes tend to be connected to other sparsely connected nodes or if they prefer to connect to highly connected nodes. In the latter case, highly connected nodes serve as *hubs* connecting lots of sparsely connected nodes.

Given a node with degree k, the average degree of its neighbors

$$\langle k_{\mathrm{ngb}} | k \rangle = \frac{1}{N_k k} \sum_{i \in \mathcal{N}} \delta_{k_i k} \sum_{j \in \mathcal{N}} a_{ij} k_j \tag{2.9}$$

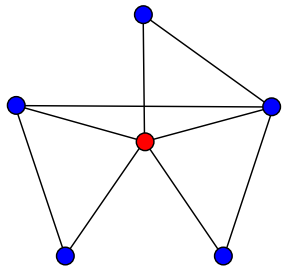


Figure 2.2: According to Eq. (2.11), the clustering coefficient of the red node $C = 0.4$ equals the number of connected neighbors over the possible number of connected neighbors $k(k-1)/2$.

represents the simplest measure for a degree correlation, where N_k is the total number of nodes with degree k. With the degree distribution $p(k)$ and the probability $p(k,k')$ of finding neighbors with degree k and k' the degree correlation $\langle k_{\text{ngb}}|k\rangle$ can be written [38]

$$\langle k_{\text{ngb}}|k\rangle = \sum_{k'} \frac{k' p(k',k)}{p_k} = \sum_{k'} k' p(k'|k). \tag{2.10}$$

On the right hand side, the term $p(k'|k)$ determines the conditional probability of finding a neighboring node with degree k', given the degree k of the other node.

The degree correlation is called *assortative* if $\langle k_{\text{ngb}}|k\rangle$ increases with k, which means that sparsely connected nodes prefer to connect with each other, as well as highly connected nodes prefer to connect to other highly connected nodes. A *disassortative* degree correlation means that $\langle k_{\text{ngb}}|k\rangle$ decreases with k. In an extremal picture, this can be imagined as a star, see also Fig. 2.1.

2.1.5 Clustering coefficient and motif-structure

The *clustering coefficient* C_i [2, 4, 5] represents the ratio of the number of direct connections between any two neighbors of node i divided by the maximum possible number $k_i(k_i-1)/2$ (see also Fig. 2.2):

$$C_i = \frac{\sum_k \sum_j a_{ij} a_{ik} a_{jk}}{k_i(k_i-1)}. \tag{2.11}$$

Averaged over all nodes in the network, the quantity $\langle C\rangle = \sum C_i/N$ is a normalized measure for the numbers of triangles in the entire network. The degree dependent clustering coefficient is defined as:

$$\langle C(k)\rangle = \frac{1}{N_k} \sum_i C_i \delta_{k_i k}. \tag{2.12}$$

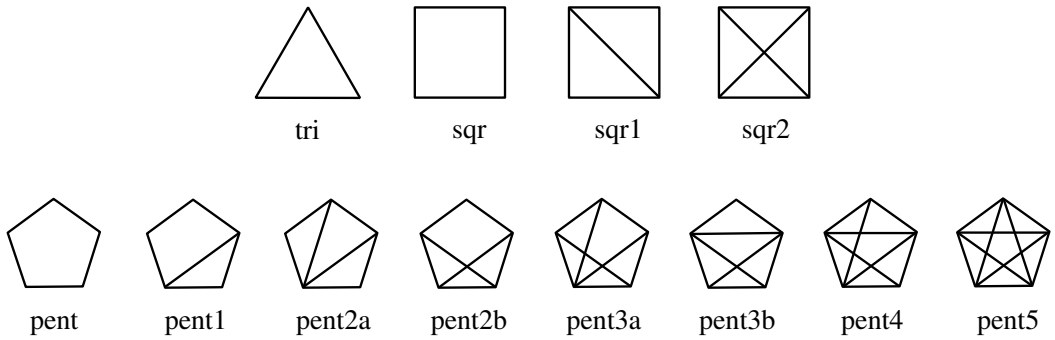


Figure 2.3: Analyzed motifs: triangles, squares and pentagons ordered according to their number and position of intra links.

Besides just counting triangles, it measures the tendency of the network to form interconnected groups of certain degrees within the network and is another measure for the hierarchy in the network. If, for example, the clustering coefficient is decreasing with the degree k, the network possesses a hierarchy with hubs, whose neighbors are not connected. However nodes with low degree are in turn strongly interconnected.

In addition, other *motifs* like squares and pentagons with different realizations of intra-links are counted (see Fig. 2.3). These closed motifs are chosen for this thesis from a variety of motif systematics (compare [39, 40, 41, 42, 43]), because loops are assumed to be especially important for processes of self-control [34].

In technical networks like integrated circuits, the smallest version of a loop, a triangle is used to fulfill functions like discriminating signals. This means a signal A is switched on or off if a signal B reaches or falls below a certain threshold. Also in some biological networks like gene expression and metabolical networks, motifs play such a regulatory role. As in technical applications, the production of a protein A can be switched on or off after reaching a certain concentration of the protein B. Hence, it is assumed that these loops play an important role in protein interaction networks as well. They do not seem to emerge by chance because they cannot be reproduced in random networks as shown later.

Motifs are found in the network by starting from a node i and by examining if two neighbors are connected for triangles, have a common neighbor for squares or have in turn neighbors that are connected for pentagons. Afterwards, inner links are determined. The total number of motifs is then divided by the multiple counts for every motif, e.g. four for a “sqr”-motif and twelve for a “sqr2”-motif.

According to [43], triangles, squares and pentagons are counted separately, i.e. a structure like “sqr1” is counted as two triangles as well, but not as a “sqr”-motif.

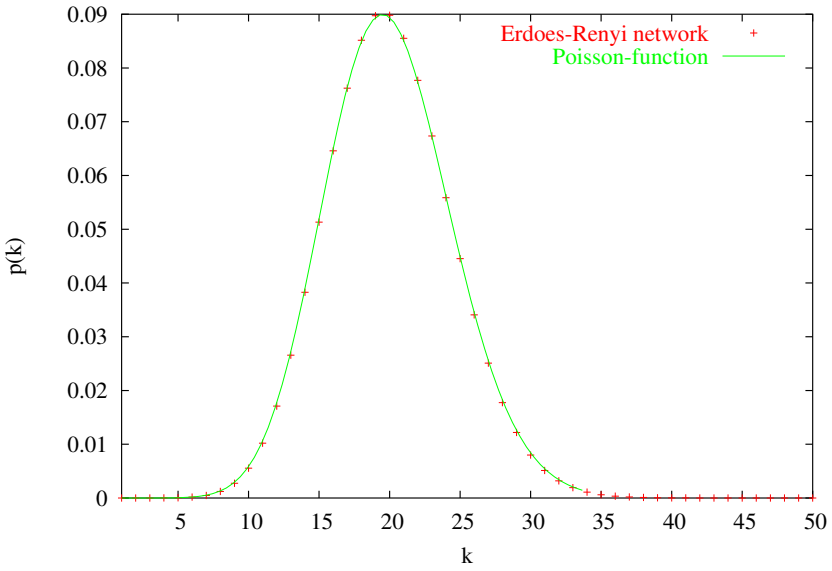


Figure 2.4: Degree distribution for Erdös-Renyi network with $N = 1000$ nodes and average degree $\langle k \rangle = 20$. An average over 10 000 network realizations has been performed.

2.2 Erdös-Renyi networks

When the study of real networks started, the common idea was that networks usually consist of a number of nodes N and randomly distributed links l_{ij} between them. Since network research started in sociology, this was first expected to be the case for friendship networks, where people have a link between each other if they have a friendly contact. Erdös and Renyi [44] proposed a model to construct these networks. If links are distributed randomly, the probability of finding a link between node i and j is

$$p(a_{ij}{=}1) = \frac{2L}{N(N-1)}. \tag{2.13}$$

Random networks can now be constructed by deciding with probability $p(a_{ij} = 1)$ for every possible link l_{ij} in the network if it is set or not.

Another almost equivalent method is to randomly choose two nodes for each of the L links. If a link between these nodes already exists, or the two nodes happen to be the same, this process is repeated. Hence, also for this method multiple links and self-links are avoided. For small networks, where $N \not\gg k$, it turns out that both algorithms do not lead to completely similar networks and that especially the degree correlation deviates (data not shown).

The degree distribution for randomly distributed links is binomial

$$p(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{2.14}$$

For $N \gg \langle k \rangle \gg 1$, it approaches a Poissonian

$$p(k) = \frac{(\langle k \rangle)^k}{k!} e^{-\langle k \rangle}. \tag{2.15}$$

The Poisson distribution in Fig. 2.4 perfectly matches a numerically simulated Erdös-Renyi network with the same average degree $\langle k \rangle$.

As links are randomly distributed to the network, there is no correlation between the clustering coefficient C_i and the degree k_i. If the average degree $\langle k \rangle$ remains constant, the total number of triangles in the network is independent of the network size N. In the simulations of this thesis, it turned out that the same holds for simple squares and simple pentagons. In contrast, the number of squares and pentagons with one inner link decreases with the network size. No conclusion can be drawn for squares with two intra-links and for pentagons with more than one intra-link because they do not occur in a sufficient number in random networks with average degree of $\langle k \rangle = 6.47$ and network sizes from 1 000 to 5 000 nodes (see Fig. 2.5).

2.3 Small world networks

In a one dimensional lattice, where every node is connected to its neighbor and with periodic boundary conditions (i.e. a ring of nodes), the average path length $\langle d \rangle$ is long compared to an Erdös-Renyi network and scales proportional to N. Even in higher dimensional lattices, the average path length $\langle d \rangle$ remains long and scales like $N^{1/D}$, where D stands for the dimension.

This is not the case for most real networks. Milgram [45] first examined over how many intermediate people two person know each other within the United States. The result was surprising. In such networks the average path length is very small with $\langle d \rangle \approx 5$.

Watts and Strogatz [46] proposed a model where in the mentioned ring of nodes also next neighbors are connected to assure good connectivity. If now only a few links are randomly rewired, the average path length decreases strongly. For further link exchange, the network becomes a random network and thus the average path length approaches the average path length of a random network, which is a small world network also due to its randomly set links. In random networks, $\langle d \rangle$ scales like $\log N$.

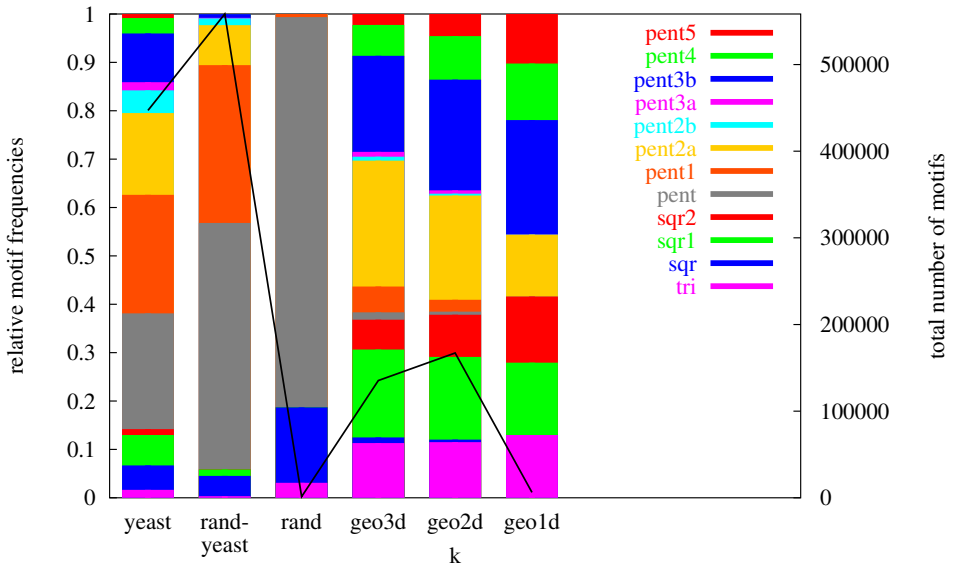


Figure 2.5: Motif structure for random networks in comparison to the yeast protein interaction network [18]: A randomized yeast network (configuration model), an Erdös-Renyi network and geometric networks in three, two and one dimensions. The colored columns represent the relative motif frequencies for the motifs depicted in Fig. 2.3 and the black line shows the total number of motifs. All networks have the same number of nodes $N = 4687$ and an average degree of $\langle k \rangle = 6.47$. Networks have been averaged over 100 network realizations in case of the randomized yeast and over 50 network realizations in case of the other simulations.

2.4 Scale-free networks

For a significant number among the real existing networks in biology, sociology and engineering, the degree distributions cannot sufficiently be described by Erdös-Renyi networks but rather follows a power law:

$$p(k) \sim k^{-\gamma} \quad ,\gamma > 0. \tag{2.16}$$

With the power law distribution, no favored degree exists and the network is self similar regarding the degree distribution. Every sub-network possesses the same degree distribution. The classification of networks as *scale-free* according to their degree distribution was introduced by Barabasi et al. [47] and derives from the absence of a reference scale in these networks.

Fig. 2.6 shows two examples of networks that are considered to be scale-free: the yeast protein interaction network, which is the subject of this work [18], and the network of linked web pages within the "nd.edu" domain of the University of Notre Dame [48]. Interestingly, also other complex systems beyond networks show this scale-free behavior. The third curve in Fig. 2.6 shows the calling habits of my former flatmate Silke. It displays the probability of the duration of her phone calls.

It is observed that all of the degree distributions of the different examples do not resemble a perfect power law. The power law only holds for an infinite network size. Hence, a modification with an exponential cutoff is more suited:

$$p(k) \sim k^{-\gamma} e^{\frac{-k}{k_c}}, \tag{2.17}$$

where k_c is the cutoff parameter.

Nevertheless, scale-free like networks appear to be favored by nature not only because of their small network diameter but also because of their robustness against random removal of links [19, 22].

To construct scale-free networks, Barabasi and Albert [47] proposed a *growth model*. In every growth step, a new node with a fixed number m of open links is introduced. The free ends of these links are connected afterwards to already existing nodes in the network preferentially with its degree:

$$p_i = \frac{k_i}{\sum_j k_j}. \tag{2.18}$$

This Barabasi-Albert-model is the most basic version of a "rich gets richer" algorithm. The scale-free degree distribution with an overwhelming number of lowly connected and very few highly connected nodes (hubs) emerge if newly introduced, lowly connected nodes preferentially choose highly connected nodes to link up with. In this way, highly connected nodes gain new links much more likely. Several other but similar algorithms are proposed in the literature [6].

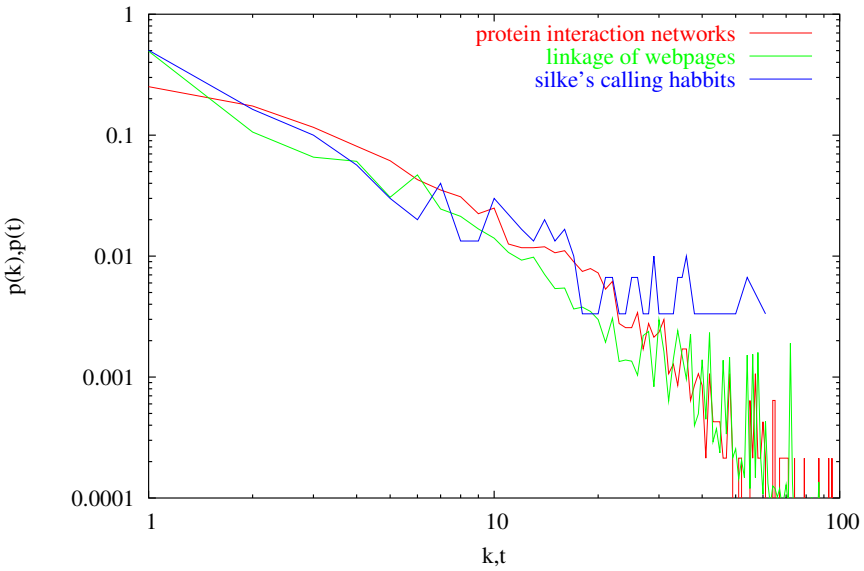


Figure 2.6: Some examples of scale-free probability distributions: degree distributions of the protein interaction network [18], of the linkage of web-sites at the University of Notre Dame [48] and the duration probability of calls of my former flatmate Silke. In this graph and in all following depictions with scale-free network data, deviations are large for nodes with high degree due to their small occurrence in the networks. Nevertheless, they have to be depicted because of their large importance for the network topology.

The hierarchical structure of scale-free like networks can be analyzed very well with the community structure introduced in Sect. 2.1.3. Fig. 2.7 shows the quality factor Q over the number of communities the network is broken into. The analysis was done for the yeast protein interaction and an Erdös-Renyi network. The latter one less hierarchical. The maximum of the quality factor is much less significant and the values of Q are rather low. In contrast, for scale-free networks a higher maximum emerges. The network can be broken in some well defined communities by removing only a few links with high betweenness centrality. By comparing several scale-free like networks that were subject to this work, no significant differences could be found. Hence, the study of the community structure was abandoned for further studies.

2.5 Random networks with given degree distribution

Random networks with a given degree distribution are also called configuration models. They are usually used to evaluate whether the topology of a given network derives only from the degree distribution. According to the given characteristic degree distribution, a degree is assigned to any individual node. Its open links are then randomly connected

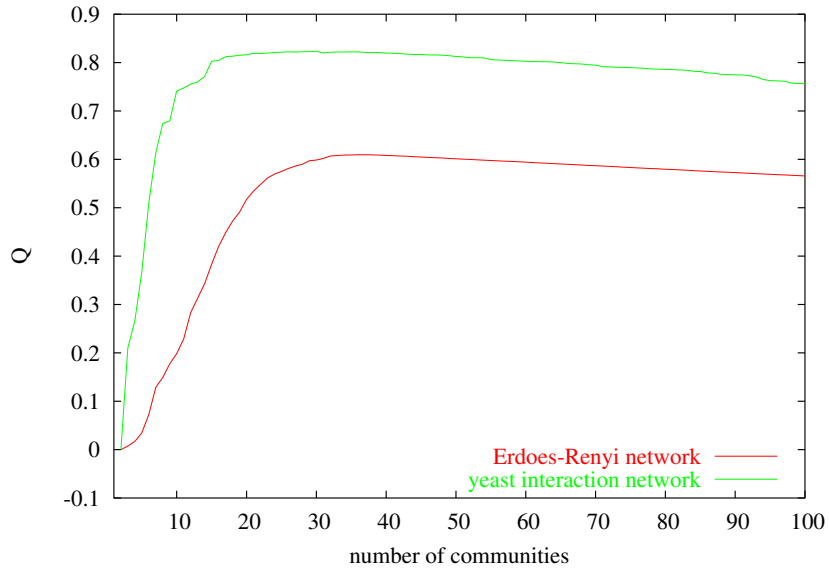


Figure 2.7: Community structure of an Erdös-Renyi network and a yeast protein interaction network. To avoid extensive calculations, only links out of the high confidential dataset of yeast [49] are regarded. This results in a network with $N = 1158$ nodes and an average degree $\langle k \rangle = 3.41$. The random network is of the same size and average degree. Furthermore, the respective curve is averaged over 20 network realizations.

to the other nodes' open links.

The resulting degree correlation and clustering coefficient are constant, as for Erdös-Renyi networks. Neither a correlation between clustering coefficient and degree nor between the degrees of a node and its neighbor is introduced by constructing these networks.

However, the total number of motifs is much higher as it is for Erdös-Renyi models and of order of the motif number in the real yeast interaction network. Also the variety of motifs is much higher for e.g. the *randomized yeast network* (see Fig. 2.5).

2.6 Geometric networks

In *geometric networks*, the spatial position potentially determines the existence of a direct link between two nodes i and j [50, 51, 52, 53]. In the simplest case, a link between two nodes is set if their mutual Euclidian distance is below a threshold. Przulj et al. [41] claim that certain properties of protein interaction networks are reproduced rather by geometric than by gene-duplication and mutation models.

In detail, geometric networks are constructed as follows: The position (x), (x,y) or

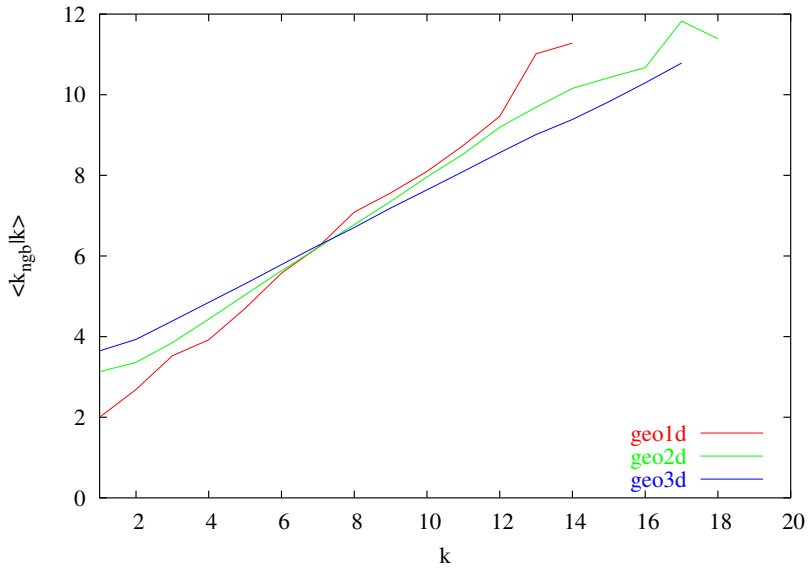


Figure 2.8: Degree correlation for geometric networks in one (geo1d), two (geo2d) and three (geo3d) dimensions. All networks have an average degree of $\langle k \rangle = 6.47$ and a size of $N = 4687$. Simulations are performed over 50 network realizations. In the one and two dimensional case, only the giant component was considered, while for one dimension no giant component arose and hence all nodes have been analyzed.

(x,y,z) of any node is randomly assigned in a one-, two- or three-dimensional cube $[0,1]^D$ respectively. For a given average degree $\langle k \rangle$, the maximum Euclidian distance between two nodes to be connected is calculated according to:

$$d_{\text{euklid}} = \frac{\langle k \rangle}{2N} \qquad \text{, in one dimension } [0,1] \tag{2.19}$$

$$d_{\text{euklid}} = \sqrt{\frac{\langle k \rangle}{N\pi}} \qquad \text{, in two dimensions } [0,1]^2 \tag{2.20}$$

$$d_{\text{euklid}} = \sqrt[3]{\frac{3\langle k \rangle}{4N\pi}} \qquad \text{, in three dimensions } [0,1]^3. \tag{2.21}$$

As for Erdös-Renyi networks, the degree distribution turns out to be Poissonian. Also the clustering coefficient is not dependent on the degree k. However, the degree correlation becomes assortative (see Fig. 2.8). This can be explained by fluctuations in the density of the spatial node distribution. If the local density is higher, nodes have a higher degree and are connected to nodes with a similar degree. In regions of lower density, the low-degree nodes are again connected to other nodes of lower degree. In regions of higher density the nodes tend to be fully connected. In this way, also motifs with a high number of intra-links are more likely to emerge and a high variety of motifs is obtained (see Fig. 2.5).

The average degree $\langle k \rangle$ is chosen to be comparable with that of protein interaction networks. In case of one dimension, no giant component is obtained. Hence, one dimensional networks are shown here for purposes of completeness only.

Facing the drawback that gene-duplication models are not able to describe the motif-structure of real yeast interactions very well, geometric networks are proposed in the literature to fit biological data much better [41]. This hypothesis may be supported by the fact that proteins whose genetic code is situated close by on the DNA strand have a higher probability to be functionally correlated and hence to interact directly [54]. Nevertheless, the results of [41] could not be verified (see Fig. 2.5), but it is remarkable that indeed only geometric networks can reproduce more complex motifs like pentagons with several intra-links in a comparable amount although they fail to fit other properties of protein interaction networks.

3 Protein interaction networks

Any life form mainly consists out of water, proteins and fatty acids, which is also applicable to the *yeast* cell. The deoxyribonucleic acid (DNA) can be seen as the data storage for the production of proteins. The proteins are produced over several intermediary steps of the transcription of the DNA. Nevertheless, from an information theory point of view, a very simple construction kit is used.

Four different base pairs exist. The DNA is built of millions of these base pairs in a strand. Three of these base pairs in a line code one amino acid. From combinatorics, 64 amino acids would be imaginable, but only 21 are used in the cell. *Proteins* in turn are macromolecules in form of a chain out of more than 100 of these proteinogenic amino acids. This makes an unimaginable number of different proteins possible, but nature is content with only a few thousands like about 5 000 in the yeast cell. Additionally some combinations of the base pairs code control sequences, e.g. where to start and to end if a protein is produced on basis of the DNA.

In Fig. 3.1, a galactose/glucose-binding protein is shown. The primary structure, which stands for the linear strand of amino acids in the protein, is not visible here. These amino acids twist this linear strand towards energy minima between each other. The resulting helices, stripes and lines (see Fig.3.1) are called secondary structure. The orientation of these secondary structures towards each other is called tertiary structure which completes the picture of the protein.

Proteins help to fulfill almost all of the complex functions in biological organisms, as reproduction or nutrition transport and processing as well as signal transduction to name a few. These functions are provided by interactions of proteins. Parts of the proteins, called *binding sites*, have the capability to bind to specific complementary binding sites of other proteins. This provides the *interaction* of proteins.

Functions in the cell are often carried out by *complexes* of proteins. These complexes are formed by many different proteins that are bound together.

In a simple picture, *evolution* takes place in unicellular organisms by errors that are made during the division of a cell into two including its DNA (mitosis). During this process, errors can occur by copying DNA sequences - the coding of one protein - twice. At the same time, errors are induced in the copy by changing one or more base pairs, altering the amino acid sequence and thus generating a new protein. In the very most of cases, these new proteins are removed or the corresponding cells extinct. But in some cases, the new life form may be able to survive and acquire an even better fitness if the new proteins fulfill new functions. Evolution is assumed to have formed the complex system of proteins and their interactions in this way, starting from a small

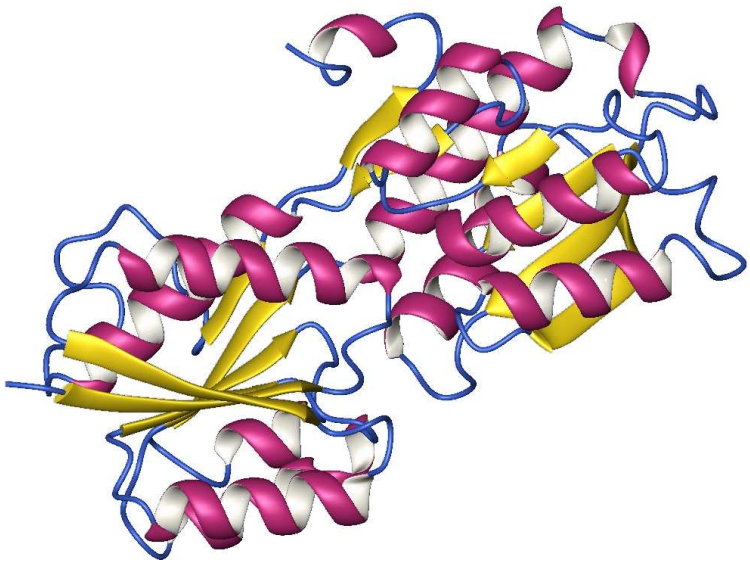


Figure 3.1: A galactose/glucose-binding protein with its secondary and tertiary structure [55].

initial number of proteins.

