Impact of Motif Content on Dynamic Function of Complex Networks

Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

der Naturwissenschaftlichen Fakultät III

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Christoph Fretter

geb. am 18. April 1983 in Frankfurt/Main

Halle, im Juli 2011

Gutachter:

Prof. Dr. Matthias Müller-Hannemann

Prof. Dr. Marc-Thorsten Hütt

Prof. Dr. Stefan Bornholdt

Die Verteidigung fand am 30.08.2011 statt.

Contents

1 Introduction **1**

1.1 Motivation: Topology and Dynamics . . . 1

1.2 Networks and Graphs . . . 2

1.3 Motifs . . . 3

1.3.1 Motif Classification . . . 3

1.3.2 Motif Significance Analysis . . . 5

1.3.3 Null Models . . . 5

1.3.4 Counting Motifs . . . 6

1.4 Outline . . . 9

2 Subgraph Fluctuations in Random Graphs **11**

2.1 Introduction . . . 11

2.2 Subgraph Statistics . . . 13

2.2.1 Subgraph categories . . . 13

2.2.2 Edge Counts . . . 14

2.2.3 Subgraph Counts . . . 16

2.2.4 Edge Fluctuations . . . 19

2.2.5 Subgraph Fluctuations . . . 19

2.3 Application . . . 21

3 Artifacts in Statistical Analyses of Network Motifs **27**

3.1 Introduction . . . 27

3.2 Methods . . . 29

3.2.1 Terminology . . . 29

3.2.2 Statistical Description of Motif Fluctuations . . . 29

3.2.3 Network Representations of Metabolic Systems . . . 30

3.3 Local Network Properties . . . 32

3.3.1 Self-Links . . . 32

3.3.2 Categorizing Bidirectional Edges . . . 33

3.4 Global Network Properties . . . 35

3.4.1 Modularity . . . 35

3.4.2 Networks with Hierarchies . . . 37

3.5 Dynamical Data on Motifs . . . 38

3.5.1 Mapping Dynamical Data onto Motifs . . . 38

3.5.2 Role Constraints . . . 40

3.6 Conclusions 42

4 Phase Synchronization in Railway Timetables 43
4.1 Introduction 43
4.2 Formalism and Numerical Experiments 46
4.3 Results 49
4.4 Avalanche Model 54
4.5 Conclusions 59

5 Co-Authorship Networks 61
5.1 Introduction 61
5.2 Results 64
5.3 Materials 68
5.4 Conclusions 70

6 Motifs as Markers of Future Success in Social Dynamics 71
6.1 Introduction 71
6.2 The model 73
6.3 Statistical Description of Network Motifs 74
6.4 Results 74
6.4.1 Motif Signature 74
6.4.2 Local Subgraph Patterns 79
6.5 Predicting the Fate of Agents on the Basis of Local Topology 80
6.6 Conclusion 82

7 Software Tools 85
7.1 State of the Art 85
7.2 Own Development 85
7.2.1 Interactive Graph Representation 87
7.2.2 Motif Dependencies 88
7.2.3 Interactive Motif Signature 89
7.2.4 Interactive Train Synchronization Visualization 92

8 Summary and Future Work 93

Bibliography 96

A Supplementary Material 109
A.1 Co-Authorship Networks 109
A.1.1 Publication Data 110
A.1.2 Citation Indices 110
A.1.3 Co-Authorship Graph Representation 111
A.1.4 Motif Analysis 111
A.1.5 Supporting Experiments 112
A.1.6 Application Areas 114
A.1.7 Generative Model 115

Chapter 1

Introduction

1.1 Motivation: Topology and Dynamics

Networks are useful abstractions of complex systems since they allow analyzing system properties by statistical and graph-theoretical methods.

In some cases the graph is known and the dynamics are not and we can attempt to predict dynamical properties from the topology. In other cases the dynamics are known but the topology is not, as in gene regulatory networks where one tries to infer the topology from the gene expression patterns (e.g. [60]).

This work will mostly focus on the cases where topology and some dynamical properties of a system are known and the goal is to understand which part of the dynamics can be explained by the topology and which part is due to other influences. As most systems are engineered or evolved to fulfill a certain function one can also ask the opposite question: To what extent is the topology functional and what part of it is random? Recent work has shown that an important aspect of the shaping of topology by dynamic requirements may be the necessity for robust systems [72, 77].

Often it is impossible to tell whether the topology shapes the dynamics or the other way round as a co-evolution is undergoing. A striking example are food webs, where the time scales of topology changes and dynamics are similar [27, 125, 166]. Another example are train networks, where on the one hand the schedule is constructed according to the available track and train capacities and the passenger demands. On the other hand, passenger demand as well as track construction / capacity adaption of tracks (allowing higher speed or higher train densities) and the acquisition of more trains is highly influenced by the load put on the system by the schedule. In such cases it is impossible to disentangle the complicated mutual relationship between topology and dynamics. Nevertheless it turned out to be fruitful to explore the topological requirements of the dynamics as well as the dynamical consequences of topology. The aim is then to transfer "building principles" and "optimization techniques" from one domain to another.

There is a history of analyzing the topological consequences of dynamical requirements. An early example is the research by Milgram on the few degrees of separation in social networks [106]. Although this study could well explain the phenomenon of short average path lengths, the question whether these are an "optimization goal" of social dynamics or just a by-product of some other requirement remains unclear.

The two seminal papers that opened up a vast avenue of research by exploring complex networks across many disciplines have been the simple model of small-world graphs by Watts and Strogatz (1998) [163] and the concept of scale-free graphs constructed via preferential attachment, formulated by Albert and Barabási (1999) [10]. The study by Watts and Strogatz is particularly remarkable for our purposes, as it constitutes the first nontrivial relationship between network topology and dynamics. Watts and Strogatz analyze the spread of infectious diseases as a function of the number of shortcuts in the network (or, more precisely, the rewiring probability). In 2001 Vespigniani described the disappearing of the epidemic threshold in scale-free networks in an analytical way [128].

In general, one can analyze the relationship between topology and dynamics on very different scales. At very weak coupling, the dynamics are governed by the dynamics of the individual nodes. Many studies in nonlinear dynamics adopt this point of view. At high coupling, the global network properties (like the connectivity, features of the degree distribution, etc.) can be expected to affect the collective dynamics. However, universal principles have only been derived for comparatively simple dynamics. The most important example are the studies on synchronization of phase oscillators by Arenas et al. [5] and Kurths et al. [132]. These findings can be seen as a "topology-refined" re-investigation of Kuramoto's formal treatment (1984) [83] of Winfree's observation (1974) [167] that synchronization sets in spontaneously, when a critical coupling is exceeded (see also [163, 152]).

The organization of dynamics on graphs can also be discussed on an intermediate scale, where small groups of nodes explain features of the collective behavior.

We will pursue this view of network motifs throughout this thesis.

1.2 Networks and Graphs

Many complex systems can be described as graphs, ranging from technical systems, as streets [32], air transportation [54] or the internet [129], over biological systems, for example metabolic networks [136], protein-protein interaction networks [142], food webs [126] or ant galleries [28], up to social networks, e.g. e-mail networks [38] or the contact network of Brazilian soccer players [124].

We use term *network* for such a system, consisting of topology and additional information such as dynamical data, annotations and context information. The term *graph* represents the underlying, purely mathematical object, consisting of *nodes* and *edges*. Throughout this work we will consider *simple graphs,* i.e. parallel edges or self-links are forbidden. Edges can be either directed or undirected. In the case of a directed graph two different types of edges are possible: i) an edge pointing from one node to another ii) a *bi-directional* edge that can be seen as the overlay of two opposing edges connecting the same two nodes. In a metabolic network this could be a reversible reaction.

Sometimes it is helpful to have graphs available that are random. For example when the impact of the degree distribution on properties of a graph is of interest.

A broad range of random graph models exists two of which are presented here:

i) The Erdős-Rényi (ER) random graphs [40] can be defined by only a number of nodes N and a number of edges M. They are random in every other property, this can be expressed by the fact that the probability p that two nodes are connected is the same for every pair of nodes. They can be created by randomly deciding for every pair of links whether they should be connected or not. The probability can be obtained with $p = \frac{M}{N(N-1)}$, for sparse graphs this is asymptotically equivalent to M times picking two random nodes and connecting them if they are not already connected, the second method being much faster (because by definition for sparse graphs M is much smaller than $N(N-1)$). This model is very simple and shows a narrow distribution of node degrees (i.e. the number of edges at a given node $P(k) = \binom{N-1}{k} p^k (1-p)^{n-1-k}$), k being the degree of a node. It is widely used because it is well suited for analytical calculations. We will also use this model in Chapter 2 to predict motif counts and motif fluctuations.

ii) The Barabási–Albert (BA) model[10] that generates random graphs with a scale-free degree distribution (i.e. following a power law of the form $P(k) \sim k^{-3}$). The algorithm works by starting with two connected nodes and then iteratively adding further nodes. The special properties of the resulting graphs come from the fact that the new nodes are not randomly attached to the existing nodes but by *preferential attachment*. The probability that the new node is connected to an existing node is positively related to the degree of that node, forming a system with positive feedback. The resulting graphs show short path lengths and a broad degree distribution, two properties that render them more realistic than ER graphs, while remaining simple enough so that some analytical calculations can be performed.

1.3 Motifs

1.3.1 Motif Classification

Generally, motifs are subgraphs that appear more often than expected in a network [109]. Similarly subgraphs that appear less often then expected can be called “anti-motifs”.

Most authors consider only induced subgraphs. Given a graph $G = (V_1, E_1)$, a subgraph $H = (V_2, E_2)$ is called *induced subgraph* if and only if $V_2 \subseteq V_1$ and if for any pair of nodes $v, w \in V_2, (v, w) \in E_1$ if and only if $(v, w) \in E_2$. This can be rephrased as any set of n nodes can only form one n-node-subgraph at a time and of all the possible subgraphs the one with maximal number of edges is selected. Subgraph number 13 in Table 1.1 contains only subgraph 13 and none of the other 12 subgraphs.

The other possible definition is non-induced subgraphs, here a set of n nodes can take part in different n-node subgraphs, for example subgraph 13 in Table 1.1 contains all the other 12 subgraphs. In a model of induced subgraphs the addition of an edge can destroy an already existing subgraph, whereas in the non-induced model an already existing subgraph cannot be destroyed by the addition of edges. This property can be exploited for the fast search for large motifs, by “growing” motifs from smaller ones by the addition of edges. It is then possible to abort unsuccessful search paths early in the process [123]. This is done by estimating the statistical over-representation of a motif occurrence, that is more or less strongly depending on the statistical over-representation of its constituting sub-motifs.

In this thesis we will mostly consider weakly connected directed 3-node subgraphs. When accounting for all automorphisms 13 different such subgraphs can be identified. It is common to refer to subgraphs by their id, which is formed by reading the adjacency matrix for the subgraph as a binary string. Of all possible node relabellings the one yielding the smallest motif id is chosen.

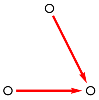
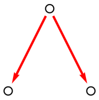
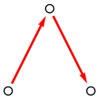
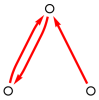
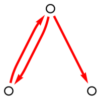
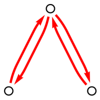
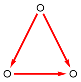
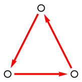
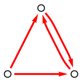
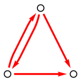
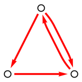
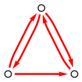
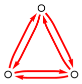
motif number	1	2	3	4	5	6	7	8	9	10	11	12	13
motif id	36 =4+32	6 =2+4	12 =4+8	74 =2+8+64	14 =2+4+8	78 =2+4+8+64	38 =2+4+32	98 =2+32+64	108 =4+8+32+64	46 =2+4+8+32	102 =2+4+32+64	110 =2+4+8+32+64	238 =2+4+8+32+64+128
adjacency matrix	0 0 1 0 0 1 0 0 0	0 1 1 0 0 0 0 0 0	0 0 1 1 0 0 0 0 0	0 1 0 1 0 0 1 0 0	0 1 1 1 0 0 0 0 0	0 1 1 1 0 0 1 0 0	0 1 1 0 0 1 0 0 0	0 1 0 0 0 1 1 0 0	0 0 1 1 0 1 1 0 0	0 1 1 1 0 1 0 0 0	0 1 1 0 0 1 1 0 0	0 1 1 1 0 1 1 0 0	0 1 1 1 0 1 1 1 0
motif													

Table 1.1: A list of all directed induced 3-node subgraphs.

1.3.2 Motif Significance Analysis

When trying to understand the impact of dynamics on topology on a statistical level, the most important question is which topological features observed in a system are non-random. This question can be answered by comparing the topology of the system to the topology of others, where we know that the dynamics does not shape the topology.

The motif signature of a network (for three-node subgraphs also called the triad significance profile, TSP) is the pattern of over- and under-representations of few-node subgraphs in a network. It has become a standard method of analyzing complex networks. More formally, it is the (normalized) z-score of the motif counts.

The z-score is defined as:

$$Z_m = \frac{c_m - \mu_m}{\sigma_m},$$

where for every subgraph c_m is the motif count in the original network, μ_m is the expectation value of c_m in a set of suitable reference networks and σ_m is the standard deviation of c_m in the reference networks. Obtaining the appropriate ensemble of reference networks is addressed in the next section.

1.3.3 Null Models

To assess the deviations from randomness of some observed features of a system we need to relate them to the features of systems that are in some sense normal. With this aim a null model is used that copies some constraints to the system but is not shaped by the dynamics.

The complexity of such constraints can range from just having the same number of nodes and edges as the original network, which can be achieved by creating an ER-graph up to a complex set of rules that ensure that for example a train schedule is feasible and does fulfill all capacity limitations. In such circumstances the fair sampling of graph instances fulfilling the constraints is highly non-trivial.

The most promising steps come from the area of mixing algorithms [134]. Here the two main problems are the accessibility of valid realizations by rewiring steps and the fair sampling i.e. that every valid realization is selected with the same probability. Another problem is that the number of mixing steps that have to be performed is not known, even for the simplest cases. Analytical estimates yield upper bounds of mixing times that are higher by many orders of magnitude (e.g $O(N^{11} log^5 N)$ in [18]) than what can be observed numerically (e.g. $O(N)$ in [108]). This big difference arises from the possibility of graphs that are difficult to mix in the sense that the state graph where every node is a network realization and every edge represents the possibility to get from one representation to the other in one mixing step contains bottlenecks. The existence of such bottlenecks could not be disproved yet.

Another approach is the stubs model or configuration model [104, 108]. Here half-edges are constructed so that all degree constrains are fulfilled. Then the task is to connect all half-edges (stubs) without creating self-links or parallel edges. Therefore a simple greedy algorithm is used. It has been shown that in this model one should not track back when a conflict occurs but rather restart, as otherwise correlations in the degrees are introduced into the graph. There have been attempts to circumvent this problem by working on classes of equally-correlated nodes [4]. In general the frequent restarts that are necessary make the stubs method very slow.

In [89] it is shown that some of the degree correlations observed are not an artifact of the graph construction method but rather intrinsic to broad degree distributions. When not only the degree sequence but also the degree-degree correlations are prescribed the problem of introduced degree-degree correlations does not exist, an algorithm performing this task is shown in [164]. It may still be possible that higher-order correlations are systematically introduced so that the ensemble of allowed configurations is not sampled in a fair way.

In all chapters of this work a (constrained) mixing algorithm is used. As no realistic minimal mixing times are available from theory case by case decisions have to be taken. When one is only interested in a z-score there is an easy way of doing this: i) Mix a set of copies of the start network for a certain number of steps (e.g. $10 \cdot N$) and compute the z-score. ii) Mix the networks again for the same number of steps and observe if the z-score has changed. If the z-score remains the same (up to a small fluctuation) the mixing time was long enough, otherwise one can simply go on and mix the networks for a longer time until the z-score converges.

Please note that the convergence of the z-score does not indicate that the mixed networks are random, they may still be biased, but in a way that does not affect motif counts anymore.

1.3.4 Counting Motifs

Conceptually there are two different methods to count motifs:

- Iterate over every triplet of nodes, compare the adjacency matrix to the adjacency matrix of every motif and increase the corresponding counter.

- Implement specific code for every motif. Properties of the wanted motif can be used to continue to expand the motif match only in situations where the motif match is still achievable.

Which method is faster depends on the number of possible motifs and on the probability of finding a motif.

The advantage of the first method is its compact code. The task of checking whether a selected subgraph is equal to a searched motif corresponds to the graph isomorphism problem, for which no algorithm is known that scales as a polynomial function or better with the graph size. But this is not relevant to the present problem as the considered subgraphs are of very small, constant size.

The advantage of the second approach is the possibility to count only the motifs one is interested in with the benefit of reduced overhead. Also the code is much simpler and easier to debug. The disadvantage lies in the necessity to write specialized code for every subgraph one is interested in.

In this work this problem is circumvented by using a meta programming approach. The specialized counting code for every motif is generated by a meta program, that has only to be written once.

Taking the adjacency matrix of the wanted subgraph two tasks are performed:

1. A search-path through the subgraph is selected (e.g. the order in which the nodes of the wanted subgraph are mapped onto the graph).

2. Automorphisms of the subgraph are identified. These have to be broken to avoid counting a motif several times.

Listing 1 The automatically generated code for the motif with the id 46 (comments added).

```
package motifs.analyzer.three;
import motifs.analyzer.Analyzer;
// auto generated
// 0->1 // 0->2
// 2->1 // 2->0
//way is
// 0->1 // 0->2
//symmetries:
// 2->0
public class m3id46Analyser extends Analyzer {

public void run() {
  motifCount=0;
  int[][] ol=network.outLinks;
  //Iterate over all nodes which have outgoing edges
  for(int node0=0;node0<ol.length;node0++){
    //Iterate over all of node0's outgoing edges
    for(int a=0;a<ol[node0].length;a++)        {
      int node1=ol[node0][a];
      if( node0 == node1||                //Do not allow a node twice in a motif
      contains(ol[node1],node0))          //Do not allow additional edges
      continue;

      //Iterate over all of node0's outgoing edges
      for(int b=0;b<ol[node0].length;b++)              {
        int node2=ol[node0][b];
        if( node1 == node2 ||             //Do not allow a node
          node0 == node2 ||               //twice in a motif
          node2< node0 ||                 //break the symmetry
          !contains(ol[node2],node1)||    //check for required edges
          !contains(ol[node2],node0)||
          contains(ol[node1],node2))      //Do not allow additional edges
            continue;

        motifCount++;                     //Found a motif instance
      }
    }
  }
}
```

For efficiency reasons and due to the used data structure (an adjacency list) it is advantageous to build a search path that does not require backward edges. Heuristically it may be of advantage to check nodes that underlay stronger constraints first, e.g. a node that must have several outgoing edges to form a subgraph. These two ideas are implemented in the current version. Further improvements depending on the network topology are possible. Here the meta programming approach is particularly helpful, as different heuristic search strategies can be quickly compared without changing large amounts of code manually. The symmetry is broken by requesting that all nodes of a automorphism class (that yield the same topology after relabeling) have increasing node IDs. An example for motif 46 is presented in the listing 1.

1.4 Outline

The aim of this work is to illuminate the interplay between topology and dynamics, with an emphasis on robustness and local network structures. To achieve this goal existing analysis methods are applied and extended. New methods are developed, implemented and applied to a set of network problems spanning a broad range of disciplines.

The thesis is organized as follows:

Chapter 2 introduces an analytical prediction of motif counts and motif fluctuations in random graphs. This is achieved by computing appropriate edge probabilities and combining them to motif probabilities. The derived equations can be used to predict motif signatures resulting from some non-standard graph properties. This is shown on the example of modular graphs, the predictions are compared to standard mixing techniques and a reference implementation of modularity-preserving mixing. The results of this chapter will be published in [46].

Chapter 3 discusses some common issues occurring while performing motif analyzes and uses the findings from Chapter 2 to give insight into the underlying combinatorical mechanisms. The false results occurring when not preserving the number of bi-directional edges in a graph are analytically predicted. Appropriate null models for some common cases including flow networks and bi-bipartite graphs are discussed. The results of this chapter will be published in [43].

Chapter 4 treats the network of long distance train connections. Especially the connection between robustness and efficiency is analyzed. The main result is a positive correlation between synchronization (the clustering of arrival/departure events in time) and delay propagation. The findings are compared to topological features of the underlying connection network, the high synchronization mainly occurs on average-sized stations. Then the connection between synchronization and dynamical robustness is further investigated in a simple model of delay propagation where the main result can be replicated. Possible implications on the process of train schedule generation are discussed. The results of this chapter have been published in [45].

Chapter 5 analyzes some real-world social networks, namely co-authorship networks. Hereby two authors are connected if they published a common paper. Additionally the corresponding citation data is obtained and the (local) motif properties of successful authors are analyzed. To achieve this different normalization schemes are discussed and tested against a set of randomized data. We find that some constellations of publications are systematically more successful than others and discuss possible reasons for this phenomenon. The results of this chapter have been published in [81].

Chapter 6 studies a model system of social dynamics. The effect of the local rewiring rules on the global topology is assessed. We show that successful and unsuccessful agents show systematically different local motif neighborhoods. Additionally novel (local) motif analysis techniques are employed to predict the success of social agents. The results of this chapter will be published in [44].

Chapter 7 introduces software tools that have been developed during the work on this thesis. Especially interactive network analysis software for some use cases is shown.

Chapter 8 finally summarizes the work and gives an outlook of future challenges.

Chapter 2

Subgraph Fluctuations in Random Graphs

Summary

The pattern of over- and under-representations of three-node subgraphs has become a standard method of characterizing complex networks and evaluating, how an intermediate level of organization contributes to network function.

Understanding statistical properties of subgraph counts in random graphs, their fluctuations and their inter-dependencies with other topological attributes is an important prerequisite for such investigations. Here we introduce a formalism for predicting subgraph fluctuations induced by perturbations of uni-directional and bi-directional edge densities. On this basis we predict the over- and under-representation of subgraphs arising from a density mismatch between a network and the corresponding pool of randomized graphs serving as null model. Such mismatches occur for example in modular and hierarchical graphs.

The results presented in this chapter have been achieved in cooperation with Matthias Müller-Hannemann and Marc-Thorsten Hütt and will be published in: “Statistical description of subgraph fluctuations in random graphs” [46].

2.1 Introduction

Network science, i.e. the discipline studying and interpreting a broad range of complex systems from a network perspective, has an enormous impact on how we perceive (and analytically approach) social, biological and technical systems. One of the most fascinating theoretical challenges of network science is the inter-dependence of network properties observed at different scales: Clustering depends on modularity, heavy-tailed degree sequences can induce degree-degree correlations [89, 127], a modular structure influences our expectations of betweenness centralities and other edge- or node-based properties. The severest impact of these inter-dependencies probably occurs when attempting to interpret the composition of a network in terms of few-node subgraphs. On this level, we can expect a very strong influence of global network properties, unless we adjust our null model (i.e. the set of random expectations) to match these global properties. It is therefore essential to understand this interplay from first principles. Here we discuss two types of correlations between network properties: (1) how single-edge fluctuations influence fluctuations in three-node subgraph frequencies; (2) how global network properties affect three-node subgraphs frequencies.

Network motifs have first been introduced as a method for analyzing transcriptional regulatory systems [109]. A comparison of the transcriptional regulatory network of the bacterium *E. coli* with random graphs has revealed that three characteristic local node/link patterns appear substantially more frequent than expected at random[148]: feed-forward loops (FFLs), single-input modules (SIMs) and densely overlapping regulons (DORs). The benefit from an identification of over-represented node/link patterns is two-fold: (i) one can formulate models of the dynamics encoded by such few-node devices; (ii) one can discuss selected examples of such motif occurrences in detail. In this way, feed-forward loops and single-input modules could, in subsequent work [3, 97, 148], be linked to specific dynamical functions (like noise buffering (FFL) and the implementation of temporal programs (SIM)).

To a certain extent, the analysis of such node/link patterns is a balance between an automatized, statistical view on a complex network and the discussion of individual cases. An interesting example of this balance is the discussion of various types of feed-forward loops in transcriptional regulatory networks. Once the statistical over-representation of this node/link pattern had been established [3, 148], the specific forms of feed-forward loops occurring in the networks could be further analyzed. One classification scheme is to enumerate all distributions of signs on the links ("activating" and "inhibitory") and see, whether the two paths (directly and via the third, intermediate node) from the top-level node to the bottom-level node in the feed-forward loop both provide the same signal (both activating or both inhibiting; coherent FFL) or provide conflicting signals (incoherent FFL). Surprisingly, not all variants of these coherent and incoherent FFLs seem to occur in equal proportions in transcriptional regulatory networks. Instead there seems to be a strong bias towards only one type of coherent FFL and one type of incoherent FFL [73].

An important debate in the study of biological systems from a network perspective is the biological relevance of statistical signals derived from graph representations (see also [110]). In order to address this question, it is interesting to explore the consistency of large-scale biological data sets with graph abstractions of biological networks. This has been done in particular for the gene regulatory network and the metabolic network of yeast and *E. coli*: (1) Luscombe et al. [96] showed that the topology of sub-network structures in yeast is specific for cellular programs triggered by environmental conditions: Slow programs (e.g. cell cycle) employ a densely interconnected sub-network structure, while programs required to act rapidly (e.g. DNA repair) employ networks with shorter path lengths and less complex motif content. (2) Using methods from point process statistics the arrangement of genes on the genome and their correspondence to the gene regulatory network have been analyzed [151]. (3) Using the method of control strengths derived from effective networks [99], the agreement of active metabolic networks (as predicted by flux-balance analysis) and gene expression data [150] has been studied. (4) With the aim of better understanding the validity of the motif perspective, the interplay between feed-forward loops and larger-scale structures (subsets formed by all nodes topologically down-stream of a reference node) in gene regulatory networks has been explored [101]. The rationale of this analysis has been to explore the interplay of two scales within the transcriptional regulatory network of *E. coli.* In particular, in [101] it was shown that when one scale dominates (high sub-net usage) few regulatory devices on the smaller scale are found (low feed-forward loop occurrence).

A strong step towards an automatized statistical view on network motifs has been the work by [107], where the over- and under-representation of three-node-subgraphs (the motif signature ore triad significance profile, TSP) compared to randomized networks is analyzed.

This analysis of the TSP has been applied to a wide range of complex networks [30, 72, 77, 81, 148]) and has become synonymous to a motif analysis. Many of the TSPs of real networks either show no significant over- or under-representation of three-node subgraphs or follow one of the four patterns (or "superfamilies") discussed in [107].

A very promising development over the last few years has been that some features of such motif signatures are found to be related to the robustness of the system (see, e.g., [77, 70])

Avetisov et al. [7] analyze the motif signatures of graphs obtained from a block-hierarchical adjacency matrix. By introducing randomness (i.e. random flips $1 \leftrightarrow 0$) in the adjacency matrix, the authors can also study the robustness of the motif pattern. They find that the motif signature persists under small amounts of such topological noise. This work is one of the few examples (together with the comment on spatial networks in [6]) of motif signatures arising from global organizational features (in this case: the block-hierarchical structure of the adjacency matrix) of the network. Remarkably, the motif signature is quite similar to one of the superfamilies from [107].

It is therefore of great interest to better understand the crosstalk between local and global network properties, as well as the inter-dependencies between the different few-node subgraphs.

For the special case of Erdős-Rényi (ER) random graphs we formulate a simple statistical description of expectation values for subgraph frequencies. Similar approaches have been formulated in [63, 74]. By grouping the possible three-node subgraphs into categories, we are able to understand differences between subgraph counts arising on purely combinatorical grounds. This description enables us to discuss for the first time the statistical fluctuations of subgraph counts arising, e.g., from fluctuations in the number of uni-directional and bi-directional edges.

Both, the expectation values and the standard deviations of subgraph counts enter the computation of subgraph z-scores, which are frequently employed for quantifying the statistical over- and under-representation of subgraphs in real networks. We can thus employ our method to the computation of a motif signature (or triad significance profile, TSP) in all cases, where fluctuations in the edge density induce a non-zero motif signature for an otherwise random graph. Modular graphs, as the most important case of this category, are discussed as an application.

2.2 Subgraph Statistics

In this part we introduce a simple model for the emergence of templates and motifs in random networks. We discuss the expectation values of motif counts and the corresponding fluctuations. This yields insight into the correlations from single node-properties to motifs.

2.2.1 Subgraph categories

Throughout this article we will only discuss simple directed graphs with a node number N and an edge number M. Simple means that parallel edges pointing in the same direction and

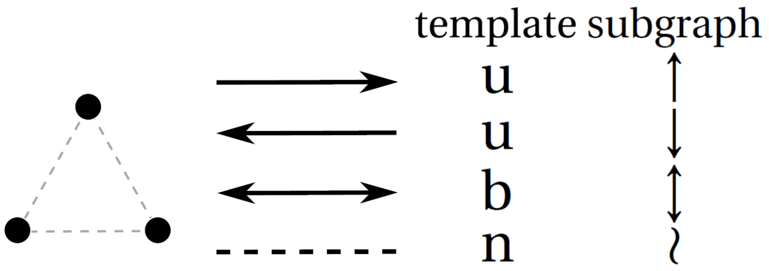


Figure 2.2.1: The three different placements for the three possible types of edges. On the right the different types of edges that can occupy the placements and their notation for templates and subgraphs.

self-links are forbidden. Because of these conditions the graph is *complete* when it contains $M = N(N-1)$ edges. When two nodes a and b are connected by a single edge they are *connected by a uni-directional edge* (u), when the opposing edge is also present they are connected by a *bi-directional* edge (b). Finally two nodes can be unconnected. Formally, this can be described by a non-edge (n).

Between global network properties on the one hand and single-node properties on the other, motifs can be used to understand networks on a mesoscopic scale. For directed networks, most studies use 3-node subgraphs, and we will here do the same although the formalism can easily be extended to higher motif sizes. In this text we distinguish between *templates* and *subgraphs*. Templates are sets of edges with an undefined position relative to each other. Some templates have two, others three edges:

two: *uu,ub,bb*

three: *uuu,uub,ubb,bbb*

These seven templates can form 13 different (*induced*) *subgraphs*. A subgraph is obtained by defining the relative orientations of the edges within a template.

Orientations are defined using $\uparrow$ for clockwise and $\downarrow$ for counter-clockwise directions of the edge. Using this shorthand notation, we can summarize the template-subgraph relationships in a tabular form, see Table 2.2.

As only relative positions and orientations matter, $\downarrow\downarrow\downarrow$ is indistinguishable from $\uparrow\uparrow\uparrow$. However $\uparrow\downarrow\wr$ is a different subgraph from $\downarrow\uparrow\wr$.

2.2.2 Edge Counts

Here we work with the Erdős-Rényi (ER) model of random graphs, where a graph is characterized by the number of nodes N and the edge probability p. A graph represented by N and p contains on average $M = p \cdot N(N-1)$edges. Here we specify a certain number of edges, M and then use the corresponding edge density (or connectivity) $p = M/(N(N-1))$ to characterize the network in our statistical assessment. For random networks we can estimate some basic probabilities and counts:

In a directed network model two edges that connect the same two nodes pointing in opposite directions form a bi-directional edge. The number of bi-directional edges can be estimated by

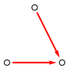
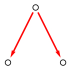
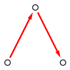
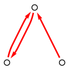
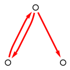
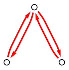
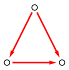
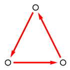
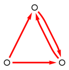
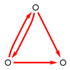
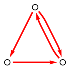
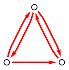
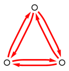
template	subgraph a	subgraph b	subgraph c
uun	↑↓ ≀ (1) 	↓↑ ≀ (2) 	↓↓ ≀, ↑↑ ≀ (3) 
ubn	↓↕ ≀ (4) 	↑↕ ≀ (5) 	
bbn	↕↕ ≀ (6) 		
uuu	↓↓↑, ↑↑↓ (7) 	↓↓↓, ↑↑↑ (8) 	
uub	↑↓↕ (9) 	↓↑↕ (10) 	↓↓↕, ↑↑↕ (11) 
ubb	↓↕↕, ↑↕↕ (12) 		
bbb	↕↕↕ (13) 		

Table 2.1: The seven templates together with the 13 subgraphs they can form, subgraph numbers in parentheses correspond to the standard ordering from [107].

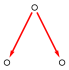
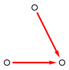
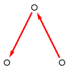
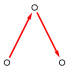
template		subgraph		r_m
uun				
↓↑ ≀	⟹		⟹	$\frac{1}{4}$
↑↓ ≀	⟹		⟹	$\frac{1}{4}$
↓↓ ≀	⟹		⟹	$\frac{1}{2}$
↑↑ ≀	⟹		⟹	

Table 2.2: This table illustrates how templates are distributed among their constituting subgraphs on the example of the template *uun*.

the probability, that a single position is selected twice p^2, times the number of possible slots $N(N-1)/2$:

$$M_{bi} = \frac{M^2}{2N(N-1)}.$$

The number of uni-directional edges is then given by:

$$M_{uni} = M - \frac{M^2}{N(N-1)}.$$

2.2.3 Subgraph Counts

In order to obtain expectation values for the subgraph counts c_m, we formulate a simple model of subgraphs, where each of the three positions between the three nodes can be in one of three states:

- uni-directional edge
- bi-directional edge
- no edge

see also Figure 2.2.1.

The edge density of the graph is defined as $p = \frac{M}{N(N-1)}$, and the probabilities for the three states are then given by:

$$\begin{aligned} p_u(p) &= \quad 2 \cdot (p - p^2) \\ p_b(p) &= \quad p^2 \\ p_n(p) &= \quad 1 - p_u - p_b = (1-p)^2 \end{aligned}$$

By denoting the numbers of uni-directional (u_m), bi-directional (b_m), and non (n_m) -edges for every subgraph m, we can write the expected number c_m of type m as

$$c_m = pl \cdot p_u(p)^{u_m} \cdot p_b(p)^{b_m} \cdot p_n(p)^{n_m} \cdot s_m$$

where the number of possible placements for a subgraph is $pl = \binom{N}{3} \cdot 3!$ and s_m are symmetry factors

$$s_m = r_m/\xi_t$$

where ξ_t accounts for the symmetries of the template t and r_m represents the ratio by which the template is split up into subgraphs.

The symmetry factor ξ_t is the ratio of possible three-symbol permutations of the distinct permutations obtained in a template. A template containing three distinct symbols exhausts the full possible 6 permutations, yielding $\xi_t = 1$, while a template with two distinct symbols allows for three permutations, yielding $\xi_t = \frac{6}{3} = 2$, and if all three symbols in the template are equal, we have $\xi_t = 6$.

As the symmetry factor r_m only accounts for the distribution of the templates on the subgraphs, see Table 2.3, there must be $\sum r_m = 1$ for every template, where the sum is over all subgraphs m in the template t.

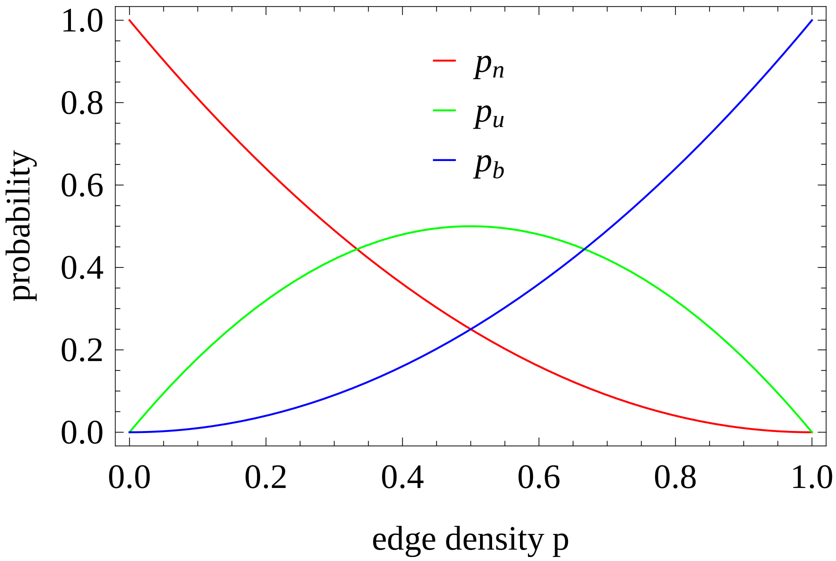


Figure 2.2.2: The probabilities p_u, p_b and p_n as functions of the edge density p.

m	1	2	3	4	5	6	7	8	9	10	11	12	13
t	1	1	1	2	2	3	4	4	5	5	5	6	7
u_m	2	2	2	1	1	0	3	3	2	2	2	1	0
b_m	0	0	0	1	1	2	0	0	1	1	1	2	3
n_m	1	1	1	1	1	1	0	0	0	0	0	0	0
r_m	4	4	2	2	2	1	$\frac{4}{3}$	4	4	4	2	1	1
ξ_t	2	2	2	1	1	2	6	6	2	2	2	2	6
s_m	8	8	4	2	2	2	8	24	8	8	4	2	6

Table 2.3: The number of uni (u_m), bi (b_m), and non (n_m) edges and the symmetry factors in all subgraphs.