3.1 Yeast data

In this study, only proteins and their interactions are considered in an abstracted way. Every type of a protein is a node $i \in \mathcal{G}$ in the network, and every possible interaction between two types of proteins is considered as a link between respective nodes $l_{ij} \in \mathcal{L}$. A network of protein interactions arises if all of the about 5 000 known yeast proteins and their interactions are mapped in one scheme. Due to the nature of mutual interactions, the protein interaction network consists of undirected links. Although protein interactions have very different strengths, links of network models discussed in literature are not weighted on this level of abstraction. Several databases of protein interactions for yeast are available [18, 49, 56, 57]. The largest datasets contain about 5 000 proteins with a maximum average degree of $\langle k \rangle \approx 7$. All of them are meta databases, including, for example, the MIPS (Munich Information Center for Protein Sequences [58]) database, and contain mostly the same sets of protein interactions. In this thesis, the first two datasets (Database of Interacting Proteins (DIP, [18]) and General Repository of Interacting Datasets (GRID, [49])) are analyzed in detail. It turns out that differences in the network properties are rather small. The only significant differences between both datasets can be found in the motif-structure. The total number of motifs is much larger in the GRID dataset than in DIP, which is probably related to the higher average

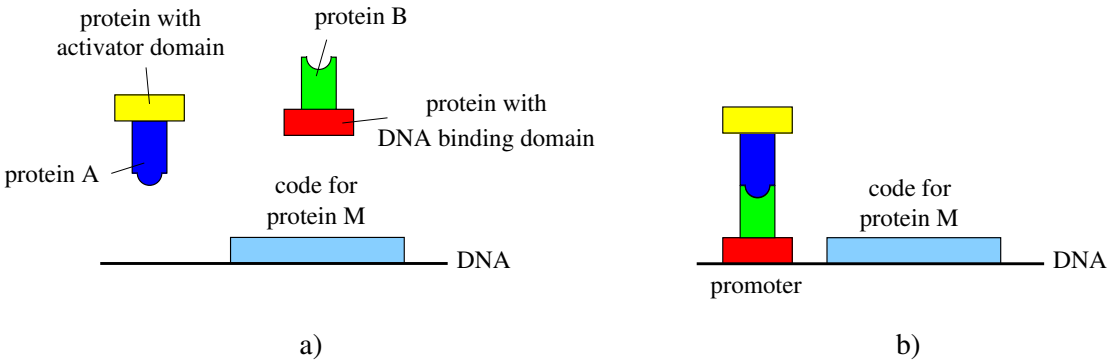


Figure 3.2: Scheme of the yeast-two-hybrid method. Two hybrid proteins are artificially constructed: A protein A (blue) is bound to another protein which possesses an activator domain (yellow) and a protein B (green) is bound to a protein with a DNA binding domain (red). If A binds to B, the entire structure forms the promoter which activates the production of a protein M. Hence, its occurrence in the dissolved cell (cell lysis) after the introduction of the two hybrid proteins is an evidence for the interaction of A and B.

degree (DIP: $\langle k \rangle = 6.47$, GRID: $\langle k \rangle = 7.22$). Furthermore, the frequency of pentagons with four intra-links is larger in the GRID dataset.

In further studies the DIP dataset was used for comparisons between models and real yeast data because the reliability appears to be larger for every single mapped interaction if the total number of links is smaller. The GRID dataset will only be used in the later analysis of different mapping methods because it documents respective mapping methods for every link.

Contained interactions derive from several mapping methods. The most important are yeast-two-hybrid, synthetic lethality and affinity isolation methods (affinity precipitation and affinity chromatography). During the last years, the amount of mapped proteins and their interactions increased largely [12, 13]: from 1 800 proteins and 2 200 interactions in the year 2001 to 5 000 proteins and more than 18 000 interactions today. This over-proportional increase of known protein interactions is due to the development of high throughput methods based on the three major methods, which will be explained in the following sections. Only about 150 interactions base on other than these three methods. All of the high throughput methods trace back to only a few groups of experimenters [59, 60, 61, 62, 63].

3.1.1 Yeast-two-hybrid

The *yeast-two-hybrid method* [9] measures the specific interaction of two proteins A and B as depicted in Fig. 3.2. These proteins are specifically selected by the experimenter. Protein A is bound to another protein, which contains an activator domain. B is bound to a protein with a DNA binding domain. Proteins - like M - are reproduced by

transcribing the DNA. The transcription is controlled by the promoter (a protein complex) at the beginning of a gene which contains a binding and an activator domain. If now protein *A* interacts with protein *B*, the entire structure is able to form the promoter with the binding domain binding to the DNA and the activator domain providing the transcription of the code of protein *M*. Promoters are specialized complexes that enable the production of one or a few proteins only. The two hybrid proteins, the activator domain combined with *A* and the binding domain combined with *B*, is placed into the cell lysis, the dissolved cell after disintegration of the cell membrane. The proximate appearance of protein *M* in the cell lysis is a proof of the interaction between *A* and *B*, because the complex can only be formed if protein *A* binds to *B*. Modern biochemistry is able to rapidly produce a large amount and variety of these artificial constructions of protein hybrids. This makes it possible to use this method in high throughput experiments although it targets single interactions. Nevertheless it can not be assured by the yeast-two-hybrid method that the interaction of the proteins *A* and *B* is not provided by an intermediary protein *C*.

3.1.2 Affinity isolation

Using complex purification methods, namely *affinity precipitation* and *affinity chromatography* [61, 63], a protein of interest is tagged and placed into the cell lysis. The tagged protein (*bait*) is then isolated with its associated proteins (*preys*). In the next step the preys are separated and analyzed. For these analyses mainly two methods are used, precipitation and chromatography.

In the commonly used spoke algorithm [10], direct links are defined between the bait and all of its preys. The possibility that the bait is not directly interacting with all of its preys but via intermediate proteins is not taken into account, neither are possible links between preys regarded.

Another common reservation to in vitro mapping methods, yeast-two-hybrid and affinity isolation, is that interactions are mapped unconditionally. It is not measured if these interactions also play a functional role in the cell. For example, an interaction between protein *A* and protein *B* can be mapped with this method, even if it never occurs because Protein *A* may be only present in the nucleus and *B* only in the rest of the cell. For the evaluation of the evolution models and to examine if the admittedly rough models can explain simple properties of the protein interaction networks, this reservation seems to be irrelevant. The examined models are not concerned with the function of protein interactions but only with the possibility of interactions due to effects of inheritance of properties.

3.1.3 Synthetic lethality

If the mutation of two proteins causes cell death while the mutation of only one of these two proteins is not lethal, it is supposed that they are functionally associated and thus

interact [62]. To measure *synthetic lethality*, cells are cultivated with different mutations of their DNA. The single mutations are characterized independently and do not lead to cell death. In a next step, mutations of two different proteins are combined in one cell. If this causes cell death there is an assumed functional connection, which is usually established by an interaction of these proteins. Nevertheless, even if these two proteins are part of the same protein complex, it is again not assured that any interaction is not provided by an intermediary protein. In addition, interactions without any functional relevance are not detected by this method.

3.2 Gene duplication and mutation models

To obtain a principle idea of how nature has evolved protein interaction networks, several models have been developed [12, 13, 14, 15]. They take into account two important principles of evolution: *duplication* of segments of the DNA and their parallel *mutation*. In this way, new genes with similar properties as the original gene emerge during the random duplication process.

Since basically proteins are transcripted from the DNA, the abstraction of gene-duplication and mutation processes to protein interaction networks leads to a model in which the proteins (nodes) are copied with their interactions (links), and mutations lead to a rearrangement of respective links. Hence, with every evolution step a new node is introduced in the network.

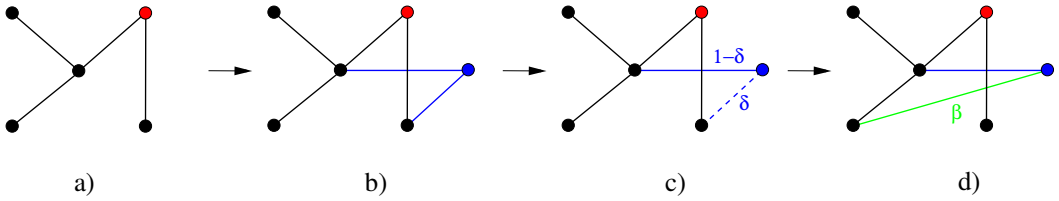
None of the models consider selection. In reality, most of the newly evolved proteins do not improve cell function and are removed either by repair processes in the cell or by the death of the whole organism. But the actual knowledge of biological processes is not sufficient to include selection in these models.

In the following, some of the proposed models will be presented in greater detail. All of the curves in this thesis that originate from simulations are averaged over 50 network realizations unless otherwise stated.

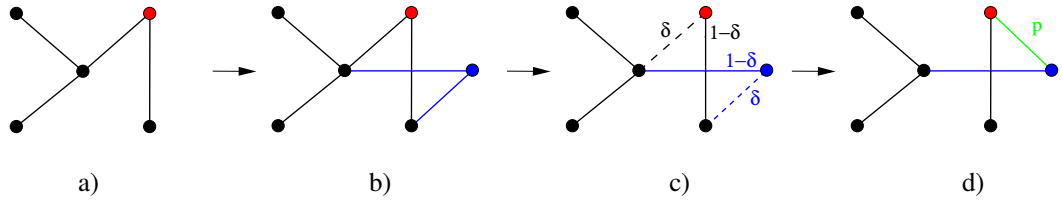
3.2.1 Gene-duplication model with random link

In the gene-duplication model with random link proposed by Solé et al. [12] a node is selected randomly in every evolution step and copied with its links. Resulting links of the new node are deleted with probability δ. New links are inserted at the copied node to another randomly chosen node with probability $\alpha = \beta/N$ (see Fig. 3.3(a)) where β is chosen arbitrarily. The resulting network is referred to as $\mathcal{G}_{\mathrm{m1}}$.

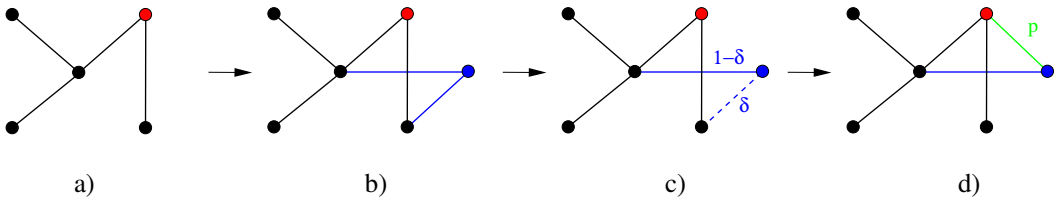
Almost all proteins in yeast interaction data are part of the giant component. Probably, nature sorts out proteins that do not interact with any other protein and thus lose their function (see Sect. 2.1.2). Hence, to keep the different models comparable with each other and with the yeast data, only the resulting giant component is examined. In



(a) The gene-duplication model with random link $\mathcal{G}_{m1}$ [12]: a) a node is selected randomly (red), b) the node is copied with its links (blue), c) new links of the new node are deleted with probability δ (dashed line), d) links between the new node and randomly chosen nodes are established with probability β (green).



(b) The gene-duplication model with homodimer-link I $\mathcal{G}_{m2}$ [13]: a) a node is selected randomly (red), b) the node is copied with its links (blue), c) links of the original and the copied node are deleted with probability δ (dashed lines), d) a link between original and copy is established with probability p (homodimer-link, green).



(c) The gene-duplication model with homodimer-link II $\mathcal{G}_{m3}$ [15]: a) a node is selected randomly (red), b) the node is copied with its links (blue), c) links of the new node are deleted with probability δ (dashed line), d) a link between original and copied node is established with probability p (homodimer-link, green).

Figure 3.3: Gene-duplication and mutation models.

the gene-duplication model with random link, the giant component has a magnitude of about 50% of the total network. Many nodes remain unconnected and thus have degree zero after the deletion of all of their links. These nodes have no significant influence on the network topology.

3.2.2 Gene-duplication model with homodimer-link I

Vazquez et al. [13] proposed a model in which links of the original or copied node are deleted with probability δ. New links are established only between the original and copied node with probability p (see Fig. 3.3(b)). This is biologically motivated by the maintenance of a self-interaction (homodimer)-link. If a homodimer protein of type A interacts with other proteins of the same type, it is very likely that after duplication and mutation this interaction is maintained and protein A also interacts with A'. The resulting network is referred to as $\mathcal{G}_{\mathrm{m2}}$.

With this model, a network with a very low connectivity and no giant component emerges. This is depicted in Fig. 3.4 in comparison to the gene-duplication model with random link $\mathcal{G}_{\mathrm{m1}}$. Parameters are chosen to match the network obtained from yeast interaction data and discussed later in greater detail. In case of the gene-duplication network with random link, a giant component emerges at $N \approx 4700$ with probability one. Additionally, smaller unconnected clusters are obtained with degrees $k = 1 \ldots 10$. Since for the gene-duplication model with homodimer-link I no giant component emerges, parameters have to be chosen to obtain a largest component within the small unconnected clusters (accordingly to the clusters on the left for $\mathcal{G}_{\mathrm{m1}}$) to be large enough to match on average the giant component of yeast data. Thus, two drawbacks have to be faced. First, the size of the largest component fluctuates strongly. The standard deviations for the size of the largest component are $\Delta N^{\mathrm{lc}}_{\mathrm{m2}} = 2000$ compared to $\Delta N^{\mathrm{gc}}_{\mathrm{m1}} = 200$ in the model with random link. Second, simulations based on this model are very extensive because the largest component consists of $\approx 20\%$ of the generated nodes only. With decreasing δ a giant component would emerge also for this model, but for all of the here proposed models it is necessary that $\delta > 1/2$, as explained later.

3.2.3 Gene-duplication model with homodimer-link II

As in the model with homodimer-link I, in the gene-duplication model with homodimer-link II proposed by Ispolatov et al. [14, 15], new links arise only when a homodimer protein is copied and the self-interaction link is conserved between original and copy. This link arises with probability p. As in the gene-duplication model with random link, these are only deleted at the copied node with probability δ (see Fig. 3.3(c)). The resulting network is referred to as $\mathcal{G}_{\mathrm{m3}}$.

In contrast to the other models, the evolution step is discarded if the new node remains unconnected because all links to the new node are deleted and no new links are

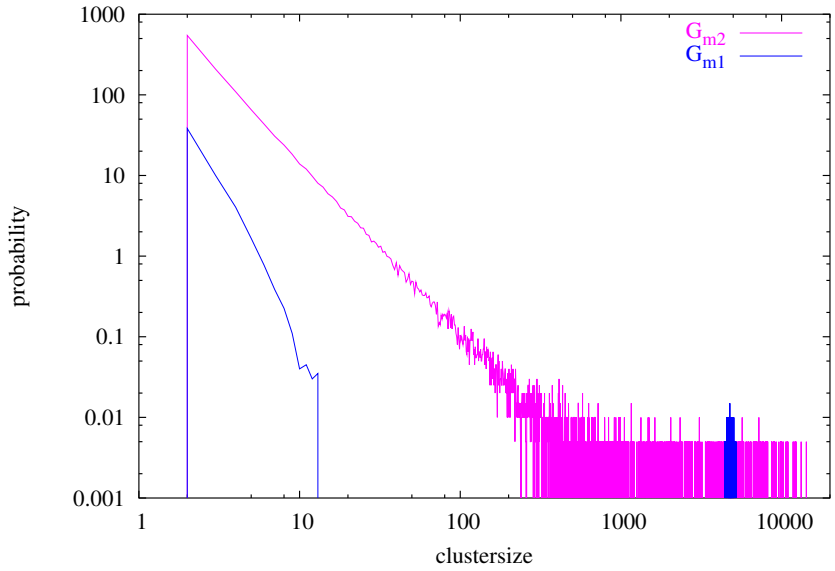


Figure 3.4: Cluster-size distributions for the networks resulting from gene-duplication models with random link $\mathcal{G}_{\mathrm{m1}}$ (blue) and with homodimer-link I $\mathcal{G}_{\mathrm{m2}}$ (magenta). The average size of the resulting giant/largest component is ≈ 4700 nodes. In the network $\mathcal{G}_{\mathrm{m2}}$, no giant component emerges but the number of evolution steps was chosen to be very large to obtain a large number of small components and hence a largest component of average size of the giant component in $\mathcal{G}_{\mathrm{data}}$. This largest component fluctuates strongly in size.

established. In this way, the network always remains fully connected. Based on the assumption of all network models that proteins always remain fully connected, it appears to be more convincing that unconnected proteins are removed during evolution like in this model. Furthermore the problem of the non-appearance of a giant component as in the model with homodimer-link I is avoided.

3.3 Comparison of models with real yeast data

With the over-proportional increase of mapped protein interactions through recently developed high throughput methods, the average degree increased from about 2.5 in the years 2001/2002 when Vazquez et al. and Solé et al. [13, 12] proposed their models to 6.5 today. Although optimal parameters are proposed in literature for all gene-duplication and mutation models, in this thesis parameters δ, p and β had to be adapted to fit the number of nodes N_{data}, the average degree $\langle k_{\mathrm{data}} \rangle$ and the degree distribution p_k^{data} of new real yeast data.

	parameters				results			
Network model	evolution steps	δ	p	β	$\Delta N_{\mathrm{gc/lc}}$	$\Delta\langle k\rangle$	$\langle C\rangle$	$\langle d\rangle$
$\mathcal{G}_{\mathrm{m1}}$ [12]	6 100	0.52	-	0.3	200	0.8	0.003	5.9
$\mathcal{G}_{\mathrm{m2}}$ [13]	14 000	0.53	0.2	-	2000	0.7	0.16	5.6
$\mathcal{G}_{\mathrm{m3}}$ [15]	4 687	0.58	0.1	-	0	0.4	0.14	5.6

Table 3.1: Parameters for gene-duplication and mutation network models with a giant component of ≈ 4700 nodes and with an average degree of $\langle k\rangle \approx 6.5$. Furthermore the respective standard deviations, the clustering coefficient $\langle C\rangle$ ($\langle C_{\mathrm{data}}\rangle = 0.13$ [18]) and the average path length $\langle d\rangle$ ($\langle d\rangle_{\mathrm{data}} = 6.1$) are shown.

3.3.1 Degree distribution

The model parameters have been chosen manually to gain a giant component of the size of the real yeast dataset N_{data}, with the same average degree $\langle k_{\mathrm{data}}\rangle$ and best coherence with the degree distribution p_k^{data} of the yeast data. The comparison of the degree distributions between model and data was always done by eye. In scale-free networks, highly connected nodes (hubs) play a decisive functional role. Hence, the application of a χ^2-evaluation or the Kullback-Leibler entropy [64] would require a weight term to incorporate the higher importance of hubs compared to their occurrence in the network. Since this weight term would be speculative and the data basis is not very strong, this was discarded.

The set of parameters is given in Tab. 3.1 and the corresponding degree distributions are shown in Fig. 3.5. Every model describes the degree distribution of real yeast data sufficiently well although slight differences do exist.

It is not surprising that all models that base on gene-duplication lead to scale-free networks like the Barabasi-Albert-Model. These models follow the so called "rich gets richer"-principle. While this is easily comprehensible for the Barabasi-Albert-model [47], where new inserted links are preferentially attached to highly connected nodes, it can be explained for gene-duplication models as follows: If degree correlations are not regarded, the probability of a node to be connected to another node is proportional to its degree k. Thus, the probability of having a neighbor that is duplicated depends again on the node's degree k. After the copying process, the degree increases to $k' = k+1$ if the respective link was not deleted. The probability of a node to gain a link at a time step in the Barabasi-Albert model is $\Pi_i = mk_i/(2mN)$, whereas in case of the gene-duplication models it is $\Pi_i = k_i/(N-1)\cdot(1-\delta)$ if the node is not copied itself and no degree correlation exists. Hence, in the case of large N and $\delta \approx 1/2$ both terms become identical.

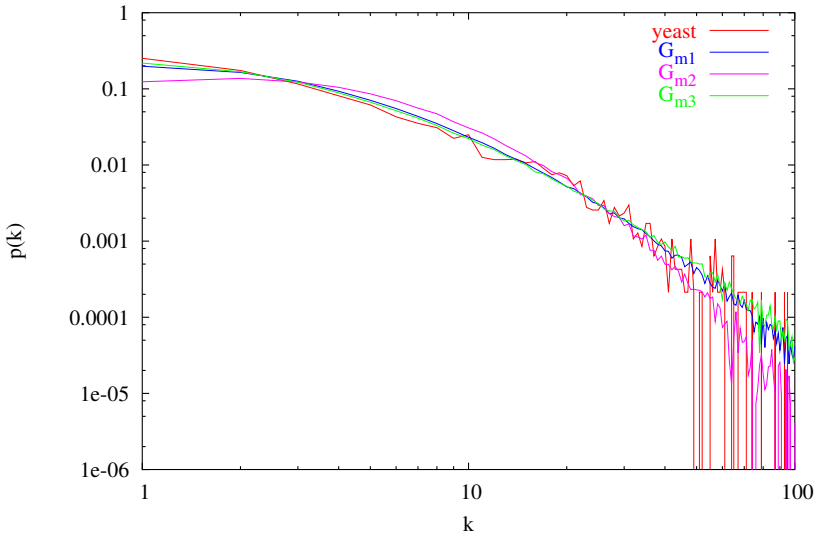


Figure 3.5: Degree distributions of the gene-duplication model with random link $\mathcal{G}_{\mathrm{m1}}$, with homodimer-link I $\mathcal{G}_{\mathrm{m2}}$ and with homodimer-link II $\mathcal{G}_{\mathrm{m3}}$. Chosen parameters are shown in Tab. 3.1. The resulting giant/largest components are of size $N \approx 4700$ and the average degree is $\langle k \rangle \approx 6.5$.

The influence of the deletion parameter δ on the degree distribution of the gene-duplication model with homodimer-link II network is shown in Fig. 3.6 (top). A larger δ diminishes the average degree and more sparsely connected nodes emerge at the cost of highly connected nodes. The choice of δ mainly determines the degree distribution and the average degree.

It is analytically shown in [13] by using a mean field approach that δ has to be larger than 1/2 to assure that $\langle k \rangle$ saturates for $N \to \infty$. Another analysis [15] showed that for $\delta < 1/2$, the network is not self-averaging anymore. Small fluctuations in the network at the beginning of the simulation would lead to completely different outcomes and no general assumptions could be made by averaging over a large number of evolution steps and for several network realizations [14].

For both gene-duplication models with homodimer-link, the parameter p has only a slight influence on the degree distribution (see Fig. 3.6 bottom). Nevertheless, it gives it a Poissonian momentum because it is no pure "rich gets richer" mechanism: A link is established between a node and its copy independent of its degree. The value of p can be estimated from biological data with the number of homodimers in the dataset: in the DIP dataset, 6 % of proteins are homodimers. The condition $\delta > 1/2$ applies also for the self-interaction links. Hence, in every evolution step the copied node has a probability of 6 % to be a homodimer. This homodimer link is maintained with a probability of $1 - \delta$, which leads to a value of $p < 0.03$. For a better significance of the

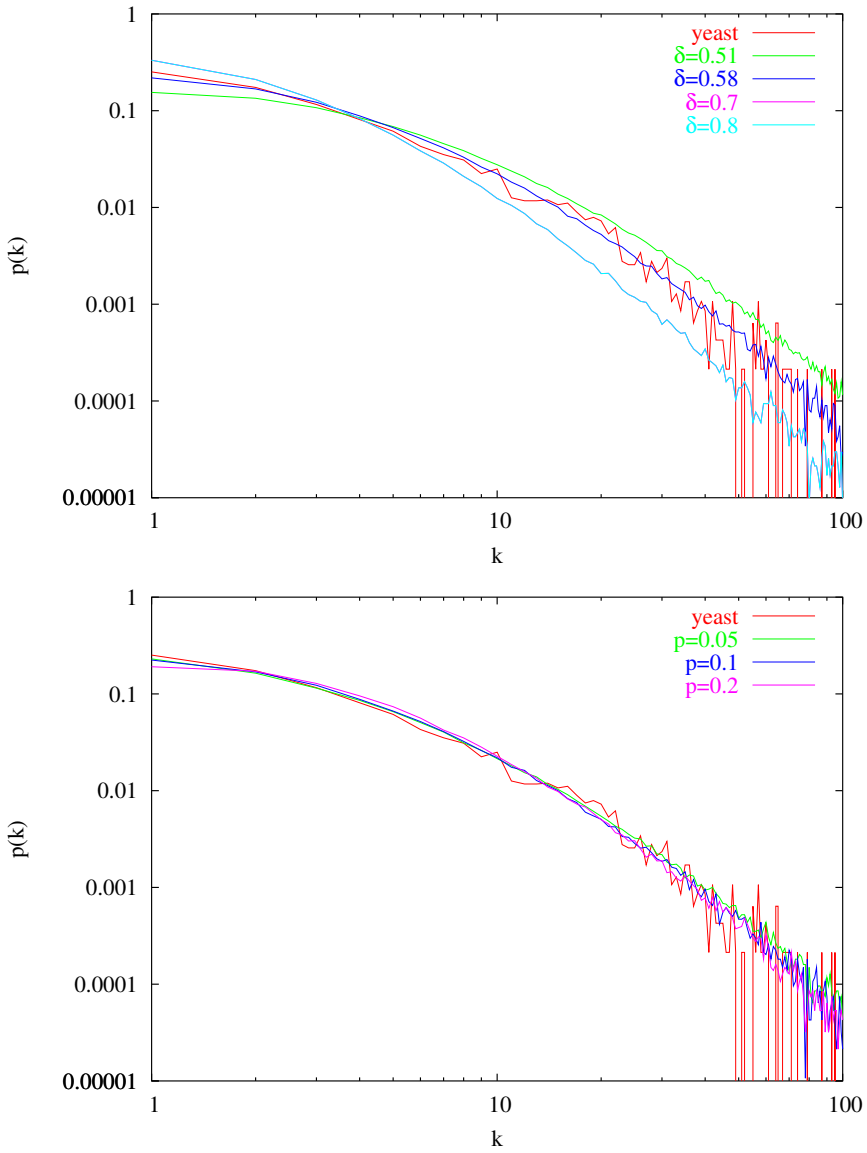


Figure 3.6: Influence of the parameters δ (top) and p (bottom) on the degree distribution of the gene-duplication model with homodimer-link II $\mathcal{G}_{\mathrm{m3}}$.