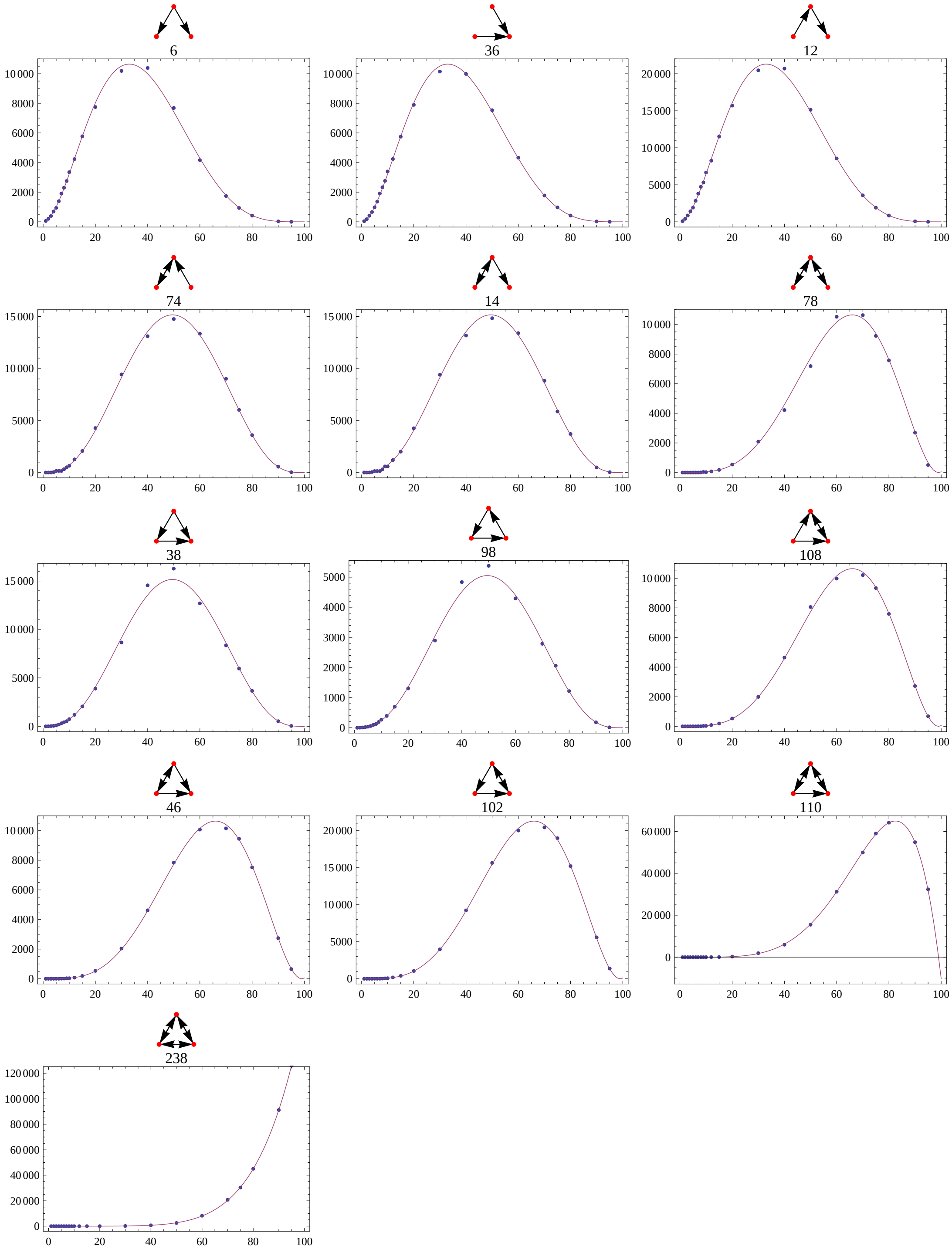


Figure 2.2.3: To illustrate the prediction quality of the subgraph counts we show the predicted (full curve) and numerically observed (dots) counts of the 13 3-node subgraphs in a random network with $N = 100$. The connectivity is varied over the whole range ($p = 0...100\%$). The numerical points are obtained by averaging over 100 random graphs.

2.2.4 Edge Fluctuations

Changing any one of p_u, p_b or p_n by a small probability Δ at fixed edge density p results in the change of the other two probabilities according to:

$$\begin{aligned}\hat{p}_u(p) = &\quad p_u(p) + 2\Delta = 2\cdot(p-p^2) + 2\Delta\\ \hat{p}_b(p) = &\qquad p_b(p) - \Delta = p^2 - \Delta\\ \hat{p}_n(p) = &\quad p_n(p) - \Delta = 1 - p_u - p_b - \Delta\end{aligned}$$

This is because the creation of a bi-directional edge needs two uni-directional edges and frees one place.

To be able to infer the fluctuations of subgraphs it is useful to first derive equations for the fluctuation of bi-directional edges.

The expectation value of the number of bi-directional edges is $c_b = \frac{p^2\cdot N^2}{2} = \frac{1}{2}(\frac{M}{N})^2$.

In order to estimate the fluctuations in the number of bi-directional edges at low edge densities (and large numbers of nodes), we take the expectation value and variance $\lambda = nP$ for the Poissonian distribution, but substitute the event number n by the number of possible sites for bi-directional edges, $\binom{N}{2}$, and the event probability P by the probability of two edges, p^2. For the standard deviation we thus obtain: $\sigma_{bl} = \sqrt{c_b}$ at low densities.

When the number of edges approaches its maximum value $M = N(N-1)$ these fluctuations decrease again, which is due to the decreasing number of places that are not yet occupied by single edges that one would have to hit to not create another bi-directional edge. In this case, the event probability is substituted by $(1-p)^2$. It is also clear that σ_{bh} must be symmetric around $p = 0.5$.

So at high densities we get:

$$\sigma_{bh} = \sqrt{\frac{N^2}{2}} - \sqrt{c_b}.$$

As both fluctuations are mutually exclusive, their reciprocal sum yields an analytical expression for the total fluctuations of bi-directional edges as a function of the edge density p, i.e.

$$\sigma_b = \frac{1}{\frac{1}{\sqrt{\sigma_{bl}}} + \frac{1}{\sqrt{\sigma_{bh}}}}$$

This situation is summarized in Figure 2.2.4. These fluctuations directly transfer to fluctuations of the uni-directional and non-edges. As every additional bi-directional edge means two uni-edges less, there is a factor of two between their fluctuations, see Figure 2.2.4.

2.2.5 Subgraph Fluctuations

In order to reduce trivial contributions to the subgraph fluctuations, we keep the number M of edges in the ER graph fixed (which makes the edge density p a secondary quantity, as described above). Otherwise, the fluctuations in the number of edges at a given p would partially mask the conceptually more important (and less trivial) contribution from fluctuations of uni-directional and bi-directional edges at fixed M.

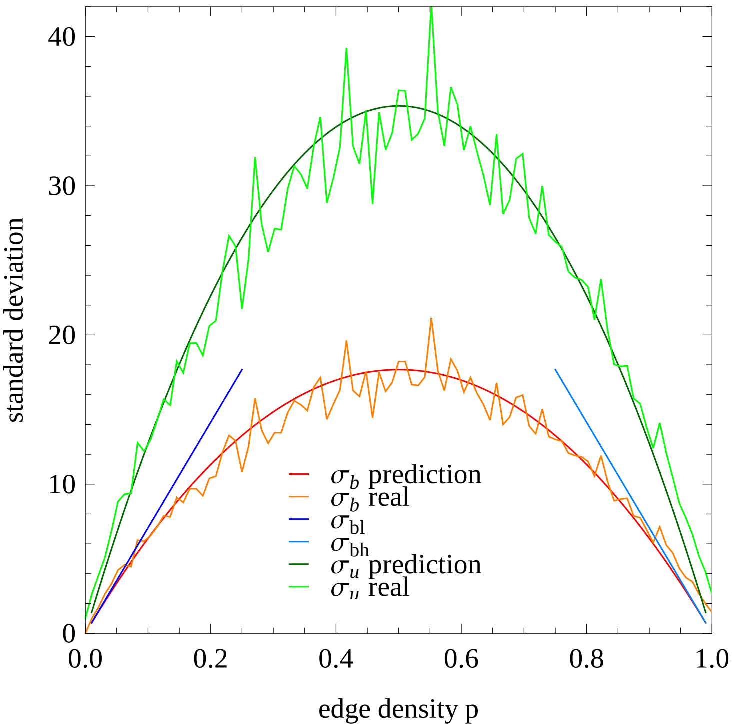


Figure 2.2.4: The fluctuations of the number of uni- and bi-directional edges in a random graph with $N = 100$ nodes and varying edge density. We show the numerical results together with the corresponding predictions.

There are two reasons for the fluctuation of a subgraph count: (i) fluctuations in the number of bi-directional edges, (ii) fluctuations due to the subdivision of templates (i.e. subgraphs with the same number of uni-, bi- and non-edges) into subgraphs. This subdivision depends on the direction of the uni-directional edges, as discussed in Tables 2.1 and 2.2.

Contribution (i): The fluctuation of the number of bi-directional edges can be translated into the fluctuation of a subgraph count by processing the normal number of subgraphs and subtracting that from the number of subgraphs one gets by changing the probabilities for uni-, bi- and no-edges by one standard deviation:

$$\sigma_{bm} = c_m(\hat{p}_u, \hat{p}_b, \hat{p}_n) - c_m(p_u, p_b, p_n).$$

Contribution (ii): The other sources of fluctuations are the fluctuations in the combination of uni- and bi-directional edges to templates and the distribution of the templates among the subgraphs. Together they can be estimated by the square root of the subgraph count:

$$\sigma_{mm} = \sqrt{c_m}$$

These sources of fluctuations need to be combined in a pythagorean sum:

$$\sigma_m = \sqrt{\sigma_{mm}^2 + \sigma_{bm}^2}$$

The resulting fluctuations are shown in Figure 2.2.5.

2.3 Application

Here we will show for a simple example how the theory presented above can be used to better understand properties of subgraph signatures. The subgraph signature of a network (or, more specifically for three-node subgraphs, the triad significance profile, TSP) is the pattern of over- and under-representations of few-node subgraphs in this network. It has become a standard method of analyzing complex networks. More formally, it is the (normalized) z-score of the subgraph counts.

To obtain the ensemble of randomized networks a randomization scheme is repeatedly applied, where typically the in- and out-degree of each node (i.e. the degree-sequence of the graph) is conserved during the randomization process, as well as the number of bi-directional edges at each node. The aim of the randomization procedure is to remove any non-random property (beyond the degree-sequence). In this way deviations of the subgraph counts (in the real network) from randomness can be detected and functionally interpreted.

Apart from the case where some kind of selective process in the evolution of a network or some other functional requirement is enriching specific subgraphs, which is the most interesting case, there are many other reasons for a non-zero z-score.

Here we discuss modularity as one such possible reason. An example of a modular network is depicted in Figure 2.3.1. If the modular structure is not taken into account during the randomization process and thereby conserved in the pool of randomized networks (i.e. eliminated from its effect on expected subgraph numbers), a false non-zero z-score appears.

We use a random graph that is composed of five strong modules, where each module is an ER-network. Additionally a certain amount of inter-module edges is introduced. As the base networks as well as the inter-module edges are constructed in a motif-blind way, correct

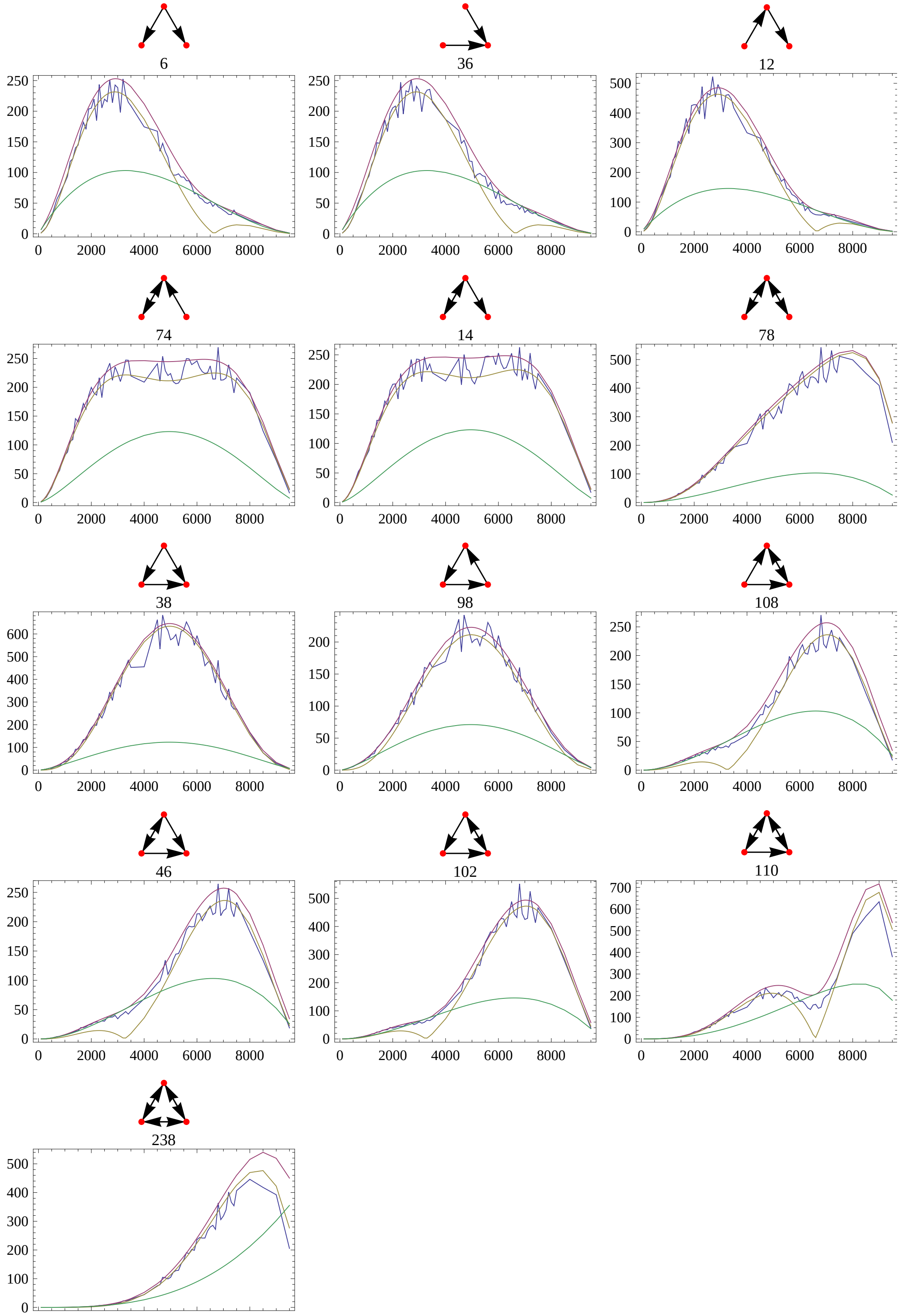


Figure 2.2.5: The fluctuations of the 13 3-node subgraphs in a random network with $N = 100$. The connectivity is varied over the whole range.($M = 0...9900$) and the contribution (i) (brown), contribution (ii) (green), as well as the total expected value (purple) and numerical results (blue) are shown.

randomization should yield a flat motif signature with z-scores close to zero. The result of the application of standard, module-blind randomization techniques can be seen in Figure 2.3.1. We also show the result of a module-aware randomization scheme that mixes only edges inside of the modules and inter-module edges. In real world networks the modular structure of a network is generally not known and it is therefore necessary to detect the modules first, before adjusting the randomization scheme accordingly.

To better understand the error made by the standard randomization scheme we will analytically predict the error signature using the formalism introduced above. To this end, it is essential to notice that, when the modules are destroyed the effective local intra-module density of the network is reduced by a factor of five.

This is because a network with $N^* = 2N$ and $M^* = 2M$ has a density of $d^* = \frac{M^*}{N^{*2}} = \frac{d}{2}$. A network with double size has to have four times the edges to have the same density.

Let N and M be the number of nodes and edges of the whole modular network. Then c_m can be estimated by $5 \cdot c_m(N/5, M/5)$, μ_m=$c_m(N, M)$ and σ_m by $\sigma_m(N, M)$. The general form when a graph consists of k modules with node counts $n_1, n_2, ..., n_k$ is

$$c_m = \sum_{i=1}^{k} c_m(n_i, \frac{Mn_i}{N}).$$

When the ratio of inter-module edges ρ is increased this can easily be taken into account:

$$c_m = \sum_{i=1}^{k} c_m(n_i, \frac{(1-\rho)Mn_i}{N}) + c_m(N, \rho M).$$

This simplification does not acknowledge for subgraph instances that contain intra- and inter-module edges. These are relevant mostly for the two-edge subgraphs. We evaluate the number of these mixed subgraphs $d_m$by taking into account the different edge densities in the module and between the modules. We therefore introduce probabilities for uni- and bi-directional edges in the components p_{uc}, p_{bc} and outside of the components p_{uo}, p_{bo}. Using these probabilities we can write the expectation value for the additional subgraphs as:

$$d_m = N^3 \begin{cases} p_{uc}p_{uo} + p_{uo}p_{uc} & \text{, where } u_m = 2 \wedge b_m = 0 \\ p_{uc}p_{bo} + p_{uo}p_{bc} & \text{, where } u_m = 1 \wedge b_m = 1 \\ p_{bc}p_{bo} + p_{bo}p_{bc} & \text{, where } u_m = 0 \wedge b_m = 2 \\ 0 & \text{, else} \end{cases}$$

These additional subgraphs are added to the inter-module and intra-module subgraphs to obtain the total subgraph counts. Figure 2.3.2 shows the quality of the prediction of the subgraph counts.

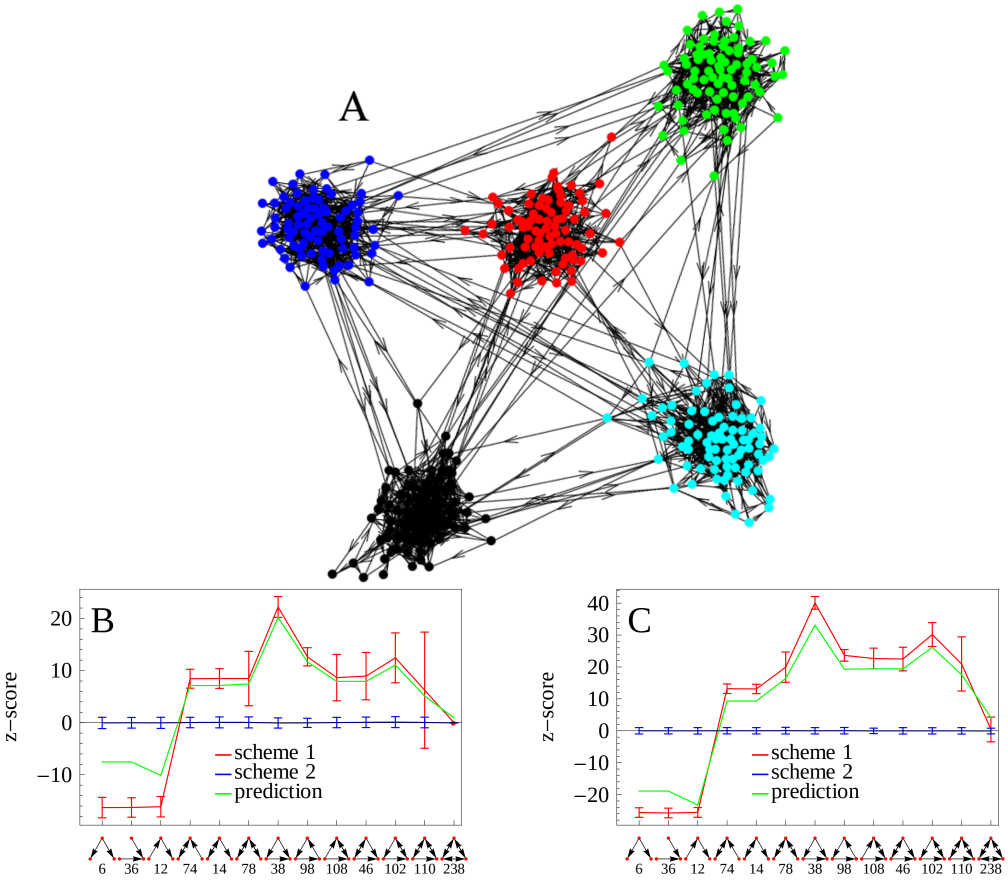


Figure 2.3.1: Motif z-scores for a network that is composed of five strong modules. (A) Example of such a random modular graph. Two different randomization-schemes are applied (1) simple flipping of two edge-endpoints, (2) flipping while preserving the module-structure. As the analyzed network is random apart from its modularity the z-score using the correct randomization scheme must be 0. This is shown for different densities: (B) $N = 500, M = 2000, \rho = 0.08$ and (C) $N = 500, M = 8000, \rho = 0.16$.

To verify the quality of the predictions of the z-score of a modular network, we compute the geometric mean of the difference of the z-score when applying the appropriate module-aware randomization scheme and the simple randomization scheme. This quantity can both be computed numerically and analytically, as in Figure 2.3.3.

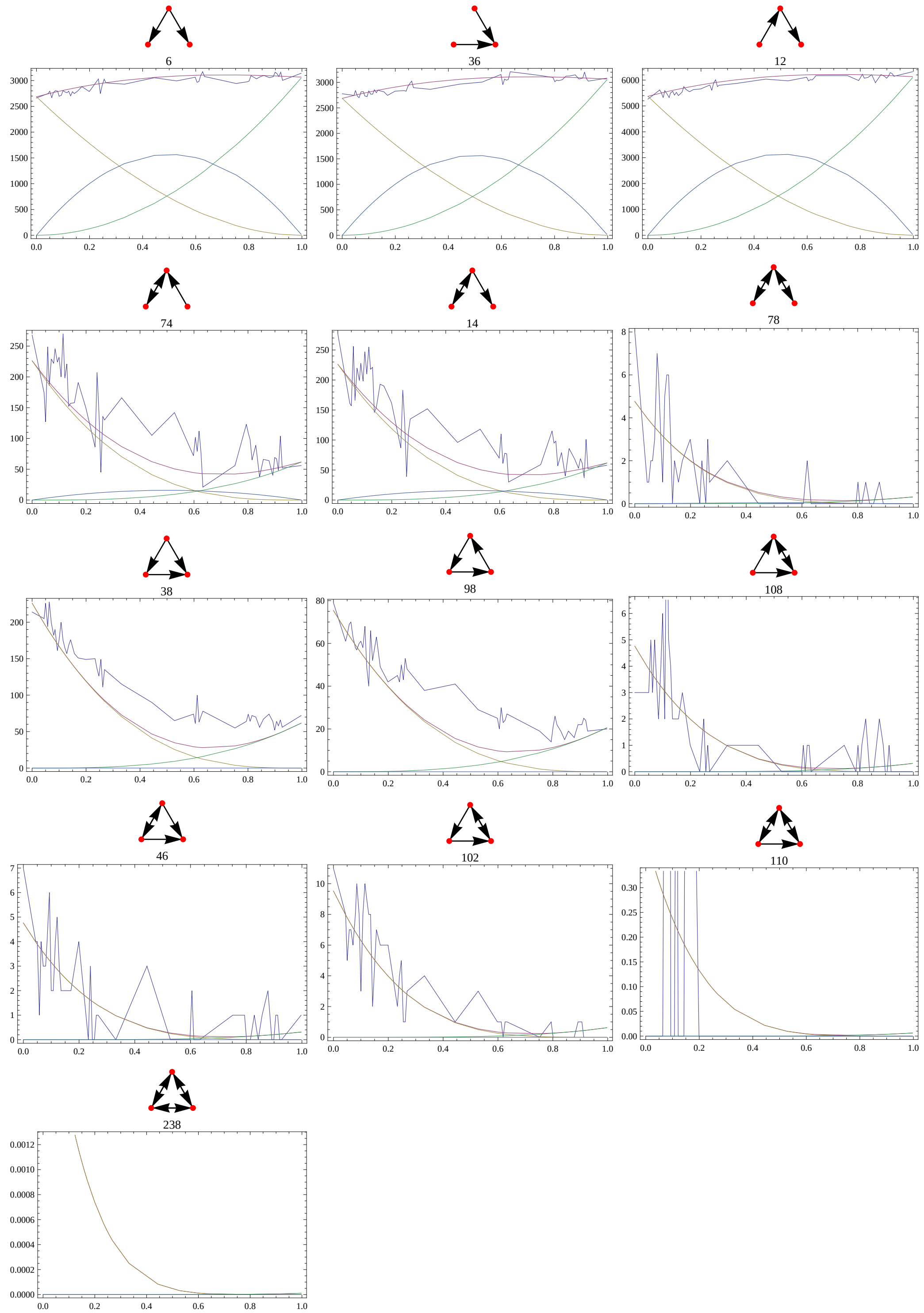


Figure 2.3.2: The composition of the 13 3-node subgraphs in a modular network. The ratio of inter-modular edges is varied and the contribution of intra-module subgraphs (yellow), inter-module-subgraphs (green), mixed subgraphs (blue) as well as the total expected value (red) and numerical results (blue) are shown (N=500,M=2000).

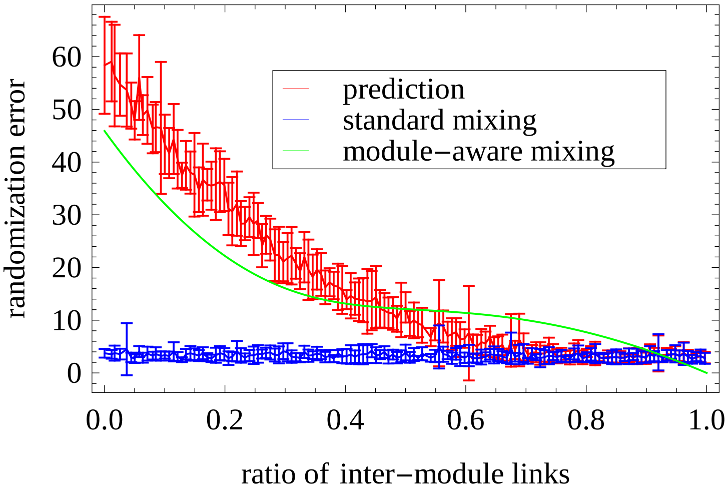


Figure 2.3.3: The sum over the squared errors that occurs from applying different randomization schemes over the ratio of inter-module edges (N=500,M=2000). We show the numerical results for two different randomization schemes together with our prediction.

Conclusion

Statistical properties of random graphs have been studied for decades in several disciplines and with a wide range of applications in mind. Here we have focused on a topic that in spite of its practical importance has received comparatively little attention so far, namely the statistical fluctuations of few-node subgraphs induced by lower-level fluctuations in the numbers of uni-directional and bi-directional edges.

In this way we can quantitatively understand some of the cross-talk between global and local network properties. As an example of such a cross-talk we have here presented the motif signature arising from the modularity of the graph. Using our analytical description of subgraph fluctuations, we can precisely predict the artefactual motif signature of this otherwise random graph. By mixing inter-module edges and the different sets of intra-module edges independently, we can additionally show that the full motif signature is a sole consequence of the modular graph structure.

Beyond a better understanding of such artifacts, we believe that the classification of three-node subgraphs into the categories introduced in Section 2.2 has the potential of unraveling the theoretical background behind the empirical observation that only four variants (or 'super-families') of three-node motif signatures are observed across a vast range of complex networks [107].

It is clear that all subgraphs within the same category will display synchronous fluctuations distributed among the participants of a category according to few well-understood combinatorical factors. This approach may constitute a solid basis for understanding correlated subgraph fluctuations and motif-motif covariations. We plan to pursue some of these questions in future work.

Chapter 3

Artifacts in Statistical Analyses of Network Motifs

Summary

Network motifs are on a mesoscopic scale between purely local and global network properties. They can be interpreted as building blocks that shape the dynamic behavior of networks. It is this promise of potentially explaining emergent properties of complex systems with relatively simple structures that led to an adaptation of motifs in an ever-growing number of studies and across disciplines. Here we discuss artifacts in the analysis of network motifs arising from incongruences between the network under investigation and the pool of random graphs serving as null model. Our aim is to provide a clear and accessible catalog of such incongruences and their effect on the motif signature. Specifically, we explore the effect of bidirectional edges, modularity, self-links, randomization of layered graphs, projections of bipartite graphs, and the mapping of dynamical data onto motifs.

The results presented in this chapter have been achieved in cooperation with Moritz Beber, Matthias Müller-Hannemann and Marc-Thorsten Hütt and have been published in: "Artifacts in statistical analyses of network motifs" [43].

3.1 Introduction

Analyzing few-node subgraphs and network motifs has become an indispensable tool for understanding complex networks. The conceptual power of network motifs lies in making accessible an intermediate scale in network organization. It is *a priori* not clear, on which topological scale a particular dynamical function is located (i.e., what a typical group size of nodes is that contributes to the function). Quite clearly (in particular for very large networks) it is implausible that functional features are a truly collective phenomenon that can only be understood on the scale of the full network.

While many studies of network topologies focus on global properties (e.g., the degree distribution — reviewed in [117], modularity [49, 55, 56, 122], degree correlations [31, 116], and hierarchical structures [12, 36, 69, 157, 135]), some of the dynamical function can be explained by small few-node subgraphs serving as devices for specific tasks organized locally in the graph. A potential signature of the functional role of few-node subgraphs is their statistical over- or under-representation (compared to a suitable ensemble of random graphs). This general concept has been developed and worked out by the Alon group [107, 109], particularly for transcriptional regulatory networks [3, 148].

Identifying a statistical over-representation of few-node subgraphs, i.e., network motifs, has two main advantages:

1. For these specific subgraphs it is then possible to develop detailed mathematical models of their dynamic function [3, 97].

2. Once the motifs are identified, the individual occurrences, e.g., in the biological network under investigation, can be discussed and analyzed experimentally [73] mostly on a technical and methodological level [6, 78] as well as in cross-validations of the proposed function [103].

Most criticism of motif analyses has focused on the mixing with a randomization procedure [6, 17, 19, 48]. It has also been argued that the crosstalk between global and local network properties may affect the statistical assessment of network motifs [65].

In [160] the strong dependence of local and global network properties for scale-free and hierarchical graphs has been formally explored. Relatedly, [74] has formulated an algorithm for generating a graph with a prescribed motif composition in the special case of very low connectivities.

In [59] the crosstalk between two graph properties (assortativity and clustering coefficient) has been studied by exploring the parameter space with the help of a biased random walk in the ensemble of all graphs with fixed degree sequence. Particularly for biological networks, it has been questioned, whether network motifs are indeed of functional relevance [61, 78, 103].

Here, we extend the topological arguments from this list and provide a clear and transparent catalog of possible artifacts arising from incongruences or mismatches between the network under investigation and the pool of randomized graphs serving as null model.

Due to the comparison between properties of a real network with a suitable set of random graphs, an argument involving motifs is a delicate balance between statements formulated about local graph properties and statistical arguments on a more global scale. More generally, it is an inescapable difficulty to disentangle the motif layer of organization from other scales within the network, if one wants to clearly attribute the system features to just the meso-scale of few-node subgraphs. Key questions are:

1. Do global graph properties distort the content of subgraphs and can a null model account for this?

2. Is the local function of a network affected by the specific architecture of a motif or does that function result from a systematic placement of that motif in its network "environment"? Is the function of isolated motifs comparable to the functioning of motifs contextualized within a graph?

3. Are absolute numbers of motifs interpretable quantities, when their scaling properties (with degree, clustering coefficient, connectivity, etc.) are taken into account?

Here we focus on the first question. The work by [48] has put forward a local argument for motif correlations induced by a property of the randomization scheme. While this argument (in the form given in [48]) is true only for graphs without bidirectional links (and therefore a reduced motif inventory) and for small densities (as otherwise a wider range of randomization steps will be available), the argument nevertheless shows the possibility of subtle intrinsic correlations influencing the result of a motif analysis. In [7] an algorithm for constructing modular hierarchical graphs is described. The authors in particular observe that their constructed graphs have a non-random motif composition, resembling one of the superfamilies from [107].

In the following, we will summarize and illustrate several technical issues in analyzing network motifs, both from a topological and a dynamical perspective. We want to systematically assess how global properties change local properties (for example, the influence of modularity on motifs). A running example throughout the different types of global-local crosstalk will be metabolic networks, where all these inter-dependent topological features are very important. Hence we will often elaborate on the application of a particular issue to such networks.

The organization of the text is as follows: A short overview on motifs, together with the theoretical framework employed here and the network representation of metabolism in the next Section; an insight into the effects of local network architecture and local decisions in Section 3.3; a detailed account of global network structure that strongly influences the meso-scale in Section 3.4; issues in mapping (dynamical) data onto motifs in Section 3.5; and in Section 3.6 some aspects of further work that will be necessary.

3.2 Methods

3.2.1 Terminology

The general idea of a motif analysis is to compare few-node subgraph counts obtained from a real network with the corresponding counts obtained from randomized versions. In the case of three-node subgraphs, the corresponding z-score can be summarized in a triad significance profile (TSP) showing the statistical over- or under-representation of each of the subgraphs. The z-score for subgraph m is defined as: $Z_m = \frac{c_m - \mu_m}{\sigma_m}$, where for every subgraph, c_m is the subgraph count in the original network, μ_m is the expectation value of c_m in the random networks and σ_m is the standard deviation of c_m in the random networks. A typical randomization scheme preserves the number of incoming, outgoing and bidirectional edges at each node.

3.2.2 Statistical Description of Motif Fluctuations

Here we briefly summarize the formalism described in Chapter 2. We will apply to the examples given in Section 3.4.1, in order to understand the effects described there also analytically.

The formalism starts from the fact that in Erdős-Rényi (ER) graphs three-node subgraph frequencies c_m can be described by the number of possible selections of three nodes $\pi = \binom{N}{3} \cdot 3!$ (where N is the number of nodes in the graph) times the probability that a particular subgraph is present at this place. This probability can be written as $P_u(p)^{u_m} \cdot P_b(p)^{b_m} \cdot P_n(p)^{n_m} \cdot s_m$ with the numbers of uni-directional (u_m), bi-directional (b_m), and absent (non) (n_m) edges for the subgraph m. The probabilities for the occurrence of a uni-directional (P_u), bi-directional (P_b),

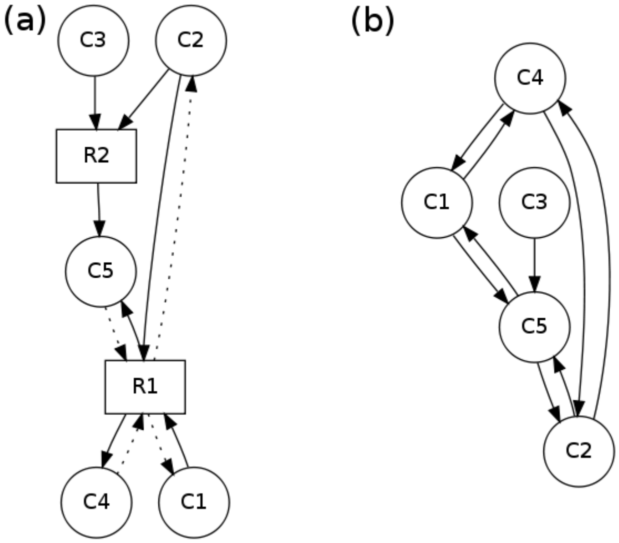


Figure 3.2.1: Projection of a metabolic bipartite network (a) onto its metabolite nodes (b). The reversibility of a reaction is internally stored as a node attribute but represented here with an additional dashed link, so that substrate and product nodes can still be identified.

and absent (P_n) edge are given by

$$\begin{aligned}P_u(p) &= 2 \cdot (p - p^2)\\ P_b(p) &= p^2\\ P_n(p) &= 1 - p_u - p_b = (1-p)^2\end{aligned}$$

where p is the link density of the graph, $p = \frac{M}{N(N-1)}$, M being the number of links in the graph. The expected number of occurrences of a subgraph m also contains a symmetry factor s_m that can be obtained from the number of permutations possible for that subgraph, as well as the number of different subgraphs formed from this constellation of u_m, b_m and n_m.

This model can be extended to take into account fluctuations of motifs and therefore to compute z-scores for some simple situations. Details can be found in Chapter 2.

3.2.3 Network Representations of Metabolic Systems

Metabolic networks reflect the sharing of metabolic compounds (metabolites) by all biochemical reactions that can occur in an organism due to the catalytic action of enzymes. Metabolites and reactions form distinct sets of nodes that are only interconnected, thus forming a bipartite network (more involved representations exist, e.g., including enzymes as another category of nodes [1], or choosing a hyper-graph structure to meticulously represent reaction-compound relations [169]). Since the mathematical formalisms developed to analyze networks have been focused on unipartite graphs, metabolic networks are customarily projected onto either one of the two sets of nodes and the projected unipartite network is then analyzed. In the

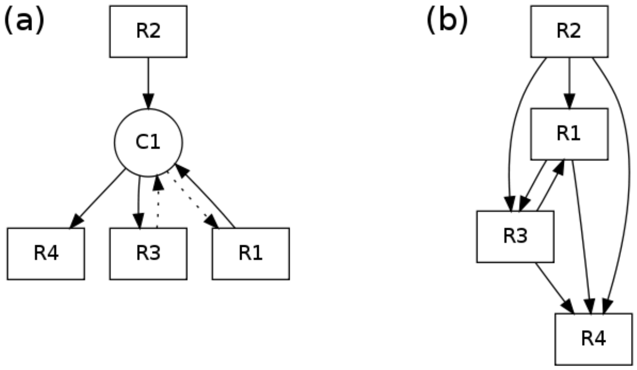


Figure 3.2.2: Projection of a metabolic bipartite network (a) onto its reaction nodes (b). As in Figure 3.2.1, the reversibility of a reaction is internally stored as a node attribute but represented here with an additional dashed link, so that substrate and product nodes can still be identified.

majority, these analyses considered metabolism as an undirected network (see, for example, [1, 68, 161] and for a critique of the projection of metabolic networks see [110]). The problem of formulating suitable null models for metabolic systems has recently been discussed in [13]. There the problem of mass balance is studied. One of the first studies on motifs in metabolic networks [39], also presents a projection method for directed metabolic networks. We retain the method presented in [39] for projecting the metabolic network onto its metabolite-centric representation, i.e., connecting each substrate with each product of a certain reaction and introducing a bidirectional link if the reaction is reversible (see the scheme in Figure 3.2.1). For a directed representation of the reaction-centric projection, we draw a link between two reaction nodes if there is a directed path of length two from a source reaction R, through exactly one metabolite M, to another reaction R'. The link is drawn in the direction of the existing path (as depicted in Figure 3.2.2).

There are two constraints on the bipartite metabolic network: If a projection onto metabolite nodes is required, the structure of the network needs to give insight into the substrates and products of each reaction at any time. That means, we cannot simply introduce bidirectional edges wherever reversible reactions occur, because then the structure of the network would not allow the unambiguous identification of their substrates and products. The second requirement is that reversible reactions form their own category and should be mixed separately from other reactions just as uni- and bidirectional links. There are at least three ways of realizing these requirements: Reversible reactions can be split up into their forward and backward direction which allows a clear identification of products and substrates. During the rewiring process, they should be regarded as one entity, though. Another approach is to give reversible reactions a special attribute, they are then mixed only with themselves, and in the projection process bidirectional links can be introduced again. The last option is to give each link an attribute that records whether it connects to a substrate or a product.

3.3 Local Network Properties

3.3.1 Self-Links

Self-links are typically not included in the definition of three-node subgraphs and are, therefore, not part of a motif analysis, and should be deleted from the graph before the analysis. In fact, the *mfinder* tool (see Section 3.2.1) will outright reject a graph containing self-links. Clearly, any rewiring method must prevent the creation of self-links or the sample of randomized graphs will have a lower connectivity (density) than the initial one. This functionality is included in all current motif analysis tools. Nevertheless, self-links can be very meaningful. Prime biological examples are auto-regulation or fast degradation.

As a first, simple example of possible artifacts in the analysis of motifs, it is instructive to explore, how an inadvertent inclusion of self-links will affect the result of the motif assessment. If self-links are considered valid edges for rewiring, they can thus increase the connectivity of the randomized graph and skew motif counts. Let us consider the following example with three nodes A, B, and C and two edges $A \to A$ and $B \to C$. A rewiring would then create the two edges $A \to C$ and $B \to A$. Upon rewiring, another edge was created that may participate in a few-node subgraph and the relevant connectivity for motif analyses has increased. The increasing error in the z-scores with an increasing number of self-links is shown in Figure 3.3.1.

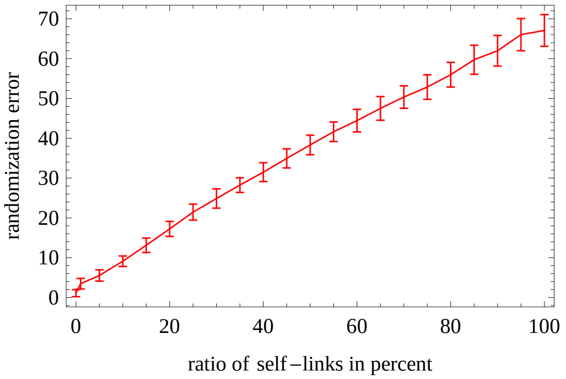


Figure 3.3.1: The increase in the sum over the squares of z-scores for random networks with 100 nodes and 400 links with an increasing number of self-links. The standard deviation of an ensemble of 100 networks is shown.

In the case of metabolic networks, the underlying biochemical reactions dictate a bipartite structure consisting of compound and reaction vertices. Currently, the motif analysis is only (convincingly) defined for unipartite graphs (see, e.g., [75] for a phenomenological extension of the motif concept to bipartite graphs). The projection schemes discussed in Section 3.2.3 may introduce self-loops that should be deleted for the above reasons. In this instance, they are not even biologically meaningful and thus have no place in the network structure.

The distortion of a motif analysis by such faulty inclusion of self-links in the randomization process is an example of the more general case of a mismatch between the network and the pool of randomized graphs in terms of link density. In Section 3.4.1 we will discuss another, quite prominent, graph property, modularity, inducing such a density mismatch.

3.3.2 Categorizing Bidirectional Edges

A topic that from our perspective has not received sufficient attention in the current debates about complex networks is the role of bidirectional links. A bidirectional link in a given network can be viewed as belonging to a unique category of links (as opposed to unidirectional links) or, alternatively, as the simultaneous occurrence of two unidirectional links with opposite orientation. It should be noted that in real networks both types of bidirectional links can in principle occur. It is plausible for example to assume that in gene regulatory networks a bidirectional link is rather the overlay (or co-occurrence) of two unidirectional links, while for example in metabolic networks reversible reactions certainly constitute an example of true bi-directionality as an individual category. In metabolism one also finds, however, the case of two distinct enzymes being responsible for the two opposing directions of a reaction and, hence, the other interpretation of bidirectional links.

Typically, the randomization procedure employed for quantifying the over- and under-representations of few-node subgraphs regard bidirectional links as an individual category of links in the network (thus applying randomization steps independently to both unidirectional sub-links and conserving in this way the number of bidirectional links at each node).

Does the distinction between two types of bidirectional links and the related distinction between two interpretations of the classical randomization scheme affect the observed motif signature?

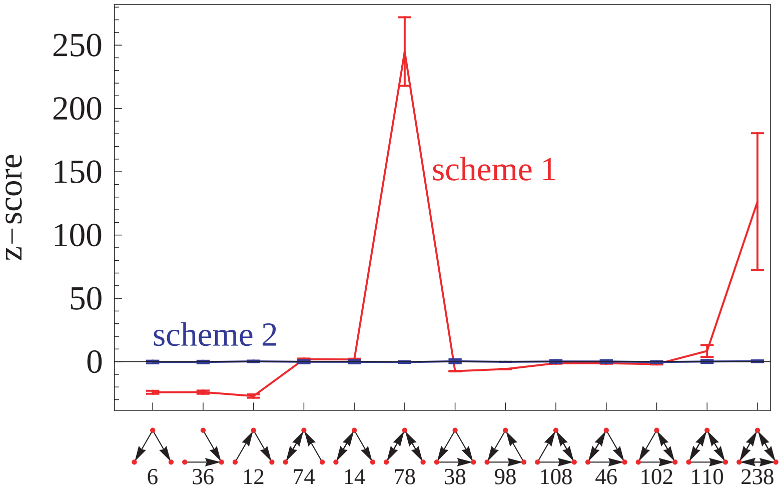


Figure 3.3.2: z-score for a graph with an elevated number of bidirectional links compared to the expected value μ. Two different randomization-schemes are applied (1) simple flipping of two edge-endpoints, (2) shuffling uni- and bidirectional links independently thus preserving the number of bidirectional links in the network. As the analyzed network is random apart from its number of bidirectional links the z-scores using the correct randomization scheme can be expected to be close to 0. Obtained from graphs with $N = 100$ and $M = 400$, the number of bidirectional links was forced to 90% instead of the 16% expected for ER graphs.

For a directed graph with N vertices, $N\,(N-1)$ is the number of possible edges and

$N(N-1)/2$ is the maximal number of bidirectional edges. If we assume sparse random graphs (i.e. no mutual exclusion of edges) with M edges, we can write $p = M/N(N-1)$ as the probability that a specific edge is present and $p^2 = M^2/(N(N-1))^2$ is the probability that a specific edge is bidirectional. For a given number of nodes N and links M the expected number of bidirectional links is thus $B = p^2N^2/2 = M^2/\left(2(N-1)^2\right)$. With the average degree $k = M/N$ we can then write $B \approx k^2/2$.

If the ratio of bi-directed edges does not correspond to this expectation value, randomization scheme (1), which consists of a bidirectional link as two independent unidirectional links, which yield a distorted triad significance profile (TSP), whereas the bi-directionality-preserving randomization scheme (2) will yield the expected z-scores close to zero. Figure 3.3.3 shows an example of this difference between the two schemes.

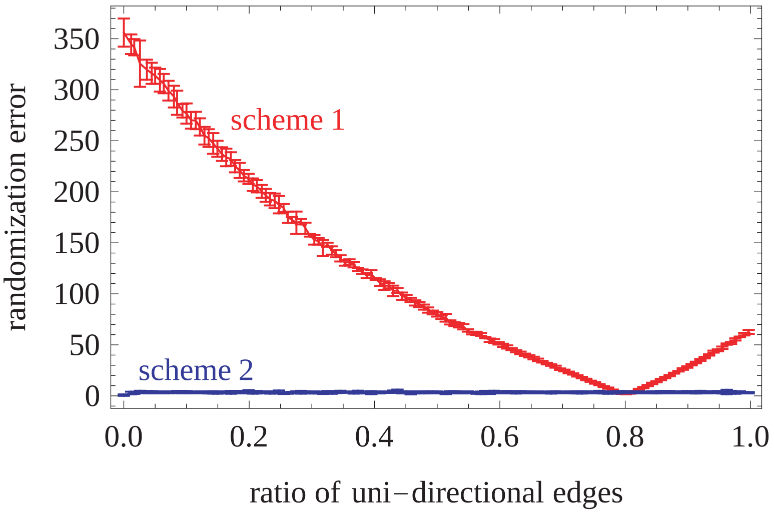


Figure 3.3.3: The sum over the squared errors that occurs from applying different randomization schemes over the ratio of bi-directional links. This network has $N = 100$ and $M = 1980$, leading to a normal ratio of unidirectional edges of 0.802. If this ratio is altered, while keeping other graph properties random, the normal randomization scheme (1) yields distorted results, whereas the bi-directionality-preserving randomization scheme (2) yields a correct z-score close to zero.

The difference between the two motif signatures depends strongly on the connectivity of the graph. In fact, the difference between the expected value of bidirectional links from connectivity and the real observed value of bidirectional links in the one scheme (regarding them as two unidirectional links) generates a gradient indicating, whether the applied randomization process will rather reduce or increase the number of bidirectional links and, hence, differently populate the two subcategories of motifs (those containing only unidirectional links and those containing also bidirectional links).

3.4 Global Network Properties

3.4.1 Modularity

The first and obvious complication in a trivial motif analysis arises, when a larger-scale pattern like modularity or a spatial embedding of the network induces a strong deviation between the original graph and a reference graph obtained by standard randomization procedures. For spatially organized graphs this has been pointed out in [6]. Here we analyze the fictitious motif signatures arising from modularity in more detail.

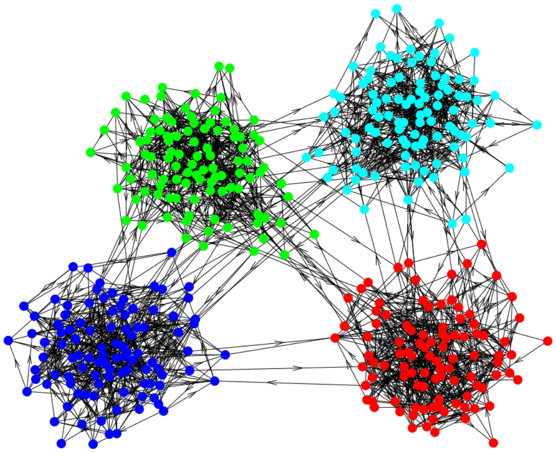


Figure 3.4.1: A graph that is composed of four strong modules.

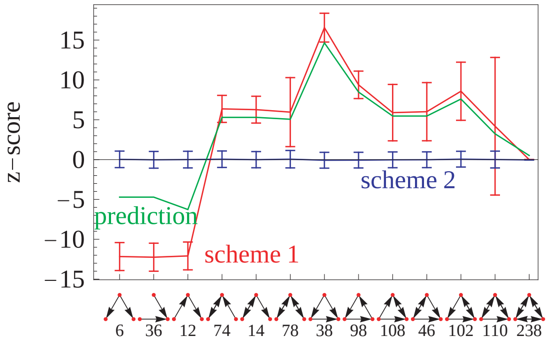


Figure 3.4.2: z-score of a graph that is composed of four strong modules. Two different randomization schemes are applied (1) simple flipping of two edge-endpoints, (2) flipping while preserving the module-structure. As the analyzed network is random apart from its modularity the z-score using the correct randomization scheme must be 0. Additionally the analytical prediction of the z-score is drawn. $N = 400$, $M = 1600$, 8% inter-modular links.

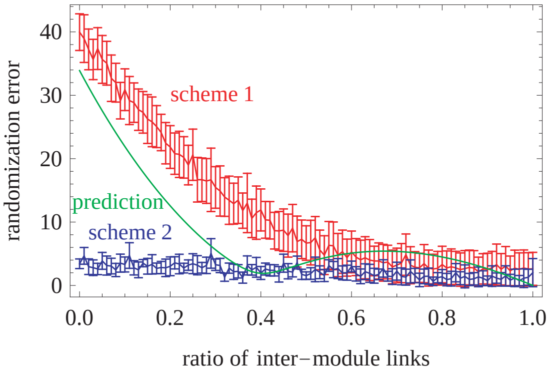


Figure 3.4.3: The sum over the squared errors that occurs from applying different randomization schemes over the ratio of inter-module links. Two different randomization schemes are applied (1) simple flipping of two edge-endpoints, (2) flipping while preserving the module-structure. As the analyzed network is random apart from its modularity the z-score using the correct randomization scheme must be 0. Additionally the analytical prediction of the errors is drawn. $N = 400$, $M = 1600$.

We assemble a graph from four dense modules, where each module is a directed ER graph. Additionally, a few inter-module links are introduced. As the elementary graphs, as well as the inter-module links are constructed in a motif-blind way, correct randomization should yield a flat TSP, with all individual z-scores close to zero. The result of the application of standard randomization techniques can be seen in Figure 3.4.2. We also show the result of a modularity-aware randomization scheme that mixes intra-module links and inter-module links separately. In real-world networks the modular structure is generally not known and thus the quality of this modularity-aware randomization scheme will depend on the quality of the module detection algorithm employed. In order to better understand the error made by a randomization scheme without any module information, we will perform an analytical calculation that yields a prediction of the error signature. To obtain predictions for the number of subgraphs c_m of type m in the original graph, the expectation value of c_m in the random graphs μ_m, and its standard deviation σ_m we use a simple model of few-node subgraphs, taken from Chapter 2.

It is essential to note, that when the modules are destroyed the effective local intra-module density of the network is reduced by a factor of the number of modules. This is because compared to a network with N nodes and M edges a network of twice the size with $N^* = 2N$ nodes and $M^* = 2M$ edges has a density of:

$$d^* = \frac{M^*}{N^*(N^*-1)} \approx \frac{d}{2}.$$

Let N and M be the number of nodes and edges of the whole modular network consisting of four modules, as shown in Figure 3.4.1. Then c_m can be estimated by $4 \cdot c_m(N/4, M/4)$, $\mu_m = c_m(N, M)$ and σ_m by $\sigma_m(N, M)$. For details see Methods and Chapter 2. This corresponds to taking the subgraph counts in the four modules, modeled as independent networks and comparing them to the total network containing all nodes and edges. The resulting prediction is plotted in Figure 3.4.2 and fits the experimental results very well.

When all links are inter-modular the effect of the modularity should vanish, Figure 3.4.3 shows that this is the case both in the experiment and in the prediction.

We would like to point out that TSPs of metabolic networks have been used in studies such as [33, 39, 67]. To our knowledge none of the studies on motifs in metabolism have accounted for the strong modularity that was discussed early on [136, 143].

3.4.2 Networks with Hierarchies

As modules in networks, hierarchies are defined by constraints on the allowed connectivity. One type of groups of nodes may not be connected to another particular group of nodes, the group of nodes may not form any intra-connections, or the network may have a clear "direction", an established order of groups of nodes determined by the distance from a set of input nodes. These restrictions on the allowed connectivity affect the global network structure dramatically and any rewiring process for producing randomized graphs must account for these specifications, otherwise the global structure is distorted.

Not only clear-cut examples of hierarchies such as in neural networks (in the sense of computer science) that have very distinct hierarchies of nodes responsible for aggregating information from lower tiers, but also less obvious examples like metabolism can be considered as layered networks. In metabolism we can distinguish an input layer provided by the set of uptake reactions (and the input then provided by the available nutrients in the environment), a middle layer where reactions process the nutrients, and an output layer consisting of all reactions directly contributing to cell growth (i.e. the "biomass vector" often encountered in constrained-based modeling of metabolic systems). This clear distinction between input, middle and output layers in metabolic networks is the basis for the parallel to the evolved flow distribution networks discussed in [70, 71, 72] that are another example of layered organization (an example flow network is shown in Figure 3.4.4).

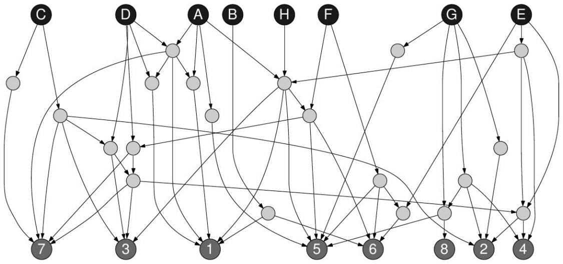


Figure 3.4.4: A sample evolved flow network as described in [70, 71, 72].

Distribution of flow has a clear direction from input to output, where input is defined by nodes with no in-degree and output by nodes with no out-degree. We will use the evolved flow distribution networks mentioned before to demonstrate the effect of one simple constraint. The evolved networks have a middle layer that may be intra-connected or connected to the output nodes, the only restriction is that input nodes may not directly be connected to output nodes. In Figure 3.4.5 we show the effect on the TSP using such a graph with three layers. Depicted are TSPs for a standard randomization scheme and for one that has the additional constraint disallowing links from input to output nodes applied to randomly initialized flow distribution networks. Current randomization methods keep the degree of nodes constant, so the only additional constraint in this case was to disallow links between the input and output layer. Since no additional motif bias (beyond the bias introduced by the layer structure) has entered the network generation, the correct null model ought to yield a flat TSP with all z-scores close to zero.

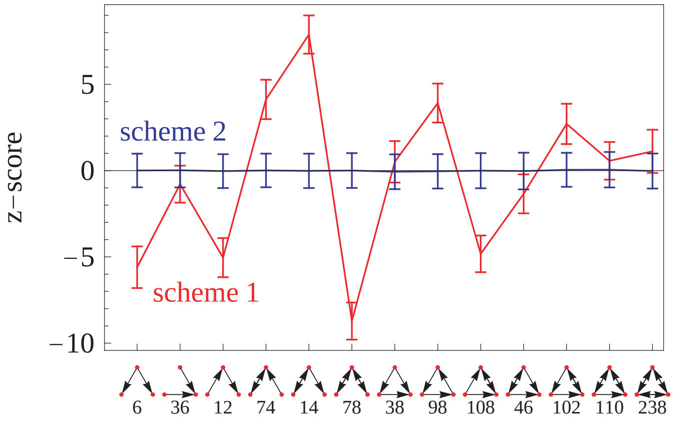


Figure 3.4.5: Average z-scores over 1000 randomly initialized flow distribution networks as described in [70, 71, 72]. The networks contain 20 input nodes, 50 middle nodes, 20 output nodes, and with the probability of a link being present of 0.3. We show z-scores resulting from a random ensemble that was created using scheme (1), a standard switch randomization method, a curve that comes from random ensembles generated with scheme (2) that has the only additional constraint of disallowing direct links between the input and the output layer.

3.5 Dynamical Data on Motifs

3.5.1 Mapping Dynamical Data onto Motifs

Beyond the purely structural issues, a completely new set of complications arises when dynamical data on graphs are discussed. The significance profile obtained from contrasting the few-node subgraph counts in the observed graph with subgraph counts from a suitable pool of randomized graphs only yields statistical indicators pointing to graph features that might be of functional relevance. In biological networks this is particularly tangible, as deviations from randomness can be interpreted as the influence of the evolutionary shaping of the network. However, this statistical observation needs to be linked back to the functional level in a more direct manner in subsequent investigations. Alon and co-workers have done this convincingly in the case of several three-node subgraphs (in particular the feed-forward loop motif) by explicitly modeling dynamical processes on such a motif and classifying the signal processing capacities of such a few-node device [73, 97].

A powerful alternative to this direct modeling (in particular in cases where the dynamics are not well known or well understood) could be to analyze the systematics of dynamical data within a motif, i.e. on each individual node in a motif or distributed across the motif occurrences in the network. This has for example been done in Chapter 2 for citation frequencies on (undirected) co-authorship networks and in [101] for gene expression data on transcriptional regulatory networks. There are several technical difficulties associated with such a study of the role of nodes within motifs. We consider here the statistical treatment of motif multiplicities in mapping dynamical data onto networks.

A convenient method to extract the motif dependence of a dynamic process is to extract the average dynamic observable (i.e. a dynamic robustness) over all nodes that are part of a specific motif. In the first case, one would add up the dynamic variable of all nodes participating in the motif at least once and then divide it by the number of unique nodes participating in the motif. In the second case, one would add up the dynamic variable of all nodes participating in the motif multiplied by the number of motif instances the considered node participates in. To obtain the average one has to divide by the total number of occurrence of the given motif multiplied by the number of nodes per motif.

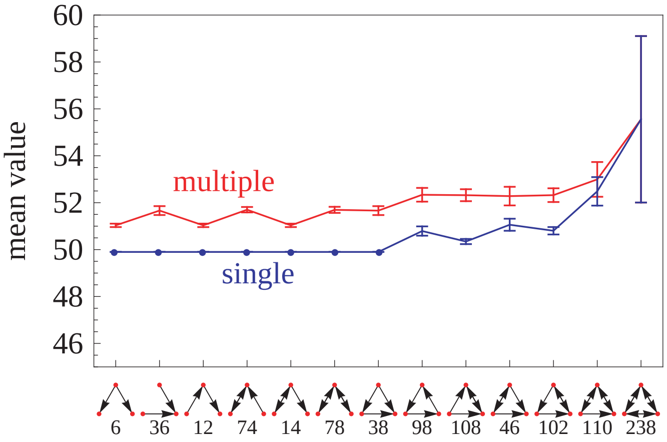


Figure 3.5.1: Mean values over three-node subgraphs for a degree-dependent value with the two different counting schemes. 100 directed ER random graphs with 1000 nodes and a connectivity of 10 % were used here. Due to the sparsity of the networks, motifs with id 110 and 238 occur extremely rarely so that there is little or no dependence on the counting scheme used.

Finding the “correct” counting scheme when dealing with degree-dependent dynamic values can be even more daunting. Even in the case of dynamics that behaves as a linear function of the degree, the number of few-node subgraphs a node participates in still depends non-trivially on the degree. In Figures 3.5.1 and 3.5.2, the difference between degree-dependent and independent dynamic values as well as single- and multiple-counting are depicted. For simple illustration purposes we used static values for the nodes in these cases. In Figure 3.5.1 we use the total degree of the node and in Figure 3.5.2 a random value between zero and one drawn from a uniform distribution. One can see that depending on the counting scheme the qualitative difference of the average dynamic value varies between motifs.

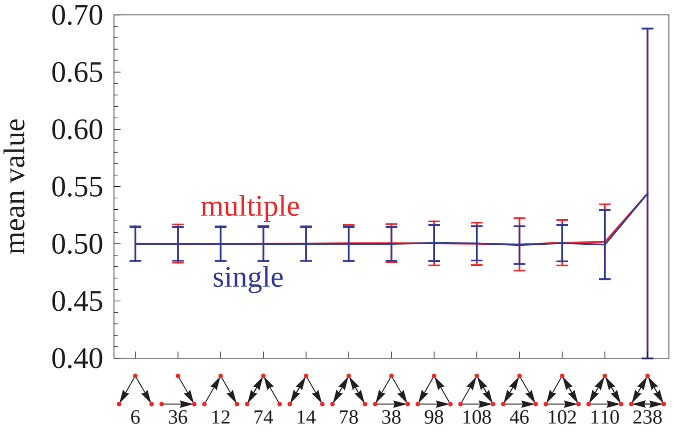


Figure 3.5.2: Mean values over three-node subgraphs for a uniform random value (between 0 and 1) with the two different counting schemes. 100 directed ER random graphs with 1000 nodes and a connectivity of 10 % were used here. Since the networks are rather sparse, the motif with id 238 occurs extremely rarely and thus does not follow the expected statistical behavior.

3.5.2 Role Constraints

Comparable to explicit modeling of a motif function it may be necessary to explore the role that individual nodes and links play (on average) within a certain motif in a network. Again, two points must be considered in such an analysis:

1. The entanglement of node position, node degree, and motif multiplicity.
2. The mean "environment" in the network.

The former point was considered in the previous Section and we only shortly elaborate on an extreme example here. The latter point is the focus of this section and is covered in more detail.

As an extreme example elucidating point (1), consider the following scenario: a directed star graph with only outgoing links from the central node. In this graph we only find the V-out triad (motif id 6). This is an idealized situation of what we might find at hub nodes in the context of larger graphs. The central node in the graph (that occupies the distributing position in the V-out triad) clearly has got the most important position, the largest degree, and is part of all of the triads in the graph. If we consider many such star graphs with a dynamic process on them and we investigate the role of the individual nodes, obviously degree and multiplicity are major factors. When considering dynamics or the function of the central node, it is difficult to determine whether the node's degree or the composition of surrounding subgraphs can be regarded as responsible for an observed behavior. In general one finds that motif multiplicities, qualitatively speaking, scale differently with node properties from motif to motif: V-in and V-out scale with $k_{in}!$ and $k_{out}!$, respectively, where k_{in} and k_{out} are the central node's in- and out-degree. Feed-forward and feedback loops will scale with degree times clustering coefficient, etc. When we determine the average dynamics of the central nodes, for example, the outcome strongly depends on whether multiplicity is regarded as an important feature and taken into the average or whether it is considered to distort the signal since nodes are weighted by the number of motifs they participate in.

In this extreme network environment there are only two types of nodes but what situations will occur in more complicated scenarios? In order to adequately describe the role of nodes and links in a motif we need to contrast the average dynamic with that of the network "environment". If we considered as the network "environment" the global network, we would simply subtract an average dynamical value from each average role in the motif. What we propose instead is to use as an "environment" only those nodes that could in principle (due to their specific degree) play a role in the considered motif, i.e., only nodes that have an in- and out-degree of at least as much as required by the role in the motif are considered. Going back to the example of multiple star graphs where we have only two types of nodes, the average dynamic value for the role in the motif and the average for the environment are exactly equal and the resulting contribution of the motif to the dynamic process is zero. When we insert a few links between nodes at the end points of the star graphs, however, we get a more variate network. More motifs occur and there will be differences between the average dynamic value on the motifs and the average of their particular environment.

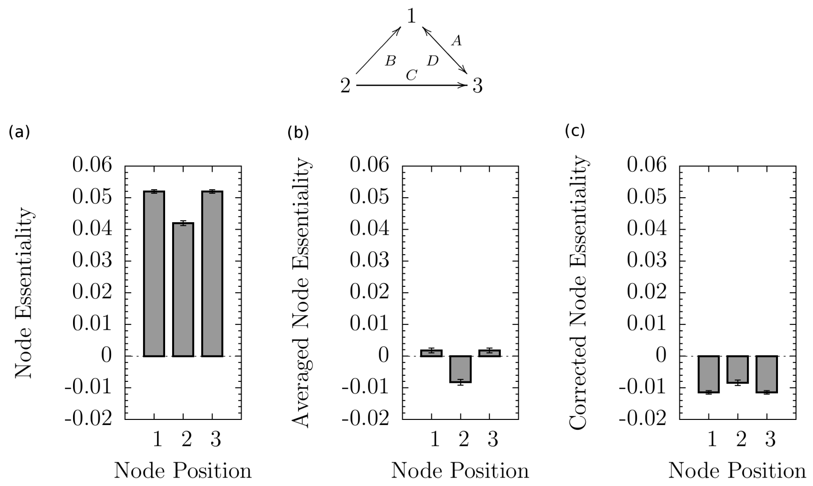


Figure 3.5.3: Analysis of node essentialities for flow networks evolved towards node robustness following the scheme from [70, 71, 72] for the motif shown in the top row: (a) node essentialities for each node position, (b) reduced average essentialities, where the global network was considered as the "environment", (c) reduced average essentialities, where suitable node-specific averages have been subtracted taking the degree constraint of each node into account. In (c) the role of position two is shown to be the more important one in a flow distribution dynamic.

The effects described are illustrated in Figure 3.5.3, showing a study of motif 108 in a flow distribution model [70, 71, 72]. The node essentiality depicted there is the result of a fit to the distribution of magnitudes of change in the output flow upon removal of that node. It depicts how through proper consideration of the network "environment", the role of node two emerges as the most important one which certainly makes sense in the context of the dynamics being flow distribution.

3.6 Conclusions

Conceptually, few-node subgraphs are a means of exploring complex networks by looking at network properties and network function at a well-defined intermediate scale (or group size of nodes involved). We have systematically explored biases introduced into motif signatures by variations of the random background, i.e., of the set of reference graphs serving as null models. Making visible the cross-talk between global and local network properties is in our opinion an important prerequisite of any interpretation of motif signatures.

We want to point out that although sometimes real artifacts can be found, most of the time the boundary between a relevant result and an artifact is rather ambiguous. General features of networks with a simple explanation will sometimes be most visible in the motif signature. For example in the case of modularity a clear motif signature is easily obtained, the reason for this result, the modular structure of the network is much more difficult to detect and to understand.

A major conceptual step in Systems Biology is to reveal systematic and significant deviations of a biological system from randomness, and subsequently relate these deviations to specific functional features. Both, our own analyses [64] and the few attempts found in the literature of exploring network motifs in metabolism [39], show the immense difficulty of disentangling contributions to motif patterns coming from the mere network construction (essentially the fact that metabolic networks are projections of a bipartite graph obtained from a list of metabolic reactions) and those coming from the evolutionary shaping of the system towards an optimized function.

Since many of the analysis techniques that are by now standard for unipartite networks have not yet been formalized for bipartite networks, projection of the bipartite networks will remain the option of choice for some time. The change of statistical properties of networks that have been projected can be quite dramatic. Formally, this has been treated in [88] and was studied in particular for metabolic networks in [110]. Key results are the higher density and degree, as well as a higher clustering due to the “all-to-all” connections formed. In future work, we will extend this to directed bipartite networks and investigate the effects on TSPs.

So, in addition to the potential of explaining functional properties of networks, motif signatures could be used as a proxy for detecting more complex properties of networks. On the other hand, some dynamical systems of networks are highly sensitive to motifs as small functional devices. A whole range of investigations have in several specific models identified a deep relationship between network motifs (or, more generally, the over- and under-representation of few-node subgraphs) and the robust functioning of systemic processes [70, 71, 72, 76, 77, 97, 103].

In order to understand the generality and fundamental nature of these links between topology and dynamics, one needs better knowledge of the intrinsic statistical properties of few-node subgraphs as well as the most minimal dynamical situations, in which such a relationship between topology and dynamics can occur. With the present investigation we want to contribute to this understanding of statistical signals obtained from motif analyses. We advocate carefully chosen null models designed with a profound grasp of the system under investigation. This also means that once more facts about the system are discovered, the chosen null model may have to be adapted.

Chapter 4

Phase Synchronization in Railway Timetables

Summary

Timetable construction belongs to the most important optimization problems in public transport. Finding optimal or near-optimal timetables under the subsidiary conditions of minimizing travel times and other criteria is a targeted contribution to the functioning of public transport. In addition to efficiency (given, e.g., by minimal average travel times), a significant feature of a timetable is its robustness against delay propagation. Here we study the balance of efficiency and robustness in long-distance railway timetables (in particular the current long-distance railway timetable in Germany) from the perspective of synchronization, exploiting the fact that a major part of the trains run nearly periodically. We find that synchronization is highest at intermediate-sized stations. We argue that this synchronization perspective opens a new avenue towards an understanding of railway timetables by representing them as spatio-temporal phase patterns. Robustness and efficiency can then be viewed as properties of this phase pattern.

The results presented in this chapter have been achieved in cooperation with Lachezar Krumov, Karsten Weihe, Matthias Müller-Hannemann and Marc-Thorsten Hütt and have been published in: “Phase synchronization in railway timetables” [45].

4.1 Introduction

Railway timetables should be designed to achieve a maximum level of utilization from a passenger’s perspective. That is, regular waiting times for connecting trains should be kept to a minimum. However, this limits the network’s robustness against perturbations: Depending on the waiting policy among connecting trains, a single delayed train may cause a cascade of further train delays in remote parts of the network. Minimal regular waiting times (minimal buffering times) cause maximal risk of such delay propagation. Understanding this trade-off and limiting the propagation of delays through the networks is a challenge of practical importance.

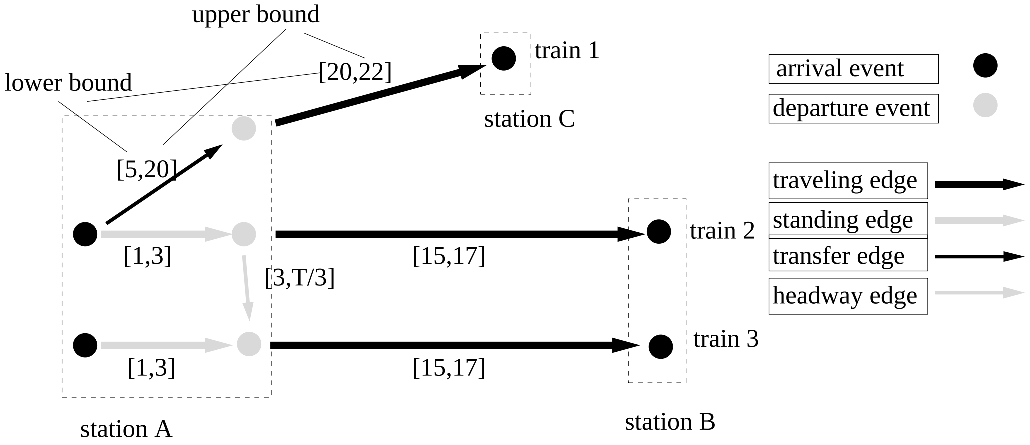


Figure 4.1.1: Small excerpt of a periodic event scheduling problem.