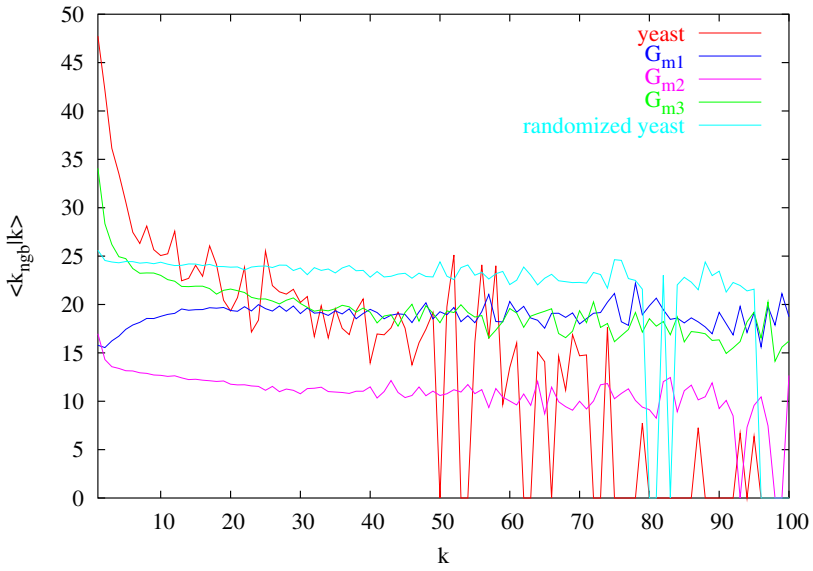


Figure 3.7: Degree correlations of the gene-duplication model with random link $\mathcal{G}_{\mathrm{m1}}$, with homodimer-link I $\mathcal{G}_{\mathrm{m2}}$ and with homodimer-link II $\mathcal{G}_{\mathrm{m3}}$. Chosen parameters are shown in Tab. 3.1. Resulting giant components are of size $N \approx 4700$ and the average degree is $\langle k \rangle \approx 6.5$. Also the randomized network of real yeast data (configuration model) is shown to depict the slight disassortative character originating from the absence of self- and multiple links.

different models, a larger value $p = 0.1$ was used here. For the gene-duplication model with homodimer-link I with $p = 0.2$, an even larger value for the parameter p had to be chosen to reach a better connectivity.

The number of evolution steps has a slight influence on the average degree but the degree distribution remains unaltered in the order of evolution steps that is of interest here. In simulations, the average degree rose from $\langle k_{\mathrm{m3}} \rangle = 5.7$ for 1 000 nodes to $\langle k_{\mathrm{m3}} \rangle = 6.8$ for 7 000 nodes using the model with homodimer-link II and $p = 0.1$.

The authors of the proposed network models [12, 13, 14, 15] used different start networks: just two, up to five fully connected nodes or a circle of five nodes, but this had no influence on the final network.

3.3.2 Degree correlation

Due to the very similar algorithms, the gene-duplication model with homodimer-link I and II show the same behavior in the degree correlation (Fig. 3.7) but the latter one fits the yeast data much better. The gene-duplication model with random link shows an assortative behavior for degrees $k < 20$. This is in contrast to yeast data but can be

explained as follows: The nodes that are selected to be copied as well the free end of the newly introduced links are chosen randomly. Both are more likely to be nodes with low degree due to their larger occurrence in the network. Hence, nodes with low degree are more likely to be connected to other lowly connected nodes.

Every network realization that does not fulfill the condition $k \ll Np_k$ shows a disassortative behavior, which is analytically shown in [65]. But it can be explained easily through the impossibility of multiple links and self-links in the network. Imagine, for example, a network with a given degree distribution of only one highly connected node and a large number of very lowly connected nodes. If no self-links are allowed, the degree correlation is per se disassortative because even if links are randomly distributed the hub can only possess a connection to lowly connected nodes. Even if two highly connected nodes are given, multiple links are necessary between them to gain an average neighbor degree $\langle k_{\mathrm{ngb}}|k\rangle = \langle k\rangle$ as for a random network with $N \to \infty$.

Hence, as observed in Fig. 3.7 for degrees $k > 20$ the disassortative degree correlation is for all models only due to this effect as it emerges in the same manner for a randomized model with the given degree distribution of the yeast data. This is in contrast to the yeast dataset, where a stronger decline of the next neighbor degree is observed.

3.3.3 Clustering coefficient

As shown in Fig. 3.8, it turns out that again the two gene-duplication models with homodimer-link do not differ significantly. In contrast in the gene-duplication model with random link almost no triangles emerge. This leads to a very low clustering coefficient over all degrees. Tab. 3.1 provides an overview over the clustering coefficients averaged over all degrees for all network models, compared to $\langle C_{\mathrm{data}}\rangle = 0.13$ in case of real yeast data.

3.3.4 Motif-structure

Figs. 3.9 show the motif-structures for the discussed models under different rates of newly introduced links. One property of all duplication and mutation models [12, 13, 14, 15] is the large number of squares in these networks compared to a rather small number in real yeast data. These squares emerge always when a node is copied and keeps more than one link. If a node with degree $k = 10$ is copied and half of its links are maintained, five squares emerge. In the gene-duplication model with random link, no triangles are formed, whereas in the two gene-duplication models with homodimer-link, triangles are obtained if the homodimer-link was set between copy and original node (compare to the clustering coefficients in Tab. 3.1). The same mechanism is responsible for the low variety of motifs in the gene-duplication model with random link. The set random link most likely does not form a new motif. The high variety in the yeast dataset cannot be reproduced by any model. Especially highly connected cliques do not emerge

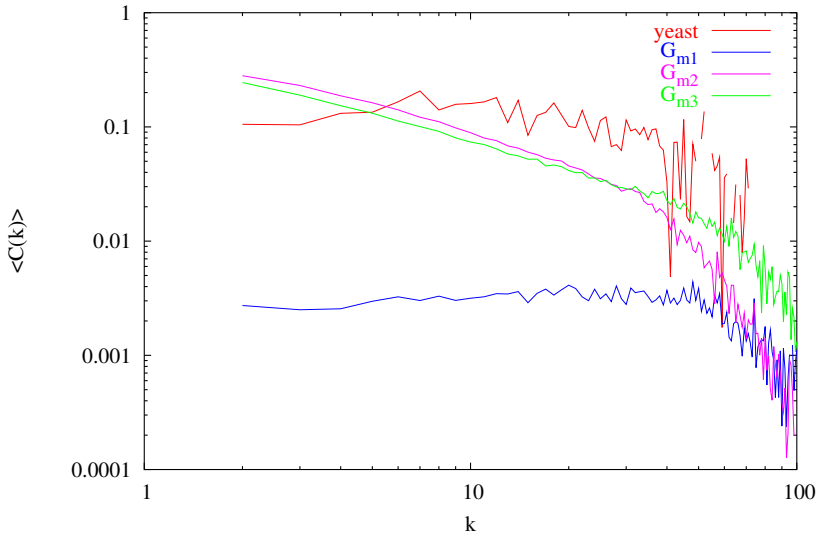


Figure 3.8: Clustering coefficient of the gene-duplication model with random link $\mathcal{G}_{\mathrm{m1}}$, with homodimer-link I $\mathcal{G}_{\mathrm{m2}}$ and with homodimer-link II $\mathcal{G}_{\mathrm{m3}}$. Chosen parameters are shown in Tab. 3.1. Resulting giant components are of size $N \approx 4700$ and the average degree is $\langle k \rangle \approx 6.5$.

in the same amount as in the yeast dataset. The number of different motifs depends on the parameters p or β respectively. With increasing β, the motif-structure improves slightly to reproduce the real yeast data network. With increasing p, frequencies of motifs improve for the two gene-duplication models with homodimer-link. But by increasing these parameters, the degree distribution shifts towards a Poissonian (compare Sect. 3.3.1) and fails to reproduce the degree distribution of real yeast data. Moreover, at least for the two gene-duplication models with homodimer-link, biological considerations suggest low values for p. The total number of motifs in the network could be well reproduced in the gene-duplication network with homodimer-link II, whereas their number is much smaller in both other models.

Compared to random networks, the motif-structure scales with the network size very differently using gene-duplication algorithms. While in random networks, the total number of motifs remains constant, it increases linearly with the network size in the orders of interest here. This is shown in Fig. 3.10 for the gene-duplication model with homodimer-link II. For the other gene-duplication models, the scaling is similar to this graph. Nevertheless, the relative frequency of motifs does hardly change with increasing network size.

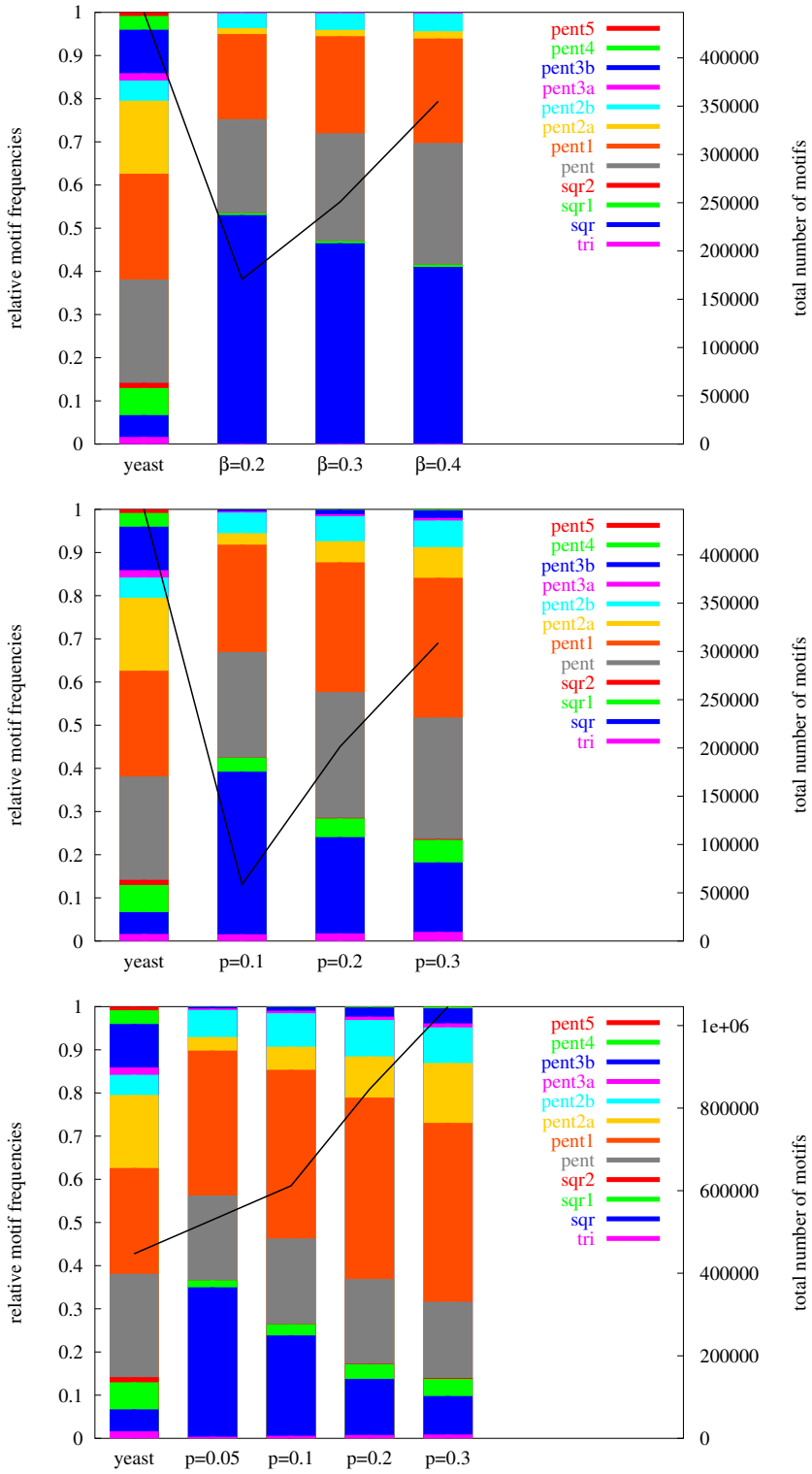


Figure 3.9: Influence of the rate of newly introduced links during mutation on the motif-structure of the three models. Respective parameters are β for the introduction of random links in the model network $\mathcal{G}_{m1}$ (top) and p for the introduction of homodimer-links in the model network $\mathcal{G}_{m2}$ (middle) and in the model network $\mathcal{G}_{m3}$ (bottom).

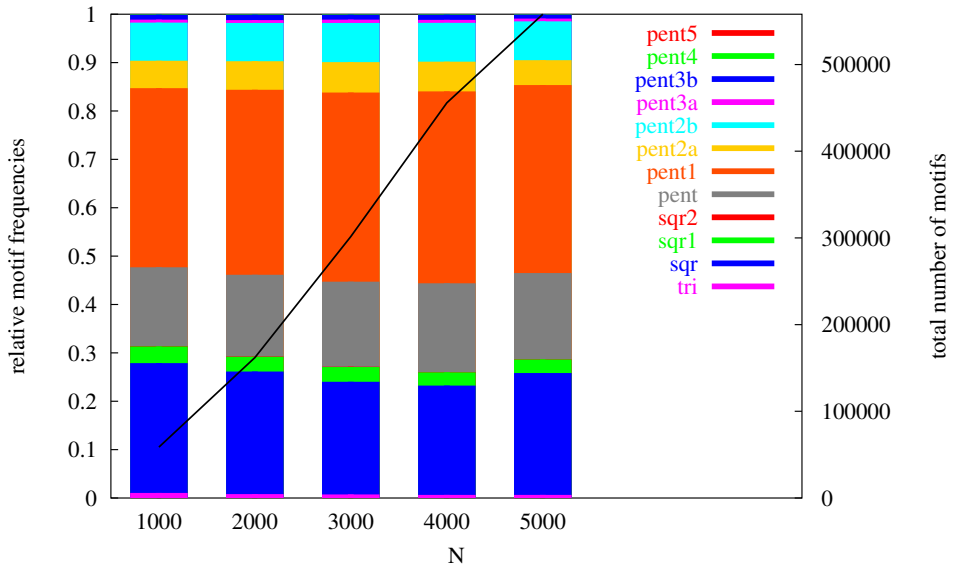


Figure 3.10: Influence of the network size N_{m3} on the motif-structure of the gene-duplication model with homodimer-link II $\mathcal{G}_{m3}$.

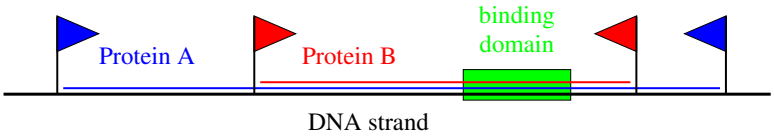


Figure 3.11: Scheme of two proteins *A* (blue) and *B* (red) encoded on the DNA strand (black). Both share the same DNA section including the coding of a binding domain (green).

3.4 Model extentions

Because of the general insufficiency of all network models to reproduce the motif-structure, two model extensions that take more biological effects into account shall be examined here. Both models ground on the gene-duplication model with homodimer-link II because it best reproduces biological data. The first model introduces a fraction of link-duplication to the gene-duplication process and the second one combines the emergence of homodimer-links and random links during gene-duplication.

	parameters					results		
Network	evolution steps	δ	p	β	ω	$\Delta\langle k\rangle$	$\langle C\rangle$	$\langle d\rangle$
link-duplication	4687	0.75	0.1	-	0.1	0.6	0.11	5.4
hybrid model	4687	0.6	0.1	0.2	-	0.4	0.11	5.3

Table 3.2: Parameters for model extentions of Sect. 3.4 to create a network with a giant component of ≈ 4700 nodes with an average degree $\langle k\rangle \approx 6.5$, their respective standard deviations, the clustering coefficient ($\langle C\rangle_{\text{data}} = 0.13$) and the average path length $\langle d\rangle$ ($\langle d\rangle_{\text{data}} = 6.1$) are shown.

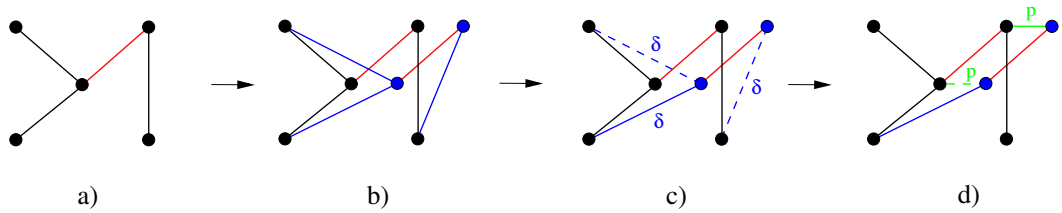


Figure 3.12: The link-duplication model: a) a link is selected randomly (red), b) the link and both nodes at its end are copied with their links (blue), c) new links are removed with probability δ (dashed line), d) links between original nodes and their copies are established with probability p (homodimer-links, green).

3.4.1 Link-duplication

In the models discussed in Sects. 3.2.1, 3.2.2 and 3.2.3, genes encoding one protein were copied and their interactions mutated by errors. But sections of the DNA that are duplicated and mutated do not necessarily encode only one protein. Fig. 3.11 illustrates the coding of two proteins that may share the same code for a binding domain. The algorithm that idealizes this mechanism is shown in Fig. 3.12. In the simulations, the applied algorithm is a mixture of a smaller fraction ω of link-duplication and the fraction $1-\omega$ of the gene-duplication model with homodimer-link II. Also in this model nodes are deleted after duplication if they remain unconnected to the network.

For the link-duplication model, estimates of the emergence of two proteins that share a same section of the DNA could not be found in literature. The chosen value of $\omega = 0.1$ seems to be reasonable to study the influences of this enhancement of the gene-duplication model with homodimer-link II. It was used to analyze the degree distribution, the degree correlation and the clustering coefficient and the motif structure (see Figs. 3.13). Parameters were chosen to fit real yeast interaction data and are shown

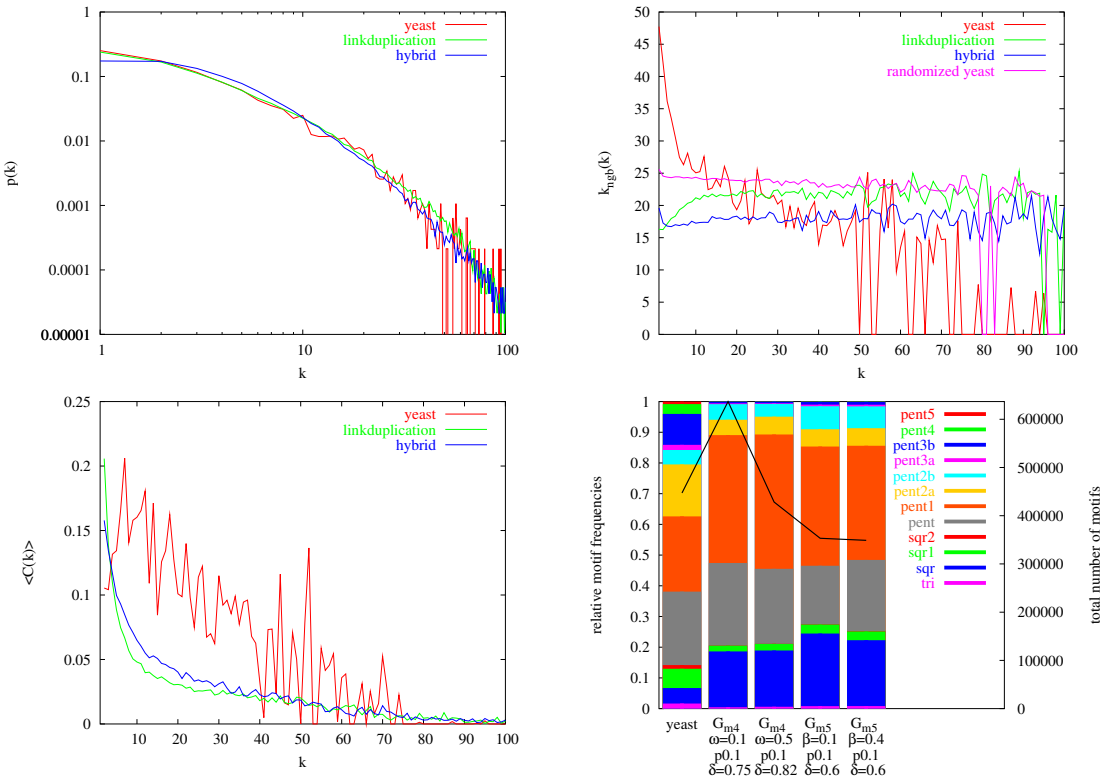


Figure 3.13: Degree distribution, clustering coefficient, degree correlation and motifs of the link-duplication $\mathcal{G}_{m4}$ and the hybrid network $\mathcal{G}_{m5}$. Chosen parameters are shown in Tab. 3.2.

in Tab. 3.2. Resulting giant components are of size $N = 4687$ and the average degree is $\langle k \rangle \approx 6.5$.

3.4.2 Hybrid model

The hybrid model is based on the gene-duplication model with homodimer-link II. Additionally, a link from the copied node to a randomly chosen other node is introduced (Compare Sect. 3.2.1). The parameters for this model δ, p and β are shown in Tab. 3.2. Resulting giant components are of size $N = 4687$ and the average degree is $\langle k \rangle \approx 6.5$.

3.4.3 Results

It turns out again that the degree distribution is always well reproduced although the influence of the additionally inserted random link in the hybrid model is visible (see Fig. 3.13 top left). The degree correlation (top right) for the hybrid model can be seen as a superposition of the degree correlations of the gene-duplication model with random-

and with homodimer link II. The degree correlation of the link-duplication model is strongly assortative for degrees $k \lesssim 15$. This is probably again explained with the strong occurrence of lowly connected nodes in the network. If a link and its nodes are duplicated, clusters emerge whose nodes are more likely to be lowly connected.

The clustering coefficient (Fig. 3.13 bottom left) has the same behavior for all gene-duplication models with homodimer-links. In the hybrid model, the additional random links have obviously a very small influence on the clustering coefficient and the motif-structure. However, not only in the hybrid model, but also in the link-duplication model it turns out that the fundamental gene-duplication model with homodimer-link II plays the major role in the formation of motifs and all extensions have no further influence. Only the total number of motifs is changed significantly for these model extensions. This is due to the fact that motifs mainly emerge through the simple duplication of a node. Hence, the total number of motifs is very dependent on the strength δ of the following deletion of links and δ had to be adjusted to ensure an average degree $\langle k \rangle = 6.47$.

The link-duplication and the hybrid model are introduced as an extension of the gene-duplication model with homodimer-link II by taking more biological effects into account that play a role during gene-duplication and mutation. Above all, that should improve the reproduction of the motif-structure in artifical protein interaction network models. As depicted in Fig. 3.13 (bottom right), the effect on network subgraphs is very small and the degree correlation cannot be described by these models.

4 Observational incompleteness

Although the simple gene-duplication and mutation mechanism disregards any selection process and does not take further regulatory mechanisms into account, it gives us a principle understanding of how evolution might have gone to work. However, it should be acted with caution when it comes to a biological interpretation of the fitted model parameter values. Attention has to be paid to the fact that the actual data of protein interactions includes a large number of false links. In Refs. [9, 10, 11], different methods are applied to give an estimate of the amount of links that are set in the real yeast datasets but do not exist (*false positives*), and those that exist but are not contained in the dataset (*false negatives*).

A vast amount of new protein interaction data has been obtained in recent years with the development of various *high throughput* methods. But this development came with its price. The reliability of this newly gained data is much lower than of older datasets obtained before high throughput methods were developed [9, 10, 11, 66, 67]. Current estimates state that about 50% of links in actual datasets are false positive links. These estimates base on a comparison of older so called *high confidence* data with high throughput datasets [9, 11]. Furthermore, the total number of links is estimated to be twice as large as todays 15 000 interactions, indicating a large number of false negative (missing) links in the actual datasets [9, 10].

4.1 Mapping methods

In Deane et al. [11], differences between high confidence interaction and high throughput interaction datasets could be explained by arguing that up to 70% of links in the yeast-two-hybrid dataset are false positive links. There it is assumed that interacting proteins show co-expression. Expression of a protein means that it is enabled to be produced over several transcription steps from the DNA (compare Sect. 3.1.1). The probability of two proteins to be functionally correlated is much larger if they are expressed at the same time and under the same conditions. Furthermore, functionally correlated proteins are very likely to interact directly. The frequency of co-expression is compared between a set of randomly chosen proteins and a set of proteins that are found to interact with high confidence, i.e. that exist in older datasets. The co-expression frequency of high throughput data leads to values in between and can be reproduced by a combination of assumed true positive and false positive interactions. False positive links are randomly chosen from the set of not existing interactions.

Another way to evaluate links in the dataset is to analyze if paralogs of interacting

proteins interact as well [11]. Paralogs are proteins that arose through gene-duplication of the same DNA section. These proteins are similar to each other and are therefore likely to have the same interaction partners because they derived from the same ancestor. It is noted that this method has the drawback that it applies only to the 50% of all proteins that actually have known paralogs. Furthermore, if two proteins interact, it is not assured that their paralogs do so as well. Nevertheless, estimates of the total number of false positive links can be given in this way. In high confidence data, this method verifies 50% of the links. In high throughput data the fraction of verified links is much lower. This leads to the conclusion, that 50% of the links in high throughput data are false positive. Of course, this holds only if errors in the dataset are induced at random. In Mering et al. [9], false positive links are also predicted to amount to 50% of known links by analyzing functional groups and the proteins belonging to them.

But not only false positive links influence the mapped data. In Mering et al. [9], the total number of links is estimated to be larger than 30 000, which is twice as many as the actual number of mapped interactions.

For an estimate of false negative links, Mering et al. [9] investigated the fraction of links that are found by more than one method compared to the respective fraction in high confidential datasets. Since this fraction is much lower in high throughput datasets, the number of 30 000 interaction could be approximated as a lower limit. A critical remark is in order here: In datasets used in [9], purified complexes were mapped as fully connected using the matrix method. This leads to a higher number of false negative links for this method and thereby to a larger overlap between different methods. Hence, the real overlap could be much smaller, implying a lower estimate less than 30 000 links if one follows the argumentation in [9]. Nevertheless, since high throughput methods find only a fraction of the high confidence interactions, there is still a significant number of false negative links.

The data achieved for protein interaction networks differs markedly between different mapping methods, see Sect. 3.1. All high throughput experiments base on one of these three. Thus, in newest network data almost all links are obtained by one or more of these methods. The overlap remains very small with about 3% of links found by more than one method. Approximately 6% of links are mentioned by more than one group of experimentalists.

This section aims to discuss how far biases within these methods have different influences on the properties of the network structure. This will become important in the following sections, where different error algorithms are discussed, which partly ground on assumptions on the errors made with single methods. Methods are analyzed in this section using the GRID dataset [49].

The dataset is separated into *sub-networks* containing only interactions found by one of the three methods. The degree distributions of these three sub-networks are shown in Fig. 4.2 (top). Apparently, they differ. If there were no biases of different methods, all sub-networks should have had the same properties.

For a further analysis of the sub-networks, it must be asked in how far it is justified to separate these sub-networks from each other. If the sub-networks can be largely separated from each other with only a few interconnections, different properties can be connected to one method. To draw conclusions from the degree distribution, e.g. whether one method is more likely to find highly or lowly connected nodes, it is important to note that the majority of links of the analyzed node is found by one respective method only. Furthermore, if the degree correlation, the clustering coefficient and the motif-structure are analyzed, also all neighbors of the questioned node should have mostly links found by the same method.

This separation is tested with a *purity measure P*, where L_m^i is the number of links at node i found by method m:

$$P^i = \sum_{m=1}^{3} \left(\frac{L_m^i}{\sum L_m^i} \right)^2 . \tag{4.1}$$

Hence, if all links of a node i are found by one method, the purity becomes $P^i = 1$, and if methods are equally distributed over all links, $P^i = 1/3$.

Fig. 4.1 shows the probability distribution of purity values for the GRID dataset. This measure gives values from $1/3$ to 1, but values of 0.66 already reflect fairly well purity with e.g. 80% of links belonging to one method and 10% to the others, respectively. For nodes with degree $k = 1$, a purity of $P = 1$ is obvious. Since many nodes in the network have degree $k = 1$ the significance of the purity distribution is tested with a randomized purity distribution. Therefore the degree of every node is kept fixed but to every link of the node, one of the three methods m is assigned with equal probabilities. This reflects the probability of one third to find a link with a certain method m in the yeast dataset and the number of links belonging to one method $\sum_i L_m^i$ is conserved. The comparison between yeast data and the randomized purity distribution shows, that the maximum at $P = 1$ can not only be explained as due to nodes with degree one. Moreover, much more nodes with low purity emerge in the randomized model. This justifies the separate analysis of the sub-networks.

The separation of all three sub-networks and the biases of every method become apparent if the overlap of the three methods is analyzed. In most cases, links could not be verified through finding them by more then one experimental group: In only 5 % of the cases, nodes were reported twice and 1 % triple to interact. Taking only citations of interactions into account where additionally different methods were used, 2.5 % of links are found twice and 0.2 % triple.

This can be explained only partly by the differences between different mapping methods because in high confidential datasets this overlap is much larger: If the more corrupted high throughput interactions are deleted from the dataset, the fractions of multiply found links rise to 13.6 % of interactions that are reported twice and to 1.9 % that are reported triple.

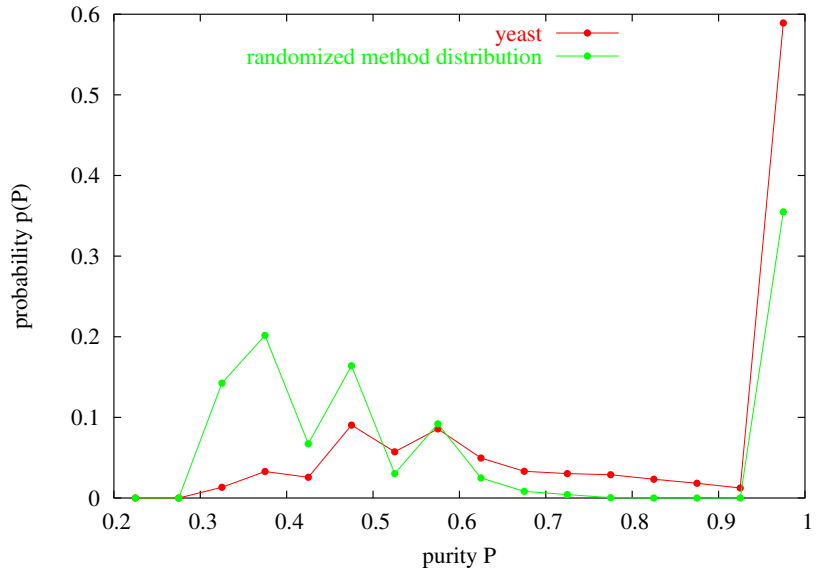


Figure 4.1: Distribution of purity values according to Eq. (4.1) for all nodes in the real yeast dataset (red) [49]. For comparison, the purity distribution is given after one of the three methods is randomly assigned to every link with the conserved the degree distribution of the network and total number of links belonging to one method (green). Values are integrated over a purity of $\Delta P = 0.05$.

In the following sections the biases of single methods will be discussed in greater detail with the degree distributions, the degree correlation, the clustering coefficient and the motif-structure.

4.1.1 Degree distribution

Although degree distributions for all sub-networks remain scale-free with exponential cut-off, differences in the slopes exist (see Fig. 4.2 top). The yeast-two-hybrid method focuses on single protein interactions. In contrast to the other methods, a network emerges that is rather large but sparsely connected with $N = 3632$ nodes and an average degree of $\langle k \rangle = 3.3$ compared to $N = 2266$, 1343 and $\langle k \rangle = 6.0$, 7.0 for affinity isolation and synthetic lethality, respectively. The same is observed in the degree distribution. For yeast-two-hybrid significantly more sparsely and less highly connected nodes are found. The two other methods focus on complexes. Thus, sparsely connected parts of the network remain undiscovered and a network emerges that is small but highly connected. For affinity isolation, this focus is obvious. For synthetic lethality sub-network the focus on complexes is assumed because proteins cannot be connected in a simple chain if the mutation of one of both proteins is not lethal. They are rather a

part of a complex that keeps its function and maybe becomes less effective after the mutation of the first and breaks down with the mutation of a second protein.

4.1.2 Degree correlation

The degree correlation differs markedly for different methods (see Fig. 4.2 middle). Surprisingly, it ranges from being constant for affinity isolation to a very disassortative behavior in case of synthetic lethality and yeast-two-hybrid. For affinity isolation and yeast-two-hybrid methods curves are not as expected.

In affinity isolation methods, links are set between a bait and all its preys according to the spoke rule (see Sect. 3.1.2) and many "stars" could be expected to emerge. This would lead to a disassortative degree correlation. But the respective degree correlation is rather constant. An explanation could be that in experiments several proteins within one protein complex are chosen as baits. This would lead to widely interconnected complexes and no disassortative degree correlation.

In contrast, the yeast-two-hybrid method focuses on single links and hence appears to be less biased in determining the degree correlation. But the degree correlation results to be much more disassortative than for every other subnetwork, including the high confidence network. The comparably smaller values over the entire degree correlation for yeast-two-hybrid data are due to the smaller average degree of the sub-network. This is comprehensible considering as example random networks, where $\langle k_{\mathrm{ngb}}|k\rangle = \langle k\rangle$. But also if there is a degree correlation that is not constant, the average neighbor degree $\langle k_{\mathrm{ngb}}|k\rangle$ depends on the average degree $\langle k\rangle$ of the entire network.

4.1.3 Clustering coefficient and motif-structure

Heavy fluctuations in the curve of the degree dependent clustering coefficient make it difficult to identify certain behaviors of the different curves (see Fig. 4.2 bottom). However, it is remarkable that for high confidential datasets the average clustering coefficient $\langle C\rangle$ is significantly larger. Thus, the real clustering coefficient for the entire network should be of the same order and larger than measured for current data. The clustering coefficient of synthetic lethality and affinity isolation methods are larger then in the yeast-two-hybrid subnetwork. This is in agreement with the assumption, that both focus on complexes.

The total number of motifs differs strongly for different methods, but can be well reproduced by randomizing the sub-networks with a fixed degree distribution (see Fig. 4.3). This casts a cloud especially over the synthetic lethality method because in this case also the relative motif frequencies could be reproduced very well. It poses the question if these measurements are that much corrupted with false links or if motifs in real yeast interaction networks indeed agree that far with randomized networks. Nevertheless, in the case of the yeast-two-hybrid and the affinity isolation methods the motif-structure

sub-network	N	$\langle k \rangle$	$\langle C \rangle$
DIP [18]	4687	6.47	0.13
GRID [49]	4814	7.2	0.12
yeast-two-hybrid	3632	3.3	0.05
synthetic lethality	1343	7.0	0.23
affinity isolation	2266	6.0	0.23

Table 4.1: Observables of two data sets [18, 49] and the sub-networks [49] to which only links of respective methods contributed are listed.

deviates noticeably from its randomized counterpart.

It is shown that properties indeed differ for different methods. Two explanations are obvious: Different methods either are better adapted to find protein interactions with certain properties or these methods are indeed so much biased that they reveal false network properties. In the following, it will be investigated which deviations may occur if several error algorithms are applied and if these algorithms reflect biases found for several mapping methods.

4.2 Random link removal, exchange and addition

The most general approach to simulate errors made during the mapping process is to generate noise on the underlying network $\mathcal{G}$ which is in the following assumed to be the exact representation of the original protein interaction network. Therefore, links are selected randomly and deleted to simulate false negative links, new links are added randomly for false positive links and with a combination of both processes, link exchange was simulated. The resulting network is then referred to as $\mathcal{G}^{\mathrm{rm}}$ (removal), $\mathcal{G}^{\mathrm{ad}}$ (addition) and $\mathcal{G}^{\mathrm{ec}}$ (exchange). The strength of the noise $\nu = \Delta L/L$ is the fraction of selected links ΔL to the total number of links L.

The impact of random link removal on the degree distribution of the gene-duplication-and-mutation network $\mathcal{G}_{\mathrm{m3}}$ of Ref. [15] is shown in Fig. 4.4 (top). Parameters for the underlying network are again $N = 4687$, $\delta = 0.58$ and $p = 0.1$. With increasing removal strength ν, the resulting degree distribution deviates more and more from its initial counterpart.

The degree distributions in Fig. 4.4 (top) with parameters $\delta = 0.58$ and $\nu > 0$ resemble those of Fig. 3.6 (top) with $\delta > 0.58$ and $\nu = 0$. In fact, the shown degree distributions with $\nu = 0.2$, 0.4, 0.6 and 0.8 perfectly match those resulting from $\delta = 0.62$, 0.66, 0.73 and 0.85, respectively. By looking at the degree distribution only, a network with

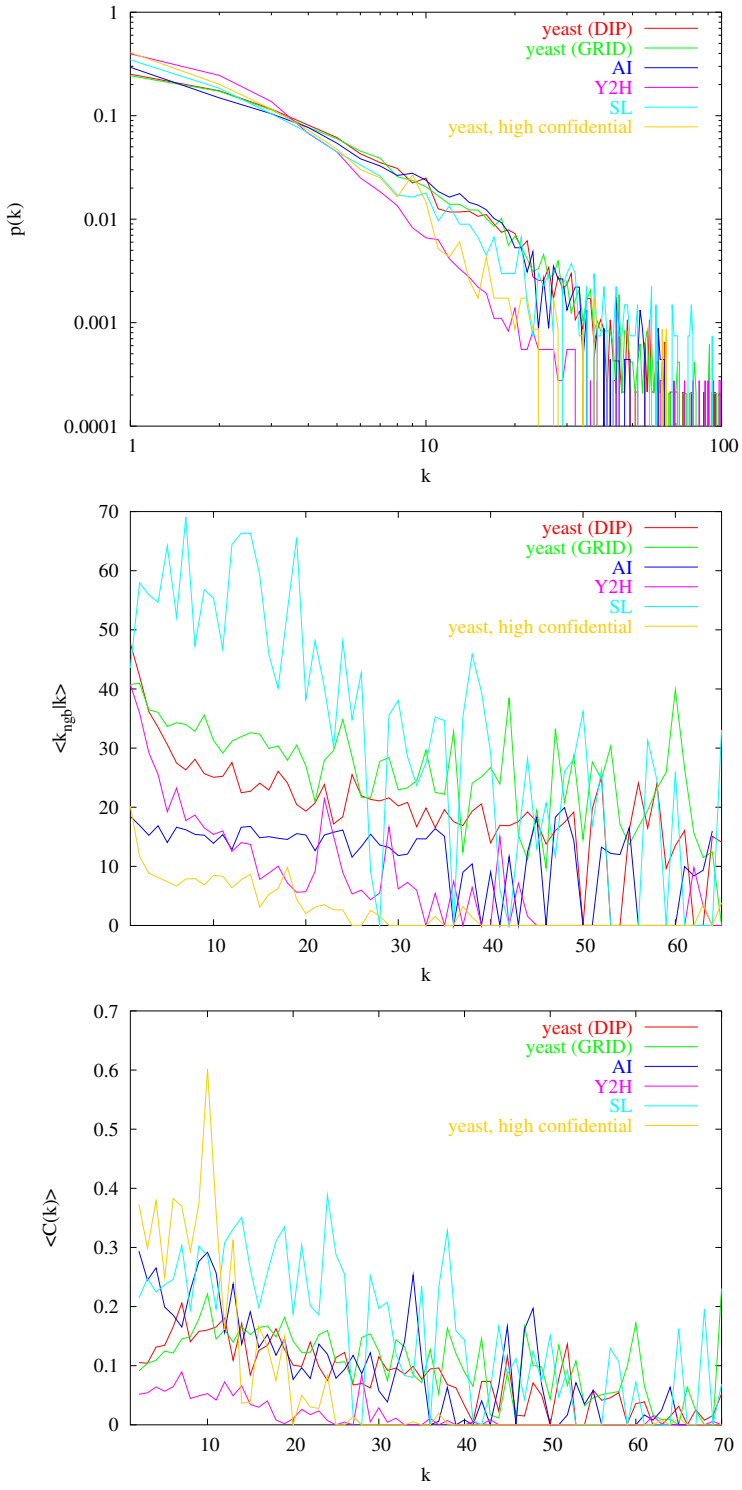


Figure 4.2: Degree distribution, clustering coefficient and degree correlation of the yeast sub-networks [49] that regard only links that are mapped by affinity isolation methods (AI), yeast-two-hybrid (Y2H) or by the synthetic lethality (SL) method. Furthermore, the respective curves for the high confidential dataset (i.e. without high throughput methods) is depicted.

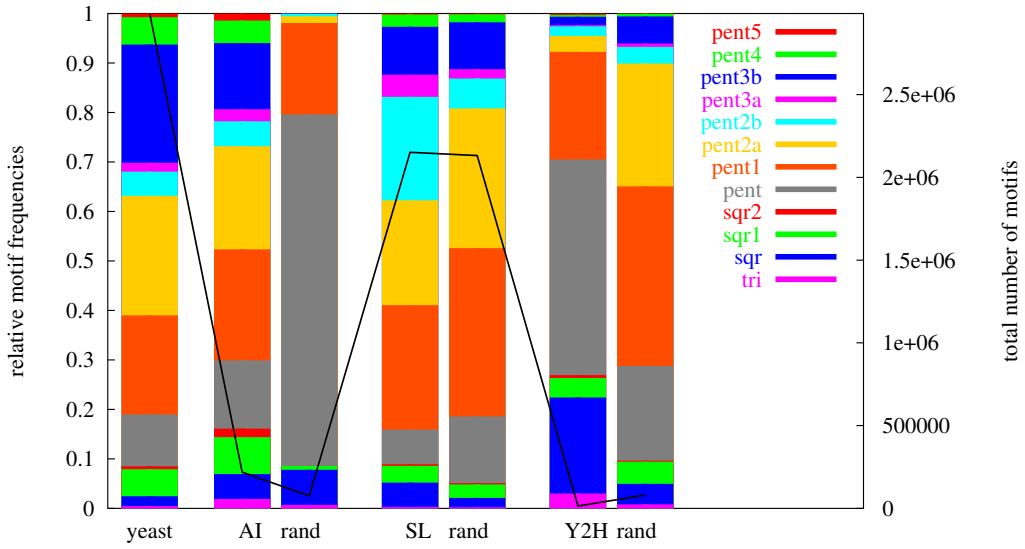


Figure 4.3: Motif structure of the yeast sub-networks [49] that regard only links that are mapped by affinity isolation methods (AI), yeast-two-hybrid (Y2H) or by the synthetic lethality (SL) method. For comparison, also the motif-structures of the respective randomized networks (rand) where the degree distribution is conserved are depicted.

specific δ, but subject to random link removal, appears like a network corresponding to a larger δ.

Note however that this comparison is not mature. With $\nu > 0$ the size $N_{\mathrm{m3}}^{\mathrm{rm}}$ of the giant network component, from which all degree distributions in Fig. 4.4 (top) have been sampled, is reduced. For $\delta = 0.58$ and $\nu = 0.2$, 0.4, 0.6 and 0.8 this results in $N_{\mathrm{m3}}^{\mathrm{rm}} \approx 4400$, 4000, 3350 and 2100 nodes, respectively. By model construction, the initial size $N_{\mathrm{m3}} = 4687$ is independent of the parameter δ. Consequently, the number of nodes contained in the giant component does not agree between the link-removed model network and the initial model network although their degree distributions match perfectly.

For a proper comparison, the model network reduced by random link removal should end up with the same average degree $\langle k \rangle$ and the same size N_{m3} for the giant component as the initial network model. For reference, it was chosen $\langle k_{\mathrm{data}} \rangle = 6.47$ and $N_{\mathrm{data}} = 4687$ as observed in the yeast data [18]. This requires the model network to have initially more nodes and links before random link removal sets in. Initial numbers of nodes and links are not independent of each other and require a careful tuning, to ensure that after random link removal a precision landing is made at the targeted $\langle k_{\mathrm{data}} \rangle$ and N_{data}. For example, for removal strengths $\nu = 0.2$, 0.31 and 0.395 the rescaled parameters are $(N, \delta) = (4950, 0.55)$, $(5100, 0.53)$ and $(5250, 0.51)$. The remaining

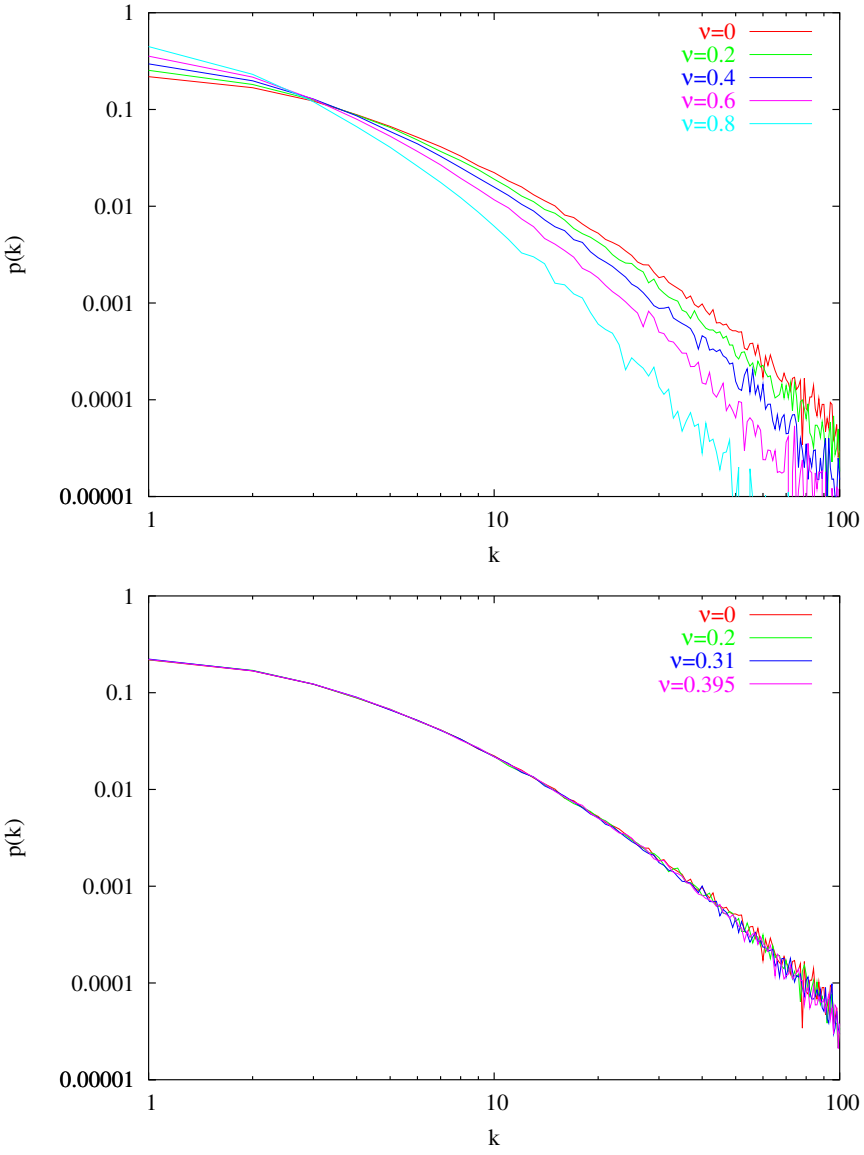


Figure 4.4: Degree distribution after link removal obtained for the giant component. Networks are not rescaled at the top ($G_{\mathrm{m3}}^{\mathrm{rm,r\acute{e}s}}$) and rescaled at the bottom ($G_{\mathrm{m3}}^{\mathrm{rm,res}}$). Start parameters for the rescaled networks are ($N = 4950$, $\delta = 0.55$), ($N = 5100$, $\delta = 0.53$) and ($N = 5250$, $\delta = 0.51$) for $\nu = 0.2, 0.31, 0.395$ respectively.

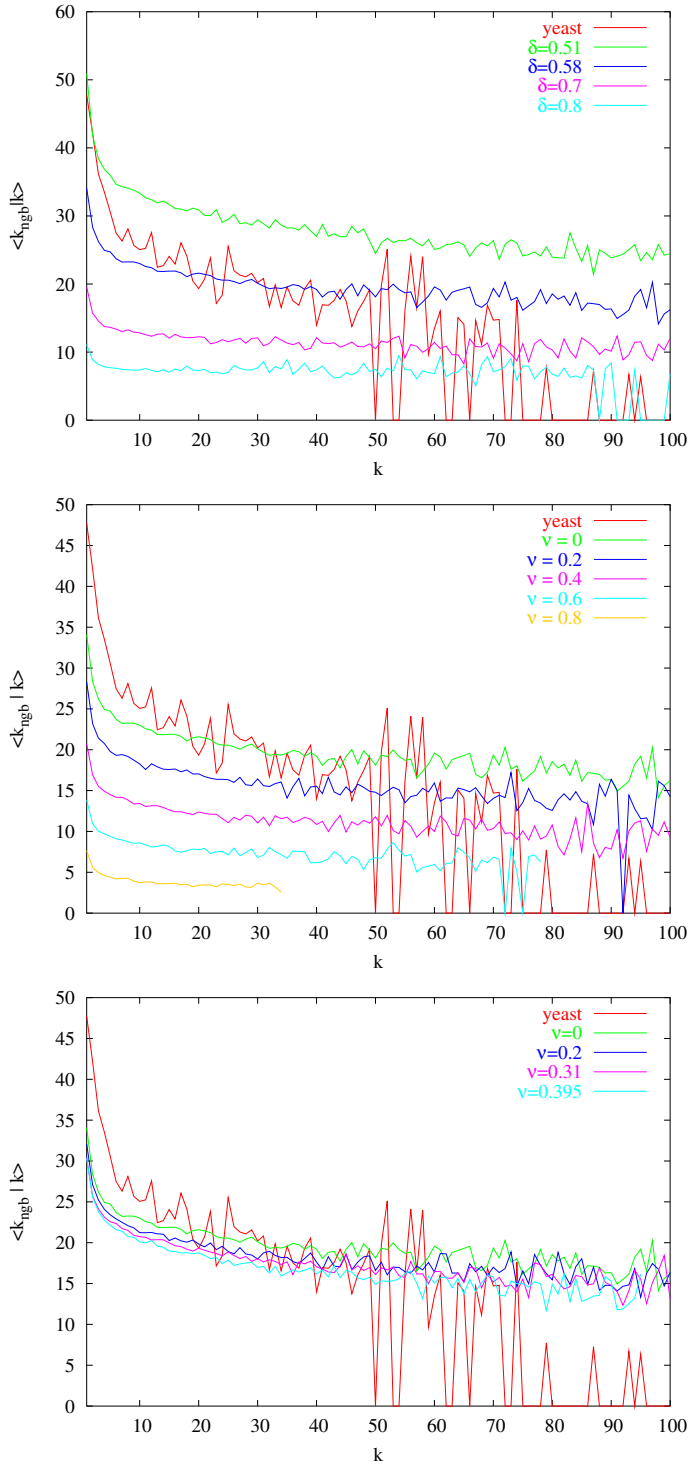


Figure 4.5: Influence of different parameters on the degree correlation during rescaling: The parameter δ is changed ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rescaled}}$) (top) and different removal strengths are applied ν ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,rés}}$) (middle). The combination of both (bottom) results in the final rescaled network under the influence of link removal ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,res}}$). Initial parameters for the rescaled networks are ($N = 4950$, $\delta = 0.55$), ($N = 5100$, $\delta = 0.53$) and ($N = 5250$, $\delta = 0.51$) for $\nu = 0.2$, 0.31 and 0.395 respectively.