The construction of periodic railway timetables is algorithmically difficult and has been intensively studied as a periodic event scheduling problem (PESP), see for example [147, 130, 90]. The technical and economical side constraints for a valid non-periodic schedule can be modeled as a feasible differential problem on a directed graph $G = (V, E)$ with lower and upper edge bounds $\ell, u \in \mathbb{Q}^E$. In a basic model, the vertex set V corresponds to departure and arrival events, while the directed edges together with the bound values model constraints (travel times, minimum headway, minimum transfer times, etc.). One seeks for a vector $\pi \in \mathbb{Q}^V$, called the timetable, which assigns to each event j a time-stamp π_j satisfying

$$\ell_e \leq \pi_j - \pi_i \leq u_e \text{ for all } e = (i, j) \in E.$$

Thus, lower and upper edge bounds restrict the difference between two time-stamps from below and above, respectively. For example, $\ell_{(i,j)} = 15 \leq \pi_j - \pi_i$ means that event j has to occur at least 15 time units after event i. See Figure 4.1.1 for a small example.

In a periodic timetable, trains are grouped into lines which are to be operated by some period T. In the periodic event scheduling problem (with one fixed period T) one searches for a vector $\pi \in [0, T)$ such that for all $e = (i, j) \in E$ there exists $k_e \in \mathbb{Z}$ with

$$\ell_e \leq \pi_j - \pi_i + T \cdot k_e \leq u_e.$$

For the local public transport in Berlin (Germany), the first optimized periodic timetable used in daily operation has been obtained using mixed-integer linear programming techniques [91]. Netherlands Railways also have recently introduced a completely new periodic timetable, generated by a number of sophisticated operations research techniques, including constraint programming [80]. For countries with a less periodic timetable, including Germany, the construction process for long-distance timetables is quite complex, and therefore still done to a large extent manually by experienced engineers. The planning process has a hierarchical component (international trains are scheduled first), and a behavioral component (keep as much as possible from the previous year's schedule).

So far, railway timetables have been studied predominantly as an algorithmic challenge with the objective of constructing optimal (or near-optimal) connection patterns, minimizing resources and overall waiting time. Only recently, there have been first computational studies aiming at delay resistant periodic timetables [79, 92, 93].

Here we adopt an opposite perspective to timetable construction and analyze the spatio-temporal patterns induced by the timetable. A suitable language for this study is a representation of the train arrival/departure events as a spatio-temporal phase pattern. We study the distribution of synchronization across stations. Synchronization phenomena have received a lot of attention in traffic modeling over the last few years, in particular for car traffic in cities and the impact of traffic light synchronization on the formation of traffic jams [26, 85, 86].

In the case of railway timetables, the situation is different in several ways: The "load" of a station is essentially determined by the number of tracks (giving the maximal number of simultaneous or nearly simultaneous arrival/departure events). The typical number of directions (which can be interpreted as the degree of a station in a suitable effective network representation), from which one can select, is higher for train stations than for typical street crossings.

If one considers a network of long-distance train connections as a mesh of routes through a planar system, where trains are started periodically at the endpoints of these routes, the spatial distances between the intersection sites of these routes determine a spatio-temporal phase pattern. The free parameters of this pattern are the relative phases of the periodically started trains. In reality, the travel time between two stations can serve as an additional degree of freedom allowing for a shaping of the phase pattern beyond this simple thought experiment.

Our main hypothesis is that the rank of the stations sorted according to size is the organizing parameter (i.e. the "control parameter" from the perspective of self-organized systems [98, 132, 162]), along which synchronization can be understood. In this chapter, we use the notion buffering time to denote the amount of time available to change between two trains (transfer time) for the planned schedule, i.e. without induced delays. Our other two observables are the average buffering time b_i at station i and the secondary delays $s_i(p)$ induced by a primary (incoming) delay p because trains have to wait for other trains.

The main result of our analysis is that a railway timetable induces a spatio-temporal phase pattern, and that properties of the phase pattern are linked to the efficiency and the robustness of the system. We observe that synchronization is highest at intermediate-sized stations.

Here we contribute two points to the general debate:

(1) We show that the current planning of railway timetables (which involves some algorithmic construction, some manual curation and the resorting to features from previous timetables) leads to an unexpected coherence on the level of the spatio-temporal phase pattern.

(2) At the same time, our analysis shows that the general concept of a spatio-temporal phase pattern is a novel and helpful view for network-based scheduling problems.

The remainder of this chapter is structured as follows. In Section 4.2, we first give a detailed description of our numerical experiment, and then discuss the results in Section 4.3. Afterwards, we introduce an avalanche model for delay propagation on graphs (Section 4.4) helping us to understand the observed relation between synchronization and robustness. Finally, we conclude with a short summary and an outlook for future work (Section 4.5).

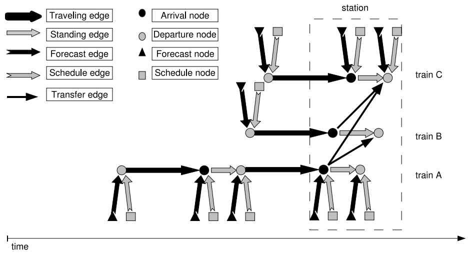


Figure 4.2.1: Illustration of the dependency graph model (taken from [112]).

4.2 Formalism and Numerical Experiments

The quality of the timetable is related to two distinct (and often conflicting) objectives: The sum of travel times over all routes should be minimal (efficiency) and typical delays should minimally increase the overall travel time (robustness). Apart from some freedom to determine the planned travel time from one station to another (i.e. the prescribed average speed of the train), the main tuning capacity lies in the interchange time between connecting trains. While efficiency requires a minimization of interchange time, robustness can be established by using the interchange time as a buffer for incoming delays.

The secondary delays s_i observed at each station i across a range of primary delays p have been obtained by a large-scale numerical experiment performed on the actual timetable of Deutsche Bahn AG, together with real passenger information. Throughout our investigation we consider only long-distance train connections (served by the train categories ICE and IC/EC). To simulate the effects of delays, we use the dependency graph model introduced in [112] and its implementation within the fully realistic multi-criteria timetable information system MOTIS [144]. The dependency graph is basically a time-expanded graph model with distinct nodes for each departure and arrival event in the entire schedule for the current and following days. In addition, the model includes two further types of nodes: forecast and schedule nodes.

Each node has a time-stamp which can dynamically change. The time-stamps reflect the current situation, i.e. the expected departure or arrival time subject to all delay information known up to this point. Schedule nodes are marked with the planned time of an arrival or departure event, whereas the time-stamp of a forecast node is the current external prediction for its departure or arrival time.

The nodes are connected by five different types of edges (see Figure 4.2.1). The purpose of an edge is to model a constraint on the time-stamp of its head node.

- *Schedule edges* connect schedule nodes to departure or arrival nodes. They carry the planned time for the corresponding event of the head node (according to the published schedule).

- *Forecast edges* connect forecast nodes to departure or arrival nodes. They represent the time stored in the associated forecast node.

- *Standing edges* connect arrival events at a certain station to the following departure

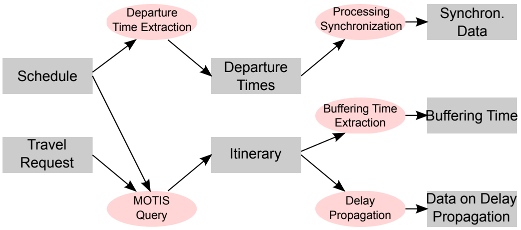


Figure 4.2.2: Data flow in the delay propagation experiment.

event of the same train. They model the condition that the arrival time of train t at station k plus its minimum standing time must be respected before the train can depart (to allow for boarding and disembarking of passengers).

- *Traveling edges* connect a departure node of some train t at a certain station k to the very next arrival node of this train at station k'.
- *Transfer edges* connect arrival nodes to departure nodes of other trains at the same station, if there is a planned transfer between these trains.

The current time-stamp for each departure or arrival node can be defined recursively, for details see [112].

The MOTIS tool can be used as follows. Given the planned train connection of a passenger and a concrete delay scenario (for example, a single primary delay of a train), we can query MOTIS for the fastest train connection towards the passenger's destination, subject to the standard waiting rules between connecting trains. In particular, the train waiting regulations of Deutsche Bahn have been used. From the difference between the planned arrival time at the destination and the calculated arrival time in the delay scenario we obtain the individual delay for each passenger.

Passenger information has been available to us for a single day in the form of all travel agency bookings for that day. While these data are certainly distorted by the fact that most tickets are sold via vendor machines at the station (and these data have not been available to us), it is nevertheless helpful to include passenger data for two reasons:

(1) Only routes, which have really been traveled, enter our analysis; in this way we avoid artifacts, e.g., from back-and-forth contributions.

(2) We can discuss both the average delay per passenger and the cumulative delay over all affected passengers (as a measure of the total systemic effect).

In Figure 4.2.2, we sketch the data flow within our numerical experiment, where we have processed 43772 train segments, 2622 stations, 130071 passenger routes, and about 1.9 million MOTIS queries.

In order to illustrate the raw data obtained from this numerical experiment, we show the station size distribution (where the station size is given by the number of arrival/departure events per day) in Figure 4.2.3; and the buffering time distribution in Figure 4.2.4. Both distributions are essentially unimodal and have a non-negligible tail at large values. The rare occurrence of low buffering times can be explained by the fact that the timetable information system does not provide connections where a (station-specific) minimal interchange time is not reached. It should be noted that this general rule is accompanied by a long list of exceptions for

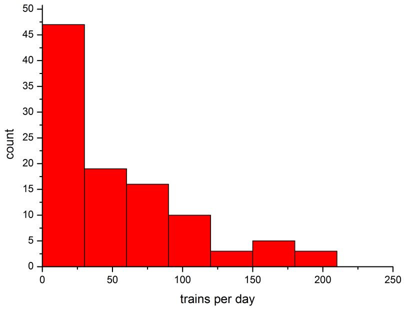


Figure 4.2.3: Distribution of daily numbers of arrival/departure (A/D) events.

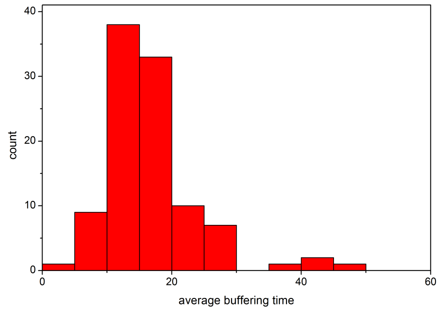


Figure 4.2.4: Distribution of the average buffering time per station.

specific trains and specific connections. All these constraints and subsidiary conditions have been included in the numerical experiment, in order to obtain realistic event data.

As a next step we compare these delays with unperturbed features of the timetable. Our approach for converting the timetable into an event pattern uses the language of phase synchronization. Let $\left\{t_j^{(k)}, j = 1 \dots T_k\right\}$ be the set of arrival/departure (A/D) times $t_j^{(k)}$ of the jth train at station k. The quantity T_k denotes the number of A/D events at station k per day. These A/D times are now translated into phases

$$\phi_j^{(k)}(\tau) = \frac{2\pi}{r}(t_j^{(k)} mod \quad \tau) \tag{4.2.1}$$

with the period length τ as a parameter. In our analysis we set this parameter to the maximal period length observed in the system, i.e., τ = 120 minutes. For each station k we can now compute the synchronization index (as known from the classical studies of synchronization in populations of phase oscillators, see [83, 152, 167]; see also the scheme depicted in Figure 4.2.5):

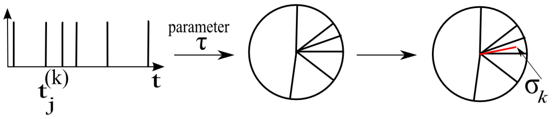


Figure 4.2.5: Conversion of the arrival/departure times at a station k into the synchronization index σ_k.

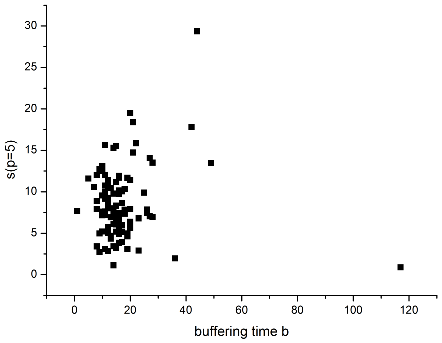


Figure 4.3.1: Secondary delay as a function of the buffering time for a fixed primary delay p = 5 minutes, raw data.

$$\sigma_k = \sigma_k(\tau) = \left| \frac{1}{T_k} \sum_{j=1}^{T_k} e^{i\phi_j^{(k)}(\tau)} \right| \tag{4.2.2}$$

The view we want to propagate here, is that the performance (in a very general sense) of a given timetable of train connections is related to its phase pattern.

4.3 Results

The large-scale numerical experiment described in the previous section in particular yields realistic values of the secondary (induced) delay $s(p)$ as a function of the primary (input) delay p. While at low p the value of $s(p)$ is mainly (but indirectly!) shaped by the buffering time b, at higher p the value is strongly influenced by the number of alternative connections.

On face value, one would expect a negative correlation of the secondary delay $s(p)$ and the buffering time b in this low-p region. In the raw data, Figure 4.3.1, there is rather a lack of correlation (or even a slight tendency towards positive correlations), which can be explained as follows: The buffering time b grows slowly with the station rank, i.e. decreases slightly with the station size (cf. Figure 4.3.2). At the same time, larger stations (i.e. more A/D events) offer more alternative routes, effectively reducing the secondary delay, even at low primary delay p.

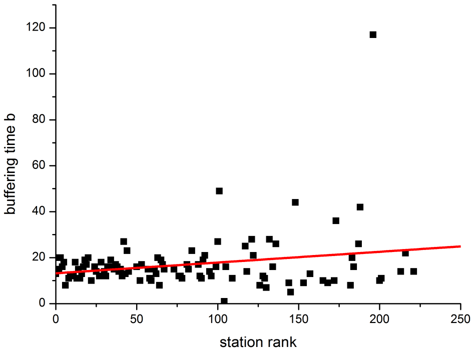


Figure 4.3.2: Dependence of buffering time b on station rank.

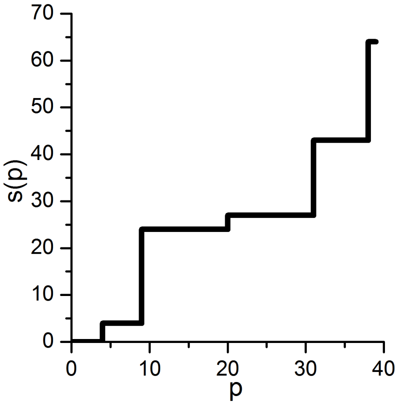


Figure 4.3.3: Secondary delay as a function of the primary delay for a single train.

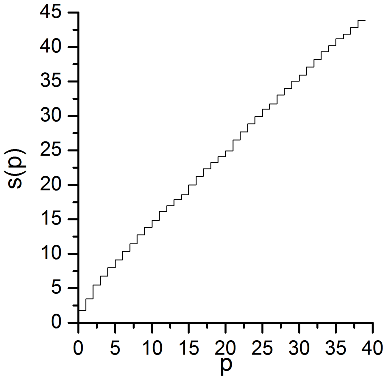


Figure 4.3.4: Average secondary delay as a function of the primary delay for a single station (here: Frankfurt (Main) central).

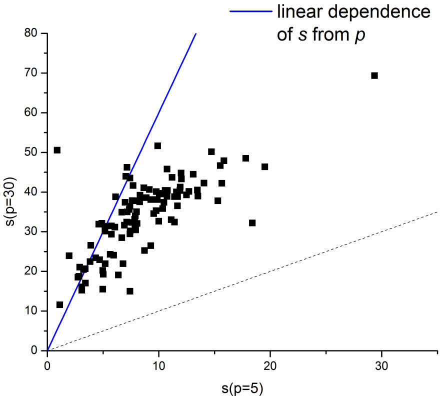


Figure 4.3.5: Correlation of the secondary delay with primary delay of 5 and 30 minutes.

In the example shown in Figure 4.3.3, there are only 4 minutes of buffering time which induce no delay. When p is in the range of $[4, 9]$ minutes, the secondary delay becomes 4 minutes, and then jumps to 25 minutes. Figure 4.3.4 shows the average secondary delay at a station, which is, as a first approximation a linear function. At higher values of p, additional effects can be expected to set in:

(1) with higher p more alternative routes become accessible,

(2) more passengers will be affected, and

(3) longer avalanches of delayed trains are triggered upon waiting.

These contributions are partially compensated by the waiting policy: Avalanche length is strongly reduced by maximal waiting times. Also, the second contribution has a smaller (but still non-zero) effect on the average secondary delay per passenger.

Figure 4.3.5 shows the correlation between the secondary delays for two different values of the primary delay, namely p=5 minutes and p=30 minutes. There is a wide spread of deviations from the solid line showing the expectation for the case of a linear $s(p)$. This is indicative of the multitude of strengths with which these additional, higher-p effects contribute.

The challenge is now to establish in detail the relations between the degree of synchronization and the performance of the system (given by low delay propagation, i.e. robustness, and low overall transfer times, i.e. efficiency).

On the level of our data, the main performance indicators of the system, namely the efficiency and robustness, are only indirectly accessible via the secondary delays and the buffering times. We expect that a small $s(p)$ is related to high robustness (a given perturbation p induces a small effect s), while a small b can be associated with high efficiency (during a full itinerary only a small amount of time, given by the local buffering times b, is accumulated upon train interchanges).

When splitting the b-$s(p)$–plane into four quadrants corresponding of contributions of high/low delays and buffering times – and, consequently, low/high ($-$/+) performance –, one can observe very different usages (i.e., frequencies of occurrence in the data) of the quadrants:

The ++ region, which is the most efficient one as those stations are both efficient (small buffering time) and robust (small secondary delays), is most densely populated, followed by

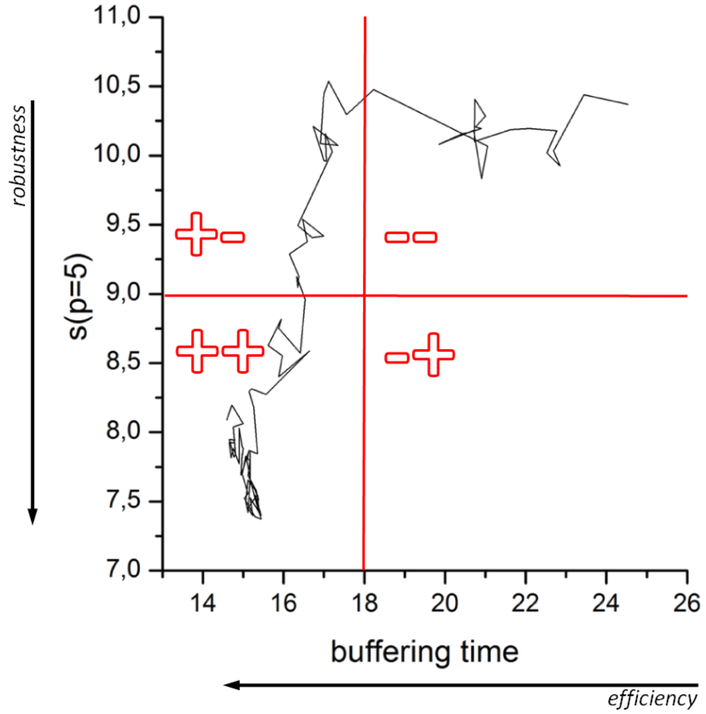


Figure 4.3.6: Secondary delay as a function of the buffering time for a fixed primary delay $p = 5$ minutes, averaged and connected along the station rank, together with a phenomenological separation into quadrants according to high/low efficiency (first quadrant label) and high/low robustness (second label).

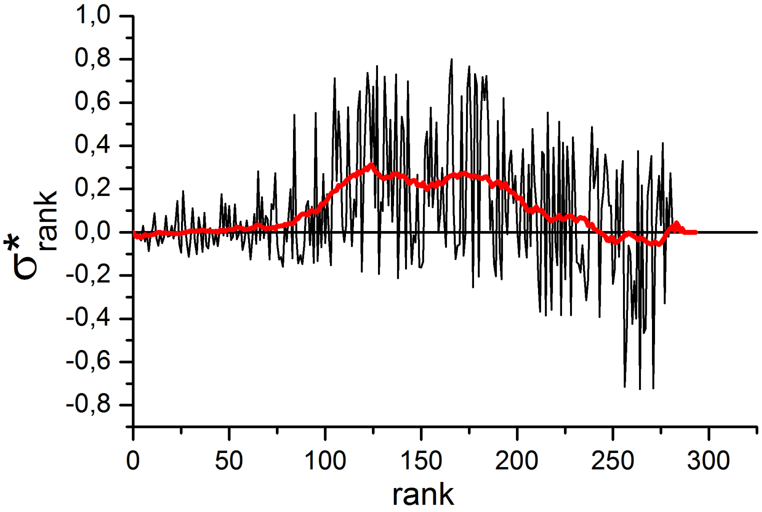


Figure 4.3.7: The synchronization indices of the stations in ascending order of the station rank in Germany with $N_R = 100$ and an averaging window of 40.

the $+-$ region (high efficiency, low robustness). Very few stations are found in the $--$ region. Interestingly we do not find stations in the $-+$ region. Probably those are quickly eliminated during the schedule building or avoided during the route search.

In order to better understand the systematic relation between $s(p)$ and b it is again helpful to use the station size as a control parameter along which local averages can be performed. The resulting curve is shown in Figure 4.3.6. This curve displays the backbone systematics of the interplay between efficiency (inverse b) and robustness (inverse $s(p)$) studied from the raw data in Figure 4.3.1, when using station size as an ordering parameter.

Figure 4.3.7 shows the synchronization index σ_k^* as a function of the station rank k. The phase data are distorted by the mere number of A/D events. In particular, at few A/D events large fluctuations of σ are induced. We therefore subtracted from each σ_k an average $\sigma_k^{(R)}$ over N_R runs of a null model, where the same number of A/D events has randomly been distributed in time. This procedure yields the reduced synchronization indices σ_k^*=$\sigma_k - \sigma_k^{(R)}$, shown as the black curve in Figure 4.3.7.

Furthermore, stations with neighboring ranks will differ (even though they are similar in size) in a variety of additional parameters. The original reduced synchronization index shows a strong local fluctuation along the rank. In order to eliminate the variation coming from these additional differences between similar-sized stations, we compute local averages over the σ_k^*. These values are shown as the red curve in Figure 4.3.7. Remarkably, synchronization is highest at intermediate station rank, decreasing towards both larger and smaller station sizes. In order to obtain this result, several processing steps of the raw data have been necessary. The systematic difference between the synchronization of large, small and intermediate train stations, respectively, is also seen, when average synchronization indices for each of these three

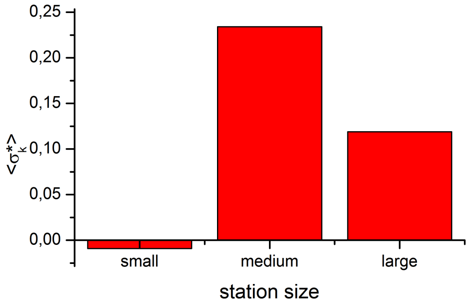


Figure 4.3.8: Average σ^* for small, medium and large stations. The rank is split at 80 and 170 A/D events per day, respectively.

categories are computed directly (Figure 4.3.8).

In order to assess, whether this elevated synchronization of intermediate-size stations is a property of train timetables beyond this individual case, we also computed the synchronization indices σ_k^* for four other counties, Austria, France, Norway, and the Czech Republic (Figure 4.3.9). France shows only a very weak signal, whereas the shapes of the synchronization curves in Austria, Norway and the Czech Republic are very similar to the one observed in Germany (Figure 4.3.7).

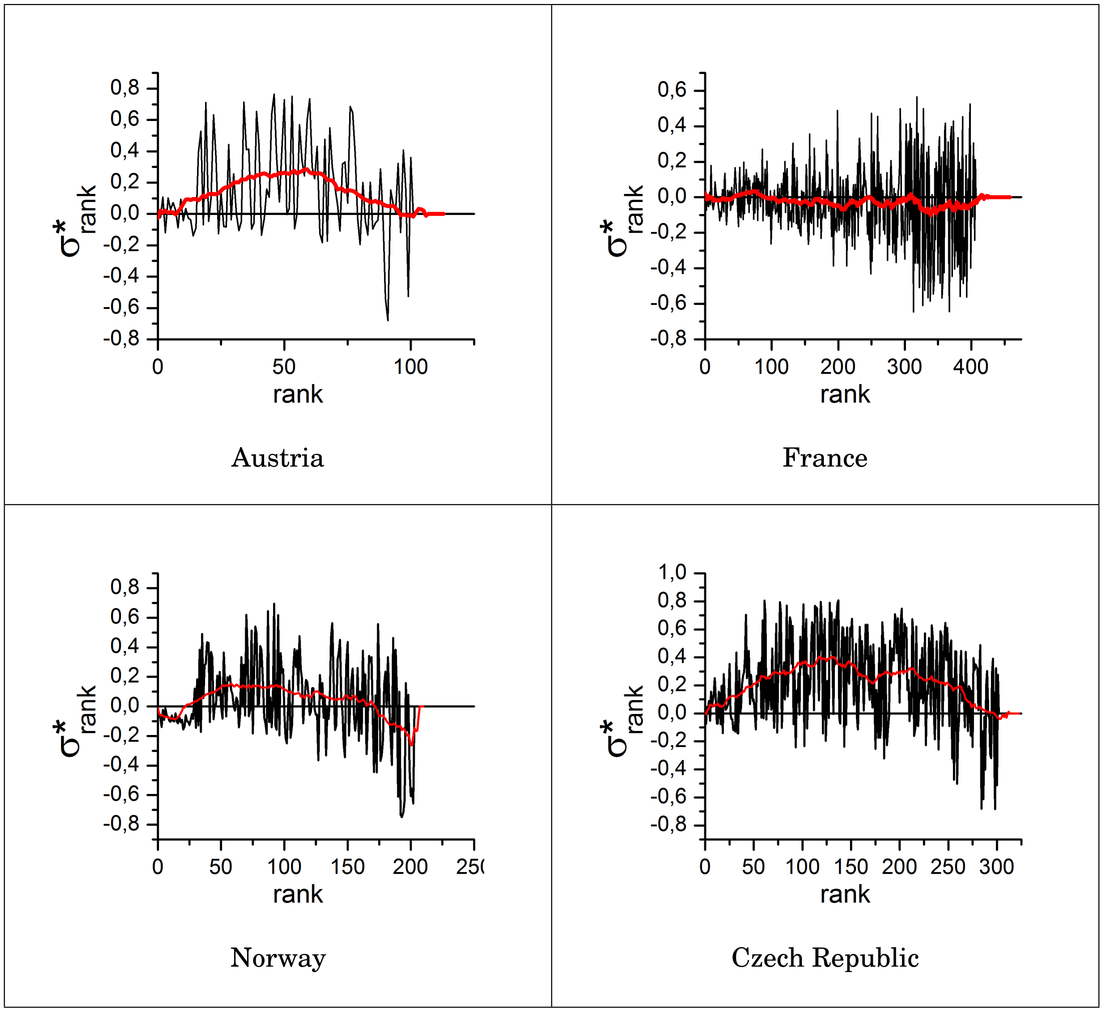
In the following we will show results for the inter-dependencies of our main quantities b_k, $s_k(p)$ and σ_k^*. In all cases, like before, we compute local averages with respect to the rank. By grouping the stations according to their position in the b-$s(p)$–plane, Figure 4.3.6, i.e. according to their robustness and efficiency, one can now study, whether stations from the same regions share a common synchronization index σ_k^*.

Figure 4.3.10 shows the average σ for the three regions containing stations. The stations from the most preferable region $++$ show extremely low synchronization, while those of the regions $+-$ and $--$ are much more synchronized.

In order to show the dependencies among these quantities more directly, Figure 4.3.11 represents all three quantities b_k, $s_k(p)$ and σ_k^*, simultaneously. The smoothing window size is set to 26. It is clearly visible that most stations are in the regions of low σ_k^*, low b_k and low $s_k(p)$. Furthermore, there is a clear correlation between $s_k(p)$ and σ_k^* and consequently, an anti-correlation between synchronization and robustness.

4.4 Avalanche Model

Can the negative correlation between synchronization and robustness that we observe in the data also be understood in some minimal model of delay propagation? The general dynamical mechanism resembles some aspects of avalanches on graphs. While avalanche models are an important focus of interest in complex systems theory and in particular in the field of self-organized criticality [66, 8], we do not expect here a power-law distribution of event sizes, as the elementary processes behind delay propagation are different from the threshold-driven re-



Austria

France

Norway

Czech Republic

Figure 4.3.9: The synchronization in the long-distance train connections of different European countries with $N_R = 100$ and an averaging window of 40.

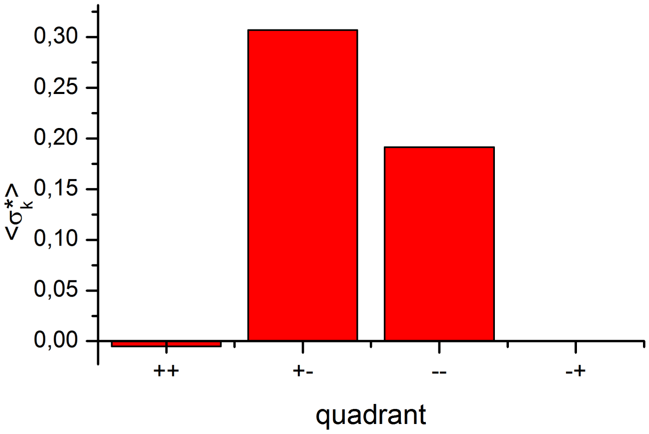


Figure 4.3.10: Average synchronization of the stations grouped together by robustness and efficiency according to the quadrants in Figure 4.3.6.

distribution schemes encoded, e.g. in the Bak-Tang-Wiesenfeld (BTW) model [9]. Therefore, we adapt the general concept of an avalanche model to the dynamical needs of delay propagation.

In our passenger delay avalanche model the dynamical variables are the accumulated delays $d_i(t)$ at node i as a function of time t. The model has three parameters:

(1) The transmission probability p is the probability that a delay propagates from one node to an adjacent node, if the threshold is crossed. This probability describes the capacity to buffer incoming delays via transfer times.

(2) The amplification factor m acknowledges the fact that a single train delay corresponds to multiple passenger delays; consequently, if a few incoming passengers cause a train with many outgoing passengers to wait, the total (passenger-based) delay is amplified. This parameter can be seen as the ratio of these passenger numbers, i.e. the average rate of additionally delayed passengers due to waiting.

(3) Delays only propagate from a node i to adjacent nodes, when the delay variable $d_i(t)$ is above a threshold T, as we assume that only incoming delays higher than this threshold are capable of triggering delay propagation. In Figure 4.4.1, the general idea of this model is schematically depicted. Figure 4.4.2 shows that the tail of the size distribution of delay avalanches is exponential.

In order to analyze the relation between synchronization and robustness within this model, we compare the average avalanche length for a system, where a single node is periodically driven, with the case of a stochastically driven node.

We assume that the gradual insertion of delay units into the system corresponds to incoming delays from other parts of the network entering the sub-network, which is here studied in detail. A periodic insertion of such delay units then corresponds to highly synchronized arrival/departure (A/D) events (as only those can give rise to periodic delays), while the stochastically driven node represents the typical pattern of incoming delays for a station with less synchronized A/D events.

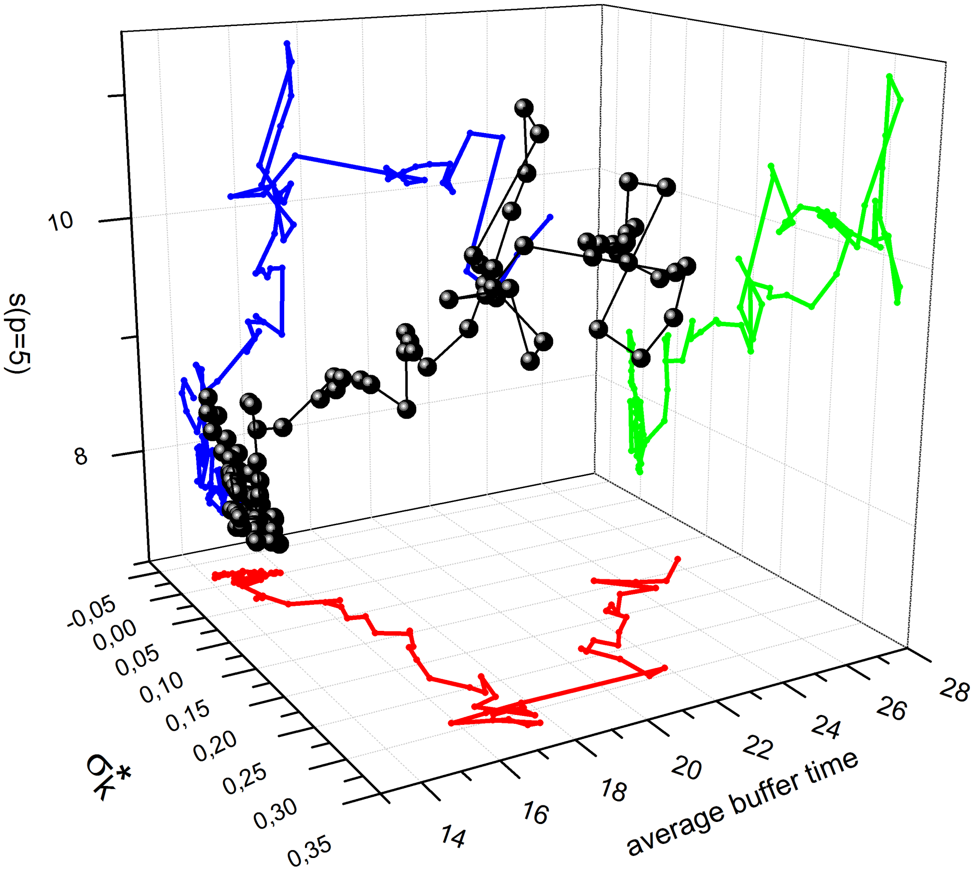


Figure 4.3.11: Dependencies of buffering time b, secondary delay $s(p)$, and synchronization index σ_k^*.

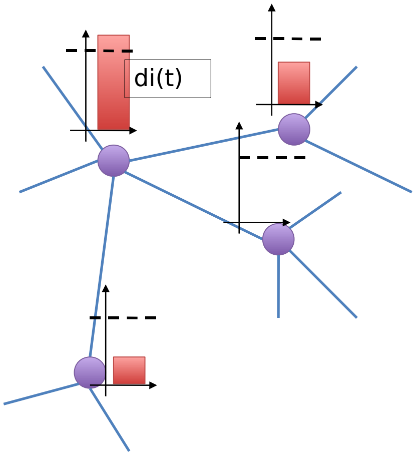


Figure 4.4.1: The passenger-delay-avalanche model (PDA-model).

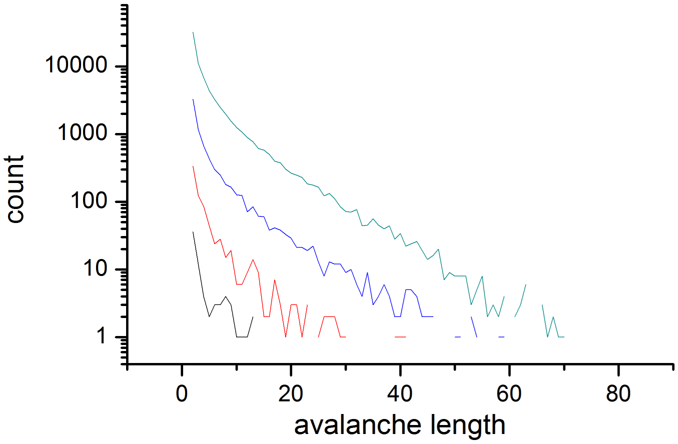


Figure 4.4.2: The distribution of the avalanche lengths for stochastic and periodic drivers with a driving period of 17, $T = 4$, $m = 0.9$, $N = 70$ nodes and $M = 240$ edges.

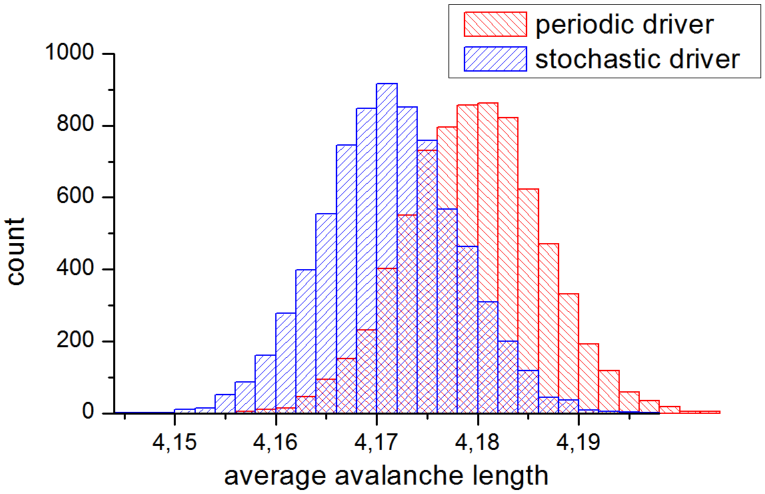


Figure 4.4.3: The average avalanche length for stochastic and periodic drivers with a driving period of 17, $T = 4$, $m = 0.9$, $N = 70$ nodes and $M = 240$ edges.

Figure 4.4.3 shows the distribution of these average avalanche lengths for the two cases. The periodically driven node (high synchronization of A/D events) coincides with a high average size of the delay avalanches (i.e. higher vulnerability or lower robustness), while the stochastically driven node (low synchronization of A/D events) displays a lower average size of delay avalanches (and therefore a higher robustness). This is in agreement with the relationship discovered in the real train connection data studied in the previous section.

4.5 Conclusions

In this chapter we have studied railway timetables from a novel and yet unexplored view, namely that of phase synchronization. For our analysis we investigated the German long-distance train timetable with respect to three distinct properties: robustness, efficiency and phase synchronization.

The robustness reflects the stability of the system to small perturbations, while efficiency is related to short accumulated waiting times per train route. These two properties have been evolved over the years by gathered experience and heuristic optimization.

When we consider the arrival and departure events of all trains at a given station over a period of time, 24 hours for example, we can translate those events into phases. Summing over all different phases we can compute a synchronization index for each station. Then, by exhaustive simulation we produce a primary delay at each station and record the induced secondary delays. Our results show a clear and surprising correlation between the synchronization index of a station, its robustness and efficiency.

In the introductory Section 4.1, we have discussed the difference between car traffic in cities and the impact of traffic light synchronization on the one side and railway timetables on the other. It would be interesting to compare these two types of traffic in detail, to quantitatively analyze the number of directions (node degrees) in the context of an effective dimension, and in particular to study the complexity (given, e.g., by the pattern of elementary decisions needed to specify the path) of a typical path in the train network compared to the car traffic case. A suitable methodology could be the framework developed in [139].

The balance between this antagonistic pair of requirements, efficiency and robustness, is of broad interest across many disciplines, ranging from industrial production to biological processes. Lack of robustness due to too high efficiency is sometimes called the systemic risk, which has recently been discussed from a theoretical perspective, for example for complex economical systems (see [15, 16, 95]).

Starting from an information-theoretical description of resilience in ecology, Ulanowicz et al. [155] could establish quantitative links between sustainability, efficiency and investments in diversity. This general framework has been employed to analyze the current bank crisis from a ecosystem perspective [94]. We believe that a quantitative view on synchronization of arrival/departure events in the network of long-distance train connections, as presented here, can similarly serve as a starting point for a theoretical understanding, and subsequently systemic optimization, of the balance between efficiency and robustness for such timetables underlying public transportation.

For biological processes this balance between efficiency and robustness has been explored in a multitude of ways resorting to both analysis of experimental data and the mathematical modeling of cellular processes. Motivated by graph theory and nonlinear dynamics, an influential trend in systems biology at the moment is to relate robustness to small regulatory devices [3, 25], serving e.g. as a noise buffer or providing a suitable amount of redundancy for maintaining systemic function even under perturbations. In particular such relations between the architecture of regulatory devices and dynamical functions have been worked out for circuits of negative feedback loops [131], for feed-forward loops as noise filtering devices in gene regulation [3, 148], for interlinked feedback loops acting on different time scales [24], for a particular composition of regulatory units [107] and their relation to robustness [70, 71, 72, 77], for number of positive and negative feedback loops in regulatory circuits [84].

It could well be that in the network of long-distance train connections such small, motif-like network components serve as mediators between synchronization, reliability and efficiency. Exploring the involvement of network topology in shaping this relationship is one of our principal goals in the continuation of the work presented here.

Chapter 5

Co-Authorship Networks

Summary

Co-authorship networks, where the nodes are authors and a link indicates joint publications, are very helpful representations for studying the processes that shape the scientific community. At the same time, they are social networks with a large amount of data available and can thus serve as vehicles for analyzing social phenomena in general.

Previous work on co-authorship networks concentrates on statistical properties on the scale of individual authors and individual publications within the network (e.g., citation distribution, degree distribution), on properties of the network as a whole (e.g., modularity, connectedness), or on the topological function of single authors (e.g., distance, betweenness).

Here we show that the success of individual authors or publications depends unexpectedly strongly on an intermediate scale in co-authorship networks. For two large-scale data sets, CiteSeerX and DBLP, we analyze the correlation of (three- and four-node) network motifs with citation frequencies. We find that the average citation frequency of a group of authors depends on the motifs these authors form. In particular, a box motif (four authors forming a closed chain) has the highest average citation frequency per link. This result is robust across the two databases, across different ways of mapping the citation frequencies of publications onto the (uni-partite) co-authorship graph, and over time.

We also relate this topological observation to the underlying social and socio-scientific processes that have been shaping the networks. We argue that the box motif may be an interesting category in a broad range of social and technical networks.

The results presented in this chapter have been achieved in cooperation with Lachezar Krumov, Karsten Weihe, Matthias Müller-Hannemann and Marc-Thorsten Hütt and have been published in: “Motifs in co-authorship networks and their relation to the impact of scientific publications” [81].

5.1 Introduction

One of the classical debates in the history of science is, whether the production of knowledge can be rather viewed as an objective, content-driven process or, conversely, is dominantly shaped by the underlying social patterns formed by the actors involved. Ever since Thomas Kuhn's groundbreaking analysis "The Structure of Scientific Revolutions", it is accepted that the social layer contributes heavily to scientific progress. Expectations of peers and the adherence to agreed-upon terminologies all have a synchronizing effect that may be considered a socially generated inertia leading to the characteristic discontinuous time course of scientific progress, Kuhn's work has become famous for [82]. While the content-driven perspective still allows for "geniuses", brilliant individuals responsible for a step-like, discontinuous advancement of knowledge, the importance of the social layer is an undeniable one.

Nowadays, due to the electronic availability of vast amounts of data on knowledge production, the study of complex networks provides a unique opportunity to quantitatively assess the social contribution to the production of knowledge. From the network perspective the strength of this social contribution can be re-phrased as follows: Does the underlying interaction network of authors and publications statistically explain parts of the output pattern of the scientific community? This question is at the core of our analysis.

A fundamental topic of interest in complex systems theory and in the analysis of complex networks is currently, how network architecture systematically shapes dynamical processes. Progress has been made over the last decade in identifying first ordering principles. One example is the synchronization of oscillators on hierarchical graphs [5]: The time course of the step-wise path towards a fully synchronized system seems to follow the pattern of gaps in the spectrum of the graph (or, more precisely, associated Laplacian matrix). Furthermore, using stylized minimal models has been helpful in revealing some other relationships between network topology and dynamics (see, e.g., [23, 100, 111]).

An interesting alternative to these simulation-driven studies is to explore the relationship between network architecture and dynamics from a data-analysis perspective, i.e. to extract this relationship from large-scale data sets, which can be expected to be produced, at least partly, by the dynamics of the network at hand. Evidences for network architecture being a clearly discernible, quantifiable component, contributing to the patterns observed in data, exist from a diverse range of fields: gene expression patterns, both on the level of whole transcriptional regulatory networks [58, 96, 99] and on the scale of small regulatory devices [109, 3], the epidemic spread of diseases [128] and attack tolerance related to a broad degree distribution [2].

Network motifs, small subgraphs with a specific interaction pattern, have been particularly successful in providing interesting, unexpected relations between network architecture and dynamical processes. In particular in systems biology, an influential trend currently relates features of network performance to such small regulatory devices [3, 25], serving e.g. as a noise buffer or providing a suitable amount of redundancy for maintaining systemic function even under perturbations. In particular such relations between the architecture of regulatory devices and dynamical functions have been worked out for circuits of negative feedback loops [131], for feed-forward loops as noise filtering devices in gene regulation [3, 148], for interlinked feedback loops acting on different time scales [24], for a particular composition of regulatory units [107] and their relation to robustness [70, 71, 72, 77], and for the number of positive and negative feedback loops in regulatory circuits [84].

Co-authorship networks are a snapshot of the knowledge production system, simultaneously shaped by the social aspects contributing to scientific activity and the topical organization of knowledge (see, e.g. [114, 118, 138]). Early studies in the mid-1970s [42], in spite of the limited access to data, already extracted some surprising statistical properties within co-authorship and citation data [87, 156]. A giant leap towards analyzing the large-scale organizational features of the system came of course with the shift towards electronically available publications (see, e.g., [20, 118, 168]).

Here we adopt a specific definition of co-authorship networks, where the nodes are authors and two authors are connected by an edge if, and only if, they have published at least one paper together. One can debate, whether this (uni-partite) graph is a suitable representation for this intricate system. Information lost in this representation is the separation of groups of authors into distinct papers (this information would be retained in the case of a bi-partite representation) and the grouping of authors beyond the two-author level (e.g., in terms of institutions; this information could be made accessible in a hyper-graph format). The uni-partite representation is particularly suited for our purposes, because of the enormous amount of graph-theoretical methods and empirical intuition available for exploring their statistical properties. This perspective has already lead to remarkable successes in understanding systems of scientific collaboration [11, 22, 27, 62, 113, 115, 118, 119, 125, 133, 137, 156, 158, 166, 168].

One of the first large-scale analyses was conducted in [115], leading to a confirmation of the small-world conjecture and to the interesting finding that, in some scientific fields, the average network properties are dominated by the many people with few collaborators (e.g., biomedical research), rather than, as in other fields (e.g., high-energy physics), by the few people with many.

A very rich topic in the discussion of co-authorship networks is the centrality of authors and the network's community structure. Repeated removal of the most central edges (sum of the betweenness values of the end nodes) is for example used in [49] to determine the community structures within the network. Alternatively, [121] applies spectral theory to analyze the community structure. In fact, co-authorship networks have frequently served as an application example for module detecting algorithms.

While the topology of co-authorship networks is an extremely interesting object of investigation, we believe that relating the topology to dynamical processes can yield outstanding insights into the functioning of the scientific system and some aspects of social dynamics.

The search for fundamental relationships between network architecture and dynamical data is the guiding principle underlying our investigation. In order to identify such relationships for co-authorship networks, we explore the distribution of impact of publications across few-node subgraphs in the co-authorship networks. The main conceptual idea of few-node subgraphs as a means of exploring complex networks is that one looks at network properties and network function at a well-defined intermediate scale between the whole network and the individual node. In this sense and for sake of brevity we in the following call features of few-node subgraphs a *local* graph property, while we denote features of the graph as a whole (modularity, degree distribution, degree correlations, etc.) as well as whole-graph averages (average clustering coefficient, average betweenness centrality, etc.) as *global* graph properties. We are aware, however, that a thorough distinction between "local" and "global" would be more involved.

To our knowledge the only study starting to (indirectly) address an intermediate scale of network organization (between individual authors, together with their role in the network, and the whole network, together with its community structure) is [57], who explore the connection between team assembly mechanisms and the structure and performance of collaboration networks (including co-authorship networks). The parameters of their team assembly model are the fraction of newcomers in a team and the probability of repeating previous collaborations.

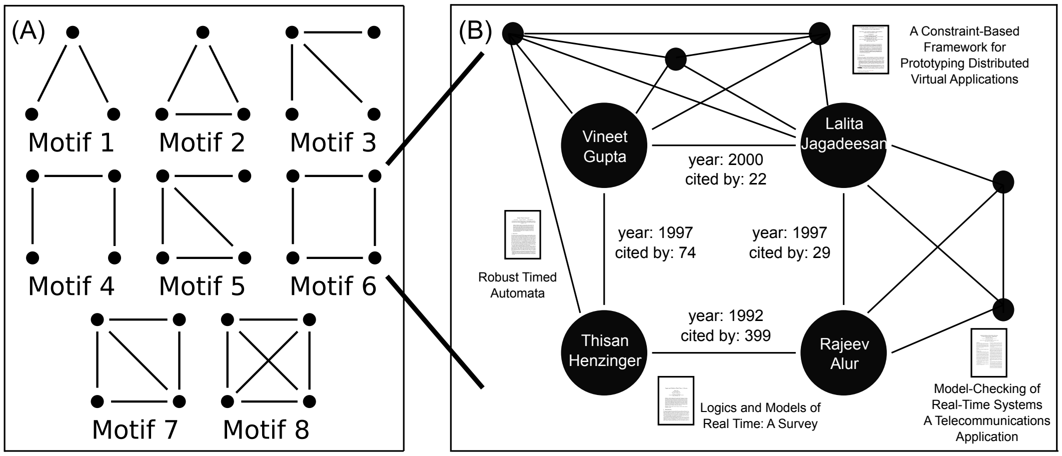


Figure 5.1.1: (A) The eight possible undirected three- and four-node motifs. (B) Example of a single occurrence of motif 6 (box motif) based on only four publications and embedded in the local network generated by these publications.

In this way they have been able to identify a phase transition towards a large connected component, as well as other structural network properties directly linked to the underlying process parameters.

The general relation between team properties and impact has also been addressed by [168], showing that teams produce more frequently cited research than individuals. This trend is increasing with time, is visible across many disciplines – from the sciences to the humanities – and includes the very high-impact research, a domain traditionally associated with the single-author "genius". Both studies, however, focus on individual publications, rather than the intermediate network scale of few-node constellations.

5.2 Results

We define the success of a motif as the average citation frequency per edge of all involved publications. From Google Scholar and CiteSeerX we extracted a database of citation frequencies for a large subset of publications entering our two co-authorship networks. These citation frequencies serve as our surrogate measure for the impact of the publication (details: see Appendix A).

It is not *a priori* clear, which of the four normalizations, defined in Appendix A, is the most natural one for the mapping of citation frequencies onto the co-authorship network. In particular, as soon as one of the normalizing quantities (the number of authors of a publication or the number of publications per edge) depends on some property of the co-authorship network (e.g., the degree of a node), which also varies among the motifs shown in Figure 5.1.1, the normalization will affect the average edge weights in a motif, even when the motif has no direct shaping influence.

For our main result shown in Figure 5.2.1, the average edge weights for the different

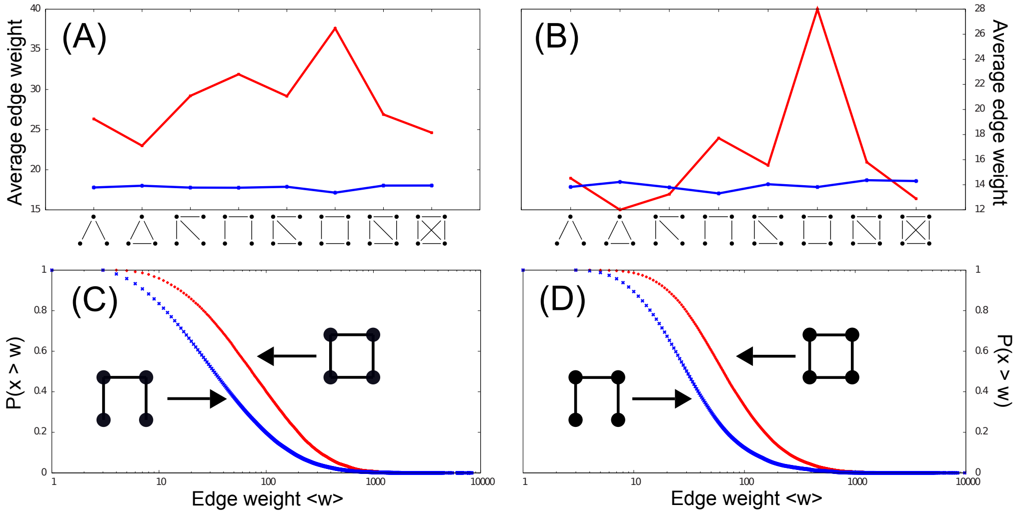


Figure 5.2.1: The average link weight per motif compared to the null model for DBLP (A) and CiteSeerX (B), according to edge weight definition from eqs. (1) and (3), respectively. In order to resolve the data behind the averages from (A) and (B), the cumulative distributions of the edge weights for two of the motifs are shown, namely the box motif (motif 6) and motif 4, for DBLP (C) and CiteSeerX (D).

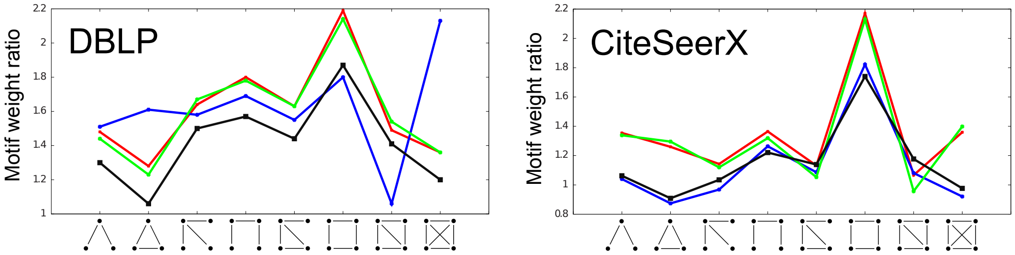


Figure 5.2.2: Ratio of average edge weights real data/null model for the edge weight definitions 1, 2, 3 and 4 in red, blue, green and black respectively, DBLP on the left side and CiteSeerX on the right side.

motifs from Figure 5.1.1, we selected the normalization that most successfully eliminates these residual dependencies. In order to understand, which of these normalizations is most suitable for a given data set, we randomize the citation frequencies of the publications, convert them into edge weights and re-compute the average edge weights of the motifs in this null-model scenario of shuffled citation frequencies. A uniform distribution of these null-model edge weights across the motifs indicates a successful elimination of the residual influences. It should be noted that in contrast to many network analyses, we do not randomize the network architecture, but rather shuffle the dynamical data on top of it. In this way we cannot discuss possible deviations of motif counts from randomness, but only the effect the motifs have in shaping the dynamical output of the network. The distributions of average edge weights across the motifs for the remaining normalization schemes is shown in the appendix, Figures A.5.1 and A.5.2.

It is clearly visible that not all normalizations yield flat distinction across the motifs for randomized citation frequencies. However, for both databases and four normalizations, the box motif (motif 6) has the highest ratio to its null model counterparts in all but one cases. This is shown in Figure 5.2.2.

Note that the average weight shown in Figure 5.2.2 is the weight *per edge* in a motif. Differences in the number of edges between the different motifs thus do not affect this quantity directly. Also, the unexpectedly high average weight observed for the box motif is not a trivial consequence of the fact that the box motif needs a minimum of four distinct publications for its construction. In fact, the box motif is no outlier with respect to the number of publications or the number of authors per edge (see in the appendix Figure A.5.7).

As the main test of robustness of our finding we construct time-truncated versions of the co-authorship networks for the past 20 years, where the network for year y includes all publications up to that year. For all the time-truncated networks the full (i.e., current-day) set of citations has been used. In Figure 5.2.3 the result from Figure 5.2.1A is thus shown for the time-truncated networks from 1990 up to 2008. The box motif clearly stands out as the motif with the highest average edge weight across all years. It should be noted that this time-resolved analysis of motif-related patterns in citation frequencies reveals some interesting additional features, for example the change in importance of motif 7 with respect to, e.g., motif 4 (probably associated with a trend towards denser motifs).

A typical example of a box motif occurrence in the co-authorship networks is shown in Figure 5.1.1B. This example helps us to look deeper into the specific mechanisms behind the box motif. Topologically, the surprising feature of the box motif is the lack of the two cross-wise links. The box motif is in this sense an “anti-clustered” motif. This “anti-clustering”, the lack of the two cross-wise links, is related to a segregation of the two pairs of authors, either geographically, temporally or with respect to the scientific disciplines. In other words, we expect that across two of the links strong gradients in space, time or discipline are observed. In the following we want to explore the nature of this separation from various angles. We can now ask, whether in those strong-gradient cases the two strong authors are linked or not.

In order to explore this, we define the weight of an author as the total number of citations of that author. In the third normalization scheme of the edge weight, equation (3), the author weight then corresponds to the sum of the weights of all edges linked to this author.

We partitioned all occurrences of the box motif into chunks of thousands. The first chunk comprises the 1,000 motif occurrences with the highest cumulative weight, the second chunk contains the 1,000 next highest ones, and so on. Figure 5.2.4 shows that, the higher the weight of a chunk, the more boxes can be found in this chunk such that the two strongest authors are

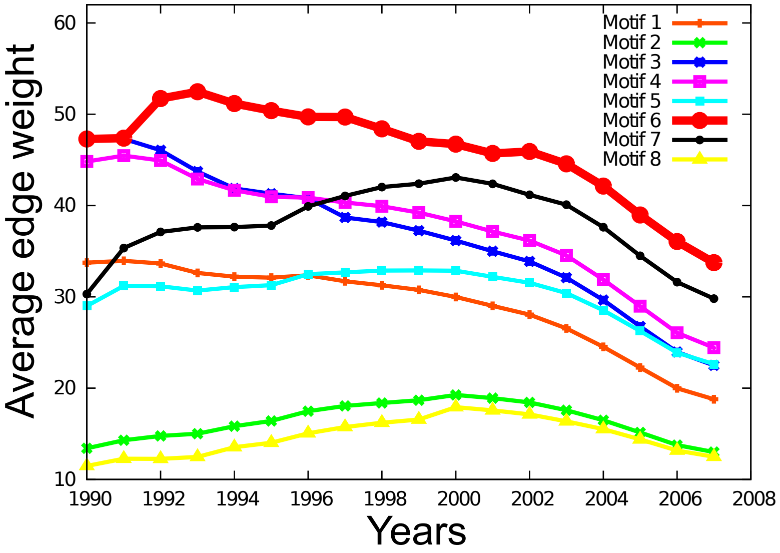


Figure 5.2.3: The average weight per motif link over the years for the DBLP database.

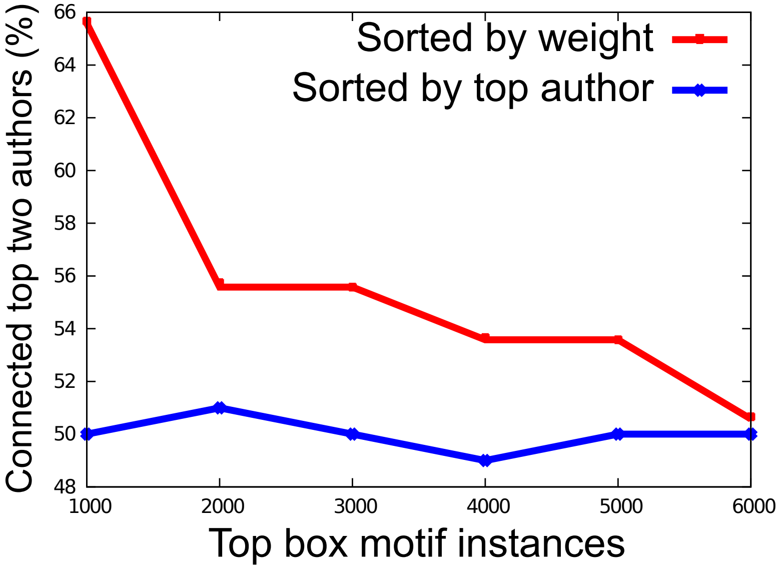


Figure 5.2.4: Percentage of box motif instances where the two top authors are connected directly within DBLP. The box motifs instances are divided in chunks of 1000 instances and sorted in descending order with respect to their weight.

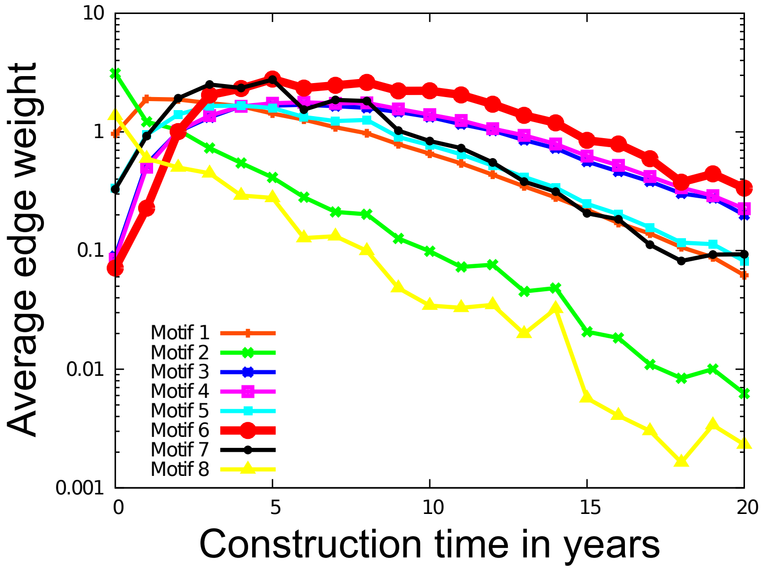


Figure 5.2.5: Relative average link weight per motif. All motif instances are distributed in bins according to their creation time.

adjacent.

The data available to us does not allow to inspect geography or discipline structure directly. To substantiate our claim, we made two further computational studies to understand how important these segregative features are for the success of the box motif.

First, we looked at the construction times of motifs. The edge initiation is given by the year of the first publication constituting this edge. For an occurrence of a motif, the construction time is the time between the earliest and the latest year of initiation of an edge within this occurrence. Figure 5.2.5 shows, for each motif and each construction time, the average weight of all occurrences of this motif that have the same construction time. It turns out that the box motif has a significantly stronger tendency than all other motifs for its heavy-weight occurrences to have high construction times. Thus, the heavy-weight occurrences of the box motif seem to span a bridge over time.

Second, we looked at the betweenness factors. For each motif, Figure 5.2.6 shows the average number of shortest paths that use edges of occurrences of a particular motif (normalized by the number of edges of this motif). Clearly, the box motif edges (together with those of motif 3) constitute high betweenness values and hence lay often on paths between larger communities within the network. This is a strong indication that the box motif is, to a certain extend, related to interdisciplinary collaborations.

5.3 Materials

Our study is based on two large-scale publication databases, DBLP and CiteSeerX as of May 2008 and October 2009, respectively, each containing several hundreds of thousands

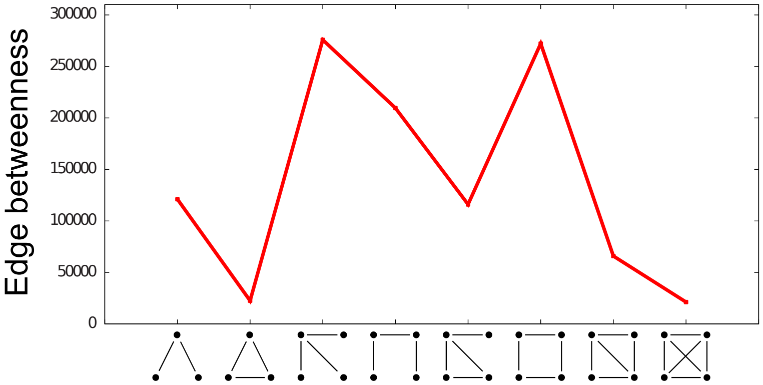


Figure 5.2.6: Average number of shortest path passing trough a motif link for the 1990 snapshot of the DBLP (all publications dating before or from 1990).

of publications. Publication lists are converted into a natural graph representation of co-authorship networks where the authors are nodes and two nodes are connected by an edge if the corresponding authors have ever published together.

We measure the success of a publication by the number of its citations by other publications. Citation indices have been acquired from the online search engines CiteSeerX and Google Scholar. A crucial step is to convert the impact of publications into edge weights in the co-authorship network. This conversion can be done in several different ways. For an edge e let $P(e)$ denote the set of publications represented by e. For a publication p, $c(p)$ denotes the citation frequency of p, and $A(p)$, the set of authors of p. The four edge weight w_e definitions are then as follows:

$$w_e := \sum_{p \in P(e)} c(p) \tag{5.3.1}$$

$$w_e := \frac{1}{|P(e)|} \sum_{p \in P(e)} c(p) \tag{5.3.2}$$

$$w_e := \sum_{p \in P(e)} \frac{c(p)}{|A(p)| - 1} \tag{5.3.3}$$

$$w_e := \frac{1}{|P(e)|} \sum_{p \in P(e)} \frac{c(p)}{|A(p)| - 1} \tag{5.3.4}$$

where $|S|$ denotes the number of elements in the set S.

The citation frequency of a publication can thus contribute to an edge weight either directly or normalized via the number of authors of that publication. Similarly, the frequencies of all publications contributing to an edge can either be summed up or averaged. These are the four variants of converting publication frequencies into edge weights given above.

5.4 Conclusions

We showed that some aspects of citation data are a consequence of the pattern of collaboration, rather than of the individual collaborators themselves. The outstanding role of the box motif, given by the highest average edge weight (derived from the citation frequencies of the underlying publications), the fact that the two authors in a box motif with the highest weight are typically adjacent, the high betweenness and long construction time, all give a first insight into the self-organization processes underlying the production of knowledge. In particular the segregative function of the box motif seems to be crucial. We believe that the lack of cross-links, the "anti-clustering", of the box motif is the main operational feature shaping the dynamical data.

In this sense the box motif, and the corresponding shaping of the data related to it, seems comparable with the "strength of weak ties" [52, 53]: High scientific success is on average associated with publications outside the densely clustered author constellations. It would be be worthwhile analyzing this also from a game-theoretical perspective, similar to the work of Goyal and Vega-Rodondo [51] on structural holes [29]. In fact, due to its "anti-clustering" feature, the box motif occurrences can be seen as small-scale versions of the structural holes distributed in the network.

There are several obvious ways of continuing this line of research. Following the general rationale of [57], we believe that a stronger connection between the motif patterns (in particular the outstanding role of the box motif) and the underlying elementary processes in the system (selecting authors for a publication, selecting articles to be cited within a publication) can only be achieved via generative minimal models. We describe a scheme for such a model in the appendix in A.8. On the level of the analysis of data a natural next quantity to explore are the conversion rates of motifs as the network evolves. The box motif might be seen as a metastable configuration as well as a decisive turning point in the individual author's scientific career.

Seeing co-authorship networks as an example of a social network and at the same time as a representative of a more generic class of production and distribution systems suggests that, via its segregative capacity similarly outstanding roles of the box motif can also be observed in other systems.

Chapter 6

Motifs as Markers of Future Success in a Model of Social Dynamics

Summary

Social networks are a fascinating field of exploration, where methods from graph theory and statistical physics can be put to use for understanding how network properties are related to the behavior of the social agents.

Here we study subgraph patterns in the Rosvall-Sneppen model of social network dynamics with the goal of predicting the network fate of individual agents. More precisely we predict the future degree of a node (which serves as a measure of social success) based on the local subgraph composition.

The results presented in this chapter have been achieved in cooperation with Marc-Thorsten Hütt and Kim Sneppen and will be published in: “Motifs as markers of future success in a model of social dynamics” [44].

6.1 Introduction

The term ’social networking’ for the task of creating connections to the right set of people with the aim of enhancing one’s own position in a social network has risen in importance over the last years, especially through the rise of online social network platforms making the already existing network property of social relations visible to the general public.

However, in most cases successful social networking reveals itself only in retrospect, when the social fates of all the players involved can be assessed. The electronic availability of data makes social network studies feasible. In particular for economical networks the last five years have seen the advent of a whole new, network-based research agenda, where a general understanding of the interplay between network architectures and collective dynamical behaviors forms the basis for risk assessment and systemic prediction (see, e.g., [95, 146]). A whole research agenda for predicting the large-scale behaviors of social systems (in particular the spread of epidemic diseases) on the grounds of detailed information on, e.g., movement of the individuals has recently been proposed [159].

An example of social dynamics studied in much detail over the last two decades are co-authorship networks and the production of scientific knowledge. Here, data availability, the clear definition of a social link and quantitative information on success (e.g., via citation frequencies) has turned this example into a model system of network applications. More precisely, the methods of complex network analysis provide a unique opportunity to quantitatively assess the social contribution to the production of knowledge. From the network perspective the strength of this social contribution can be re-phrased as follows: Does the underlying interaction network of authors and publications statistically explain parts of the success pattern of the scientific output?

While many studies of network topologies focus on global properties (e.g., the degree distribution — reviewed in [117], modularity [49, 55, 56, 122], degree correlations [31, 102, 116, 154], and hierarchical structures [12, 36, 69, 135, 157]), some of the dynamical function can be explained by small few-node subgraphs serving as devices for specific tasks organized locally in the graph. A potential signature of the functional role of few-node subgraphs is their statistical over- or under-representation (compared to a suitable ensemble of random graphs). This general concept has been developed and worked out by the Alon group [107, 109], particularly for transcriptional regulatory networks [3, 148].

Recent results statistically comparing network motifs in co-authorship networks with citation frequencies revealed an unexpectedly strong influence of the interaction network on a local scale, see Chapter 5. The key result is that the success of individual authors or publications correlates strongly with (three and four-node) network motifs. In particular, the box motif (four authors forming a closed chain) has the highest average citation frequency per link. This result is robust across two databases, across different possibilities of mapping the citation frequencies of publications onto the (uni-partite) co-authorship graph, and over time, and is markedly different from observations in suitably randomized data sets.

Prompted by these findings, we here explore the relationship between social success and network motifs for a simple model of social dynamics, the Rosvall-Sneppen model, introduced in [140] and further studied in [141].

We investigate the relationship between subgraph composition and dynamics in the Rosvall-Sneppen model in three steps: (1) As a first, global analysis we explore, whether the subgraph composition of the evolved social networks is non-random; (2) we then systematically screen the subgraph composition in the neighborhoods of nodes with increasing and decreasing social success (represented by their connectedness), respectively; (3) we use subgraph compositions obtained in step (2) from training data to predict, whether a node's degree will increase or decrease over time, i.e., whether a node is on its way of becoming socially more or less successful in the near future.

As the subgraph composition is a consequence of an intricate interplay of local and global network properties (see, e.g., [46, 59]) we formulate an analytical model of both the expected subgraph numbers and their composition. The model extends the formalism from Chapter 2 to four-node motifs in undirected graphs.

6.2 The model

In this chapter we use the model of social dynamics from [140] and [141]. The model is based on the concept that social rewiring is guided by communication events between social agents about other social agents. Neighbors of those neighbors that have frequently provided the most useful (i.e. most recent) information about other agents are selected as the most likely candidates for establishing a new link.

At every communication step two neighboring agents a and b communicate about a randomly selected other agent c. The agent with the older information a updates the age of its information (on c) to the age of the information of its partner b (on c). It will also remember that it got this information about c from b. After the system is initialized these 'information pointers' indicate for every agent who provided the information on this agent. At every rewiring step a random link in the network is removed and a randomly selected agent, say d has the opportunity to create a new link. This new link will connect the selected agent d with a neighbor of its neighbors.

Socially, these model rules mimic the process of 'social networking'. As the agents strive for better access to information, they will try to enhance their relative positioning in the network. The main parameter of the model is the communication rate C, which indicates how many communication steps are made per social (rewiring) step. Striking features of the model are the emergence of a scale-free degree distribution at not too low communication rates (and a transition from an Erdős-Rényi graph to a scale-free graph with increasing communication rate) and a turnover of hubs as a function of time.

In Figure 6.2.1 both features are clearly seen. The four sample networks shown in Figure 6.2.1A show a trend from rather random (Erdős-Rényi like) networks to networks with a broader degree distribution with increasing communication rate. In [140] it was numerically shown that, indeed, at high communication rate the degree distribution approximates a power law with slope of approximately -2.2. For a subset of nodes, the time courses of the nodes' degrees are shown for a sample run of the model at high communication rate. Essentially, one can discern certain time windows ('dynasties'), where a single high-degree node, stands out. Eventually this node's degree decays over time with some other node taking over as the system's dominant hub (Figure 6.2.1B).