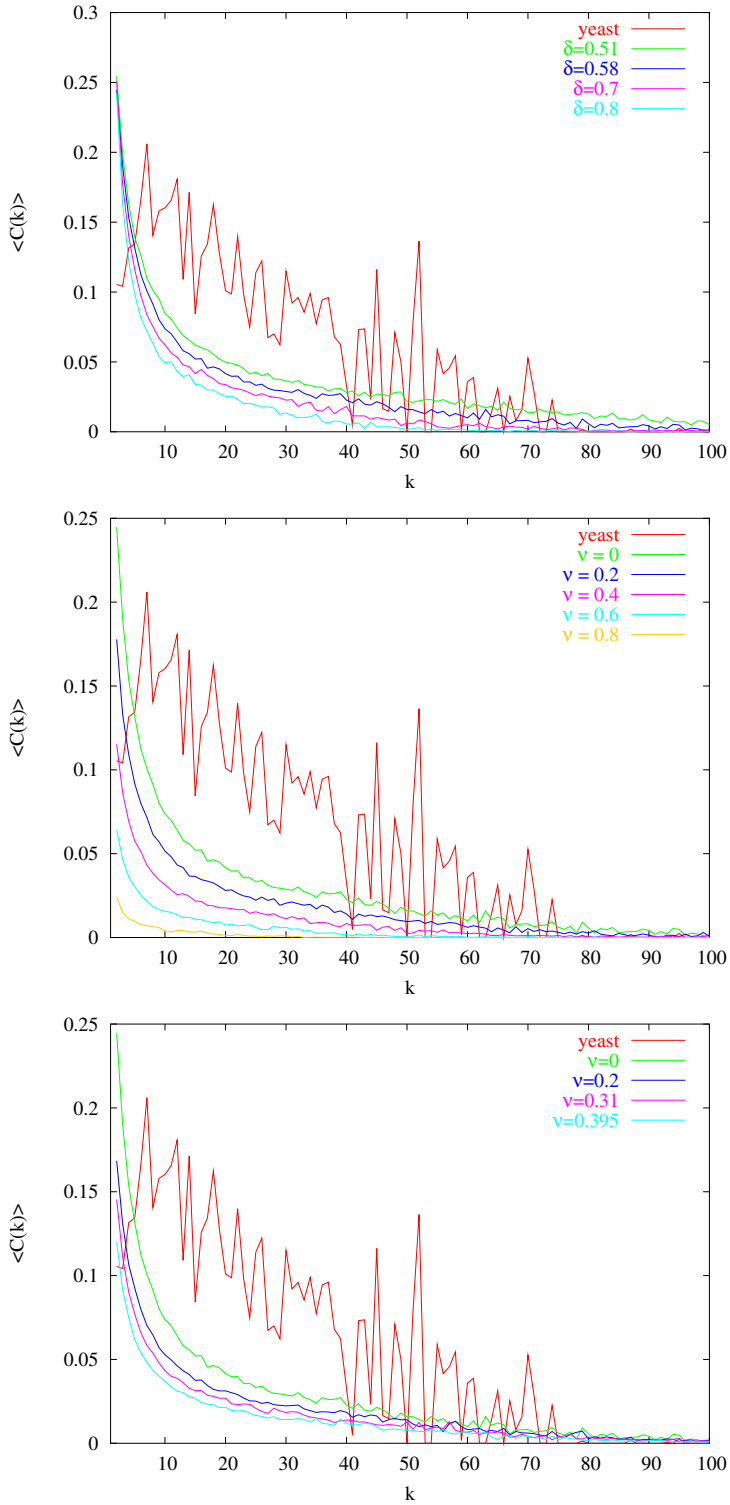


Figure 4.6: Influence of different parameters on the clustering coefficient during rescaling: The parameter δ is changed ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rescaled}}$) (top) and different removal strengths are applied ν ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,r\acute{e}s}}$) (middle). The combination of both (bottom) results in the final rescaled network under the influence of link removal ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,res}}$). Initial parameters for the rescaled networks are ($N = 4950$, $\delta = 0.55$), ($N = 5100$, $\delta = 0.53$) and ($N = 5250$, $\delta = 0.51$) for $\nu = 0.2$, 0.31 and 0.395 respectively.

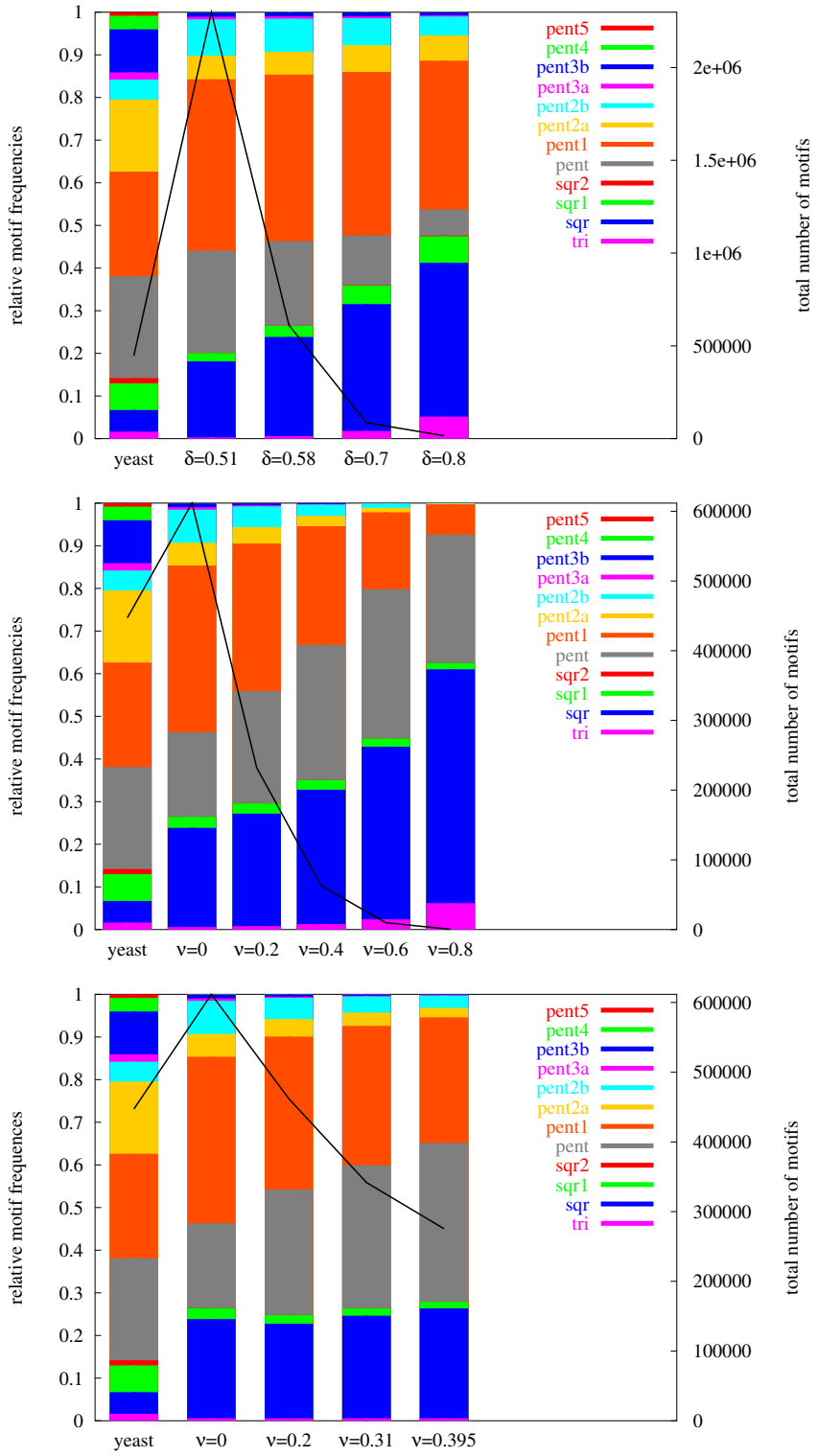


Figure 4.7: Influence of different parameters on the motif-structure during rescaling: The parameter δ is changed ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rescaled}}$) (top) and different removal strengths are applied ν ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,r\acute{e}s}}$) (middle). The combination of both (bottom) results in the final rescaled network under the influence of link removal ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,res}}$). Initial parameters for the rescaled networks are ($N = 4950$, $\delta = 0.55$), ($N = 5100$, $\delta = 0.53$) and ($N = 5250$, $\delta = 0.51$) for $\nu = 0.2$, 0.31 and 0.395 respectively.

parameter $p = 0.1$ has been kept fixed. Note that even larger removal strengths are not feasible for the chosen network model. It would require $\delta < 0.5$. In this regime, the model is not self-averaging any longer [13, 14]. In the following, non-rescaled networks will be referred to as $\mathcal{G}^{\text{rés}}$ and rescaled networks as $\mathcal{G}^{\text{res}}$.

Fig. 4.4 (bottom) illustrates the degree distributions obtained after random link removal has been applied to the parameter-rescaled model realizations. All distributions corresponding to different removal strengths collapse to one single curve. This curve collapse is somewhat surprising because by construction only the size of the resulting giant network component and the resulting average degree have been set the same. Each curve results from the interplay of two effects: initially, i.e. before random link removal sets in, a smaller δ leads to a flatter degree distribution (see again Fig. 3.6 top), which is turned into a steeper distribution once random link removal is applied (see again Fig. 4.4 top).

If the resulting degree distributions of the gene-duplication models had all been Poissonians, the curve collapse would have been straightforward to understand: The rate equation for random link removal [22]

$$\frac{\mathrm{d}p_k}{\mathrm{d}\nu} = \frac{k+1}{L(1-\nu)}p_{k+1} - \frac{k}{L(1-\nu)}p_k \tag{4.2}$$

is solved by $p_k = (\lambda^k/k!)e^{-\lambda}$ with $\lambda = \langle k\rangle = 2L(1-\nu)/N$. A Poissonian degree distribution remains Poissonian in spite of the rescaled parameter λ. It has been tested with simulations that this result, where all N nodes of the network enter, also carries over to an analysis based on the giant network component only.

Also for pure scale-free distributions $p_k \sim k^{-\gamma}$, the curve collapse can be reconstructed with a rescaling of model parameters. In the case of a growth process with preferential attachment $\pi \sim k+\lambda$, the model parameters are the number m of open links, with which a new node enters the network, and the attractiveness λ [4]. They determine the scale-free exponent $\gamma = 3+\lambda/m$. In Ref. [22] it has been shown that during random link removal, where the initial average degree $\langle k\rangle = 2m$ is reduced, the scale-free exponent is conserved. This implies that after random link removal with strength ν, the resulting network appears as one that has been grown with rescaled model parameters $m_{\text{rescaled}} = (1-\nu)m$ and $\lambda_{\text{rescaled}} = (m_{\text{rescaled}}/m)\lambda = (1-\nu)\lambda$. Since for not too large removal strengths the size of the giant network component remains very close to the total number of nodes N, this result also carries over to an analysis based on the giant network component only.

Although the small excursions to Poissonian and scale-free networks have shed some light on the nature of the curve collapse, its appearance in connection with gene-duplication networks remains without a deeper explanation. Nevertheless, from a pragmatic point of view it can be said: If a gene-duplication-and-mutation network is considered as the "true" network and false negatives are introduced in the form of random link removal, the resulting degree distribution appears as one obtained from the

same gene-duplication-and-mutation process, but with different parameters. Therfore, it would be inappropriate to give a biological interpretation of the magnitude to the extracted parameters.

The curve collapse motivates to look at observables beyond the degree distribution. Figs. 4.5, 4.6 and 4.7 provide an overview over the degree correlation, the clustering coefficient and the motif-structure, respectively. They compare $\mathcal{G}_{\mathrm{m3}}^{\mathrm{rescaled}}$ for various parameters δ (top), $\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,r\acute{e}s}}$ for different removal strengths ν (middle) and the final rescaled network $\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,res}}$ (bottom) for different removal strengths ν.

A similar finding as for the degree distribution is obtained for the degree correlation $\langle k_{\mathrm{ngb}}|k\rangle$. The form of the curve does change neither for different δ nor if the strength of link removal ν is changed. Nevertheless, the curve increases for all degrees k with lower deletion parameters δ and decreases with higher removal strengths ν. In this way, both effects compensate mostly. After rescaling, the degree correlation deviates only slightly under random link removal. Hence, the differences in the curves in Figs. 4.5 (top and middle) are mostly due to the different average degree of resulting networks and fit real yeast data very well.

The clustering coefficient behaves in similar directions (see Figs. 4.6). The number of triangles in the network is dependent on the average degree $\langle k\rangle$. After rescaling, the shift of the clustering coefficient is partly compensated if the parameters δ and ν are changed simultaneously. Nonetheless, also here a small deviation remains. Besides these findings, the clustering coefficient fails to reproduce real yeast data. This is not changed after random link removal.

A similar result as for the clustering coefficient is obtained for the total number of motifs. It is strongly dependent on the average degree but the decline of the motif number is not fully compensated if the network is rescaled (Figs. 4.7). A shift towards more complex motifs is observed with decreasing δ and towards simpler motifs with increasing ν. Changes in the motif frequency are partly compensated through rescaling and link removal. After rescaling, the three dominant contributions come from the motifs "sqr", "pent" and "pent1" (see Sect. 2.1.5). The relative frequency of "sqr" basically remains independent of ν . With increasing removal strength, the relative frequency of "pent" increases slightly, whereas that of "pent1" decreases to some small extend.

Random link addition was realized by distributing $L^{\mathrm{ad}} = \nu L$ links to all possible links. The resulting network $\mathcal{G}^{\mathrm{ad}}$ can be considered as the superposition of the underlying noise-free network and an added Erdös-Renyi network $\mathcal{G}_{\mathrm{ER}}$. Also the observed properties degree distribution, clustering coefficient and degree correlation appear as a superposition of the scale-free and a random graph, compare Figs. 4.8.

The same holds for random link exchange. Here a link is removed according to the link removal algorithm, and a new link is inserted according to the link addition algorithm. This is repeated $L^{\mathrm{ec}} = \nu L$ times, and a superposition of the graph under random removal $\mathcal{G}^{\mathrm{rm}}$ and random addition emerges $\mathcal{G}^{\mathrm{ad}}$. Results are also shown in Figs. 4.8. For random link addition a rescaling was abandoned because it would not

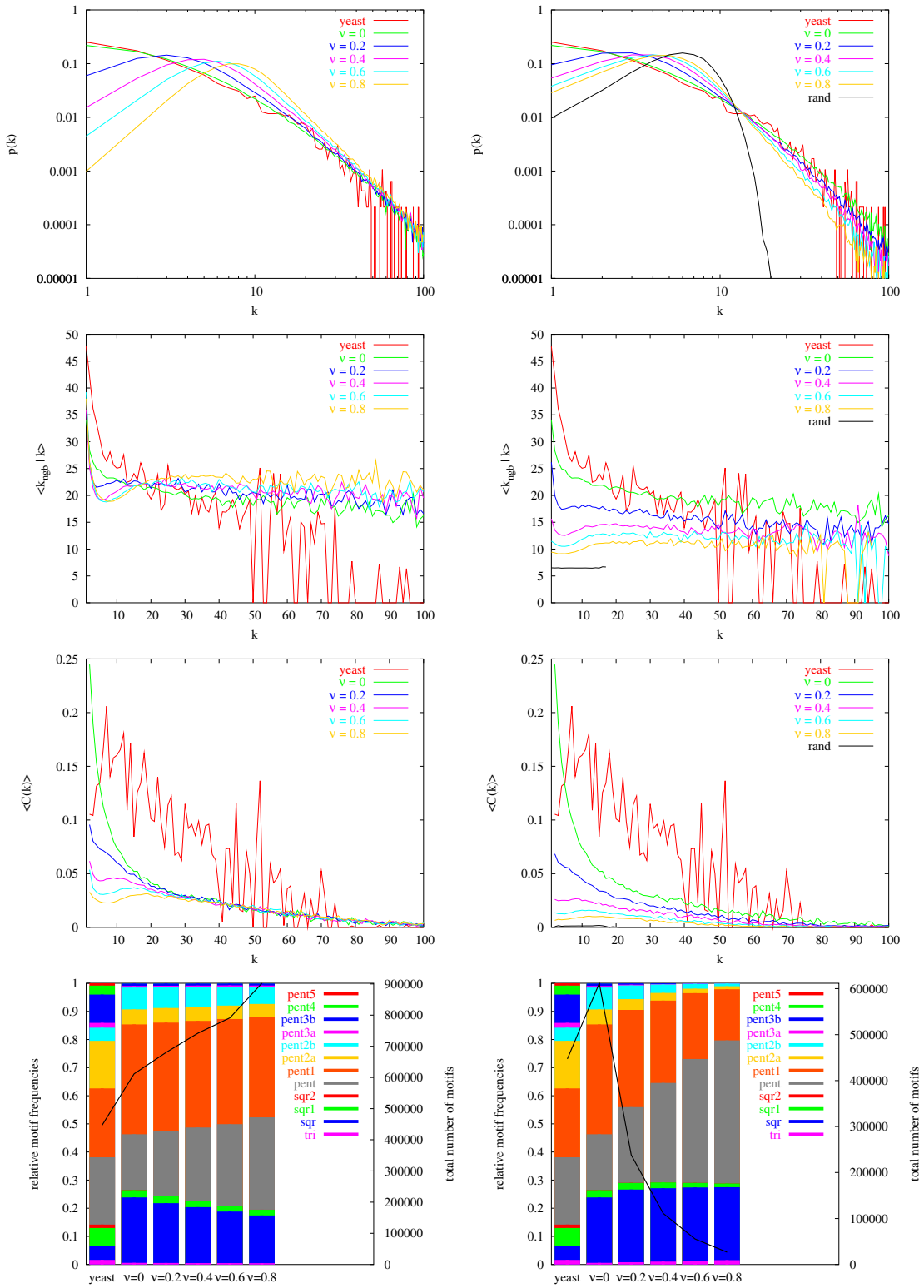


Figure 4.8: Influence of random link addition (first column, $\mathcal{G}_{\mathrm{m3}}^{\mathrm{ad,rés}}$) and random link exchange (second column, $\mathcal{G}_{\mathrm{m3}}^{\mathrm{ec}}$) on the degree distribution, degree correlation, clustering coefficient and motif-structure (from top to bottom row) of the underlying network $\mathcal{G}_{\mathrm{m3}}$.

lead to new conclusions. The difference between results for link addition and link removal derive mostly from the different average degree for a fixed noise strength ν.

The degree distributions are shown in Figs. 4.8 (top). With increasing strength of link addition (left) or exchange (right), the curve shifts from being scale-free with exponential cut-off towards a Poissonian. In between the degree distribution is a mere superposition of the extremes. In case of link exchange, the curve approaches the Poisson distribution much faster while in the case of link addition, highly connected nodes remain highly connected. The resulting degree distribution for link addition can be described as

$$p^{\mathrm{ad}}_{\mathrm{m3}}(k) = \frac{1}{1+\nu} p_{\mathrm{m3}}(k) + \frac{\nu}{1+\nu} p_{\mathrm{ER}}(k) \tag{4.3}$$

In turn for link exchange the degree distribution changes like

$$p^{\mathrm{ec}}_{\mathrm{m3}}(k) = \nu p^{\mathrm{rm}}_{\mathrm{m3}}(k) + (1-\nu) p_{\mathrm{ER}}(k) \tag{4.4}$$

if the possibility of exchanging an already exchanged link is disregarded, which is possible for small ν.

Also the limiting case $\nu \to \infty$ of the degree correlation after link addition (Fig. 4.8 second row left) or exchange (right) corresponds to an Erdös-Renyi network, where degree correlation is constant and equal to the average degree $\langle k_{\mathrm{ngb}}|k\rangle = \langle k\rangle$. Note that $\langle k\rangle$ increases with ν during link addition while it remains constant during link exchange.

The emerging minima can be explained as follows: During link addition or exchange, affected nodes are chosen randomly. Link addition and exchange influence much stronger nodes with low degree. Hence, their neighbor-degree is much faster shifted towards the limit for random networks, the average degree of the entire network. Especially for link addition, since no links are removed, highly connected nodes are significantly modified only for $\nu \gg 0$. For link addition, nodes with degree one do not gain a new link in contrast to nodes they are connected with. Thus, the degree correlation increases for nodes with degree one.

The same explanation holds for the clustering coefficient. The limiting case is again a constant and very low clustering coefficient like for Erdös-Renyi networks, where e.g. $\langle C_{\mathrm{ER}}\rangle = 0.001$ for $\langle k_{\mathrm{ER}}\rangle = 6.47$ and $N_{\mathrm{ER}} = 4687$. If a new neighbor is gained, it is very likely not connected to any other. A minimum emerges for $\nu < \infty$ and low degrees k.

With the addition of links and thus with the increase of the average degree $\langle k^{\mathrm{ad}}_{\mathrm{m3}}\rangle$, the total number of motifs increases. Looking at the motif structure again, a superposition of the $\mathcal{G}_{\mathrm{m3}}$ and a $\mathcal{G}_{\mathrm{ER}}$ network is visible since mainly the total number of "pent"-motifs increases, which is the motif that mainly emerges in Erdös-Renyi networks (compare Fig. 2.5). In contrast, under link exchange the total number of motifs decreases strongly. Its decrease due to link removal cannot be compensated by the small number of motifs that emerge in random networks. Again "pent"-motifs are increased in their frequency and the entire motif-structure approaches the limit of an Erdös Renyi network.

To summarize, the effects of random link removal on the underlying network model $\mathcal{G}_{\mathrm{m3}}$ are small. No improvement in describing real yeast interaction networks could be achieved. Only the motif-structure changes but it is shifted away from real yeast interactions. Also through the insertion of random links during random link addition and exchange, all observables are shifted towards Erdös-Renyi networks and hence away from real yeast data.

4.3 Random walk and avalanche subnetwork sampling

Using random link removal (Sect. 4.2), false negative links are simulated by erasing single links from the network model $\mathcal{G}_{\mathrm{m3}}$. Another way to simulate false negative links is to subsample links and nodes from the underlying network $\mathcal{G}_{\mathrm{m3}}$. The same occurs when real yeast interactions are mapped by finding single interactions in experiments and inserting them into the protein interaction map.

Measurements may be biased through already known proteins and their interactions. In this section, such errors are simulated by two different sampling algorithms: random walk and avalanche exploration. The first method identifies nodes and links on the underlying network by walking on it in random manner from one node to the other. All traversed links are taken over into the corrupted counterpart. In contrast, the avalanche exploration algorithm discovers a certain fraction of neighbors of a node and respective links. This is repeated continuously for every found node in the network.

The random walk starts from only one node or from multiple nodes (number of start-nodes N_S). As a second parameter, for every single walk the number of hops (walking length l_p) is fixed. The random walk algorithm walks from one node to the next. Going back to any previously found node is allowed as well as walking over the same link several times. Every link and every node that has been walked over contributes to the sampled network.

The exact realization is represented by the following algorithm:

for the number of random start-nodes N_S **do**
  for the walking length l_p **do**
    choose one of the neighbors $n_j \in \mathcal{N}_i$ of n_i at random with uniform probabilities
    map respective link l_{ij}
    hop to this neighbor n_j
  end for
end for.

The exploration of the underlying network is studied for different parameters N_S and l_p. In Figs. 4.9, the dependence of the total number of nodes $N_{\mathrm{m3}}^{\mathrm{rw}}$ of the sampled network on

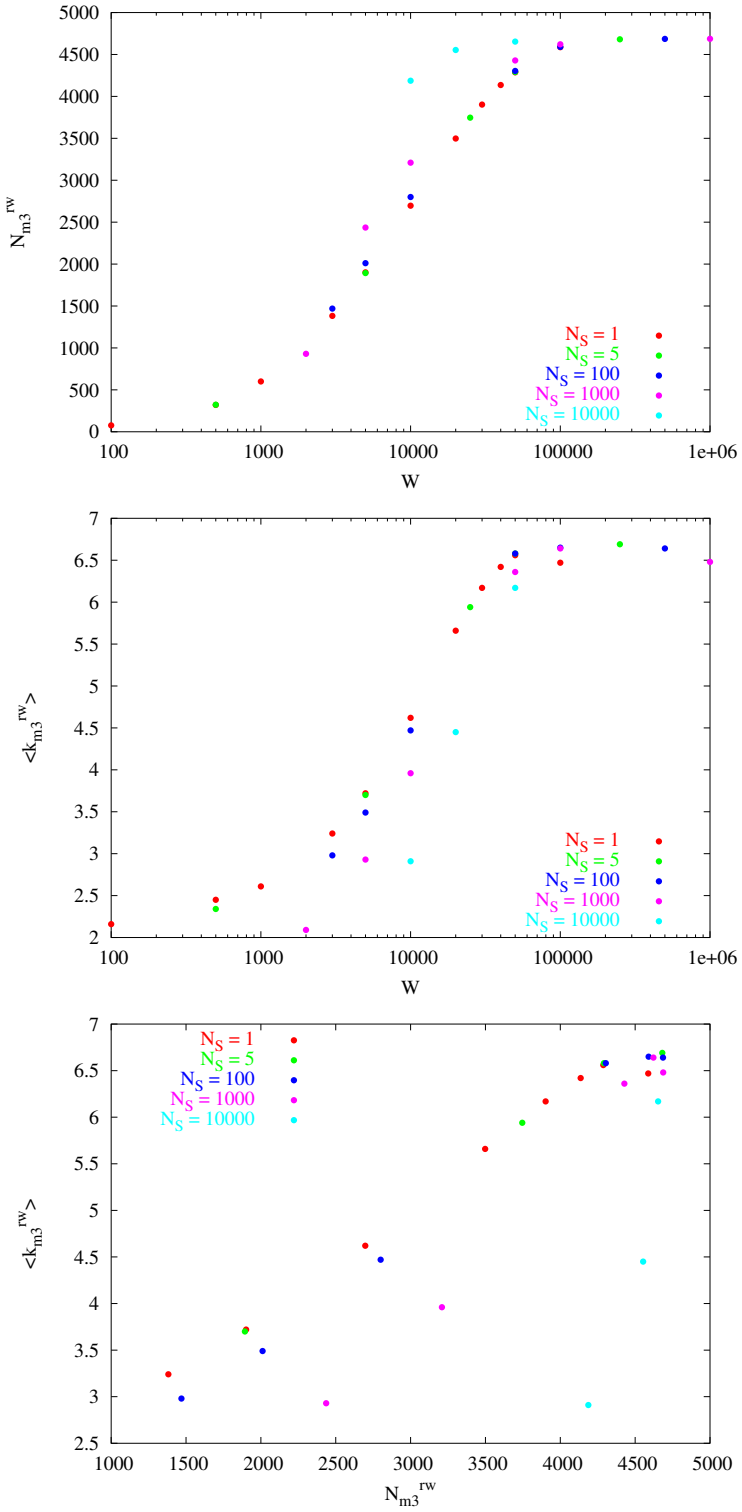


Figure 4.9: Exploration of the underlying network $\mathcal{G}_{\mathrm{m3}}$ with the random walk algorithm. The number of nodes $N_{\mathrm{m3}}^{\mathrm{rw}}$ that are explored after $W = N_S \cdot l_p$ exploration steps (top). The average degree $\langle k_{\mathrm{m3}}^{\mathrm{rw}} \rangle$ as a function of W (middle) and $N_{\mathrm{m3}}^{\mathrm{rw}}$ (bottom). The colors of the points represent the chosen number of start-nodes N_S.

the exploration length $W = N_S \cdot l_p$ (top) and of the average degree $\langle k^{\mathrm{rw}} \rangle$ on W (middle) is shown. It turns out that the number of nodes in the final network is fairly independent from the chosen number of start nodes N_S, as long as the product $W = N_S \cdot l_p$ remains constant. As expected, with increasing W the number of explored nodes increases until the network becomes fully explored.

The average degree of the sampled network $\mathcal{G}_{\mathrm{m3}}^{\mathrm{rw}}$ increases with increasing exploration length until it reaches a maximum for $W \approx 300\,000$. Note that this is already much larger than the number of links in the underlying network with $L \approx 15\,000$. This average degree $\langle k_{\mathrm{max}}^{\mathrm{rw}} \rangle = 6.7$ is larger than the average degree of the underlying network $\langle k_{\mathrm{m3}} \rangle = 6.47$. For even larger exploration lengths the average degree declines to the value of the underlying network. During the exploration, sparsely connected parts of the network are more likely to remain undiscovered, which increases the average degree $\langle k^{\mathrm{rw}} \rangle$ of the network. This is also visible in the degree distribution of Fig. 4.10 (top left).

For a larger number of start-nodes (see $N_S = 10\,000$ in Fig. 4.9 bottom), the network becomes faster explored in respect to the number of nodes, but the average degree of the resulting network is smaller. This is easily comprehensible because the more the rather short random walk procedures are distributed over the network, the better is the exploration of nodes, but not all links in a cluster are explored that way.

When avalanche exploration has been applied $\mathcal{G}_{\mathrm{m3}} \rightarrow \mathcal{G}_{\mathrm{m3}}^{\mathrm{ep}}$, the idea that protein interactions are found only through previously found proteins is followed again. It is assumed that for every exploration step a fraction of the total number of interacting neighbors is discovered. The respecting algorithm works as follows: A node is selected at random, and the fraction σ of its neighbors is tagged and put into a First In First Out (FIFO) queue. Links towards the tagged nodes are copied into the sampled network. This procedure is repeated for the first node in the FIFO queue, which is removed from it afterwards and cannot be explored again. The algorithm ends if all nodes in the queue are processed. In the case $N_S > 1$ a new node is chosen at random and the entire algorithm is repeated again in the same manner for the number of start-nodes N_S. This is represented by the following code:

for the number of start-nodes N_S **do**
  choose a node n_i at random and put it in the FIFO queue
  repeat
    choose first node in FIFO queue
    choose neighbors $\mathcal{N}_i^\sigma$ of the node n_i randomly with probability σ
    for these neighbors **do**
      map respective links $l_{ij} : n_j \in \mathcal{N}_i^\sigma$ to these neighbors
      if these neighbors $\mathcal{N}_i^\sigma$ have not been explored already before **then**
        put them at the end of the FIFO queue
      end if
    end for
  until FIFO queue empty
end for.

If the actual analyzed node has a link to an already explored neighbor, the link is mapped, but this neighbor is not inserted into the FIFO queue twice. The FIFO queue processes the nodes in order of their insertion starting with the first one. Other ways of dealing with the queue are imaginable, e.g. a last in first out queue. Previous studies [32] suggest that also for this algorithm the order of analyzed nodes does not significantly affect the outcome.

Fig. 4.11 shows the influence of different parameters N_S and σ on the exploration of the network. The groups of points with $N_S = 1, 3, 5$ and 10 belong to values of $\sigma = 0.2, 0.4, 0.6$ and 0.8 from the bottom left to the top right, respectively. With increasing σ, the number of nodes N and the average degree $\langle k \rangle$ in the sampled network increases. The ratio between N and $\langle k \rangle$ is constant and there is no significant dependence on the number of start-nodes. Since points in the graph for $N_S = 1, 3, 5$ and 10 and constant σ match very well, the exploration is not dependent on the number of start-nodes, but only on σ. Deviations emerge for $N_S = 100$ start-nodes.

Both algorithms are applied here to gene-duplication networks with homodimer-link II. Parameters have been chosen again to match best real yeast interaction data: $N_{\mathrm{m3}} = 4687$, $\delta = 0.58$ and $p = 0.1$. The resulting degree distribution, the degree correlation, the clustering coefficient and the motif-structure of the obtained networks are shown in Figs. 4.10.

Networks are rescaled in the average degree and the number of nodes to ensure a network of the same average degree and size as in real yeast data. Start parameters in case of random walk are $N_S = 1$ and ($N = 6050, \delta = 0.55$), ($N = 6400, \delta = 0.53$) and ($N = 6600, \delta = 0.51$) for $l_p = 29\,000, 25\,500$ and $23\,500$, respectively and in the case of avalanche exploration $N_S = 1$ and ($N = 5900, \delta = 0.55$), ($N = 6350, \delta = 0.53$) and ($N = 6625, \delta = 0.51$) for $\sigma = 0.46, 0.36$ and 0.298, respectively. Since the underlying networks are limited to a maximum average degree $\langle k_{\mathrm{m3,\,max}} \rangle \approx 10$, the exploration level in respect to the discovered nodes and links is always rather high. There-

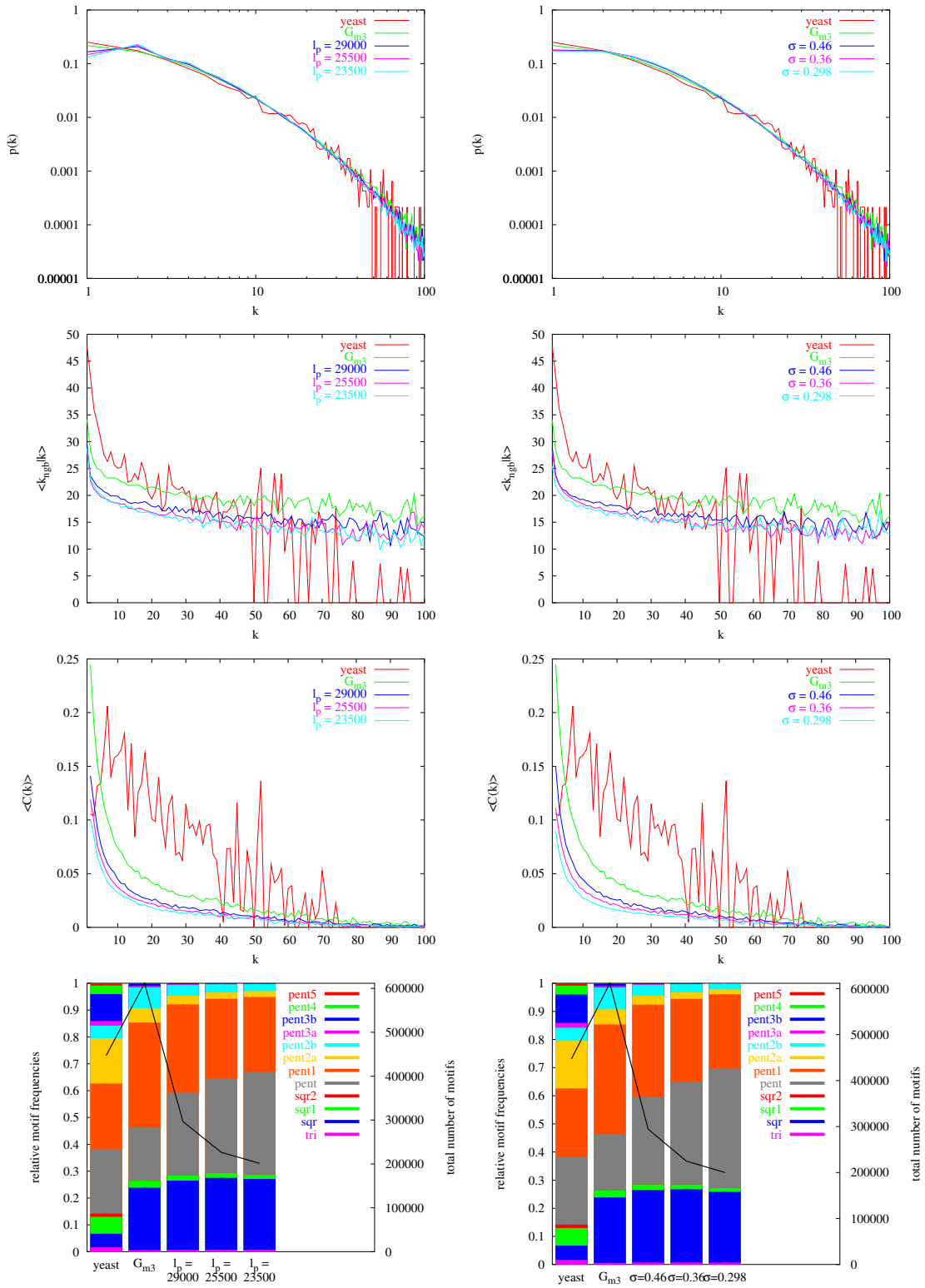


Figure 4.10: Degree distribution, degree correlation, clustering coefficient and motif-structure (from top to bottom row) of the rescaled gene-duplication network with homodimer-link II $\mathcal{G}_{\mathrm{m3}}^{\mathrm{rescaled}}$ subject to random walk ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{rm,res}}$, left) and avalanche exploration ($\mathcal{G}_{\mathrm{m3}}^{\mathrm{ep,res}}$, right).

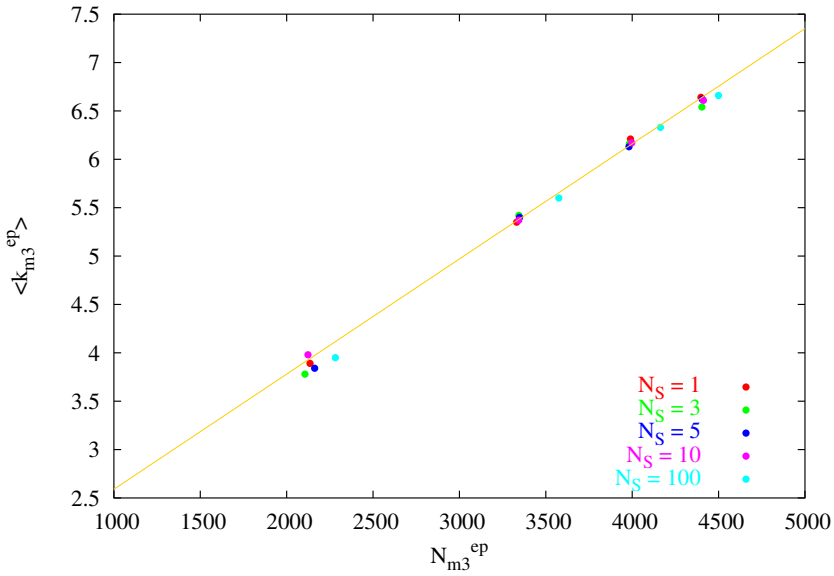


Figure 4.11: Exploration of the underlying network $\mathcal{G}_{\mathrm{m3}}$ with the avalanche exploration algorithm. The dependence of the average degree $\langle k^{\mathrm{ep}}_{\mathrm{m3}} \rangle$ on the size of the explored network $N^{\mathrm{ep}}_{\mathrm{m3}}$ is depicted. The colors of the points represent the chosen number of start-nodes N_S. Groups of points at $N_S \approx 2200, 3400, 4000$ and 4400 (without the cyan points) derive from explorations with $\sigma = 0.2, 0.4, 0.6$ and 0.8 respectively. The constant ratio between $\langle k^{\mathrm{ep}}_{\mathrm{m3}} \rangle$ and $N^{\mathrm{ep}}_{\mathrm{m3}}$ is depicted with a linear fit $\langle k^{\mathrm{ep}}_{\mathrm{m3}} \rangle = a N^{\mathrm{ep}}_{\mathrm{m3}} + b$ with $a = 0.0012$ and $b = 1.4$.

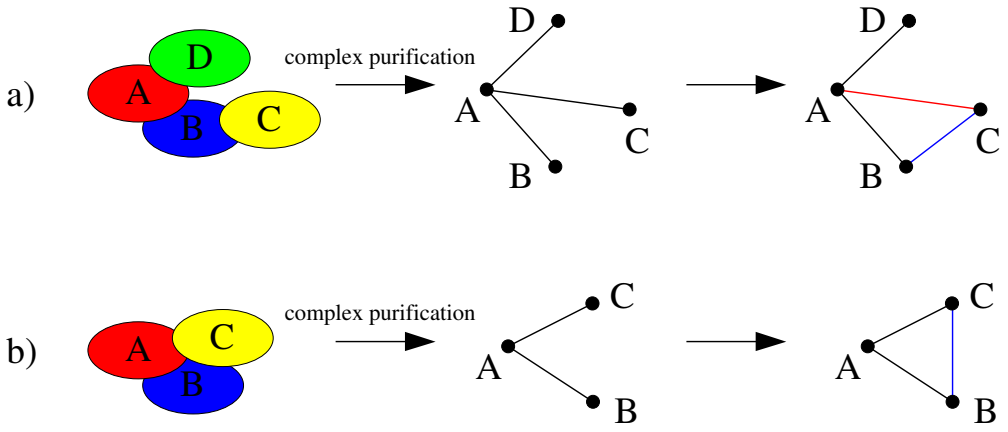


Figure 4.12: Affinity isolation methods may lead to wrong link assignments. (a) Bait protein A and prey proteins B, C, D bind for a complex. Assigned links reflect the bait-prey relationship. However, A does not directly bind to C (false positive, red). It is B, which binds to C (false negative, blue). (b) For the complex ABC links are only assigned between bait A and preys B, C. Link B-C is missing, resulting in a false negative (blue).

fore, rescaled simulations where for example l_p is low and only a small fraction of the network is explored could not be carried out. But the analysis for random walk and avalanche exploration without rescaling shows that there is no qualitative difference in the resulting network if the underlying network is either slightly or almost fully explored (data not shown).

Besides the small deviations for nodes with low degree, in the degree distribution, the same curve collapse is observed as after link removal if the underlying networks are rescaled (see Sect. 4.2). Also the results for the degree correlation, clustering coefficient and motif-structure, are identical to findings for random link removal, which is a surprising result. Obviously, due to randomness, the exact realization of the sub-network sampling algorithm is not important.

4.4 Spoke link rearrangement

So far, the modeling of observational incompleteness has taken into account random errors only. In this section, a specific link rearrangement will be discussed, which is directly motivated from the shortcomings in the generation and interpretation of protein-interaction data.

In Sect. 3.1, three methods to identify protein interactions have been explained. The link rearrangement, discussed here, is motivated by uncertainties that occur if affinity isolation methods are used. The central problem is outlined in Fig. 4.12. For a protein A, which is used as bait in the cell lysis, a certain number of prey proteins (BCD)

are identified that form a complex with A. Even if all proteins in the complex are identified, there is no information provided which of the proteins interact directly. In the commonly used "spoke" rule [10], direct links are assigned between the bait and all its preys. This method does not take into account the possibility that the bait is not directly interacting with all the preys but via intermediate proteins. This results in false positive and negative links (Fig. 4.12a). Moreover, possible interactions between the prey proteins themselves are also neglected, resulting in even more false negative links (Fig. 4.12b).

Similar effects occur with the yeast-two-hybrid [9] and the synthetic lethality method [62]. Although the yeast-two-hybrid method characterizes the interaction between two target proteins, no assurance can be given that this interaction is not provided by an intermediate protein. With regard to the synthetic lethality method, an interaction is assumed between two functional correlated proteins, but even if they are part of the same complex, it is not clear if a direct interaction exists.

To study the influence of this effect on the network topology, a local random link rearrangement is proposed, which is referred to as spoke link rearrangement. After selection of an initial node $i \in \mathcal{N}$ (bait), one of its direct neighbors $j \in \mathcal{N}_i$ (prey) is chosen at random. The latter continues to choose randomly one of its first neighbors $k \in \mathcal{N}_j \setminus i$, excluding, of course, the initial node. Afterwards two cases have to be distinguished: If the last node k is a second neighbor of the bait node i, a false-positive link l_{ik} between these two nodes is introduced, and the old link between the two prey nodes l_{jk} is removed to gain false-negative status (see again Fig. 4.12a). In the other case, the second prey node k turns out to be a first neighbor of the bait node i (i.e. $k \in \mathcal{N}_i$), upon which only the link between the two prey nodes l_{jk} is removed and becomes false-negative (see again Fig. 4.12b).

To apply this algorithm, a rule needs to be defined to choose possible bait proteins. In yeast data [18], it turns out that $\approx 25\%$ of the proteins have been used as baits. Their degree distribution p_k^{bait} is slightly different from the overall degree distribution p_k. Fig. 4.13 shows the degree distribution of the baits in real yeast data (red) [49]. The green curve is identical to the degree distribution of the entire dataset.

It is now assumed that baits are preferentially chosen from the entire set of proteins according to

$$p_k^{\text{bait}} \sim k^{\alpha} p_k^{\text{yeast}}. \tag{4.5}$$

This indicates that a bait node i with degree k_i might be picked with the preferential bias

$$\Pi_{i,\alpha}^{\text{bait}} = \frac{k_i^{\alpha}}{\sum_{j=1}^{N} k_j^{\alpha}}\,. \tag{4.6}$$

In blue, magenta and cyan in Fig. 4.13, degree distributions are fitted according to Eq. (4.5) with $\alpha = 0.3, 1$ and 3. The green curve corresponds to an $\alpha = 0$. This suggests values $0 \lesssim \alpha \lesssim 1$.

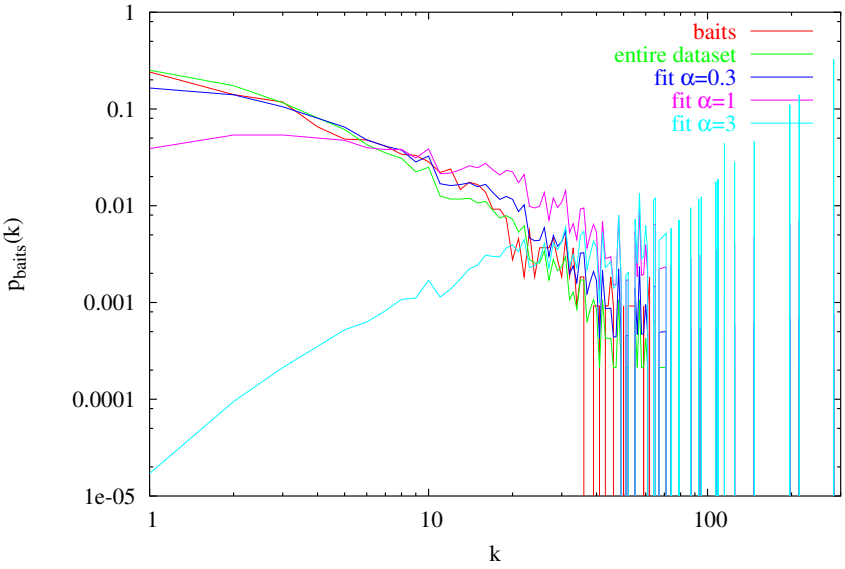


Figure 4.13: Comparison of the real bait degree distribution [49] to the bait distributions that result from the degree distribution of the entire yeast dataset according to Eq. (4.5) and a different choice of α.

When it comes to an estimate of the exponent α, it has to be considered that the degree distribution p_k^{yeast} of the real protein interaction network enters in Eq. (4.5), but only the degree distribution of yeast data p_k^{data} is available. Since yeast interaction data is corrupted by many false positive and false negative links, also the degree distribution extracted from yeast data is corrupted. Hence, the estimate of α by means of the corrupted degree distribution in turn influences the degree distribution and thus the choice of α itself recursively. To cover all possible changes of the degree distribution, probabilistic bait selection with $\alpha = 0$, 1 and 3 is discussed in the following.

Application of the spoke algorithm on the network model $\mathcal{G}_{\text{m3}}$

The combination of the biased bait selection (Eq. (4.6)) and the spoke link rearrangement process are applied to the network structure obtained with the gene-duplication-and-mutation model $\mathcal{G}_{\text{m3}}$ of Ref. [15]. Resulting networks are referred to as $\mathcal{G}_{\text{m3}}^{\text{sp}}$. Again, model parameters are chosen to match the yeast data, i.e. $N_{\text{m3}} = 4687$, $\delta = 0.58$ and $p = 0.1$. The rearrangement strength $\nu = \Delta L/L$ counts the relative number of bait selections implying link rearrangement or removal.

Note that the giant component of the network does not change with ν and remains at its initial value $N_{\text{m3}} = N_{\text{m3}}^{\text{sp}}$. Both subprocesses of the spoke link rearrangement always keep the three involved nodes (bait-prey-prey) connected to the overall network.

Since link removal is always included in the spoke link rearrangement (see Fig. 4.12b), the average degree decreases with increasing ν from its initial value $\langle k_{\mathrm{m3}} \rangle = 6.47$. For the already very large rearrangement strength $\nu = 0.8$, we arrive at $\langle k^{\mathrm{sp}}_{\mathrm{m3}} \rangle = 6.40, 5.97$ and 3.12, for $\alpha = 0$, 1 and 3, respectively. This α-dependence is easily explained: Due to gene-duplication, many square-motifs are built into the initial network structure (see Sect. 3.3.4). Upon application of the spoke link rearrangement, the square-motifs disappear and a lot of triangles emerge instead. Consult also Figs. 4.17 and 4.18 which show the clustering coefficient as an increasing function for low values of ν and the relative frequency of square-motif as a decreasing function of the rearrangement strength. Note, that the further increase of the clustering coefficient in case $\alpha = 3$ is for other reasons as explained later. If a previously picked bait node is selected again, which does rarely happen for a small α, but more often for a large α, the bait and the two preys find themselves more likely in a triangle. Only the link between the two preys is removed, but no new link is introduced. This explains the ν- and α-dependence of the average degree.

The strong decrease of the average degree for $\alpha = 3$ makes rescaling necessary to keep the network comparable. Due to the limit of the deletion parameter $\delta > 1/2$ during network evolution, the level of link rearrangement is limited to $\nu = 0.51$ because the underlying initial network is limited to a maximum average degree of $\langle k_{\mathrm{m3}} \rangle \approx 10$. Initial parameters are $\delta = 0.55, 0.53$ and 0.51 for $\nu = 0.27, 0.35$ and 0.51 respectively.

Degree distribution

The resulting degree distribution is discussed first for the three different outcomes for $\alpha < 1$, $\alpha \approx 1$ and $\alpha > 1$, followed by an analytical solution for the stationary limit that could be found for $\alpha = 0$ and $\alpha = 1$. The following paragraphs which base again on simulations give a deeper view into the network topology.

Figs. 4.14 show the dependence of the degree distribution on the rearrangement strength ν for $\alpha = 0$, 1, 3. If $\alpha < 1$ (see $\alpha = 0$), nodes with a small degree are favored to be chosen as baits and gain links. The degree of highly connected nodes is constantly diminished when they are chosen as first neighbors of the baits and their links are reconnected to the lowly connected baits. Lowly connected nodes disappear, because they gain links. In the limit of $\nu \to \infty$ an equilibrium emerges, where on average an equal number of links is removed from and attached to all nodes. Very highly and lowly connected nodes have widely disappeared. The degree distribution becomes a Poissonian and a random network appears. This is shown in Fig. 4.14 (top) for $\nu = 0.1, 0.3, 0.5, 0.7$ and 5. Only the first four values of ν are of biological relevance. $\nu = 5$ is depicted to illustrate the limiting case.

For $\alpha > 1$ (see $\alpha = 3$, Fig. 4.14 bottom), the few highly connected nodes gain more and more links by attaching second neighbors directly to them. Note that with $\nu = \Delta L/L$ even for a low $\nu = 0.27$ the same high-degree bait is selected again and again. After

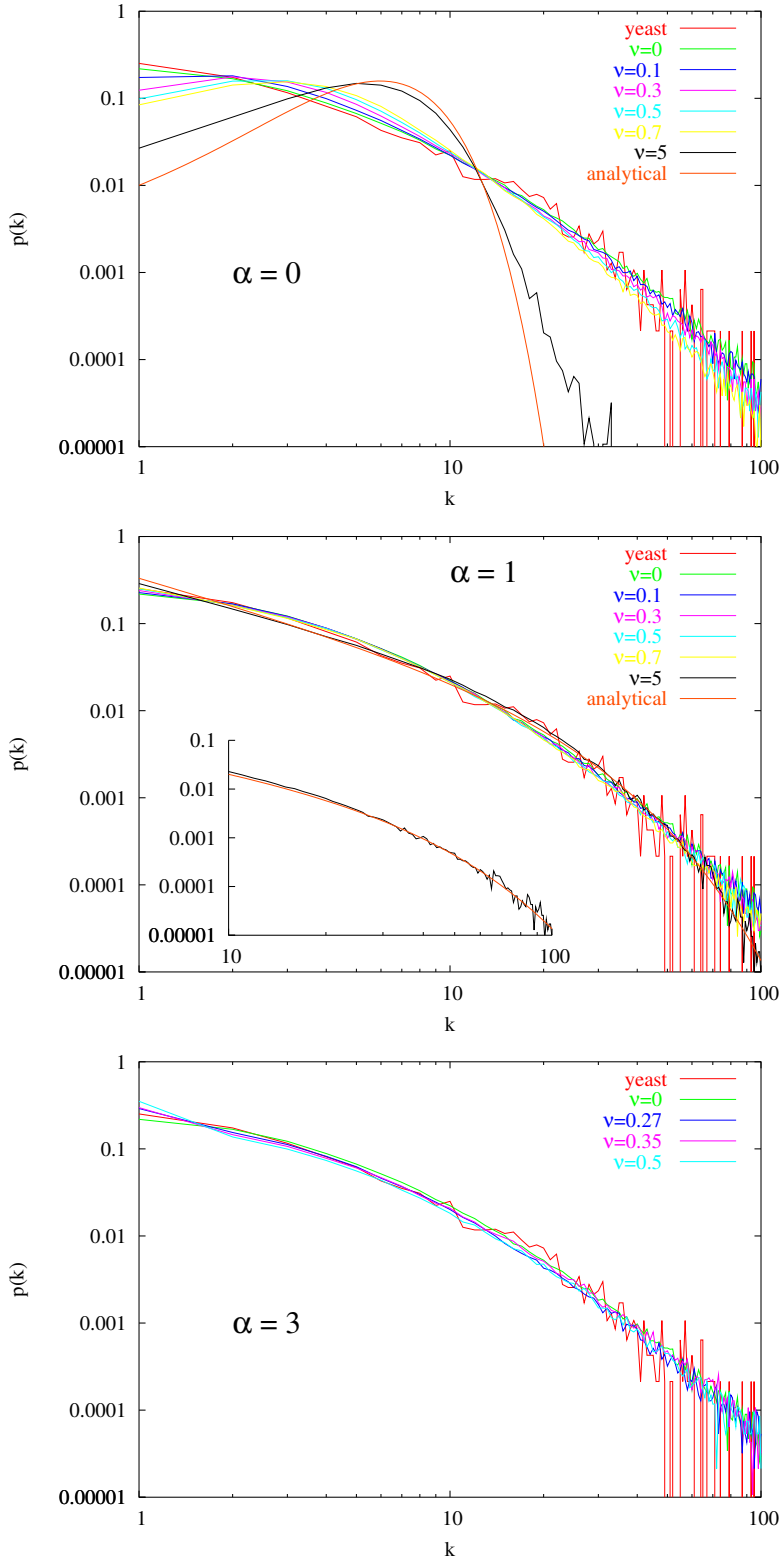


Figure 4.14: Degree distribution for the gene-duplication model with homodimer-link II network ($\mathcal{G}^{sp}_{m3}$) after spoke link rearrangement with $\alpha = 0, 1, 3$. Networks are rescaled in the case $\alpha = 3$ and for $\nu = 5$, $\alpha = 1$.