In our analysis we focus on local network properties, i.e. the frequencies of few-node subgraphs. In this way we generalize the findings from [141], where, e.g., the clustering coefficient has been discussed. Analyzing the over- and under-representation of three-node and four-node subgraphs in this system as a function of the communication rate reveals the first non-trivial property of the network architecture on this scale of observation: As we will discuss in Section 6.4.1 social rules described above yield a clear motif signature that becomes more pronounced with increasing communication rate and does slightly resemble one of the superfamilies for undirected networks described in [107]. A major methodological problem in interpreting such a motif signature is the shaping of the pattern by the asymmetry in the three-node subgraphs that induces already a certain pattern of up- and down-regulations on the four-node level. We therefore construct a statistical theory allowing us to predict the number of four-node subgraphs given the three-node subgraph frequencies. More details are given in Section 6.3.

Figure 6.5.1 shows the quality of our prediction for an ER random graph. We use this formalism to verify that the signal observed in the four-node subgraphs is not induced by the signal observed in the three-node subgraphs.

In addition to the motif signature of the full graphs, we also analyze the local subgraph environment of each individual node. This is discussed in Section 6.4.2.

6.3 Statistical Description of Network Motifs

Similarly to the formalism developed in Chapter 2 for directed graphs, we here introduce a simple formalism that allows us to predict motif counts in undirected graphs. More specifically, in a first step, the 3-node motif counts are predicted from 2-node motif counts, edges and non-edges. The 4-node motif counts can be predicted in the same way, or by using the 3-node motif counts and the 2-node motif counts together. This last step makes it possible to analyze the intrinsic correlations between 3- and 4-node motifs.

In this study we are especially interested in the question whether the over-representation of triangles and the suppression of 3-chains is enough to explain the whole motif signature or if there is another higher-order effect of the social dynamics on the scale of 4-node motifs.

The maximal number of edges m_{max}in a motif with n nodes is $m_{max}(n) = \binom{n}{2}$. Every one of these positions can either be occupied by an edge or not. The probability for a place to be occupied is $p = \frac{M}{2N(N-1)}$ when M is the total number of edges and N is the total number of nodes in an undirected graph. The probability that a place is not occupied is denoted by $q = 1 - p$. To extend this formalism to include 3-node motifs we introduce the probability $r = p^2 q$ that three randomly chosen nodes form a 2-chain and $s = \frac{p^3}{3}$ that three randomly chosen nodes form a triangle.

The next step is to 'build' four-node motifs from a three-node motif and additional edges. We 'grow' the four-node-motifs from three-node-motifs by adding edges. The expressions for the growing motif theory can now be used to predict the number of 4-node motifs based on the number of 2- and 3-node motifs. This is done by estimating $p = M/(2N(N-1))$, $q = 1 - p$, $r =$ 2-chains$/((N-2)(N-1)N)$, $s =$ triangles$/(2(N-2)(N-1)N)$.

Note that $r + s \neq 1$, as "incomplete" motifs are not considered. The resulting motif counts from applying this theory can bee seen in Figure 6.4.2.

6.4 Results

6.4.1 Motif Signature

When the communication rate is low, the "information pointers" are most of the time outdated, making the social step more of a random process. On the other hand when the communication rate is high, one approaches the limit of infinite communication, where every decision is locally the best to be made.

We perform a motif analysis on the resulting networks after 10^5 social steps with varying communication rate. Figure 6.4.1 shows that the social dynamics represented by the model lead to a clear non-random pattern of over- and under-representations of few-node subgraphs. The pattern is already visible at comparatively low (but non-zero) communication rates, increases strongly with increasing communication rate and then saturates, yielding a pronounced and stable motif signature.

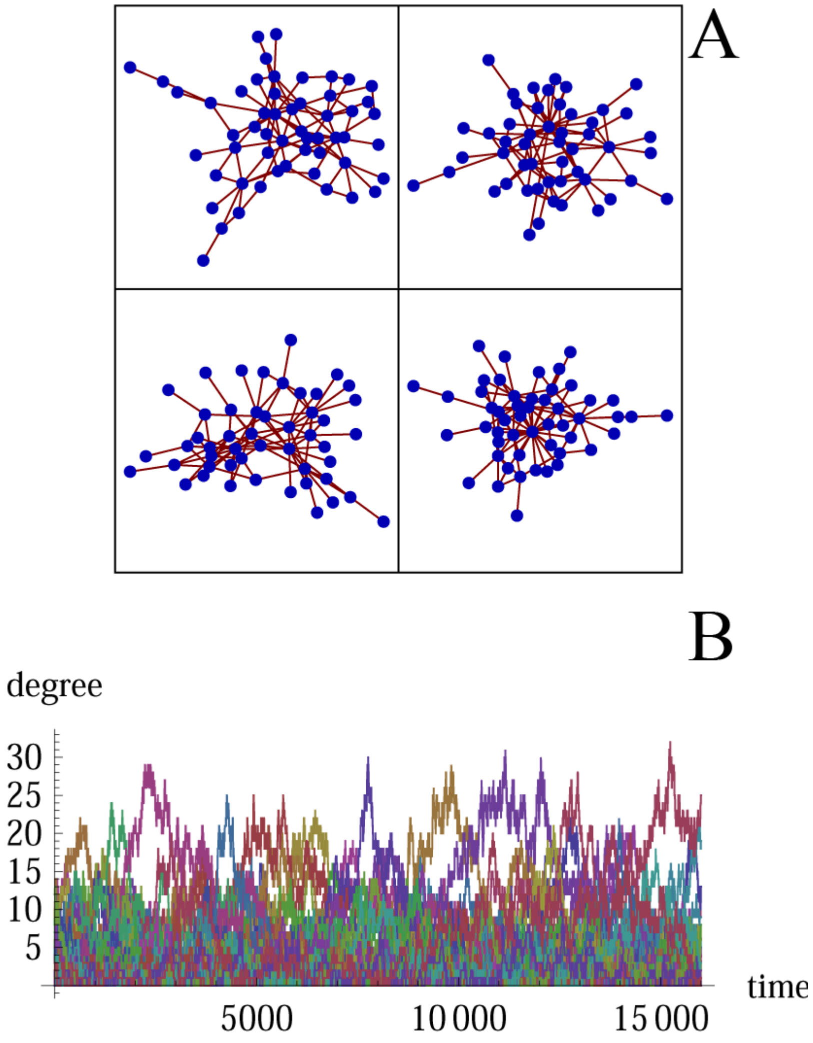


Figure 6.2.1: Basic features of the model. (A) Typical networks for different values of the communication rate, namely (a) $C = 8$ (b) $C = 79$ (c) $C = 500$ (d) $C = 794$ chats per rewiring. (B) Time series of the degree for a randomly picked and randomly colored subset of nodes at high communication rate ($C = 794$ chats per rewiring).

motif	1	2	3	4	5	6	7	8
nodes	3	3	4	4	4	4	4	4
edges	2	3	3	3	4	4	5	6
triangle	0	1	0	0	1	0	1	1
3chain	1	0	1	1	0	1	0	0
symmetry factor	1	1	1/3	1	3	1/4	3/2	1/4
additional edges	0	0	1	1	1	2	2	3
anti-edges	0	0	2	2	2	1	1	0
simple	$\frac{p^2qn!}{2(n-3)!}$	$\frac{p^3n!}{6(n-3)!}$	$\frac{p^3q^3n!}{6(n-4)!}$	$\frac{p^3q^3n!}{2(n-4)!}$	$\frac{p^4q^2n!}{2(n-4)!}$	$\frac{p^4q^2n!}{8(n-4)!}$	$\frac{p^5qn!}{4(n-4)!}$	$\frac{p^6n!}{24(n-4)!}$
growing	$\frac{rn!}{2(n-3)!}$	$\frac{sn!}{2(n-3)!}$	$\frac{pq^2rn!}{6(n-4)!}$	$\frac{pq^2rn!}{2(n-4)!}$	$\frac{3pq^2sn!}{2(n-4)!}$	$\frac{p^2qrn!}{8(n-4)!}$	$\frac{3p^2qsn!}{4(n-4)!}$	$\frac{p^3sn!}{8(n-4)!}$

Figure 6.3.1: The symmetry factors and probabilities for all 3- and 4-node motifs in an undirected graph. It is important to note that while technically we just substituted $\frac{p^3}{3}$ by s, this can only be done if the 4-node motif really contains a triangle. Otherwise the resulting information about the correlations will be wrong as soon as the triangle probability does not correspond to $\frac{p^3}{3}$ (which is often the case).

The dominant signal in the motif signature observed here is an over-representation of triangles and motifs containing triangles and an under-representation of sparser motifs. Also, in the 4-node motifs, significant additional deviations from randomness can be observed. Disentangling the one effect from the other requires the formalism developed in Section 6.3. We start from the prevalences of edges and three-node motifs and compute expectations values for four node-motifs. Figure 6.4.2 demonstrates that the information contained in the three-node subgraphs does not trivially induce the pattern of over- and under-representation observed in the four-node subgraphs. If the four-node-motif signature was only the result of the three-node motifs the curve '3-motifs' should be overlapping with the 'real' curve, incidentally the two signals even seem anti-correlated.

Several studies have recently attempted to understand the distinct categories of motif compositions (the "superfamilies" from [107]) from a functional perspective. In particular for directed graphs, some relationships with systemic robustness could be established for different dynamical processes (see, e.g., [72, 77]). In the light of these investigations, we would like to emphasize that the motif signature of the evolved networks at high communication rate resembles the one observed for contact networks derived from protein structures.

In order to better understand the universality of this pattern we measure the distance of the motif signatures to the signature of a random (initialization) and from the signature of a network that was obtained from a simulation with very high communication. This is shown in Figure 6.4.3. For high communication rates the motif signature seems to converge, when the communication is low the social steps rather act like random mutations, as the basis for decisions (the "information pointers") do not reflect reality anymore as they are outdated.

We can now use the formalism introduced in Section 6.3 to predict expected motif prevalences and by this proxy motif z-scores for larger motifs. This makes it possible to assess the impact of dynamics on 3- and 4-node motifs separately.

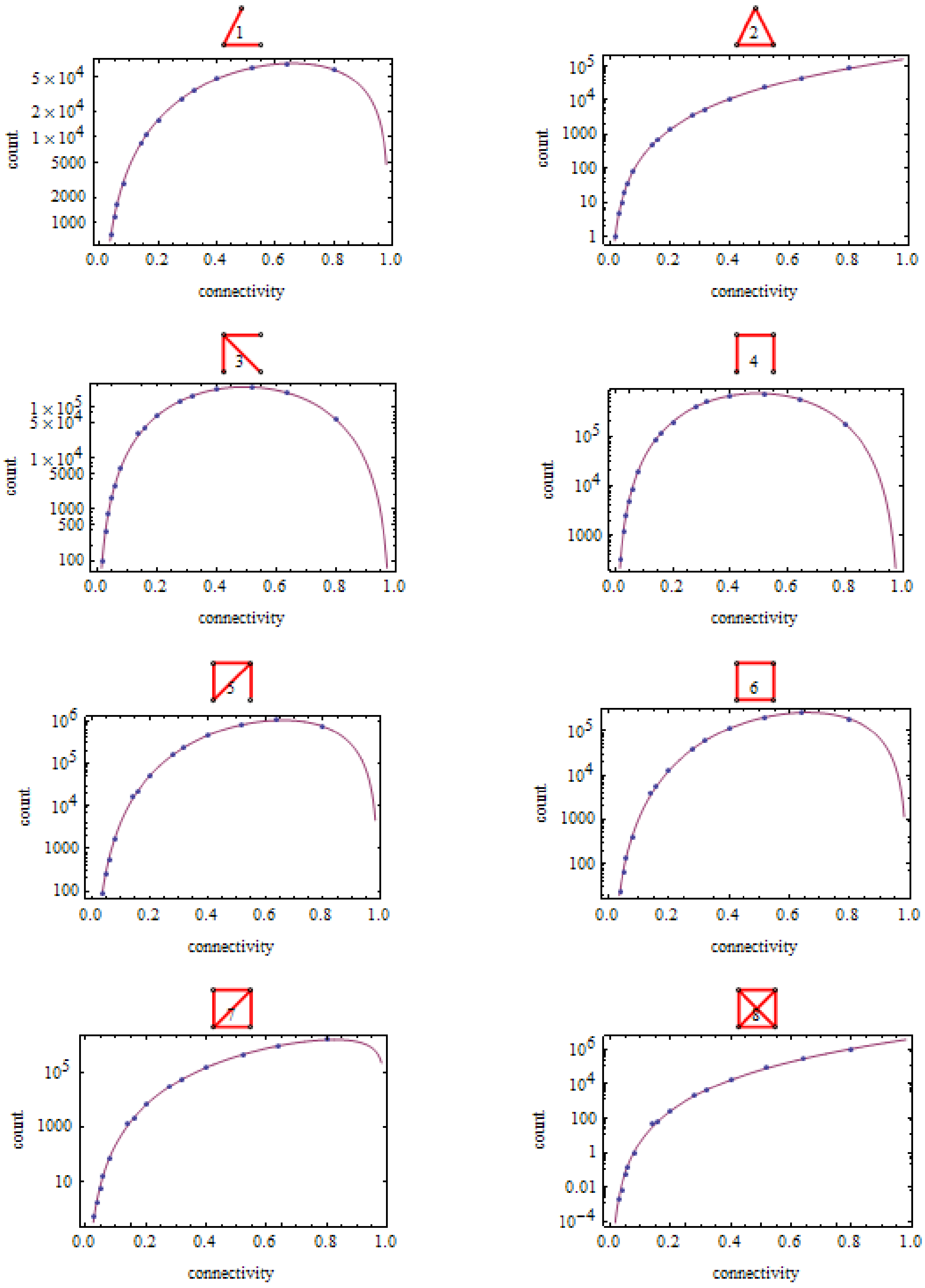


Figure 6.3.2: Comparison of predicted with numerically obtained motif counts. The networks used had $N = 100$ nodes and connectivities varying over the whole range.

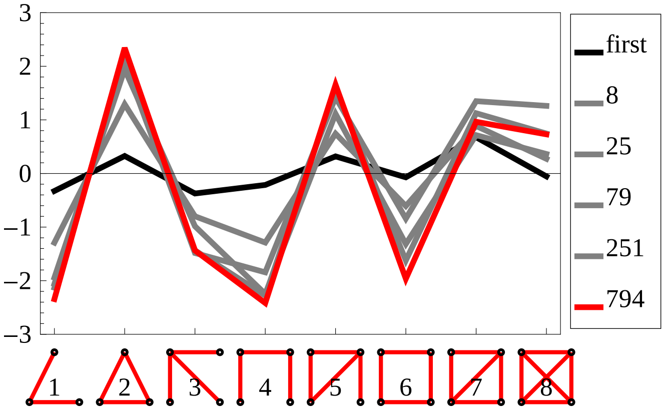


Figure 6.4.1: The motif signatures of the connection network after 10^5 social steps for varying communication rates. The networks were generated with $N = 70$ nodes and $M = 100$ edges. The communication rate is given in chats per rewiring event.

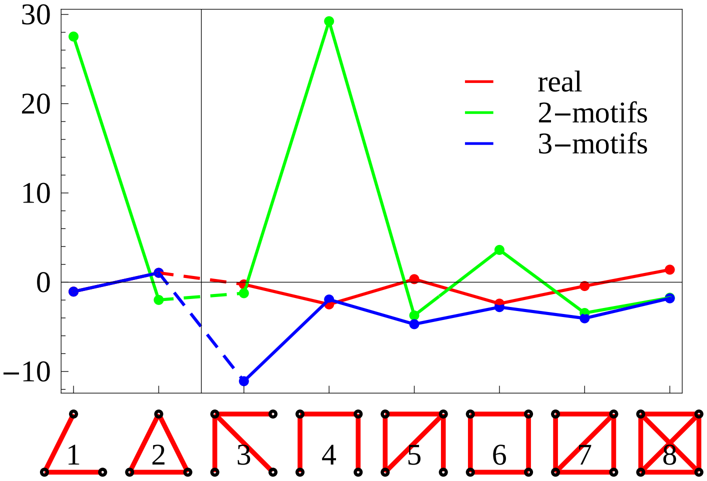


Figure 6.4.2: The z-score of all undirected motifs with an analytical curve based on 2- and 3-node motifs. The two first points of the signature are equal, as they are the basis for all other points.

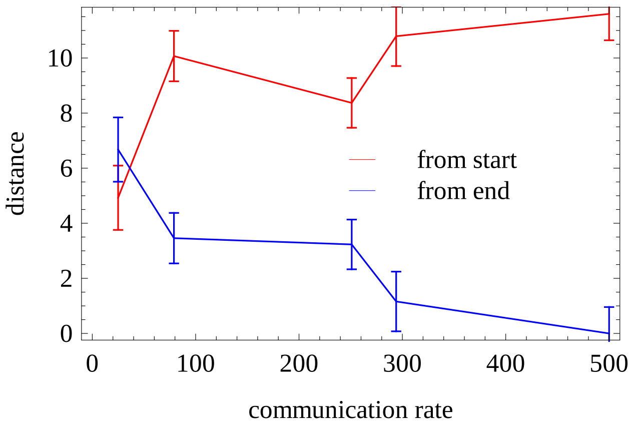


Figure 6.4.3: To better show that the motif signature actually converges for high communication rates we here show the sum of the squares of the distances between the motif signatures for different communication rates towards the start (un-evolved network) and to the resulting network at very high communication.

6.4.2 Local Subgraph Patterns

Next, we investigate, how the motif signature obtained in Section 6.4.1 is distributed across the graph. As all nodes retain their identity under randomization, we can compute a motif signature at each node individually (i.e. the z-score for each subgraph at each node). This is shown in Figure 6.4.2. It is surprising, how robustly this very local quantity resonates a common property of the networks.

In order to test the predictability of future hubs in the system, we have converted a set of simulation runs performed for various communication rates into a database containing the following information: For several reference degrees (namely k^* = 18, 20, 22, 24) we enumerate all pairs (node, time), where this degree is transversed, together with the corresponding slope (dk/dt) (computed at a window size of ΔT=$200/C$ time steps) and the subgraph composition around this node at this moment in time. Next, we select the pairs with the 50 highest and 50 lowest (i.e. highest negative) slopes dk/dt as our reference events for the rising and falling nodes, respectively.

When processing the simulated data in such a way, we can now perform averages of the local motif composition for subsets of nodes selected according to the nodes' future fate in the social system. Figure 6.4.6 shows such averages for three subsets of nodes, each analyzed at a fixed degree of t = 24, (a) only rising nodes, (b) only falling nodes, and (c) both sets of nodes together.

The local motif composition of rising nodes differs strikingly from those of falling nodes for three subgraphs: the triangle, the four-node star and the box. A typical rising node has a lower suppression of triangles, a comparatively small number of four-node stars and, quite pronouncedly, a very low number of box motifs in its local neighborhood. The route towards social success (represented by a high future degree) is therefore characterized by an unexpectedly careful balance between not-too-low clustering (less triangle suppression) and

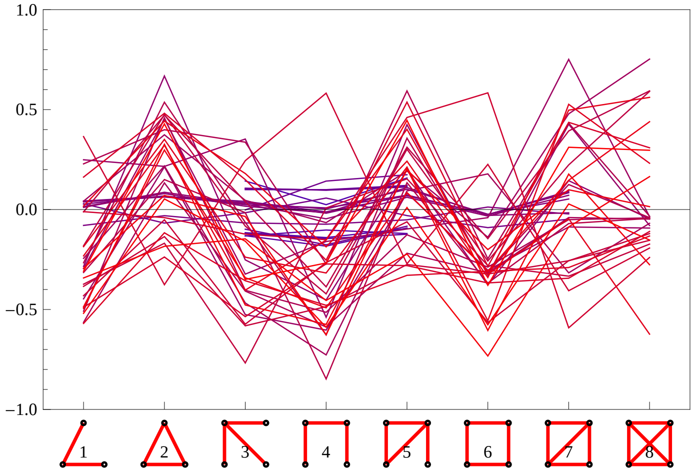


Figure 6.4.4: Local z-scores for all nodes of a network after 10^5 rewiring steps at high communication rate ($c = 1000$ chats per rewiring). The z-score is obtained by generating 1000 randomized networks and computing the motif count for every node in every random graph. Then the mean value and the standard deviation of these motif counts are used to calculate the z-score for each node, similarly to the z-score of a whole network. The color goes from red (high degree) to blue (low degree).

the avoidance of a very low-clustering motif (the box motif, which is 'anti-clustered' in the sense of a systematic lack of cross-links among the four social actors involved; see also Chapter 5).

6.5 Predicting the Fate of Agents on the Basis of Local Topology

We will now try to use motifs as markers for predicting future success, therefore we use the tagged agents from Figure 6.4.6 and try to predict solely based on there motif environment to which group they belong.

The prediction has been performed for the ratio of subgraph counts, e.g., $r_m^{(r)} = N_m^{(r)}/(N_m^{(a)})$, where $N_m^{(r)}$ denotes average number of subgraphs m at the degree k^* in the 'rising' category and $N_m^{(a)}$ denotes average number of subgraphs m in 1000 randomly picked agents at the degree k^*. The agent is assumed to be rising if $r_m > s_m$ and falling otherwise.

Establishing a quality measure of a binary classification test in general requires the simultaneous assessment of the specificity and sensitivity of the method, as well as a monitoring of true positives and true negatives. Here we ensure that both categories (the "positives" and "negatives") contain the same number of events. This is done by choosing the top 100 events for each category, in the sense that the trend towards increasing/decreasing degree is the strongest. We can therefore use just one measure to assess the specificity and sensitivity at once. In this case we choose the

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{number of events}}.$$

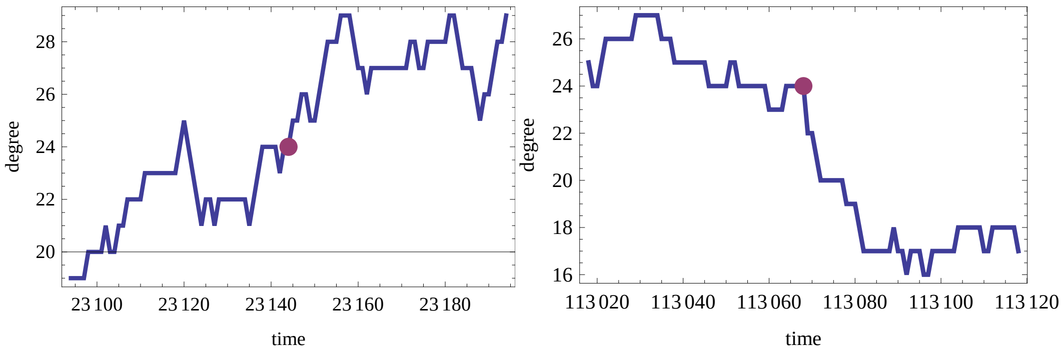


Figure 6.4.5: Typical time-series of events that are considered as rising or falling, respectively. The threshold is set at 24 on a system with $N = 70$ and $M = 100$.

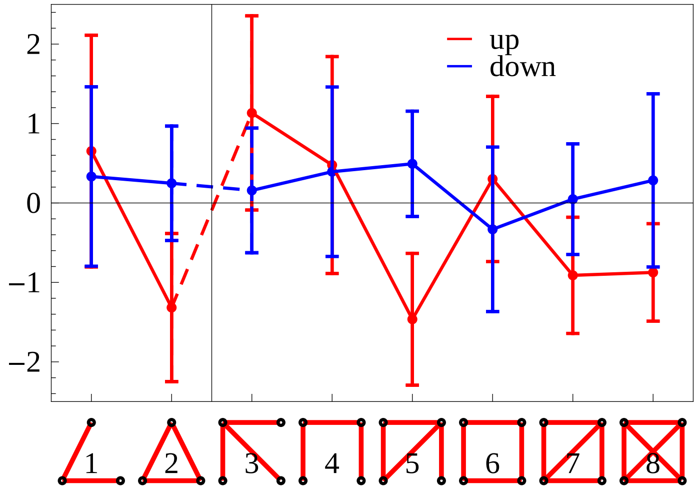


Figure 6.4.6: Motif signatures for nodes with increasing and decreasing degree. The motif environment of the top 100 of the rising and falling agents of the same degree ($k^* = 17$). This data is extracted from a system with (N=70,M=200,C=794 chats per rewiring) and averaged over five time-series. The z-score of the rising/falling agents is obtained by extracting the motif counts and standard deviations of 1000 randomly selected agents of the same degree.

We use a s_m that was obtained by an evolutionary optimization process (as only one variable has to be optimized and the optimization goal, the accuracy is not very rugged the optimization method does not matter, an optimal s_m is achieved in only a few steps, we use a simple simulated annealing.

In spite of the clearly visible large error bars, it is quite remarkable that the two sets of nodes (increasing and declining social leaders) show very distinct local subgraph compositions. On the basis of this observation, we next attempt to predict purely on the basis of the subgraph environment, whether a node is bound to become the next social leader.

When an individual four-node motif is chosen as a marker for future success, the results from Figure 6.5.1 suggest that after the triangle the box motif and the semi-clique provide the most reliable information (highest prediction quality, when the information is taken only from a single motif). It is intuitively clear that combining the information from several motifs may increase the prediction quality further. At the same time, motif-motif correlations (in particular between three-node and four-node motifs) can distort this result substantially. It is therefore indispensable to explore such correlations more quantitatively, in order to assess the pure information about future success contained in such a motif (undiluted by, e.g., smaller patterns in the network).

6.6 Conclusion

We have shown that the simple model of social dynamics induces a specific motif content into the network, and explained how this motif signature is formed. Additionally, we could show that successful agents can be distinguished from less successful agents by means of their local motif neighborhood. We used this knowledge to predict solely on the local motif structure whether an agent will be rising or falling in degree. The analysis of the prediction quality also yielded some interesting information about the importance of several motifs to the agent's success.

On the level of four-node subgraphs, the segregative function of the box motif and the semi-clique seem to be crucial. We believe that the lack of cross-links (in particular, the "anti-clustering", of the box motif, as discussed in Chapter 5) is the main operational feature shaping the dynamical data. A difficulty in interpreting subgraph patterns is the crosstalk between statistically related subgraphs. In our investigation we observe that the triangle-plus-line motif has a similar predictive power as the box motif. This is due to its antagonist role (to the box motif) of measuring isolated triangles, as opposed to the densely packed triangles in a clique. In this sense, the triangle-plus-line motif also has an "anti-clustered" structure, just like the box motif.

In this sense the box motif, and the corresponding shaping of the data related to it, seems comparable with the “strength of weak ties" [52]. Looking at social systems as complex networks and analyzing them with graph-theoretical tools has the potential of providing important building blocks for a quantitative theory of social processes.

By using our understanding of dynamical processes on graphs as a template for analyzing patterns in data we can in this way uncover unexpected patterns in data and extract fundamental design principles underlying social dynamics in a formal way. Dramatic recent examples include suggestions of optimized vaccination strategies (based on the concept that a random walk will map out the degree sequence of the social network) [34], and the view of obesity as a

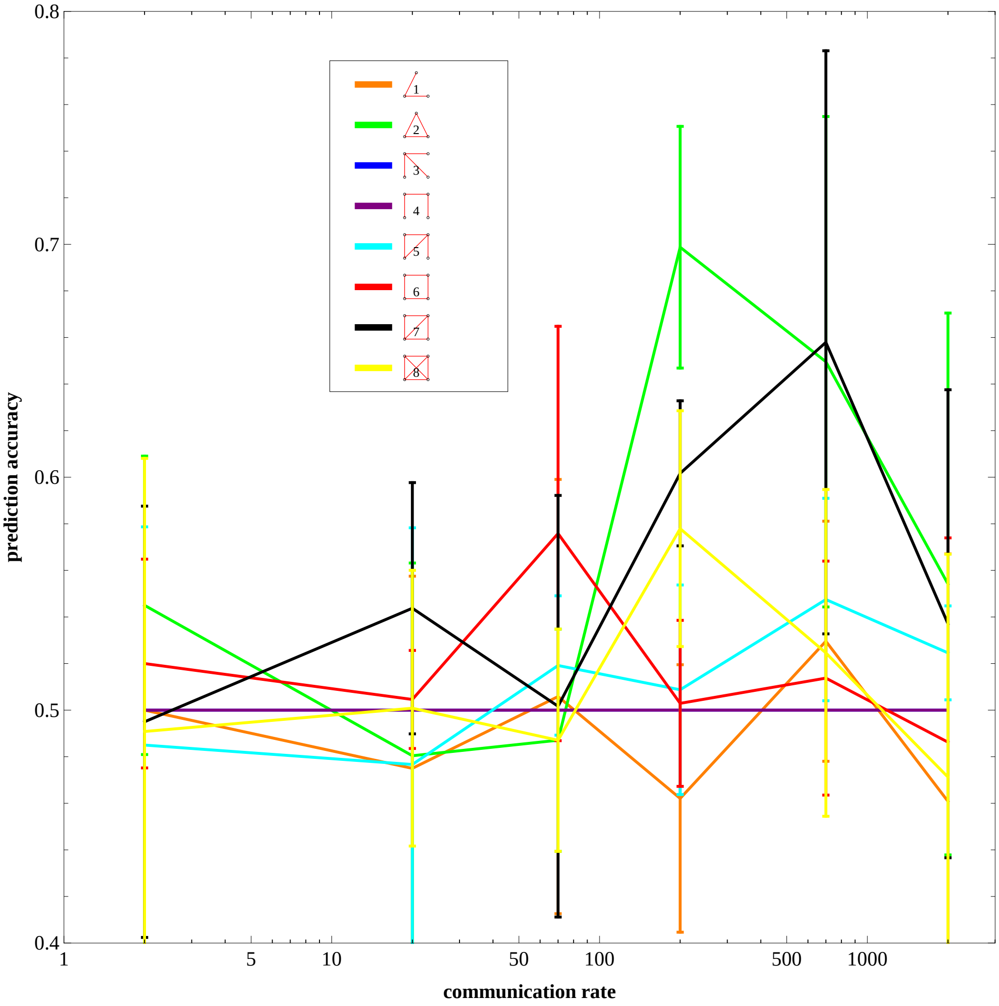


Figure 6.5.1: The prediction quality depending on which motif is used for the prediction. The communication rate is given in chats per rewiring event and and the prediction is performed at a degree of $k^* = 17$.

(social) epidemic propagating on the network of human interactions [35, 41].

The main result of this study is that, socially, there is no additional benefit from forming densely packed network environments and from striving towards high clustering. Instead, a pronounced anti-clustering (as represented by the box motif and, less strongly, by the semi-clique) is indicative of the future social fate.

Chapter 7

Software Tools

7.1 State of the Art

The most important software package for a motif analysis is MFINDER, the software used in the original works on motifs (e.g., [109]). Motif finding was also implemented for the PAJEK network analysis software [14].

Graphcrunch [105] is a tool that is specialized in the automatic detection of the best fitting graph model for a given input network. For this it implements five different graph models, some of which have tunable parameters. The fitting criterion can be chosen among a series of graph properties, including the spectrum of shortest paths and subgraph counts. The best fitting network model (together with the tuned parameter values if it has any) as well as a measure indicating the quality of the fit is returned.

NetworkWorkbench [153] is a more general framework for network analysis. It implements many of the standard methods for network generation, analysis and visualization, interfaces to other graph analysis software packages are provided. Some authors have described and implemented sampling algorithms that are, especially for larger subgraphs, faster by several orders of magnitude than total enumeration algorithms. *FANMOD* [165] was the first software package implementing *unbiased* sampling, MAVISTO [145] additionally has the ability to highlight motifs in labeled networks. Finally, [123] describes and implements a sampling algorithm that uses a pattern growth approach to efficiently detect larger subgraphs.

7.2 Own Development

During my thesis work, a software package specialized on motifs and networks has been developed, as the existing packages were not easily extendable. Especially the systematic manipulation of networks is not part of any of the described software tools and would be difficult to add. The standard way to use i.e. mfinder is to run it on a network and then parse the output for the z-score number. It would be possible to extend NetworkWorkbench to include motif analysis, but at the current state really interactive applications are not possible, also a way of batch processing large numbers of graphs does not exist at the moment.

The core modules of our software package are:

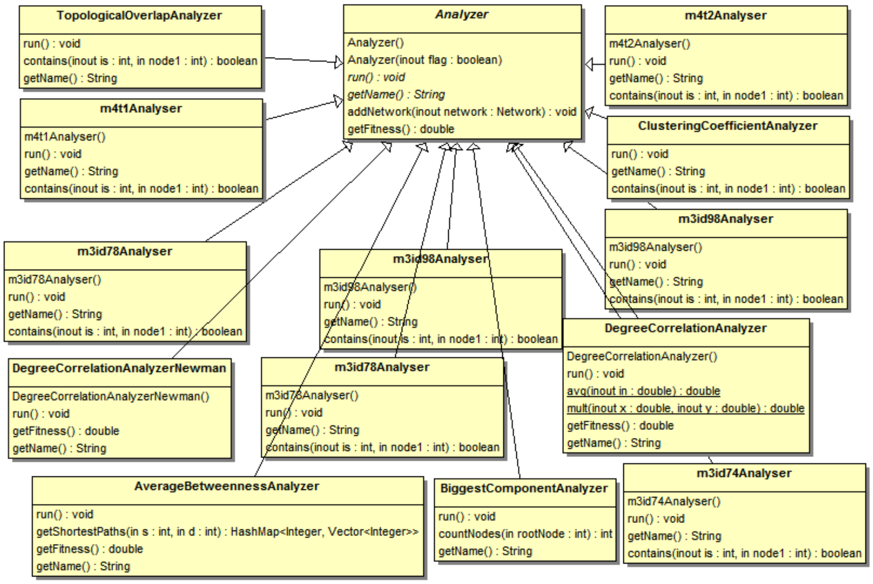

TopologicalOverlapAnalyzer
run() : void
contains(inout is : int, in node1 : int) : boolean
getName() : String
Analyzer
Analyzer()
Analyzer(inout flag : boolean)
run() : void
getName() : String
addNetwork(inout network : Network) : void
getFitness() : double
m4t2Analyser
m4t2Analyser()
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean
m4t1Analyser
m4t1Analyser()
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean
ClusteringCoefficientAnalyzer
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean
m3id98Analyser
m3id98Analyser()
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean
m3id78Analyser
m3id78Analyser()
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean
m3id98Analyser
m3id98Analyser()
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean
DegreeCorrelationAnalyzer
DegreeCorrelationAnalyzer()
run() : void
avg(inout in : double) : double
mult(inout x : double, inout y : double) : double
getFitness() : double
getName() : String
DegreeCorrelationAnalyzerNewman
DegreeCorrelationAnalyzerNewman()
run() : void
getFitness() : double
getName() : String
m3id78Analyser
m3id78Analyser()
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean
AverageBetweennessAnalyzer
run() : void
getShortestPaths(in s : int, in d : int) : HashMap<Integer, Vector<Integer>>
getFitness() : double
getName() : String
BiggestComponentAnalyzer
run() : void
countNodes(in rootNode : int) : int
getName() : String
m3id74Analyser
m3id74Analyser()
run() : void
getName() : String
contains(inout is : int, in node1 : int) : boolean

Figure 7.2.1: An extract of the class diagram surrounding *motifs.analyzer.Analyzer,* some of the 3- and 4-node analyzers are omitted for space reasons. The method getFitness() will return a single number. The class diagram for *motifs.analyzer.AnalyzerInd* is almost identical except for the naming and the return type of getFitness().

- the generation of graphs by a number of standard algorithms
- the counting of motifs in directed and undirected graphs
- the randomization of graphs with various constraints (preservation of bi-directional edges, preservation of modules...)
- computing other global graph-theoretical measures (diameter, edge density, degree correlation)
- computing local graph-theoretical measures (centrality betweenness, clustering coefficient, local degree correlation) as well their distribution over the graph
- analyzing the inter-dependencies of such measures

All modules can easily be reconnected to quickly fulfill new tasks. For example all global network measures (i.e. that return only a single number) are derived from the same super class *motifs.analyzer.Analyzer*, (see also 1.3.4) so that a goal-directed walk that increases the number of triangles in a graph can instantly be changed into decreasing the centrality betweenness. An extract of the class diagram is shown in Figure .