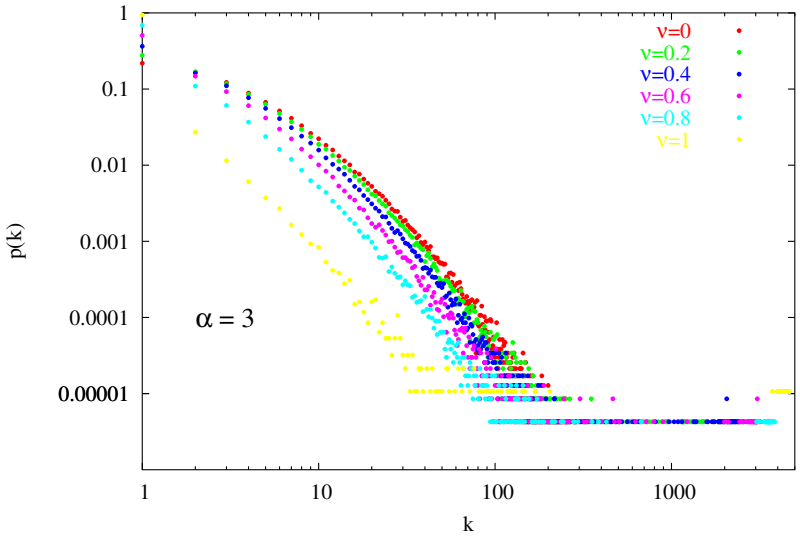


Figure 4.15: Degree distribution for the gene-duplication model with homodimer-link II network ($\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$) after spoke link rearrangement with $\alpha = 3$ ($\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$). Networks are not rescaled. Hence, average degrees evolve like 6.10, 5.25, 4.15, 3.12 and 2.14 for $\nu = 0.2, 0.4, 0.6, 0.8, 1$, respectively. Additionally the nodes with highest degree (hubs) are depicted.

a certain number of time steps of the spoke algorithm, a few highly connected nodes remain with a direct link between them and surrounded by lots of nodes with degree one. Note that this is not very significant in the depiction of Fig. 4.14 (bottom) because ν was limited to $\nu_{\max} = 0.5$ to make rescaling possible. Hence, the outcome of the spoke algorithm without rescaling is depicted additionally in Fig. 4.15. The process comes to an end when only one node that forms a "star" with all other nodes with degree $k = 1$ remains. In the case of the assumed process of mapping yeast interactions, the strength of link rearrangement represents the strength of the attachment of the proteins. If $\nu \to \infty$, one bait is picked out with all other proteins attached to it. According to the spoke rule, all interactions are directly assigned to the bait and a prey. Hence, a "star" is formed with a hub in the middle. This explains the increase of nodes with degree $k = 1$ in the degree distribution. For larger perturbations ν, also the number of highly connected nodes declines since for $\nu \to \infty$ only one highly connected node remains. The highest connected nodes are not shown in Fig. 4.14 bottom) but in Fig. 4.15 it is visible that for $\nu = 0$ the maximum degree is $k_{\max,\mathrm{m3}} \approx 200$, while for $\nu = 0.8$ it becomes $k^{\mathrm{sp}}_{\mathrm{m3,max}} \approx 3500$. These hubs are connected to almost all nodes in $\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$.

For $\alpha \approx 1$, the resulting degree distribution changes very slowly (see Fig. 4.14 middle). Hence, for values of $\nu \lesssim 1$ that are of biological relevance, it can be considered as unchanged by the spoke algorithm.

Analytical solution

For an analytical solution of the degree distribution in the limit $\nu \to \infty$, it is assumed that there are no degree correlations and no links are removed according to the case b) of Fig. 4.12. Thus, the spoke rearrangement algorithm is described by the rate equation:

$$\begin{aligned} p_k(t+1) - p_k(t) =& (1-\delta_{k1})\frac{1}{N}\frac{(k-1)^\alpha}{\langle k^\alpha\rangle}p_{k-1}(t) - (1-\delta_{k0})\frac{1}{N}\frac{k^\alpha}{\langle k^\alpha\rangle}p_k(t) \\ &+ (1-\delta_{k0})\frac{1}{N}\frac{k+1}{\langle k\rangle}p_{k+1}(t) - (1-\delta_{k1})\frac{1}{N}\frac{k}{\langle k\rangle}p_k(t). \end{aligned} \tag{4.7}$$

It gives the average change of the number of nodes $\Delta p_k(t)$ with degree k at every time step t. The first two terms on the right describe the average gain and loss in the p_k-bin when a bait is selected and its degree is increased by one as $k-1 \to k$ or $k \to k+1$. The bait is selected with probability according to Eq.(4.5). The last two terms represent the neighbor of the bait that loses a link and thus increases or decreases p_k for every time step. The probability of choosing a neighbor is proportional to the number of its links. Within all these terms, the first term like $(1-\delta_{k1})$ assures that the baits have a degree $k \geq 1$ and the preys that lose a link have $k \geq 2$. The term $1/N$ enters due to the fact that in every rearrangement step only one node is selected as bait and prey.

The stationary solution

$$p_k(t) = p_k(t+1) = p_k^\infty \tag{4.8}$$

is found with the insertion of degrees $k = 1, 2, \ldots$ into the rate equation. This leads for $\alpha = 0$ to

$$p_k^\infty = \frac{\langle k \rangle^{n-1}}{k!} p_1^\infty. \tag{4.9}$$

Note, that

$$p_{k=0}^\infty = p_{k=0}(t=0) = 0. \tag{4.10}$$

With the condition

$$\sum_{k=1}^{\infty} p_k(t) = 1, \tag{4.11}$$

p_1^∞ can be resolved as

$$\begin{aligned} p_1^\infty &= \left(\sum_{k=1}^{\infty} \frac{\langle k \rangle^{k-1}}{k!} \right)^{-1} \\ &= \left(\frac{1}{\langle k \rangle} \left(\sum_{k=0}^{\infty} \frac{\langle k \rangle^k}{k!} - 1 \right) \right)^{-1} \\ &= \frac{\langle k \rangle}{e^{\langle k \rangle} - 1}. \end{aligned} \tag{4.12}$$

Thus for $\alpha = 0$, the rate equation is solved by

$$p_k^\infty = \begin{cases} 0 & (k = 0) \\ \\ \dfrac{\langle k \rangle^k}{k!} \dfrac{1}{e^{\langle k \rangle} - 1} & (k \geq 1). \end{cases} \tag{4.13}$$

This Poissonian solution confirms the simulation results for $\nu \to \infty$, see Fig. 4.14 (top).

For the case $\alpha = 1$, the rate equation Eq. (4.7) is solved by

$$p_k^\infty = p_1^\infty k^{-1}. \tag{4.14}$$

This solution cannot be normalized. But the rate equation describes infinite networks. To approach finite network sizes, a cutoff term must be introduced. Thus, the solution becomes

$$p_k^\infty = p_1^\infty k^{-1} e^{-\frac{k}{k_c}}, \tag{4.15}$$

where k_c is the cutoff parameter. In Fig. 4.14 (middle), the solution for the rate equation with $p_1 = 0.35$ and $k_c = 18$ illustrates very well the limiting case for $\nu \to \infty$.

For $\alpha > 1$, no simple solution of the rate equation could be found. Simulation results suggest that the degree distribution in the stationary limit must be $p_1 = N - 1/N$ and $p_{N-1} = 1/N$ with all other values zero to represent the star structure. The average degree then becomes:

$$\langle k_{\nu\to\infty}\rangle = \frac{N-1}{N}\cdot 1 + \frac{1}{N}\cdot(N-1) = \frac{2(N-1)}{N} \approx 2. \tag{4.16}$$

This suggests that for a solution in the $\alpha > 1$ regime, link removal must be included in the rate equation.

Simulation results for degree correlation, clustering coefficient and motif-structure

In case of $\alpha = 0$ and 1 the degree correlation $\langle k_{\rm ngb}|k\rangle$ and the clustering coefficient $\langle C(k)\rangle$ become independent of the node degree k for very large ν, which is characteristic to fully randomized networks. See Figs. 4.16 and 4.17. For $\alpha = 0$, the overall clustering coefficient declines like $\langle C^{\rm sp}_{\rm m3}\rangle = 0.14, 0.11, 0.08, 0.06$ and 0.002 for $\nu = 0.1, 0.3, 0.5, 0.7$ and 5, respectively. Also for $\alpha = 1$ it declines as $\langle C^{\rm sp}_{\rm m3}\rangle = 0.20, 0.19, 0.16, 0.12$ and 0.008 for $\nu = 0.1, 0.3, 0.5, 0.7$ and 5, respectively. This decline is again a clear signature of the convergence towards randomized networks. Also the motif-structures converge to the randomized limit. This is documented in Figs. 4.18 (top and middle) with the motif-structures of equivalent randomized network models, which match very well.

Qualitative differences between the cases $\alpha = 0$ and $\alpha = 1$ are obtained for smaller ν. Compared to $\alpha = 1$, the k-dependent clustering coefficient declines for $\alpha = 0$ much faster for low degrees $k \lesssim 10$. The same holds for the degree correlation. Both can be explained as follows: If nodes with low degree are chosen as baits in a larger amount, they underlie rearrangement much stronger than nodes with higher degree. Furthermore, link rearrangement affects lowly connected nodes much more then higher connected nodes, compare also random link exchange in Sect. 4.2. Thus, their observables shift much faster towards the limiting case of an Erdös-Renyi network.

Nevertheless, for low rearrangement strengths $\nu < 0.3$ the clustering coefficient increases in both cases $\alpha = 0$ and $\alpha = 1$. As discussed in Sect. 3.3.4, many squares emerge in gene-duplication models. If a node with degree k_i and clustering coefficient $C_i = 0$ is duplicated, on average $k'_i(k'_i-1)/2$ squares emerge where $k'_i = (1-\delta)k_i$. The spoke rearrangement algorithm compensates this effect when the copied node is attached directly to the original or vice versa and an intermediary link is removed. This leads to the emergence of $k'_i - 1$ triangles and explains the increase in the clustering coefficient. Note that the number of emerging triangles is much smaller than the number of squares that disappear.

Results are very different for $\alpha = 3$. As mentioned before, nodes are repeatedly reattached to highly connected nodes. This converges to a "star" with a hub in the middle. Consequently, the degree correlation becomes strongly disassortative and the average neighbor degree of lowly connected nodes becomes extremely large. For $\nu = 0.8$ a maximum degree $k^{\rm sp}_{\rm max} \approx 3500$ comes with a very low average neighbor degree of

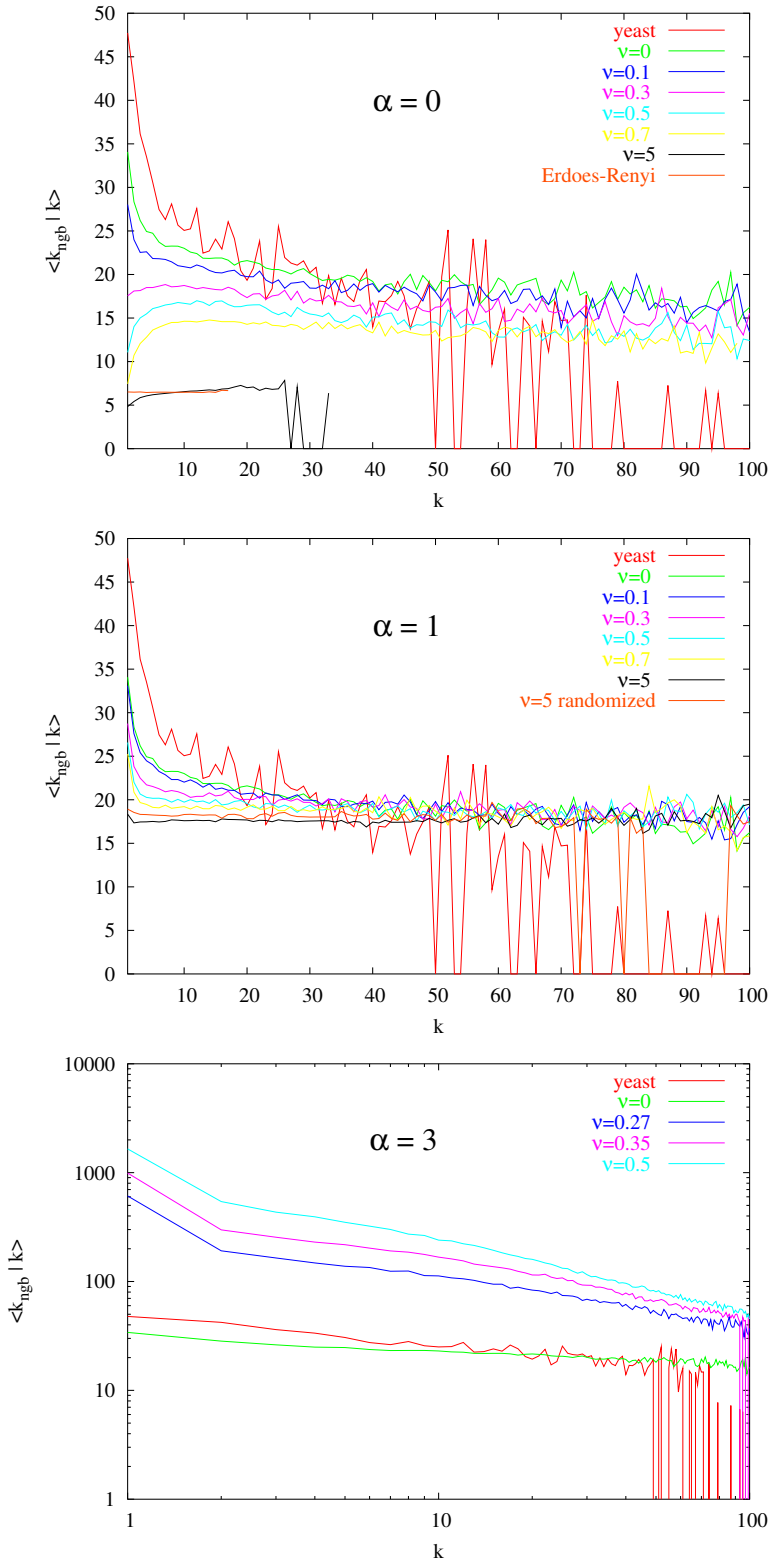


Figure 4.16: Degree correlation for the gene-duplication model with homodimer-link II network ($\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$) after spoke link rearrangement with $\alpha = 0$, 1 and 3 ($\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$). Networks are rescaled in the case $\alpha = 3$ and for $\nu = 5$, $\alpha = 1$.

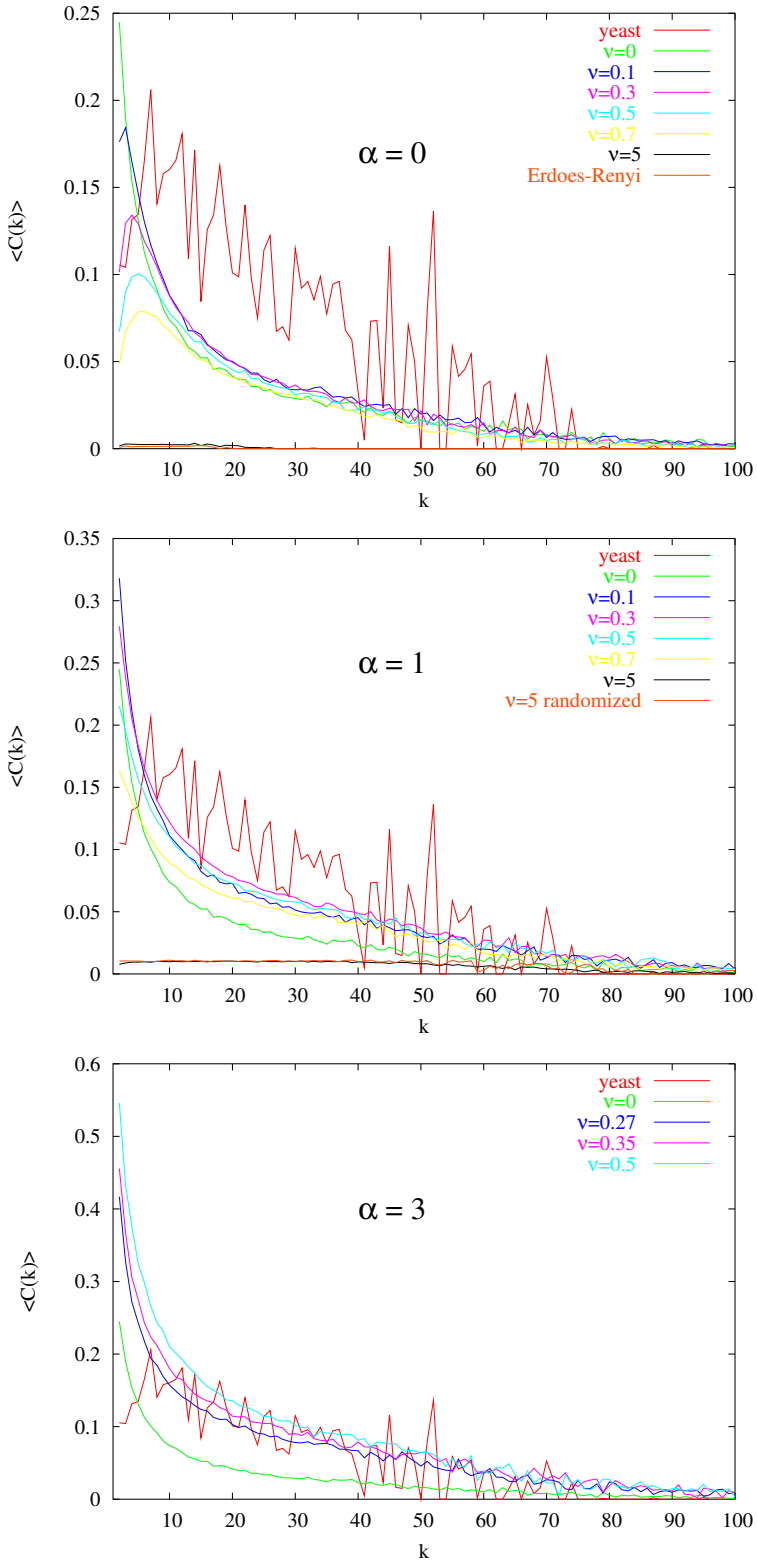


Figure 4.17: Clustering coefficient for the gene-duplication model with homodimer-link II network ($\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$) after spoke link rearrangement with $\alpha = 0$, 1 and 3 ($\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$). Networks are rescaled in the case $\alpha = 3$ and for $\nu = 5$, $\alpha = 1$.

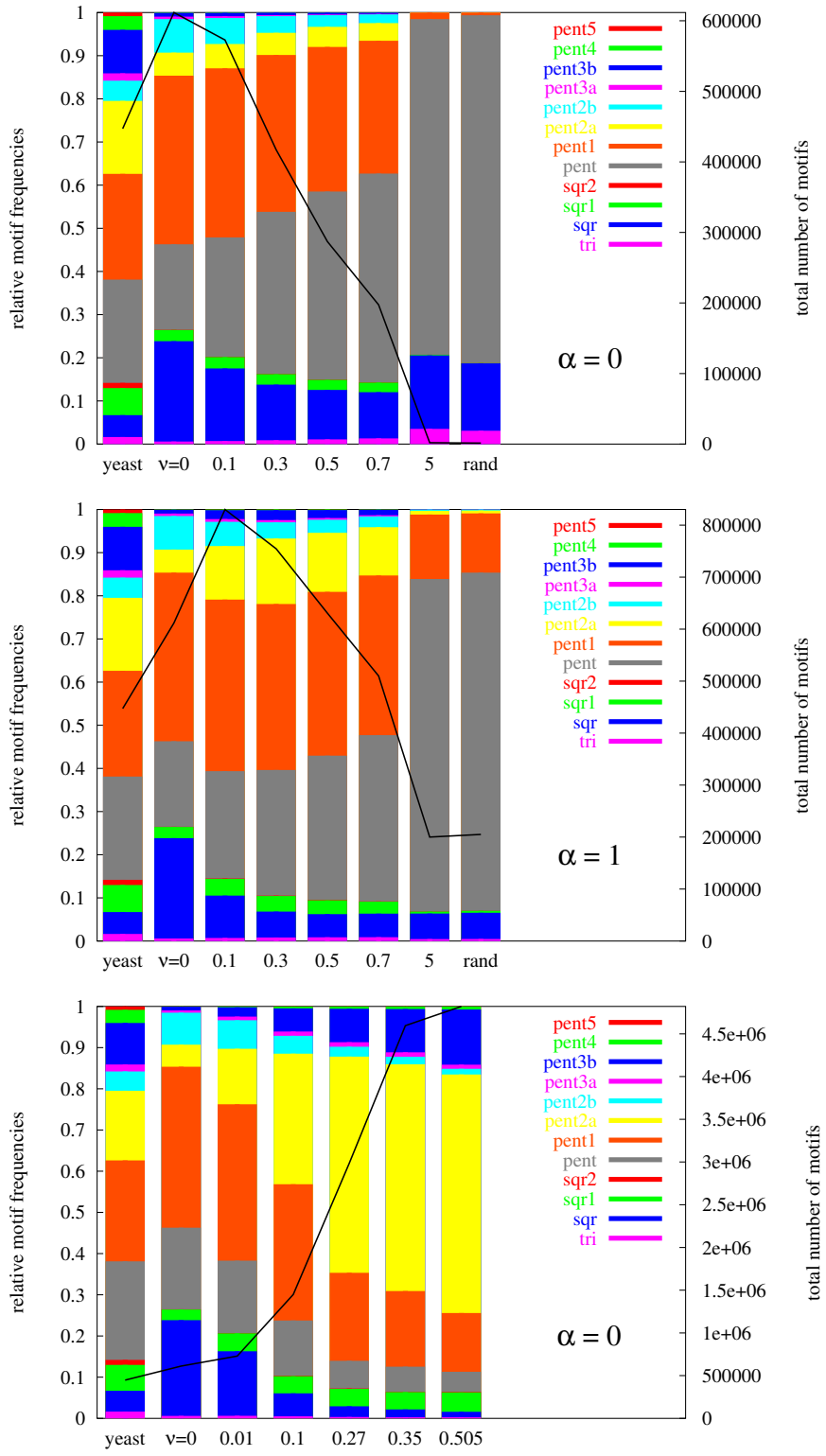


Figure 4.18: Motif structure for the gene-duplication model with homodimer-link II network ($\mathcal{G}_{\mathrm{m3}}$) after spoke link rearrangement with $\alpha = 0$, 1 and 3 ($\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$). Networks are rescaled in the case $\alpha = 3$ and for $\alpha = 1, \nu = 5$.

$\langle k^{sp}_{ngb}|k^{sp}_{max}\rangle \approx 1.2$, compared to $k_{max} \approx 200$ and $\langle k_{ngb}|k_{max}\rangle \approx 15$ for $\nu = 0$. The clustering coefficient increases for all degrees k nearly proportional to the rearrangement strength ν. For a lower ν, the same explanation holds as in the $\alpha \approx 1$ case: Squares are rearranged to triangles. Consult Fig. 4.18 (bottom), in which the decline of the number of squares is detected. For a higher ν, e.g. $\nu = 0.8$, the highest connected node (hub) with $k_{max} \approx 3500$ is already connected with almost any other node in the network. Given this hub h, any connected pair of nodes $i \neq h$ and $j \neq h$ is very likely to be connected to h and thus form a triangle. In the same way, more complex motifs appear and regarding the entire motif-structure, the total number as well as the complexity of motifs increases strongly.

In contrast to the other noise and sampling algorithms the spoke rearrangement indeed changes the properties of the network significantly. For the here used gene-duplication and mutation model the regime of choosing nodes as baits, determined by $\alpha \lesssim 1$, $\alpha \approx 1$ or $\alpha \gtrsim 1$, is crucial. The exact value within these ranges are of minor importance for the network properties.

Comparison with biases of mapping methods

To study biases of several mapping methods, in Sect.4.1 different sub-networks consisting of links found by only one method were analyzed. These findings can now serve to prove if results of the spoke algorithm are reflected in the affinity isolation sub-network. But it must be pointed out that the problem of the introduction of false links within a protein complex exists also in the synthetic lethality method and that in general intermediary proteins may not be recognized in the yeast-two-hybrid method neither (see Sect. 3.1).

The analysis of the degree distribution of the baits suggest an $\alpha < 1$ (see Fig. 4.13). Thus, a shift towards a Poissonian degree distribution between the distribution of all nodes or the high confidential dataset and the distribution of the affinity isolation sub-network could be expected, which is not the case (see again Fig. 4.2 top). Rather the degree distribution of the affinity isolation sub-network is more scale-free when compared to the degree distribution of the entire dataset, which would actually suggest $\alpha \gtrsim 1$.

Looking at the degree correlation of the affinity isolation sub-network (Fig. 4.2 middle) suggests $\alpha \lesssim 1$ because here the degree correlation is constant in contrast to the entire network. This difference between the entire network and the sub-network is very well reflected if the model and the outcome after spoke rearrangement are compared.

The preference of $\alpha < 1$ may be substantiated by the observed maximum in the clustering coefficient at degree $k \approx 5$ (Fig. 4.17 top). A similar maximum can be observed in yeast data (see Fig. 4.2 bottom). In turn for degrees $k \geq 7$, the clustering coefficient can be much better reproduced in the $\alpha \approx 1$ and even better in the $\alpha > 1$ regime (Fig. 4.17 middle). The sub-networks can hardly be consulted to draw any conclusion because

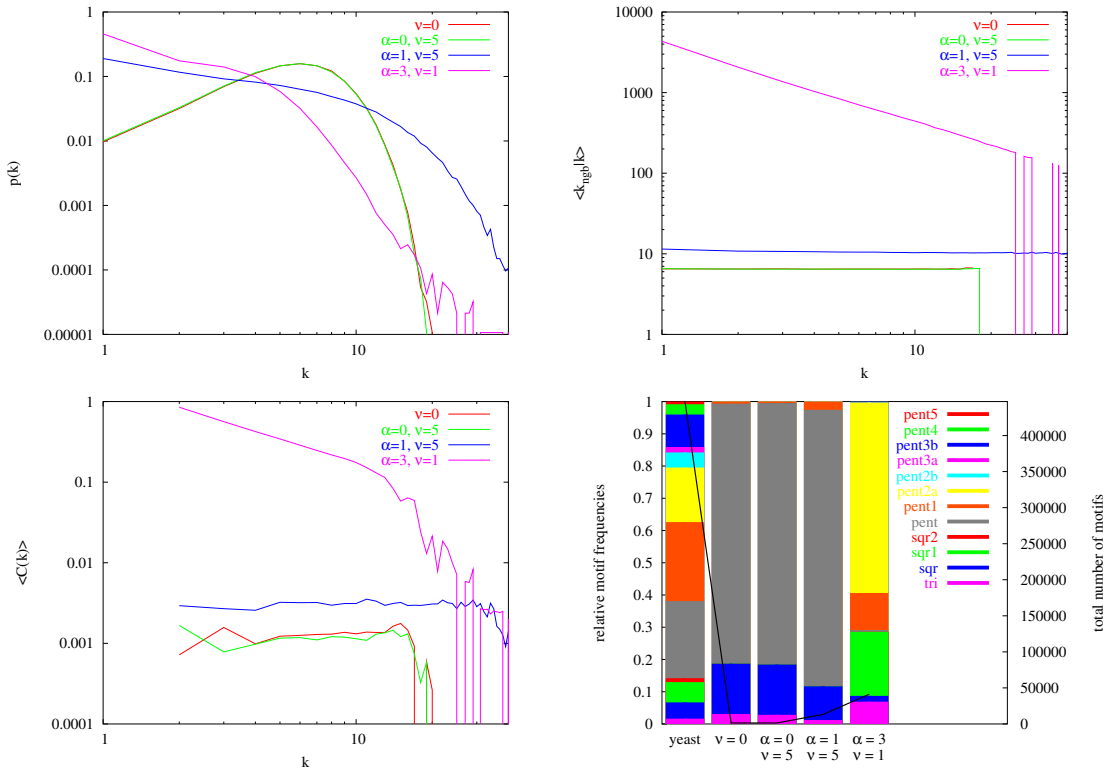


Figure 4.19: Degree distribution, degree correlation, clustering coefficient and motif-structure after spoke link rearrangement on an Erdös-Renyi network $\mathcal{G}_{\mathrm{ER}}$ for different choices of the baits with $\alpha = 0, 1, 3$ and constant $\nu = 5$ in the first two cases and $\nu = 1$ in the latter one. The different choice of ν is due to the fact that the network converges faster to the limiting "star" structure in the case $\alpha = 3$.

.

the clustering coefficient fluctuates strongly. However, the average over all degrees $\langle C \rangle$ is much larger for affinity isolation and synthetic lethality methods than for yeast-two-hybrid (see Tab. 4.1). That may substantiate the findings of this section, where $\langle C \rangle$ increases in all cases for smaller ν. On the other hand, the clustering coefficient might be per se larger because both methods focus on complexes.

Again the motif structure of the entire network can be best reproduced in the $\alpha \approx 1$ case for $\nu = 0.3$ (Fig. 4.17 middle) but no result for $\alpha < 1$, $\alpha = 1$ or $\alpha > 1$ matches well with the sub-networks for different methods (see Fig. 4.3).

Influence of spoke link rearrangement on other network models

Additionally, the spoke rearrangement has been applied to Erdös-Renyi networks (see Figs. 4.19) as well as to gene-duplication and mutation models that have been intro-

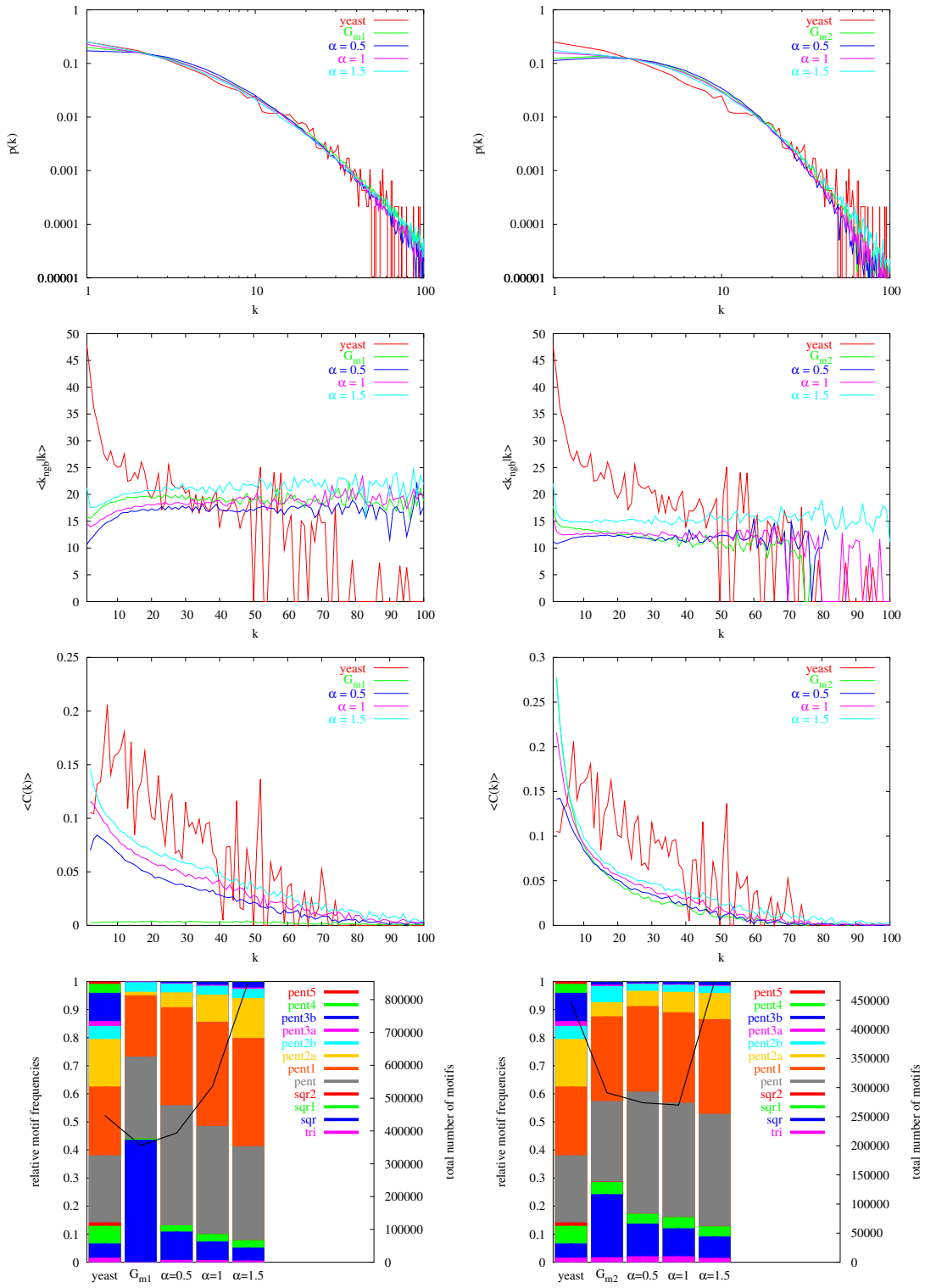


Figure 4.20: Degree distribution, degree correlation, clustering coefficient and motif-structure after spoke link rearrangement on the duplication and mutation network with random link $\mathcal{G}_{\mathrm{m1}}$ (left) and on the duplication and mutation network with homodimer-link I $\mathcal{G}_{\mathrm{m2}}$ (right) for different choices of the baits with $\alpha = 0.5$, 1 and 1.5 and constant $\nu = 0.4$.

duced in Sect. 3.2 (Figs. 4.20) to obtain a general view on the properties of the algorithm.

In the case of an Erdös-Renyi network $\mathcal{G}_{\mathrm{ER}}$, the analytical results for $\alpha = 0$ and $\alpha = 1$ are confirmed. Also in the case of $\alpha = 3$, the network shifts towards a "star"-structure. Compare Fig. 4.19 (top left) where lots of nodes with degree $k = 1$ and some hubs emerge.

Other results for the degree distribution, degree correlation, clustering coefficient and motif structure are as one would expect if the network converges either to a randomized network ($\alpha \lesssim 1$) or to a "star" structure ($\alpha > 1$). The average neighbor degree and the clustering coefficient are independent of the node degree but somewhat larger in the $\alpha = 1$ case. Most significant is the outcome in the $\alpha = 3$ case: Mostly nodes with degree $k = 1$ and a hub remain. The hubs have degrees of $k^{\mathrm{sp}}_{\mathrm{max,rand}} \approx 4500$, compared to $k_{\mathrm{max,rand}} \approx 20$ for $\nu = 0$. As for the gene-duplication and mutation network with homodimer-link II $\mathcal{G}^{\mathrm{sp}}_{\mathrm{m3}}$, the degree correlation is strongly disassortative. Rescaling was abandoned here although the average degree declines to $\langle k^{\mathrm{sp}}_{\mathrm{ER}} \rangle = 3.38$.

Looking at the influence of the spoke rearrangement on other gene-duplication networks $\mathcal{G}_{\mathrm{m1}}$ and $\mathcal{G}_{\mathrm{m2}}$, above stated results for $\mathcal{G}_{\mathrm{m3}}$ are confirmed and networks converge to the same topology.

Conclusion

In general, the spoke algorithm changes the structure of any network towards either a randomized network with Poissonian degree distribution, a scale-free distribution with $\gamma = 1$ and exponential cutoff or a "star"-like structure depending on the choice of the baits. In the limit $\nu \to \infty$ this outcome is independent of the underlying network.

Thus, the choice of the specific gene-duplication and mutation model is of minor importance if the influence of this inaccuracy that is due to mapping experiments is evaluated. But the question which regime of choosing the baits is the most probable one remains difficult to be answered. Most observables point to an $\alpha \lesssim 1$ regime, in which all observables are shifted towards pure randomized networks. Regarding all observables, the spoke rearrangement with $\alpha = 1$ and $\nu = 0.3$ brings the curves closest to real interaction data.

5 Conclusion and outlook

Analyses of the available evolution models for protein interaction networks showed that difficulties with respect to the reproduction of further topological observables exist, although these models are able to reproduce the degree distribution very well. Especially the motif structure with their rich variety of motifs is not described very well by any model. Even the inclusion of some biological details - such as more freedom in the mutation of interactions after gene-duplication or the possibility of copying not only one but two correlated proteins in one evolution step - could not improve the outcome significantly.

Hence, it appears necessary to include also selection and biochemical repair mechanisms in the actual models to obtain larger improvements in the modeling of protein interaction networks. The exclusion of unconnected proteins during the evolution of the gene-duplication model with homodimer-link II might be a first step into this direction. Another aspect of the same question is: Is it legitimate to consider the average network topology or did nature select a very special network realization? In this as well as in all earlier studies only averaged network realizations were regarded. But a model that includes selection based on a fitness model would probably much more suited. As long as this is not possible, further studies should focus on the entire variety of networks that emerge with the actual models.

Main target of this work was to propose and analyze several error algorithms to enable an appropriate comparison between models and real yeast data since the data is corrupted by a very large number of false links. It could be shown that topological network properties are very robust against randomly introduced false negative links. In simulations it turned out that thereby it is not important if these links are removed from the underlying (true) network or if a corrupted network is randomly sampled from the true one.

False positive links were also simulated in a random manner. The comparison between real yeast interaction data and the resulting networks showed that randomly distributed false positive links are unlikely to be contained in the real protein interaction dataset in such a large number as predicted [9, 11].

Furthermore a very specific (spoke) rearrangement algorithm was applied that is based on one type of the three major mapping methods. With the specific exchange of links in the neighborhood of selected nodes, this algorithm includes the simulation of false positive as well as false negative links.

For all other simulations, the degree distribution is not largely questioned: It appears to be robust against perturbations through random false negative errors and the influence

of random link addition might be small. But depending on the choice of the baits, with the spoke link rearrangement the degree distributions can be changed into any direction. Unfortunately, different observables point to different regimes of choosing the baits.

Most convincing is the regime, in which baits are chosen fairly equally with all degrees ($\alpha = 1$). This would not change the degree distribution significantly. All other observables would change towards the limit of a randomized network and for a lower perturbation ν come closer to protein interaction data.

For a better description of topological errors it might be promising to consider even more errors that are characteristic for certain mapping methods. But not only in biology, also in any other, especially social networks, empirical data is largely corrupted by false links. Hence, in topological investigations the influence of false links should be regarded, either by general or by specific error assumptions.

Appendix - Symbols

symbol	meaning
$\mathcal{G}$	network
$\mathcal{N}$	set of nodes
N	number of nodes
n_i	node i
$\mathcal{N}_i$	set of neighbors of node n_i
$\mathcal{L}$	set of links
L	number of links
l_{ij}	link between i and j
k	degree
$\langle k \rangle$	average degree
$p(k)$	degree distribution
d_{ij}	shortest path length between nodes i and j
$d_{\max}$	network diameter
Q	quality factor during the evaluation of the community structure
$\langle k_{\mathrm{ngb}} \vert k \rangle$	degree correlation
$\langle C \rangle$	average clustering coefficient
$\langle C(k) \rangle$	k dependent average clustering coefficient
$\mathcal{G}_{\mathrm{ER}}$	Erdös-Renyi network
$p(a_{ij} = 1)$	probability of a link between nodes i and j
D	dimension
γ	scale-free exponent
k_c	cutoff parameter
$\mathcal{G}_{\mathrm{m1}}$	gene-duplication and mutation network with random link [12]
$\mathcal{G}_{\mathrm{m2}}$	gene-duplication and mutation network with homodimer-link I [13]
$\mathcal{G}_{\mathrm{m3}}$	gene-duplication and mutation network with homodimer-link II [14, 15]
δ	parameter for the deletion of links during evolution in $\mathcal{G}_{\mathrm{m1}}$, $\mathcal{G}_{\mathrm{m2}}$ and $\mathcal{G}_{\mathrm{m3}}$ as well as in the link-duplication and the hybrid model
β	parameter for the insertion of random links during evolution in $\mathcal{G}_{\mathrm{m1}}$ and the hybrid model

symbol	meaning
p	parameter for the insertion of maintained homodimer-links during evolution in $\mathcal{G}_{\mathrm{m2}}$ and $\mathcal{G}_{\mathrm{m3}}$ as well as in the link-duplication and the hybrid model
ω	fraction of link-duplication in the link-duplication model
P^i	purity of node i
L_m^i	number of links at node i detected by method m
$\mathcal{G}^{\mathrm{rm}}$	network after random link removal
$\mathcal{G}^{\mathrm{ad}}$	network after random link addition
$\mathcal{G}^{\mathrm{ec}}$	network after random link exchange
$\mathcal{G}^{\mathrm{rw}}$	network after random walk
$\mathcal{G}^{\mathrm{rw}}$	network after avalanche exploration
$\mathcal{G}^{\mathrm{sp}}$	network after spoke rearrangement
$\mathcal{G}^{\mathrm{res}}$	rescaled network
$\mathcal{G}^{\mathrm{rés}}$	not rescaled network
ν	fraction $\nu = \Delta L/L$ of removed, added or exchanged links during random link removal, addition and exchange as well as during spoke link rearrangement
N_S	number of start-nodes during random walk and avalanche exploration
l_p	walk length during random walk
W	exploration length $W = N_S l_p$ during random walk
σ	probability to discover neighbors during avalanche exploration
$\mathcal{N}_i^{\sigma}$	set of discovered neighbors of node n_i during avalanche exploration
α	exponent of preferential bait picking during spoke link exchange

Acknowledgment

I would like to express my deepest gratitude to Prof. Martin Greiner, who offered me the opportunity to work in this fascinating field and for his engagement which made it possible to closely work together even on the distance Dresden - Munich. He was always an inspiration and enriched this thesis very strongly with his knowledge and ideas.

I am also deeply indebted to Ingmar Glauche, who not only shared his profound biological knowledge with me but also raised my enthusiasm to biology. Furthermore, I thank him for his engaged support of this thesis and for many discussions.

Prof. Gerhard Soff deserves special thanks for giving me the opportunity to write this theses at his chair at the Dresden University of Technology. Much to our regret, he passed away during my work on this thesis. I am also thankful to Prof. Sigismund Kobe for taking over the responsibility as an official advisor.

I am thankful to my friends Michael Graupner, Jakob Schweizer, Monique Rust and Isabel Raabe and my colleagues at the Institute for Theoretical Physics, to Michael for allowing me to use some of his processors in Paris and to Jakob, Monique and Gernot Schaller for their careful reading of my manuscript. Isabel helped me to find out, what biologists actually do when they map protein interactions.

All my friends and my family must be thanked for supporting me and for their interest in my work.

Bibliography

[1] J. L. MORENO. *Who shall survive?* Beacon House Inc., Beacon, N.Y., 1953.

[2] M. E. J. NEWMAN. *The structure and function of complex networks*. SIAM Review, 45:167–256, 2003.

[3] S. WASSERMAN AND K. FAUST. *Social network analysis: methods and applications*. Cambridge University Press, Cambridge, 1994.

[4] R. ALBERT AND A.-L. BARABÁSI. *Satistical mechanics of complex networks*. Review of Modern Physics, 74:47–97, 2002.

[5] A.-L. BARABASI AND Z. N. OLTVAI. *Network biology: understanding the cell's functional organization*. Nature Reviews Genetics, 5:101–113, 2004.

[6] S. N. DOROGOVTSEV AND J. F. F. MENDES. *Evolution of Networks - From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[7] T. S. EVANS. *Complex networks*. Contemporary Physics, 45:455–474, 2004.

[8] S. H. STROGATZ. *Exploring complex networks*. Nature, 410:268–276, 2001.

[9] C. MERING, R. K. B. SNEL, S. CORNELL, S. G.OLIVER, S. FIELDS AND P. BORK. *Comparative assessment of large-scale data data sets of protein-protein interactions*. Nature, 417:399–403, 2002.

[10] G. D. BADER AND C. HOGUE. *Analyzing yeast protein-protein interaction data obtained from different sources*. Nature Biotechnology, 20:991–997, 2002.

[11] C. M. DEANE, L. SALWINSKI, I. XENARIOS AND D. EISENBERG. *Protein interactions: Two methods for assessment of the reliability of high-throughput observations*. Molecular & Cellular Proteomics, 1:349–356, 2002.

[12] R. V. SOLÉ, R. PASTOR-SATORRAS, E. D. SMITH AND T. KEPLER. *A model of large-scale proteome evolution*. Advances in Complex Systems, 5:43–54, 2002.

[13] A. VÁZQUEZ, A. FLAMMINI, A.MARITAN AND A. VESPIGNANI. *Modeling of protein interaction networks*. Complexus, 1:38–44, 2003.

[14] I. ISPOLATOV, P. L. KRAPIVSKY AND A. YURYEV. *Duplication-divergence model of protein interaction network*. arXiv, (q-bio.MN/0411052), 2004.

[15] I. ISPOLATOV, P. L. KRAPIVSKY, I. MAZO AND A. YURYEV. *Cliques and duplication-divergence network growth.* arXiv, (q-bio.MN/0502005), 2005.

[16] J. BERG, M. LÄSSIG AND A. WAGNER. *Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications*. BMC Evolutionary Biology, 4:51, 2004.

[17] J. KIM, P. KRAPIVSKY, B. KAHNG AND S. REDNER. *Infinite-order percolation and giant fluctuations in a protein interaction network.* Physical Review E, 66:055101, 2002.

[18] L. SALWINSKI, C. S. MILLER, A. J. SMITH, F. K. PETTIT, J. U. BOWIE AND D. EISENBERG. *The database of interacting proteins: 2004 update*. Nucleic Acids Research, 32:D449–451, 2004.

[19] R. ALBERT, H. JEONG AND A.-L. BARABÁSI. *Error and attack tolerance of complex networks*. Nature, 406:378–382, 2000.

[20] V. ÁGOSTON, P. CSERMELY AND S. PONGOR. *Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example.* Physical Review E, 71:051909, 2005.

[21] P. CSERMELY, V. ÁGOSTON AND S. PONGOR. *The efficiency of multi-target drugs: the network approach might help drug design.* Trends in Pharmacological Sciences, 26:178–182, 2005.

[22] J. SCHOLZ, M. DEJORI, M. STETTER AND M. GREINER. *Noisy scale-free networks*. Physica A, 350:622–642, 2004.

[23] L. K. GALLOS. *Random walk and trapping processes on scale-free networks*. Physical Review E, 70:046116, 2004.

[24] J. SARAMÄKI AND K. KASKI. *Scale-free networks generated by random walkers*. Physica A, 341:80–86, 2004.

[25] A. VÁZQUEZ. *Knowing a network by walking on it: emergence of scaling*. arXiv, (cond-mat/0006132), 2000.

[26] L. DALL'ASTA, I. ALVAREZ-HAMELIN, A. BARRAT, A. VÁZQUEZ AND A. VESPIGNANI. *Statistical theory of internet exploration*. Physical Review E, 71:036135, 2005.

[27] M. E. J. NEWMAN. *Ego-centered networks and the ripple effect*. arXiv, (cond-mat/0111070), 2001.

[28] D.-H. KIM, J. D. NOH AND H. JEONG. *Scale-free trees: The skeletons of complex networks*. Physical Review E, 70:046126, 2004.

[29] M. P. H. STUMPF, C. WIUF AND R. M. MAY. *Subnets of scale-free networks are not scale-free: Sampling properties of networks*. Proceedings of the National Academy of Sciences, 102:4221–4224, 2005.

[30] M. P. H. STUMPF. *Sampling properties of random graphs: the degree distribution*. arXiv, (cond-mat/0507345), 2005.

[31] S. H. LEE, P.-J. KIM AND H. JEONG. *Statistical properties of sampled networks*. arXiv, (cond-mat/0505232), 2005.

[32] A. CLAUSET AND C. MOORE. *Why mapping the internet is hard*. arXiv, (cond-mat/0407339), 2004.

[33] A. CLAUSET AND C. MOORE. *Traceroute sampling makes random graphs appear to have power law degree distributions*. arXiv, (cond-mat/0312674), 2004.

[34] R. ALBERT. *Scale-free networks in cell biology*. Journal of Cell Science, 118:4947–4957, 2005.

[35] M. E. J. NEWMAN. *Scientific collaboration networks. II. Shortest paths, weighted networks and centrality*. Physical Review E, 64:016132, 2001.

[36] M. E. J. NEWMAN AND M. GIRVAN. *Finding and evaluating community structure in networks.* Physical Review E, 69:026113, 2004.

[37] M. E. J. NEWMAN. *Assortative mixing in networks*. Physical Review Letters, 89:208701, 2002.

[38] A. BARRAT, M. BARTHÉLEMY AND A. VESPIGNANI. *Modeling the evolution of weighted networks*. Physical Review E, 70:066149, 2004.

[39] S. ITZKOVITZ, R. MILO, N. KASHTAN, G. ZIV AND U. ALON. *Subgraphs in random networks*. Physical Review E, 68:026127, 2003.

[40] A. VÁZQUEZ, R. DOBRIN, D. SERGI, J.-P. ECKMANN, Z. N. OLTVAI AND A.-L. BARABÁSI. *The topological relationship between the large-scale attributes and local interaction patterns of complex networks in cellular networks*. Proceedings of the National Academy of Sciences, 101:17940–17945, 2004.

[41] N. PRŽULJ, D. G. CORNEIL AND I. JURISICA. *Modeling interactome: Scale-free or geometric*. Bioinformatics, 20:3508–3515, 2004.

[42] E. ZIV, R. KOYTCHEFF, M. MIDDENDORF AND C. WIGGINS. *Systematic identification of statistically significant network measures*. Physical Review E, 71:016110, 2005.

[43] R. MILO, S. SHEN-ORR, S. ITZKOVITZ, N. KASHTAN, D. CHKLOVSKII AND U. ALON. *Network motifs: simple building blocks of complex networks*. Science, 298:824–827, 2002.

[44] P. ERDÖS AND A. RENYI. *On random graphs I*. Publicationes Mathematicae, 6:290–297, 1959.

[45] S. MILGRAM. *The small world problem*. Psychology Today, 2:60–67, 1967.

[46] D. WATTS AND S. STROGATZ. *Collective dynamics of 'small-world' networks.* Nature, 393:440–442, 1998.

[47] A.-L. BARABASI AND R. ALBERT. *Emergence of scaling in random networks*. Science, 286:509–512, 1999.

[48] R. ALBERT, H. JEONG AND A.-L. BARABÁSI. *Diameter of the world wide web*. Nature, 401:130–131, 1999.

[49] B.-J. BREITKREUTZ, C. STARK AND M. TYERS. *The GRID: The General Repository for Interaction Datasets*. Genome Biology, 4:R23, 2003.

[50] S. ITZKOVITZ AND U. ALON. *Subgraphs and network motifs in geometric networks*. Physical Review E, 71:026117, 2005.

[51] M. KAISER AND C. C. HILGETAG. *Spatial growth of real-world networks*. Physical Review E, 69:036103, 2004.

[52] M. T. GASTNER AND M. E. J. NEWMAN. *The spacial structure of networks*. arXiv, (cond-mat/0407680), 2004.

[53] I. GLAUCHE, W. KRAUSE, R. SOLLACHER AND M. GREINER. *Continuum percolation of wireless ad hoc communication networks*. Physica A, 325:577–600, 2003.

[54] A. J. LEVINE. Lecture at: *22nd Jerusalem winter school in theorertical physics on: Biological networks and evolution*, December 27, 2004 - January 7, 2005.

[55] *http://160.114.99.91/astrojan/protein/pictures/galgbpr.jpg*.

[56] C. ALFARANO, C. ANDRADE, K. ANTHONY, N. BAHROOS, M. BAJEC, K. BANTOFT *et al. The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Research, 33:D418–424, 2005.

[57] A. ZANZONI, L. MONTECCHI-PALAZZI, M. QUONDAM, G. AUSIELLO, M. HELMER-CITTERICH AND G. CESARENI. *MINT: a Molecular INTeraction database.* FEBS Letters, 513:135–140, 2002.

[58] *http://mips.gsf.de/genre/proj/yeast/.*

[59] P. UETZ, L. GIOT, G. CAGNEY, T. MANSFIELD, R. JUDSON, J. KNIGHT *et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 403:623–627, 2000.

[60] T. ITO, T. CHIBA, R. OZAWA, M. YOSHIDA, M. HATTORI AND Y. SAKAKI. *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* PNAS, 98:4569–4574, 2001.

[61] A.-C. GAVIN, M. BÖSCHE, R. KRAUSE, P. GRANDI, M. MARZIOCH, A. BAUER *et al. Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 415:141–147, 2002.

[62] A. TONG, M. EVANGELISTA, A. PARSONS, H. XU, G. BADER, N. PAGÉ *et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants.* Science, 294:2364–2368, 2001.

[63] Y. HO, A. GRUHLER, A. HEILBUT, G. D. BADER, L. MOORE, S.-L. ADAMS *et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 415:180–183, 2002.

[64] J. HONERKAMP. *Statistical physics:an advanced approach with applications*. Springer-Verlag, Berlin, Heidelberg, 1998.

[65] M. CATANZARO AND R. PASTOR-SATORRAS. *Analytic solution of a static scale-free network model*. The European Physical Journal B, 44:241–248, 2004.

[66] J. S. BADER, A. CHAUDHURI, J. M. ROTHBERG AND J. CHANT. *Gaining confidence in high-throuput protein interaction networks*. Nature Biotechnology, 22:78–85, 2004.

[67] L. SALWINSKI AND D. EISENBERG. *Computational mehtods of analysis of protein-protein interactions*. Current Opinion in Structural Biology, 13:377–382, 2003.

Declaration

Hereby I declare that this diploma thesis is the result of my own work except where explicit reference is made to the work of others. This work has not been submitted for another qualification to this or any other examination board.

Mathias Kuhnt Dresden, 28th February 2006

Erklärung

Hiermit bestätige ich, dass ich diese Diplomarbeit ohne unzulässige Hilfe Dritter angefertigt und alle Quellen als solche gekennzeichnet habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde in dieser oder ähnlicher Form vorgelegt.

Mathias Kuhnt Dresden, 28.Februar 2006