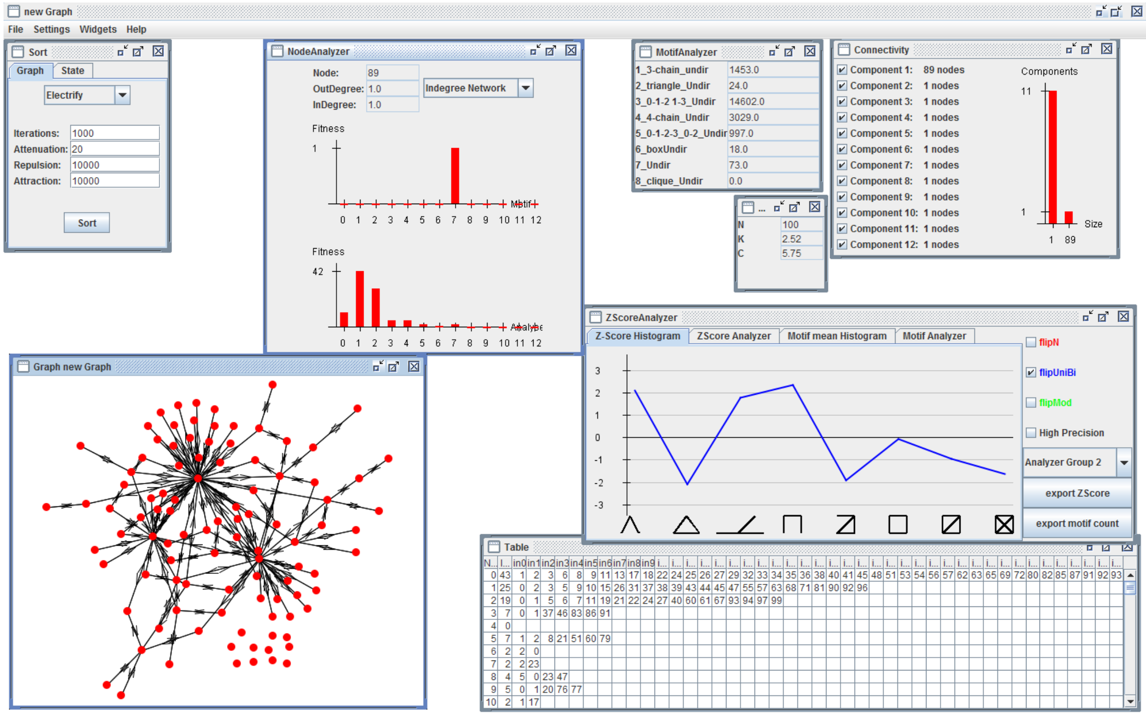
The same is true for local graph measures that can be evaluated for every node. Here they return an array of numbers with the length being the number of nodes in the analyzed graph and are derived from the super class *motifs.analyzer.AnalyzerInd*. The rapid prototyping is further simplified by the use of class loaders, so that for example a specific motif count can be requested via a command-line interface. All registered classes derived from *motifs.analyzer.Analyzer* are scanned for their static field *name* and the matching class is then instantiated. This removes the need for maintenance-intensive glue logic.

All modules can be used via the command line, either via files or pipes. Additionally for some use cases where an interactive inspection is helpful user interfaces were developed.

Here, the most important of these user interfaces are presented in detail.

7.2.1 Interactive Graph Representation

Our interactive general-purpose graph analysis tool is programmed in Java and optimized for portability and easy extendability.



Electrify
Iterations: 1000
Attenuation: 20
Repulsion: 10000
Attraction: 10000

Node: 89
OutDegree: 1.0
InDegree: 1.0
Indegree Network

1_3-chain_undir	1453.0
2_triangle_Undir	24.0
3_0-1-2 1-3_Undir	14602.0
4_4-chain_Undir	3029.0
5_0-1-2-3_0-2_Undir	997.0
6_boxUndir	18.0
7_Undir	73.0
8_clique_Undir	0.0

N	100
K	2.52
C	5.75

☑ Component 1: 89 nodes
☑ Component 2: 1 nodes
☑ Component 3: 1 nodes
☑ Component 4: 1 nodes
☑ Component 5: 1 nodes
☑ Component 6: 1 nodes
☑ Component 7: 1 nodes
☑ Component 8: 1 nodes
☑ Component 9: 1 nodes
☑ Component 10: 1 nodes
☑ Component 11: 1 nodes
☑ Component 12: 1 nodes

☐ flipN
☑ flipUniBi
☐ flipMod
☐ High Precision
Analyzer Group 2

N...	I...	in0	in1	in2	in3	in4	in5	in6	in7	in8	in9	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...	i...
0	43	1	2	3	6	8	9	11	13	17	18	22	24	25	26	27	29	32	33	34	35	36	38	40	41	45	48	51	53	54	56	57	62	63	65	69	72	80	82	85	87	91	92	93
1	25	0	2	3	5	9	10	15	26	31	37	38	39	43	44	45	47	55	57	63	68	71	81	90	92	96																		
2	19	0	1	5	6	7	11	19	21	22	24	27	40	60	61	67	93	94	97	99																								
3	7	0	1	37	46	83	86	91																																				
4	0																																											
5	7	1	2	8	21	51	60	79																																				
6	2	2	0																																									
7	2	2	23																																									
8	4	5	0	23	47																																							
9	5	0	1	20	76	77																																						
10	2	1	17																																									

Figure 7.2.2: The Interactive Graph Representation showing a generated BA-graph with its most important widgets open.

The widgets shown in Figure 7.2.2 from top left to bottom right have the following functions:

1. The Sorting widget. It is possible to randomly place the nodes on the panel, to order them in a grid, arrange them in a circle or to sort them by a spring embedding model, similar to [47].

2. The Node Analyzer widget, set to show the in-degree of a selected node (upper panel) and the corresponding distribution (lower panel).
3. The Motif Analyzer widget, set to show undirected 3- and 4-node subgraph counts.
4. A quick view of node count, average degree, and clustering coefficient.
5. An overview of the weakly connected (disregarding the edge directions) components of the analyzed graph. Every component is shown with its size, additionally a histogram of component sizes is displayed.
6. The graph panel, showing the graph currently under analysis. Nodes can be rearranged by drag and drop. When a node is clicked on a window displaying the neighbors of the node is shown. This list is editable edges can be removed or added.
7. The z-score analyzer, here set to show undirected 3- and 4-node motif z-scores. It can be switched to directed 3-node motifs and the randomization method can be set to:

- Edge flipping, not preserving the number of bi-directed edges
- Edge flipping, preserving the number of bi-directed edges
- Edge flipping, preserving the number of bi-directed edges and the modular structure

  Additionally, the current motif counts, as well as the average motif count and standard deviation in the random ensemble can be shown. The high precision mode will compute a complete z-score several times and then show the average z-score with error bars derived from the ensemble of obtained z-scores.

8. The (editable) adjacency list of the graph.

All widgets and the graph view are updated in real time when any change on the graph is performed, this works well up to some 100 nodes. Larger graphs can also be processed, but the graph representation will be slow. CPU-intensive calculations (especially the z-score and the spring embedding) are run in parallel background processes, so that the interface always remains responsive and CPU resources are optimally used.

Graphs can be imported and saved via a simple file format (the adjacency list separated by tab-stops).

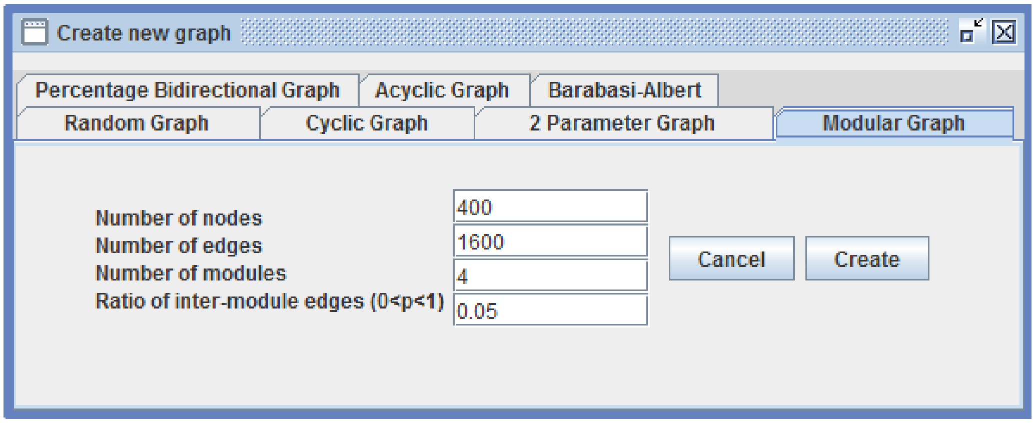
Graphs can also be generated by a number of standard algorithms, see Figure 7.2.3.

7.2.2 Motif Dependencies

For a more general investigation of motif properties, especially motif inter-dependencies another software tool was implemented.

The concept is to have an ensemble of graphs and to observe their multidimensional drift in motif space. Additionally the random walk can be biased, transforming it into a goal-directed walk. This is performed with two vectors, the goal vector g_m, representing a point in motif space, and a weight vector w_m, representing the relative importance of the corresponding dimensions.

Motifs that have zero at their place in the weight vector are ignored for the random walk, when every entry in the weight vector is set to zero a random walk is the result. Positive



Number of nodes 400
Number of edges 1600
Number of modules 4
Ratio of inter-module edges (0<p<1) 0.05

Figure 7.2.3: The options for generating random graphs, the tab for modular graphs is selected.

values at a place in the weight vector makes the networks drift away from the goal point in the corresponding dimension. With c being the current motif count vector a goal function d is defined:

$$d = \sum_m (g_m - c_m)^2 w_m$$

Every edge flipping event is accepted if d decreases or remains constant. This corresponds to accepting every enhancement as well as neutral mutations.

The control panel of the tool is depicted in Figure 7.2.4.

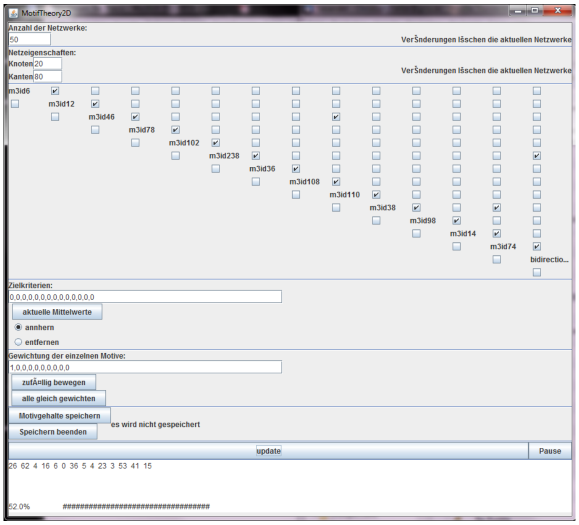
The motif projections selected in the controls are then shown in a new window, an example is shown in Figure 7.2.5. The individual scatter plots are automatically scaled and updated in real time as the network evolution goes on.

7.2.3 Interactive Motif Signature

When using the standard procedure, the calculation of motif z-scores is computationally expensive, as a whole ensemble of random graphs has to be created, each of them requiring a large number of randomization steps. This makes it virtually impossible to observe the inter-dependencies of the z-scores of individual motifs in an interactive way.

$$Z_m = \frac{c_m - \mu_m}{\sigma_m}$$

When constraining the analysis to graphs with a fixed degree sequence a much faster approach is viable. Even in a graph with fixed degree sequence large variations of the motif content are possible. On the other hand, the ensemble of random graphs used for z-score calculations is the same for every graph of the same degree sequence. This fact can be exploited by pre-computing the large random ensemble and only manipulating few graphs of the same degree sequence. The random ensemble is then fixed, σ_m and μ_m do not change during the simulation and the c_m of the few graphs that are under manipulation can be processed very



MotifTheory2D

Anzahl der Netzwerke:
50
VerŠnderungen lšschen die aktuellen Netzwerke

Netzeigenschaften:
Knoten 20
Kanten 80
VerŠnderungen lšschen die aktuellen Netzwerke

m3id6 ☑ □ □ □ □ □ □ □ □ □ □ □ □
□ m3id12 ☑ □ □ □ □ □ □ □ □ □ □ □
□ m3id46 ☑ □ □ □ □ ☑ □ □ □ □ □
□ m3id78 ☑ □ □ □ □ □ □ □ □ □
□ m3id102 ☑ □ □ □ □ □ □ □ □
□ m3id238 ☑ □ □ □ □ □ □ ☑
□ m3id36 ☑ □ □ □ □ □ □
□ m3id108 ☑ □ □ □ □ □
□ m3id110 ☑ □ □ □ □
□ m3id38 ☑ □ ☑ □
□ m3id98 ☑ □ □
□ m3id14 ☑ □
□ m3id74 ☑
□ bidirectio...
□

Zielkriterien:
0,0,0,0,0,0,0,0,0,0,0,0,0,0
aktuelle Mittelwerte
● annhern
○ entfernen

Gewichtung der einzelnen Motive:
1,0,0,0,0,0,0,0,0,0
zufÃ¤llig bewegen
alle gleich gewichten

Motivgehalte speichern
Speichern beenden
es wird nicht gespeichert

update
Pause

26 62 4 16 6 0 36 5 4 23 3 53 41 15

52.0% ##################################

Figure 7.2.4: The controls for the interactive motif dependency analysis. Random networks can be generated, the motif projections can be selected and the two vectors directing the evolution can be set. Additionally the motif-trajectories can be recorded for later analysis and the current acceptation rate of edge-flips is shown.

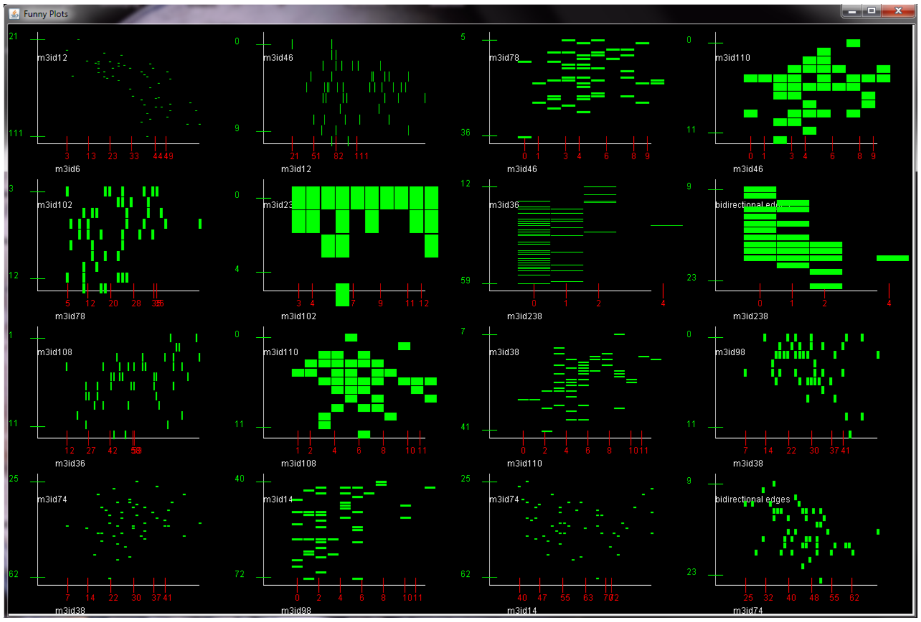
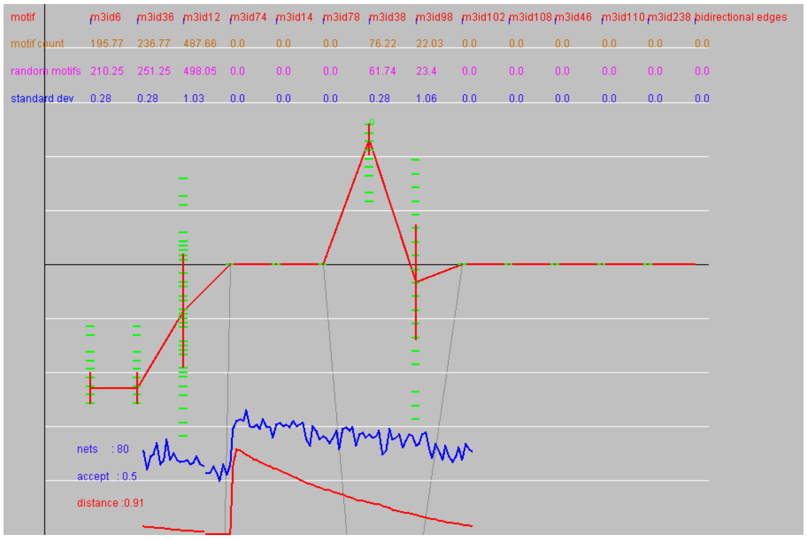


Figure 7.2.5: The main view of the interactive motif dependency analysis.



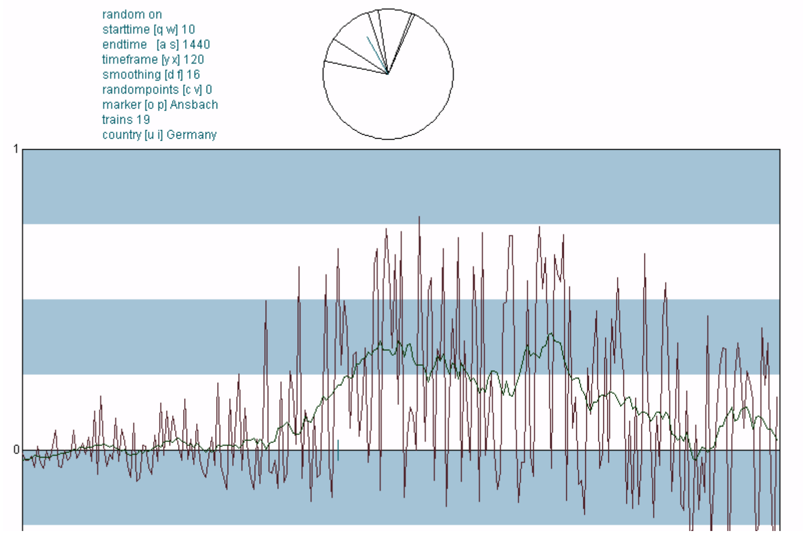
motif	m3id6	m3id36	m3id12	m3id74	m3id14	m3id78	m3id38	m3id98	m3id102	m3id108	m3id46	m3id110	m3id238	bidirectional edges
motif count	195.77	236.77	487.66	0.0	0.0	0.0	76.22	22.03	0.0	0.0	0.0	0.0	0.0	0.0
random motifs	210.25	251.25	498.05	0.0	0.0	0.0	61.74	23.4	0.0	0.0	0.0	0.0	0.0	0.0
standard dev	0.28	0.28	1.03	0.0	0.0	0.0	0.28	1.06	0.0	0.0	0.0	0.0	0.0	0.0

nets : 80

accept : 0.5

distance :0.91

Figure 7.2.6: A screen shot of the interactive z-score demonstration. The z-score is shown in real time, the driven graphs (in this case 80 graphs of the same degree sequence are used) are very sparse and motif id 6 is selected to go to a z-score of -2.5. These settings yield a motif signature very similar to super family 1 in [107]. The blue line on the bottom shows the acceptance ratio of the mutation steps and the red line shows the distance of the current average z-score to the selected goal.



```
random on
starttime [q w] 10
endtime   [a s] 1440
timeframe [y x] 120
smoothing [d f] 16
randompoints [c v] 0
marker [o p] Ansbach
trains 19
country [u i] Germany
```

Figure 7.2.7: A screen shot of the train synchronization analysis tool. Germany is selected, the null model subtraction is switched on and the smoothing (a moving average) is set to 16 stations.

quickly. The user interface of an implementation of this method can be seen in Figure 7.2.6. By clicking onto the panel a goal point can be selected, first performing the z-score calculation backwards (to obtain the necessary motif count for the desired z-score, $c_m = Z_m \sigma_m + \mu_m$). Then the adaptive walk is started similarly to the method described in the previous section. The current motif count, the average motif count and the standard deviation of the motif count in the random ensemble are shown for every subgraph. The standard deviation of every motif z-score is shown as error bars and every one of the graphs is shown by a small horizontal green line.

To asses the evolution, the value of the goal function and the acceptance rate are shown at the bottom of the screen together with their temporal development in red and blue, respectively.

7.2.4 Interactive Train Synchronization Visualization

To perform the analysis presented in Chapter 4, a specialized tool was implemented.

It reads in arrival/departure event data obtained from MOTIS (for details see Chapter 4). The synchronization index for every station is computed, a null model is subtracted and the data is smoothed in real time.

All parameters can be interactively varied. A marker is implemented, one can select a station and the wrapped-up arrival/departure events are shown together with the synchronization index. Additionally the country that should be analyzed can be selected. The stations are place along the x-axis and sorted along their size. The synchronization index is shown on the y-axis. Figure 7.2.7 contains a screen shot of the user interface.

Chapter 8

Summary and Future Work

This work has contributed to the field of network and motif analysis in several ways:

1. By the theoretical and phenomenological description of motif signatures that appear as artifacts of too coarse-grained null models.
2. By the introduction of advanced null models for several systems, e.g. module-aware mixing in Chapter 3, local z-scores in Chapter 6, the station-size aware expected synchronization indexes for train departure times and a model of scientific co-operations for Chapter 5.
3. By applying motif analysis to new network-based systems, namely co-author networks and artificial social networks.
4. By investigating synchronization in a network-based logistical system, namely long distance passenger train connections.
5. By extending the set of tools available, especially by different counting schemes in Section 3.5, the application of synchronization measures on not strictly periodical, time-discrete systems, the analysis of weighted motifs in Chapter 5 and characterizing the predictive strength of motifs in Chapter 6.
6. By the introduction of an analytical theory that enables us to predict subgraph counts, subgraph fluctuations and finally motif signatures for a number of situations (e.g. modular networks in Chapter 3 and networks with a distorted distribution of smaller subgraphs in Chapter 6) for directed as well as for undirected graphs.

Chapter 2 introduced an analytical theory of motifs, that can be used to predict subgraph counts. This is achieved by combining the probabilities for every edge of a subgraph and properly accounting for its symmetries. Additionally, fluctuations of subgraph counts can be predicted. By grouping subgraphs into *templates* we can disentangle different reasons for subgraph count fluctuations. The counts and fluctuations of templates are mostly related to the density of the network. The fluctuations inside of the templates (i.e. the distribution of templates into subgraph bins) are anti-correlated as the subgraphs belonging to one template compete for the same resources. Together, the predictions of the subgraph counts and their fluctuations can be used to predict motif signatures, also called triad significance profiles (TSPs). This can help to better understand the impact of some large-scale topological properties on motif signatures.

There are two logical next steps:

i) investigate the impact of small scale properties (i.e. a node's degree) on motif signatures. This is probably possible when assuming statistical independence between the degrees of adjacent nodes. It is known that this assumption of independence is violated in the case of broad degree distributions [89]. Also in many practical situations it is known that degrees are not uncorrelated [116].

ii) explore the intrinsic correlations of motifs, i.e., the reactions of the other motif counts when one motif is artificially enhanced or reduced, keeping all other graph properties random. This research could, by analyzing the systematic motif count shifts occurring during the randomization procedure, yield a deeper understanding of the amount of independent information contained in the superfamilies described in [107] (e.g. Figure 3.4.2 contains only one bit of information, namely modularity instead of 13 independent bits of information).

Chapter 3 discusses some issues that can arise when performing advanced motif analyses. Especially the problem of finding appropriate null models is addressed. For the case where strong global features like modularity are not considered in the null model, analytical predictions for the resulting error are given. This is an important step towards disentangling real local motif deviations from deviations that result from larger-scale features. Still the next goal remains to find appropriate null models for such complex situations. A general language describing the topological constraints that compose a reasonable null model is most probably not possible, for example because of overlapping modules, but an important source for motif deviations can be summarized as density fluctuations. This ranges from very small-grained fluctuations that are not much larger than the considered motifs up to a network consisting of few modules. The effect has two origins: i) The non-linear scaling of density with the network size. ii) The scaling of motif counts with the node degree that is fundamentally different for different classes of motifs. For the case of a known modular structure a randomization method that preserves modules and their connection pattern could be implemented. There are several community detection algorithms that could be used for this (e.g. [122, 37, 149]). The necessary input to the null model is which node belongs to which module. Then a constrained randomization could be applied, only flipping two links when they are a) both inside the same module (when all nodes adjacent to the selected links are part of the same module A) or b) if they both connect the same two modules, (i.e. both links originate in module A and point into module B). Every algorithm that attempts to partition a graph will need some method to decide whether further partitioning should be performed. When looking for an "optimal" partitioning this can, for example, be the increase in *modularity,* a quantity defined as "the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random" [120].

Such attempts may in some cases be useful, but here they are not really necessary. Most algorithms work by iteratively removing edges and observing the connected components. As every edge removal step can at most create one more component also the number of modules can only increase by one. So during one run of the algorithm N increasingly fine partitionings for a graph with N nodes can be obtained, containing $1...N$ modules. For every one of these partitionings randomization can be performed and z-scores can be obtained.

Two limiting cases can easily be understood: i) when the network is partitioned in only one module the randomization is not constrained, the result is exactly the same as with the standard randomization scheme. ii) when the network is partitioned in N modules the randomization is totally constrained, every network in the random ensemble is identical to the original network. This will then yield a zero z-score.

If a network has no relevant motif over- or under-representations but only a modular structure, when refining the partitioning the z-score of every motif should quickly approach zero. Opposed to that a network with true, distributed motif anomalies would show no such behavior with motif z-scores remaining far from zero even for relatively high partition counts. How to deal with overlapping modules (like they can for example be observed in social networks) remains an open question.

Another topic addressed in this chapter is the counting scheme for motif occurrences. There are two different ways to count:

i) the standard method: How many occurrences of a subgraph are present in a network. This number can be very high, in the extreme case at very high density $N(N-1)(N-2)$ subgraphs can be identified in a network with N nodes.

ii) the "single counting" method: How many nodes take part in a subgraph at least once? This number is bounded by the number of nodes in the graph. Especially when counting local motifs or when connecting dynamics with topology the results of the two methods are very different. Method ii) shows a much weaker reactions to the subgraph multiplicities, that is the scaling of subgraph counts with the local node degree.

Chapter 4 has shown interesting connections between robustness and synchronization both in a real system (the long distance train connections of Germany) and in a simple model of delay propagation. The robustness was extracted from a fully realistic model of delay propagation. The synchronization is extracted similarly to the synchronization index in [83]. These findings could be related to topological features of the underlying network, especially the "size" of a station (as a proxy the number of departures per day is used).

The next step would be to look into local network properties that take the neighborhood of a station into account, namely network motifs. The challenge is to find a way to disentangle the apparent motif effect, that arises from the fast scaling of local motif counts with the degree of a node from the real, functional motif effect. Two paths of investigation seem possible:

i) building appropriate null models for train networks, hereby the challenge lies in properly taking into account all the boundary conditions like station and track capacities as well as scheduling and passenger flow constraints. Then one could compare the local motif signatures in the "real" network to those in the artificial network.

ii) developing a method for computing self-consistent local z-scores without recurring to a null model or a set of "motif-blind" networks. For example the expected numbers of subgraphs in a small neighborhood around a single node could be directly computed taking into account the connectivities of all connected nodes. Here a deeper understanding of the motif scaling laws and the interplay of different motifs is necessary. A starting point for this could be the motif theory presented in Chapter 2.

Chapter 5 deals with success in social networks. The social structure is observed by the proxy of co-authorships and success is measured by the number of citations a publication gets. A dependence of the success on the local motif structure could be shown. Especially the box-motif seems to be strongly associated with success. This is probably the case because it marks locations in the network that show some kind of separation, which could be temporal, spacial or on the level of scientific domains / communities. An artificial "toy model" of scientific collaborations is implemented where every topic is located on a 2D topic-plane (see Appendix A.1.7). Some features of the real networks can be reproduced, especially the high edge weight of the box motif. It was also attempted to understand the effect of temporal separation by relating the time it needs to form a motif to its later success. Here further investigation is necessary.

Chapter 6 also deals with success in social networks, but follows another approach. We analyzed how a model of social dynamics shapes the topology of the underlying network. An over-representation of triangles has been observed before [140], but we show that additionally a non-trivial motif-signal exists for four-node-motifs, that can not be explained only based on the deviation of the three-node motifs. The gained insights should be used to look for similar patterns in real social network data.

Additionally it was possible to predict the future success of a social agent based on the local topology alone. Here different motifs show different predictive strength, with some motifs being beneficial to social success and others diminishing it. Is social success also predictable in real systems? Here the large quantities of dynamic (in the sense of time-resolved) data produced by online communities may be the key to a deeper understanding of social processes.

The clear advantage of the box motif over other motifs that was found in Chapter 5 can not be observed in Chapter 6. One reason for this is the homogeneity of the information flow in the social model analyzed. There is no difference in cost between connecting to an agent that is “far” or “close” to the connecting agent. Because of this the network structure is very homogenous. In an advanced model with refined connection costs, an agent that “bridges” the gap between two communities would be rewarded as he would have access to a larger information base, therefore have more “recent” information and because of this collect more and more links.

Over all this work has applied motif analysis to a set of network problems, in every case going further than pure application. We have questioned and extended the existing methods for two very different reasons. On the one hand to adapt the methods to account for the special conditions that are different for every system. On the other hand we advanced the understanding of the underlying systematics inherent to every single network problem.

Bibliography

[1] R. Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957, Nov. 2005.

[2] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, July 2000.

[3] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[4] M. Ángeles Serrano and M. Boguñá. Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E*, 72(3):036133, Sep 2005.

[5] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Physical Review Letters*, 96(11):114102, Mar. 2006.

[6] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone. Comment on "network motifs: Simple building blocks of complex networks" and "superfamilies of evolved and designed networks". *Science*, 305(5687):1107c–, 2004.

[7] V. Avetisov, S. Nechaev, and A. Shkarin. On the motif distribution in random block-hierarchical networks. *Physica A: Statistical Mechanics and its Applications*, 389(24):5895–5902, Dec. 2010.

[8] P. Bak. *How Nature Works: The Science of Self-Organised Criticality*. Copernicus Press, New York, 1996.

[9] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.*, 59(4):381–384, Jul 1987.

[10] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[11] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, Aug. 2002.

[12] A. Barabási, E. Ravasz, and T. Vicsek. Deterministic scale-free networks. *Physica A: Statistical Mechanics and its Applications*, 299(3-4):559 – 564, 2001.

[13] G. Basler, O. Ebenhöh, J. Selbig, and Z. Nikoloski. Mass-balanced randomization of metabolic networks. *Bioinformatics*, 27(10):1397 –1403, May 2011.

[14] V. Batagelj and A. Mrvar. Pajek-program for large network analysis. *Connections*, 21(2):47–57, 1998.

[15] S. Battiston, D. Delli Gatti, and M. Gallegati. Trade credit networks and systemic risk. In D. Helbing, editor, *Managing Complexity: Insights, Concepts, Applications*, pages 219–239. Springer, 2008.

[16] S. Battiston, D. Delli Gatti, M. Gallegati, B. Greenwald, and J. Stiglitz. Credit chains and bankruptcy propagation in production networks. *Journal of Economic Dynamics and Control*, 31(6):2061–2084, 2007.

[17] A. Berger and M. Müller-Hannemann. Uniform sampling of digraphs with a fixed degree sequence. In *Graph Theoretic Concepts in Computer Science*, volume 6410 of *Lecture Notes in Computer Science*, pages 220–231. Springer Berlin / Heidelberg, 2010.

[18] I. Bezáková, N. Bhatnagar, and E. Vigoda. Sampling binary contingency tables with a greedy start. *Random Structures & Algorithms*, 30(1-2):168–205, 2007.

[19] E. Birmele. Detecting local network motifs. *ArXiv e-prints*, July 2010.

[20] J. Bollen, H. van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, L. Balakireva, and A. Ruttenberg. Clickstream data yields high-resolution maps of science. *PLoS ONE*, 4:4803, Mar. 2009.

[21] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *arXiv:cond-mat/0502147*, 2005.

[22] K. Börner, J. Maru, and R. Goldstone. The simultaneous evolution of author and paper networks. *PNAS*, 101(Suppl 1):5266, 2004.

[23] S. Bornholdt. Less is more in modeling large genetic networks. *Science*, 310(5747):449–451, 2005.

[24] O. Brandman, J. E. Ferrell, R. Li, and T. Meyer. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science*, 310:496–498, Oct. 2005.

[25] O. Brandman and T. Meyer. Feedback loops shape cellular signals in space and time. *Science*, 322:390–, Oct. 2008.

[26] E. Brockfeld, R. Barlovic, A. Schadschneider, and M. Schreckenberg. Optimizing traffic lights in a cellular automaton model for city traffic. *Phys. Rev. E*, 64(5):056132, 2001.

[27] U. Brose, A. Ostling, K. Harrison, and N. Martinez. Unified spatial scaling of species and their trophic interactions. *Science*, 292:1525–1528, 2001.

[28] J. Buhl, J. Gautrais, R. Solé, P. Kuntz, S. Valverde, J. Deneubourg, and G. Theraulaz. Efficiency and robustness in ant networks of galleries. *European Physical Journal B*, 42(1):123–129, 2004.

[29] R. S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA, 1992.

[30] L. Buzna, K. Peters, and D. Helbing. Modelling the dynamics of disaster spreading in networks. *Physica A: Statistical Mechanics and its Applications*, 363(1):132–140, 2006.

[31] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz. Are randomly grown graphs really random? *Phys. Rev. E*, 64(4):041902, 2001.

[32] S. Chan, R. Donner, and S. Lämmer. Urban road networks–spatial networks with universal geometric features? *European Physical Journal B*, pages 1–15.

[33] X. Chang, Z. Wang, P. Hao, Y. Li, and Y. Li. Exploring mitochondrial evolution and metabolism organization principles by comparative analysis of metabolic networks. *Genomics*, 95(6):339–344, 2010.

[34] Y. Chen, G. Paul, S. Havlin, F. Liljeros, and H. E. Stanley. Finding a better immunization strategy. *Phys. Rev. Lett.*, 101(5):058701, Jul 2008.

[35] N. Christakis and J. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370, 2007.

[36] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Phys. Rev. E*, 65(6):066122, June 2002.

[37] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72(2):027104, 2005.

[38] H. Ebel, L. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66(3):035103, 2002.

[39] Y. Eom, S. Lee, and H. Jeong. Exploring local structural organization of metabolic networks using subgraph patterns. *Journal of Theoretical Biology*, 2006.

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

[41] J. Fowler, J. Settle, and N. Christakis. Correlated genotypes in friendship networks. *PNAS*, 108(5):1993, 2011.

[42] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, vol. 1:215–239, 1978.

[43] C. Fretter, M. E. Beber, M. Müller-Hannemann, and M.-T. Hütt. Artifacts in statistical analyses of network motifs. *submitted*, 2011.

[44] C. Fretter, M.-T. Hütt, and K. Sneppen. Motifs as markers of future success in a model of social dynamics. *submitted*, 2011.

[45] C. Fretter, L. Krumov, K. Weihe, M. Müller-Hannemann, and M.-T. Hütt. Phase synchronization in railway timetables. *Eur. Phys. J. B*, 77:281–289, 2010.

[46] C. Fretter, M. Müller-Hannemann, and M.-T. Hütt. Subgraph fluctuations in random graphs. *Phys. Rev. E, submitted*, 2011.

[47] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[48] R. Ginoza and A. Mugler. Network motifs come in sets: Correlations in the randomization process. *Phys. Rev. E*, 82(1):011921, July 2010.

[49] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99:7821–7826, June 2002.

[50] K.-I. Goh, E. Oh, B. Kahng, and D. Kim. Betweenness centrality correlation in social networks. *Phys. Rev. E*, 67(017101):017101, 2003.

[51] S. Goyal and F. Vega-Redondo. Structural hole in social networks. *Journal of Economic Theory*, 137(1):460–492, 2007.

[52] M. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[53] M. Granovetter. *Getting a job: A study of Contacts and Careers*. The University of Chicago Press, Chicago, 2 edition, 1995.

[54] R. Guimera, S. Mossa, A. Turtschi, and L. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *PNAS*, 102(22):7794, 2005.

[55] R. Guimerà, M. Sales-Pardo, and L. A. Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70(2):25101, 2004.

[56] R. Guimerà, M. Sales-Pardo, and L. A. Amaral. Module identification in bipartite and directed networks. *Phys. Rev. E*, 76(3):036102, Sept. 2007.

[57] R. Guimerà, B. Uzzi, J. Spiro, and L. A. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308:697–702, Apr. 2005.

[58] M. J. Herrgård, M. W. Covert, and B. O. Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Research*, 13(11):2423–2434, 2003.

[59] P. Holme and J. Zhao. Exploring the assortativity-clustering space of a network's degree sequence. *Phys. Rev. E*, 75(4):046111, apr 2007.

[60] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature genetics*, 31(4):370–378, 2002.

[61] P. Ingram, M. Stumpf, and J. Stark. Network motifs: structure does not determine function. *BMC Genomics*, 7(1):108, 2006.

[62] A. Inzelt, A. Schubert, and M. Schubert. Incremental citation impact due to international co-authorship in hungarian higher education institutions. *Scientometrics*, 78(1):37–43, 2009.

[63] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon. Subgraphs in random networks. *Phys. Rev. E*, 68(2):026127, Aug 2003.

[64] S. Jain, M. E. Beber, and M.-T. Hütt. Motif signatures of metabolic networks: The null model problem. *in preparation*, 2011.

[65] A. Jamakovic, P. Mahadevan, A. Vahdat, M. Boguna, and D. Krioukov. How small are building blocks of complex networks. *ArXiv e-prints*, Aug. 2009.

[66] H. J. Jensen. *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems*. Cambridge University Press, 1998.

[67] H. Jeong. Analysis of e. coli network. In S. Y. Lee, editor, *Systems Biology and Biotechnology of Escherichia coli*, pages 113–132. Springer Netherlands, 2009.

[68] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct. 2000.

[69] S. Jung, S. Kim, and B. Kahng. Geometric fractal growth model for scale-free networks. *Phys. Rev. E*, 65(5):056101, Apr. 2002.

[70] P. Kaluza, M. Ipsen, M. Vingron, and A. Mikhailov. Design and statistical properties of robust functional networks: A model study of biological signal transduction. *Phys. Rev. E*, 75(1):015101, 2007.

[71] P. Kaluza and A. S. Mikhailov. Evolutionary design of functional networks robust against noise. *Europhysics Letters*, 79(4):48001, 2007.

[72] P. Kaluza, M. Vingron, and A. S. Mikhailov. Self-correcting networks: Function, robustness, and motif distributions in biological signal processing. *Chaos*, 18(2):026113, June 2008.

[73] S. Kaplan, A. Bren, E. Dekel, and U. Alon. The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Mol Syst Biol*, 4, July 2008.

[74] B. Karrer and M. E. J. Newman. Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E*, 82(6):066118, Dec 2010.

[75] P. Kharchenko, G. M. Church, and D. Vitkup. Expression dynamics of a cellular metabolic network. *Mol Syst Biol*, 1, 2005.

[76] M. Kittisopikul and G. M. Süel. Biological role of noise encoded in a genetic network motif. *PNAS*, 107(30):13300 –13305, July 2010.

[77] K. Klemm and S. Bornholdt. Topology of biological networks and reliability of information processing. *PNAS*, 102:18414–18419, 2005.

[78] A. Konagurthu and A. Lesk. On the origin of distribution patterns of motifs in biological networks. *BMC Systems Biology*, 2(1):73, 2008.

[79] L. Kroon, R. Dekker, and M. Vromans. Cyclic railway timetabling: a stochastic optimisation approach. In F. Geraets, L. Kroon, A. Schöbel, D. Wagner, and C. Zaroliagis, editors, *Algorithmic Methods in Railway Optimization*, volume 4359 of *Lecture Notes in Computer Science*, pages 41–66. Springer, 2007.

[80] L. Kroon, D. Huisman, E. Abbink, P. Fioole, M. Fischetti, G. Maróti, A. Schrijver, A. Steenbeek, and R. Ybema. The new Dutch timetable: The OR revolution. *Interfaces*, 39(1):6–17, 2009.

[81] L. Krumov, C. Fretter, M. Müller-Hannemann, K. Weihe, and M.-T. Hütt. Motifs in co-authorship networks and their relation to the impact of scientific publications. *European Physical Journal B*, pages 1–6, 2011. 10.1140/epjb/e2011-10746-5.

[82] T. Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1996.

[83] Y. Kuramoto. *Chemical Oscillations, Waves, and Turbulence*. Springer, Berlin, 1984.

[84] Y.-K. Kwon and K.-H. Cho. Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. *Bioinformatics*, 24:987–994, 2008.

[85] S. Lämmer and D. Helbing. Self-control of traffic lights and vehicle flows in urban road networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(04):P04019 (34pp), 2008.

[86] S. Lämmer, H. Kori, K. Peters, and D. Helbing. Decentralised control of material or traffic flows in networks using phase-synchronisation. *Physica A: Statistical Mechanics and its Applications*, 363(1):39 – 47, 2006.

[87] P. O. Larsen and M. von Ins. Lotka's law, co-authorship and interdisciplinary publishing. *WIS*, 2008.

[88] M. Latapy, C. Magnien, and N. D. Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31 – 48, 2008.

[89] J. Lee, K. Goh, B. Kahng, and D. Kim. Intrinsic degree-correlations in the static model of scale-free networks. *European Physical Journal B*, 49(2):231–238, 2006.

[90] C. Liebchen. *Periodic timetable optimization in public transport*. PhD thesis, Technische Universität Berlin, 2006.

[91] C. Liebchen. The first optimized railway timetable in practice. *Transportation Science*, 42:420–435, 2008.

[92] C. Liebchen, M. Schachtebeck, A. Schöbel, S. Stiller, and A. Prigge. Computing delay resistant railway timetables. *Computers and Operations Research*, 37:857–868, 2010.

[93] C. Liebchen and S. Stiller. Delay resistant timetabling. *Public Transport*, 1:55–72, 2009.

[94] B. Lietaer, U. Ulanowicz, and S. Goerner. Options for managing a systemic bank crisis. *SAPIENS*, 2:1–15, 2009.

[95] J. Lorenz, S. Battiston, and F. Schweitzer. Systemic risk in a unifying framework for cascading processes on networks. *European Physical Journal B*, 71:441–460, 2009.

[96] N. M. Luscombe, M. Madan Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, Sept. 2004.

[97] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 100(21):11980–11985, 2003.

[98] S. C. Manrubia, A. S. Mikhailov, and D. H. Zannette. *Emergence of Dynamical Order, Synchronization Phenomena in Complex Systems*. World Scientific, 2004.

[99] C. Marr, M. Geertz, M. Hütt, and G. Muskhelishvili. Dissecting the logical types of network control in gene expression profiles. *BMC Systems Biology*, 2(1):18, 2008.

[100] C. Marr and M.-T. Hütt. Outer-totalistic cellular automata on graphs. *Physics Letters A*, 373:546–549, Jan. 2009.

[101] C. Marr, F. J. Theis, L. S. Liebovitch, and M.-T. Hütt. Patterns of subnet usage reveal distinct scales of regulation in the transcriptional regulatory network of escherichia coli. *PLoS Comput Biol*, 6(7):e1000836, 07 2010.

[102] S. Maslov. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, May 2002.

[103] A. Mazurie, S. Bottani, and M. Vergassola. An evolutionary and functional assessment of regulatory network motifs. *Genome Biology*, 6(4):R35, 2005.

[104] B. McKay and N. Wormald. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11(1):52–67, 1990.

[105] T. Milenković, J. Lai, and N. Pržulj. Graphcrunch: a tool for large network analyses. *BMC bioinformatics*, 9(1):70, 2008.

[106] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[107] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[108] R. Milo, N. Kashtan, S. Itzkovitz, M. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *Arxiv preprint cond-mat/0312028*, 2003.

[109] 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, Oct. 2002.

[110] R. Montañez, M. A. Medina, R. V. Solé, and C. Rodríguez-Caso. When metabolism meets topology: Reconciling metabolite and reaction networks. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 32(3):246–256, Mar. 2010. PMID: 20127701.

[111] M. Müller-Linow, C. Hilgetag, and M. Hütt. Organization of excitable dynamics in hierarchical biological networks. *PLoS Comput. Biol*, 4(9):e1000190, 2008.

[112] M. Müller-Hannemann and M. Schnee. Efficient timetable information in the presence of delays. In R. K. Ahuja, R. H. Möhring, and C. D. Zaroliagis, editors, *Robust and Online Large-Scale Optimization: Models and Techniques for Transportation Systems*, volume 5868, pages 249–272. Springer, 2009.

[113] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(2):025102, Aug. 2001.

[114] M. E. J. Newman. From the cover: The structure of scientific collaboration networks. *PNAS*, 98:404–409, Jan. 2001.

[115] M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.

[116] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Oct. 2002.

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

[118] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101:5200–5205, Jan. 2004.

[119] M. E. J. Newman. *Complex Networks*, volume 650/2004, pages 337–370. Springer Berlin / Heidelberg, 2004.

[120] M. E. J. Newman. From the cover: Modularity and community structure in networks. *PNAS*, 103:8577–8582, June 2006.

[121] M. E. J. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.

[122] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb. 2004.

[123] S. Omidi, F. Schreiber, and A. Masoudi-Nejad. Moda: An efficient algorithm for network motif discovery in biological networks. *Genes & genetic systems*, 84(5):385–395, 2009.

[124] R. Onody and P. de Castro. Complex network study of brazilian soccer players. *Phys. Rev. E*, 70(3):037103, 2004.

[125] S. B. Otto, B. C. Rall, and U. Brose. Allometric degree distributions facilitate food-web stability. *Nature*, 2007.

[126] R. Paine. Food web complexity and species diversity. *The American Naturalist*, 100(910):65–75, 1966.

[127] J. Park and M. Newman. Origin of degree correlations in the internet and other networks. *Physical Review E*, 68(2):026112, 2003.

[128] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, Apr. 2001.

[129] R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge Univ Pr, 2004.

[130] L. Peeters. *Cyclic railway timetable optimization*. PhD thesis, Erasmus University Rotterdam, Rotterdam School of Management, The Netherlands, 2003.

[131] S. Pigolotti, S. Krishna, and M. H. Jensen. Oscillation patterns in negative feedback loops. *PNAS*, 104:6533–6537, Apr. 2007.

[132] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, 2003.

[133] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras. Self-organization of collaboration networks. *Phys. Rev. E*, 70(3):036106, Sept. 2004.

[134] A. Rao, R. Jana, and S. Bandyopadhyay. A markov chain monte carlo method for generating random (0, 1)-matrices with given marginals. *Sankhyā: The Indian Journal of Statistics, Series A*, 58(2):225–242, 1996.

[135] E. Ravasz and A. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67(2):026112, Feb. 2003.

[136] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[137] S. Redner. How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B*, 4:131–134, Aug. 1998.

[138] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105:1118–1123, Jan. 2008.

[139] M. Rosvall, A. Grönlund, P. Minnhagen, and K. Sneppen. Searchability of networks. *Phys. Rev. E*, 72(4):046117, Oct 2005.

[140] M. Rosvall and K. Sneppen. Modeling self-organization of communication and topology in social networks. *Phys. Rev. E*, 74(1):016108, Jul 2006.

[141] M. Rosvall and K. Sneppen. Reinforced communication and social navigation generate groups in model networks. *Phys. Rev. E*, 79(2):026111, Feb 2009.

[142] J. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. Berriz, F. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[143] A. Samal, S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC bioinformatics*, 7(1):118, 2006.

[144] M. Schnee. *Fully Realistic Multi-Criteria Timetable Information*. PhD thesis, Technische Universität Darmstadt, Germany, 2009.

[145] F. Schreiber and H. Schwöbbermeyer. Mavisto: a tool for the exploration of network motifs. *Bioinformatics*, 21(17):3572–3574, 2005.

[146] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White. Economic networks: The new challenges. *Science*, 325(5939):422–425, 2009.

[147] P. Serafini and W. Ukovich. A mathematical model for periodic event scheduling problems. *SIAM Journal on Discrete Mathematics*, 2:550–581, 1989.

[148] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31(1):64–68, 2002.

[149] S. Smyth. A spectral clustering approach to finding communities in graphs. In *Proceedings of the fifth SIAM international conference on data mining*, volume 119, page 274. Society for Industrial Mathematics, 2005.

[150] N. Sonnenschein, M. Geertz, G. Muskhelishvili, and M.-T. Hütt. Analog regulation of metabolic demand. *BMC Systems Biology*, 5(1):40, 2011.

[151] N. Sonnenschein, M.-T. Hütt, H. Stoyan, and D. Stoyan. Ranges of control in the transcriptional regulation of escherichia coli. *BMC Systems Biology*, 3(1):119, 2009.

[152] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[153] N. Team. Network workbench tool. Northeastern University, and University of Michigan, 2006.

[154] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen. Hierarchy measures in complex networks. *Phys. Rev. Lett.*, 92(17):178702, Apr 2004.

[155] U. Ulanowicz, S. Goerner, B. Lietaer, and R. Gomez. Quantifying sustainability: Resilience, efficiency and the return of information theory. *Ecological Complexity*, 6:27–36, 2009.

[156] A. F. J. Vanraan. Fractal dimension of co-citations. *Nature*, 347:626, Oct. 1990.

[157] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *Phys. Rev. E*, 65(6):066130, June 2002.

[158] T. Velden and C. Lagoze. Patterns of collaboration in co-authorship networks in chemistry - mesoscopic analysis and interpretation. *ISSI 2009*, 2009.

[159] A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, Jul 2009.

[160] 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. *PNAS*, 101(52):17940–17945, 2004.

[161] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proceedings: Biological Sciences*, 268(1478):1803–1810, 2001.

[162] J. Walleczek, editor. *Self-organized biological dynamics and nonlinear control*. Cambridge University Press, 2000.

[163] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[164] S. Weber and M. Porto. Generation of arbitrarily two-point-correlated random networks. *Phys. Rev. E*, 76(4):046111, 2007.

[165] S. Wernicke. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, Nov 2005.

[166] R. Williams and N. Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–183, 2000.

[167] A. T. Winfree. *The Geometry of Biological Time*. Springer, New York, 1980.

[168] S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316:1036–1038, May 2007.

[169] W. Zhou and L. Nakhleh. Properties of metabolic graphs: Biological organization or representation artifacts? *BMC Bioinformatics*, 12(1):132, 2011.

Appendix A

Supplementary Material

This Appendix contains supplementary materials to Chapter 5 that provides detailed insight into the analyzed data, performed analysis and supporting experiments.

It has already been published as supplementary material in the paper [81], co-authored with Lachezar Krumov, Karsten Weihe, Matthias Müller-Hannemann and Marc-Thorsten Hütt.

A.1 Co-Authorship Networks

Since co-authorship networks include temporal data (via the year of publication), they allow for the retrospective observation of dynamically growing networks. In [113], evidence for preferential attachment and a degree distribution approximating a power law were found in a large range of node degrees (up to 150 cumulated co-authors in physics and up to 600 in biomedicine). Moreover, it is observed that both phenomena start vanishing at roughly the same number of authors. In addition to an extensive empirical analysis of a variety of topological measures varying over time, [11] incorporates the fact that edges are also inserted between old nodes, not only between a new and an old node, respectively (thus deviating from the formal preferential attachment protocol). They discover that the probability of inserting an edge is roughly proportional to the product of the node degrees. They also give an explanation why, nonetheless, a power law distribution of the node degree is observed. In [22], a generative model is introduced, which explains deviations from the power law distribution of citations by typical phenomena of scientific collaboration such as distinct topics and the partitioning of scientific disciplines into subdisciplines.

Central, influential authors are analyzed in [50], where the influence of an author is measured in terms of the betweenness. Basically, it is shown that influential authors do not collaborate more often with each other than average authors.

An alternative way of looking at patterns in knowledge production in a network-like fashion is to analyze the citation of a publication in another work. Such citation networks have for example been studied in [21]. The work of [21] goes one step beyond the bipartite graph representation of authors and publications of [133] and equips the graph with additional directed edges between publications to indicate citation. On this basis, they discuss four measures for the performance of individual authors: the betweenness measure and three observables based on the number of publications and citations. The discrepancies between the four top-ten lists reveal that the measures evaluate different qualities of the authors. Moreover, they introduce a new variant of entropy to measure how uniformly the impact of an author is divided among his/her co-authors, and a strong tendency towards few high-impact and many low-impact collaborative efforts was found.

A.1.1 Publication Data

We have investigated two publication databases: CiteSeerX and DBLP.

Dumps of the DBLP database in XML format are provided on regular bases by the DBLP official website: http://dblp.uni-trier.de. Each publication entry provided in the XML dump contains at least the publication title, the publication year and the list of authors who have co-authored the corresponding publication. The dump investigated in this work is as of May 2008, which contains 599,734 authors and 978,786 publications.

The CiteSeerX is available for download and synchronization through the CiteSeerX official website: http://citeseerx.ist.psu.edu. The data download and synchronization is available through Open Archive Initiative Harvesters. We developed our own harvester to acquire the publication data available in CiteSeerX as of October 2009. Each publication entry contains at least the publication title, the publication year and the list of authors. That snapshot of the online database contains 999,856 authors and 1,247,732 publications.

A.1.2 Citation Indices

In our work we investigate the *success* of collaboration patterns. We project the success of a given publication as the number of citations by other publications. That is, successful innovative and ground breaking publications attract interest by other scientists, who then later on refer to those publication in their own work.

Hence, the next step for our analyses was to acquire citation indices for the the publications within the two investigated publication databases. For this purpose we deployed 107 web crawlers compatible with the online publication search engines CiteSeerX and GoogleScholar (http://scholar.google.com). All acquired data is publicly available through web interfaces of both search engines. The web crawlers obeyed the time out policies and request frequencies provided by the search engines and ran as background processes, being even less intrusive than a human user.

We requested the title of each publication within the two acquired databases and stored the responses by both search engines (the responses are provided with citation indices by other publications). A response is considered a match, if the title (by trimming white spaces and special characters), the publication year and the list of authors (by trimming white spaces and special characters) were identical to a publication within one of the acquired databases.

The title of each publication was requested on both search engines. If both of them returned a match, then the citation index for that publication was set to the maximum of both responses.

We were able to acquire non-empty citation indices for 192,688 of the papers within the DBLP database, which is around 19% of all publications. That number excludes publications which were found on the search engines, but still have not been cited by other papers. For the CiteSeerX databases we found 434,794 papers with citation index of at least one, which corresponds to 34% of all publication within the acquired database.

The lower match success by DBLP comes from the fact that it is actually a third party with respect to the search engines. On the other side, we requested the publications provided by CiteSeerX by using the Open Archive Initiative (OAI) protocol directly from the CiteSeerX search engine, leading to a better match ratio.

Note that the citation indices considered are as of their time of acquisition, which for both databases was short after acquiring the publication data.

A.1.3 Co-Authorship Graph Representation

We use the natural graph representation of co-authorship networks where the authors are the nodes and two nodes are connected if they have ever published together.

We parsed the publication lists in both databases with the following assumptions:

- A publication is considered unique through its title (trimmed from white spaces) and publication year. Publications without specified publication year are expelled from our analysis.
- Multiple publication entries with different publication years, but the same title and authors, are considered as distinct publications.
- An author is considered unique through her/his first and family names.
- Authors with identical first and family names (as they appear in the database) are considered the same author.

The above assumption may lead to considering two real world authors as the same author in our database if they have the same names. On the other side, if an author uses different signatures on her/his publications, one real world author may be considered as two distinct authors in our database. There is no way around this problem and its impact was already investigated by related work cited, see Chapter 5. The number of authors and publications within both databases presented in the first section are the result of the above assumptions.

Furthermore, both databases contain entries representing online reports and websites, listed with several hundred authors. To clean up the databases from such entries, we excluded from our analysis all publications with more than 8 authors. For DBLP those were less than 0.6% of all publications and 1.7% for CiteSeerX.

As we were interested in citation frequencies and their interplay with topology, we also excluded all publications with none or zero citations, i.e. publications that are not in the databases or have yet no citations respectively. Thus, our analyses was performed on 190,893 of all 978,786 publication entries within the DBLP snapshot and 430,233 of all 1,247,732 publication entries within CiteSeerX.

After acquiring both databases and available citation indices, we build a graph representation of each database based on the publication entries with citation index of at least one and less than 9 authors. Each distinct author is represented by a node and two nodes are connected if they have ever coauthored a publication.

A.1.4 Motif Analysis

We projected the citation indices as edge weights in four different ways and counted the average link weight per motif. A motif is considered an induced connected subgraph, i.e. any four nodes

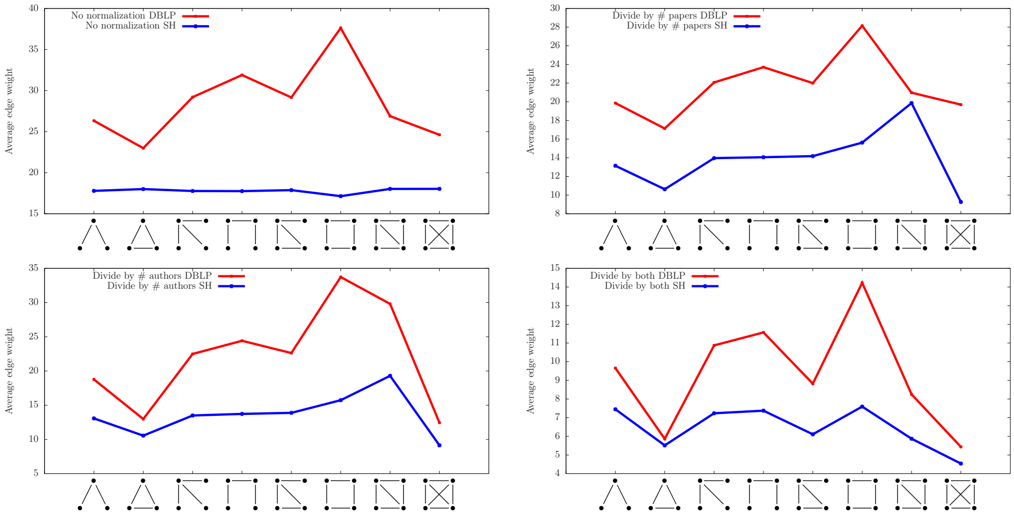


Figure A.1.1: The average link weight per motif for all four edge weight definitions compared to the null model, denoted by SH, for DBLP.

can form only one of the six four-node undirected motifs. The weight of a motif is the sum of the weights of its edges. The average link weight per motif is the sum of the weights of all instances of a particular motif divided by the number of instances and then divided by the number of edges within that particular motif. Hence, the average link weight per motif is comparable across the different three- and four-node undirected motifs.

The average link weight per motif for both databases, DBLP and CiteSeerX, and for all four normalization schemes are shown in Figures A.1.1 and A.1.2.

Furthermore, each co-authorship has an year of appearance, namely the year their work was published. Naturally, we define the creation time of an edge as the year of the first publication among a pair of authors. Intuitively, motifs are build up by edges and each edge has a creation time, hence a motif also has a creation time. We define it as the difference of the creation year of its first and its latest edge.

A.1.5 Supporting Experiments

To assure that our both databases comply with already investigated co-authorship networks, we computed a set of network properties usually discussed in related work. These include degree distribution, citation distribution, clustering coefficient, papers per author and authors per paper. None of the computed network measures show deviation from already published results on other collaboration databases, see Figure A.1.3 and Table A.1.

All four distributions follow the power-law form already investigated in many other co-authorship networks.

To assure that the average link weight per motif presented are long term results, we retrospectively investigated their evolution over time for the DBLP database. For that purpose 17 snapshots of the database we created, one for each of the years between 1990 and 2007

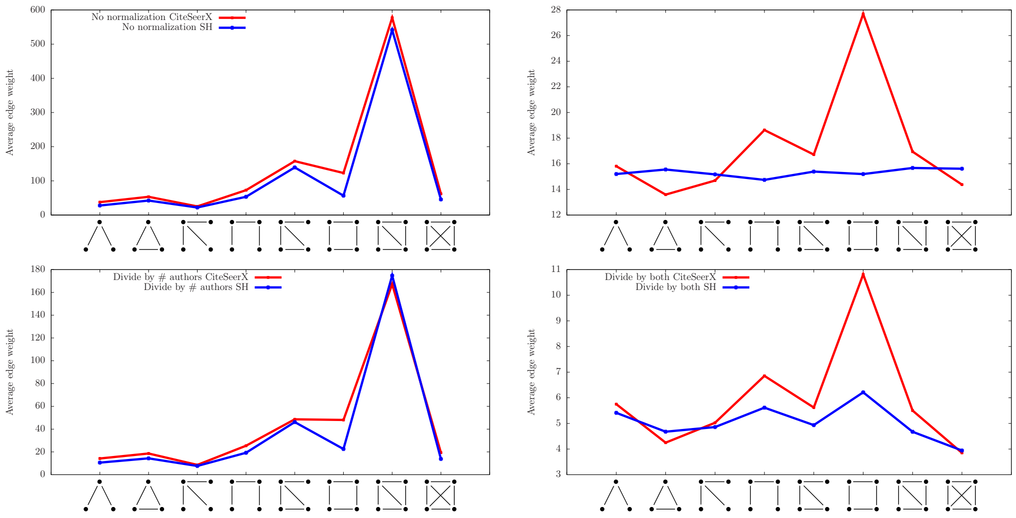


Figure A.1.2: The average link weight per motif for all four edge weight definitions compared to the null model, denoted by SH, for CiteSeerX.

Network	Authors per Paper	Papers per Author	Clustering Coefficient
DBLP	2.74	4.04	0.658
CiteSeerX	2.69	3.26	0.667

Table A.1: Average authors per paper, papers per author and clustering coefficients for the DBLP and CiteSeerX database. All values comply with results from related work.

respectively. Each snapshot contains only publications published prior or within the year of the snapshot. We then computed the average weight per motif link for definition 3. One observes that the box motif prevails over all snapshots, see Figure A.1.4.

The motif weights change slowly and only slightly over the years. Their values drop for more recent years as the number of *fresh* publications with none or few citations increases. Note that the citation indices of the publications for all snapshots are as of 2008, because of the lack of information which citation in which year was acquired.

Furthermore, we investigated the whole motif weight distributions instead of just looking at their average values. All eight distributions are monotone and governed by the box motif as can be seen from Figure A.1.5. The motif weights were computed over the whole database and with respect to edge weight definition 3.

It is obvious from the commulative distributions that the average values over the motif weights are well defined and justified. Our next step was to investigate how they change when one consistently disregards the heaviest box motif instances when computing the mean values. Again we investigated the whole DBLP database under edge weight definition 3 and take as a reference motif 4, see Figure A.1.6.

One observes that the average motif link weight reduces gradually for the box motif as well as the reference motif 4. Hence, the high average value of the box motif is not a result of a few

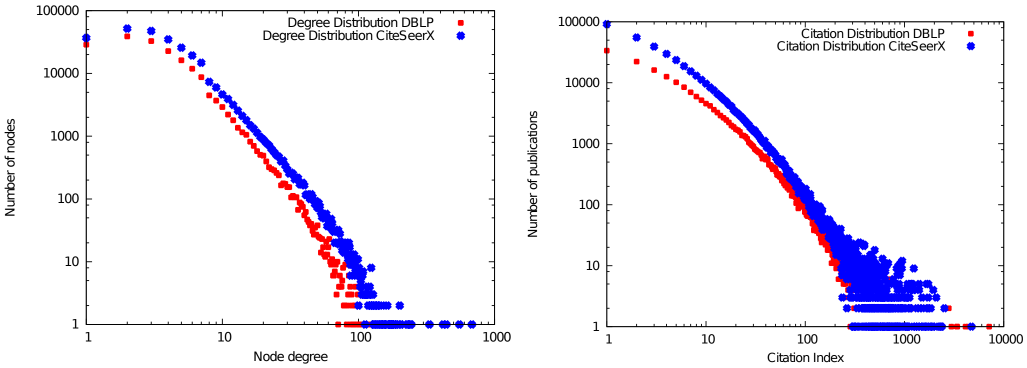


Figure A.1.3: Degree and citation distributions of DBLP and CiteSeerX.

extremely heavy instances, but rather constitutes a high number of relatively heavy instances.

Up to now we have shown that our two databases comply with related work on co-authorship networks, that the results are stable over the years and that computed mean values are justified and are not influenced by a few extreme values. To conclude, we want to tackle one last question, namely the number of papers and the number of their authors that constitute the box motifs.

Note that the four edge weight definitions implicitly address that issue, as they integrate the number of papers between a pair of authors, the number of co-authors on those publication, or both effects simultaneously. Otherwise, one can assume that the high average value of the box motif comes from one of those two effects. Recall, that independently of the edge weight definition, the box motif was still the most *successful* one. To exclude any doubt, we have calculated the average number of publications between a pair of authors in all motifs, as well as the number of co-authors on those publications. The results are displayed in Figure A.1.7. One clearly sees that the box motif neither profits from high number of papers running through its edges, nor those publications have significantly few authors. Thus, its prevailing weight is not a result of any trivial effects one could suspect.

To conclude, in this supplementary work we carried out a set of sanity checks of the analyzed data, as well as a deeper look on the presented results. We observed that the properties of the investigated co-authorship networks comply with related work. Furthermore, we showed that the presented results are well defined and justified, as well as that they do not come from certain trivial effects.

A.1.6 Application Areas

One clear result of the statistical analysis presented so far is that the box motif is a functionally interesting – and so far not discussed – building block of complex network. Beyond co-authorship networks, we believe that there are several areas of application, where box motifs may contribute similarly significantly to function. Table A.2 lists some of those areas.

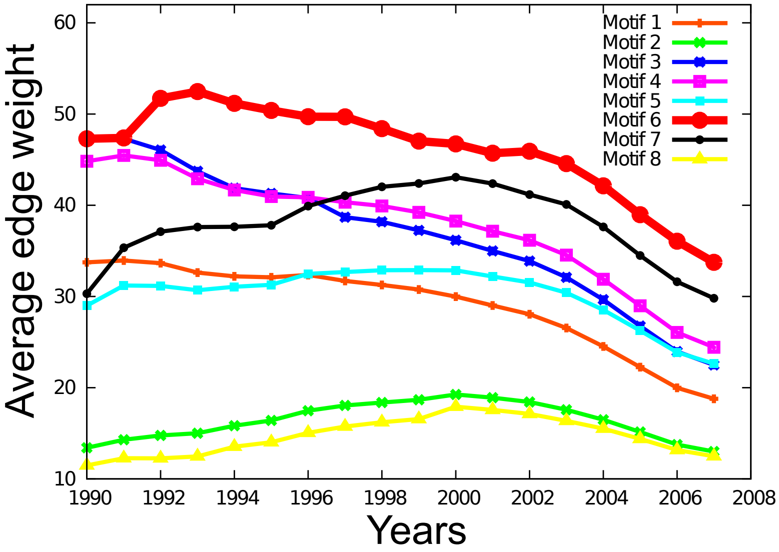


Figure A.1.4: The average weight per motif link over the years for the DBLP database.

Network Type	Dynamical Observable	Potential Box Motif Role
Acquaintance networks	Gossip	Sites with maximal re-organization
Metabolic networks	Metabolic fluxes	New category of enzyme essentiality
Trust networks	Recommendations	Double reassuring of reliability
Peer-to-Peer	Data exchange	Alternative paths to target peer
Train Connections	Passenger flow	Alternative connections to destination
P2P Live Streaming	Video/Music/TV on demand	Concurrent frame exchange
Routing	Package delivery	Bandwidth separation along routing paths

Table A.2: Expected applications of the box motif in diverse technological and social networks.

A.1.7 Generative Model

In this section we briefly sketch a possible generative model, which may serve as a convenient framework for exploring the relation between impact of a publication and topological properties of the co-authorship network as a function of the underlying elementary processes.

We assume the authors and publications to be distributed on a plane (the “content proximity plane"). The two elementary processes in the model are the writing of a scientific publication (paper production) and the citing of already existing publications in new ones (citing articles). In the case of paper production, the content proximity plane is used with a probability α_1 to select authors from. With probability $(1-\alpha_1)$ authors are selected at random for the publication at hand. In the case, when the plane is used, another parameter, β_1, regulates, whether authors are selected according to impact or proximity.

For the second process, citing articles, the parameters α_2 and β_2 have the same function for selecting publications to be cited in the publication at hand, as their counterparts have in the

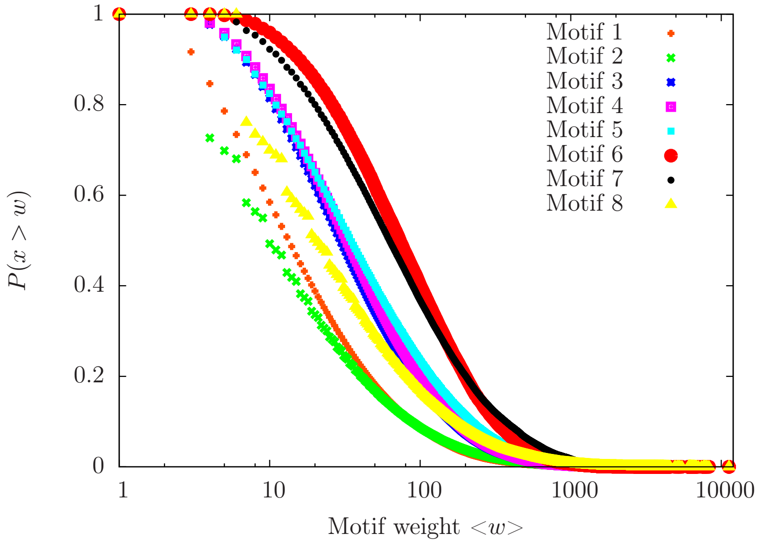


Figure A.1.5: The motif weight distributions for all eight motifs within the DBLP database.

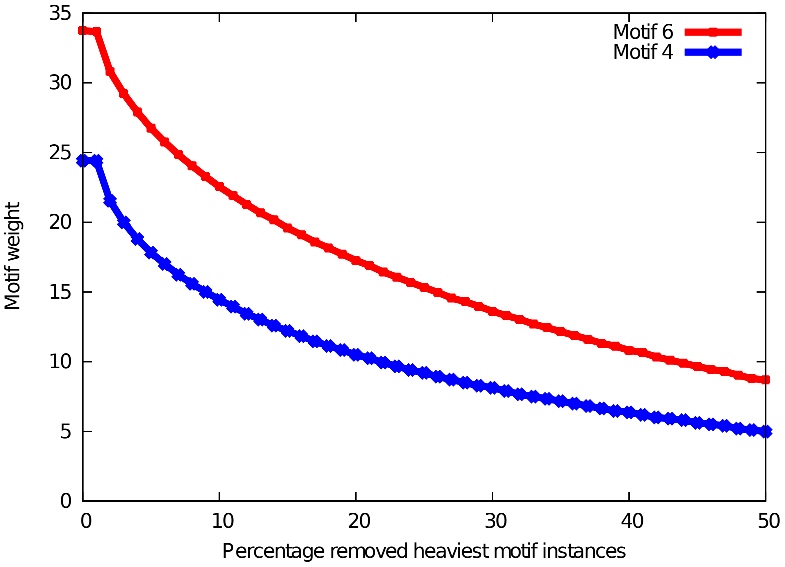


Figure A.1.6: The effect on the average motif link weight when one gradually removes the heaviest instances of that motif.

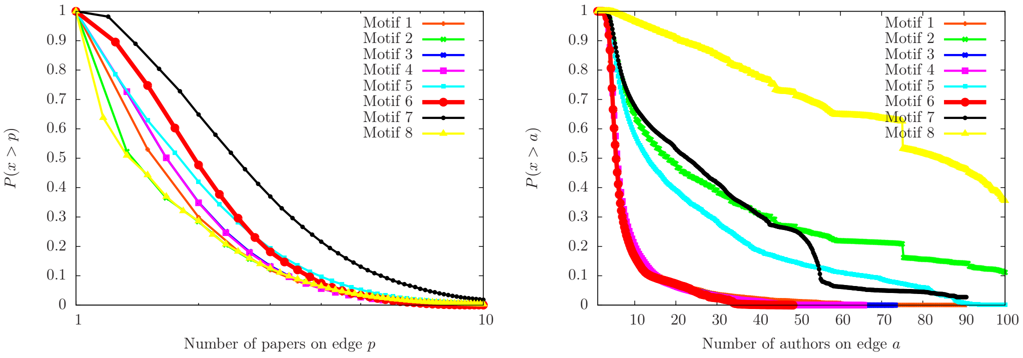


Figure A.1.7: The number of papers respectively co-authors per motif edge for the DBLP database.

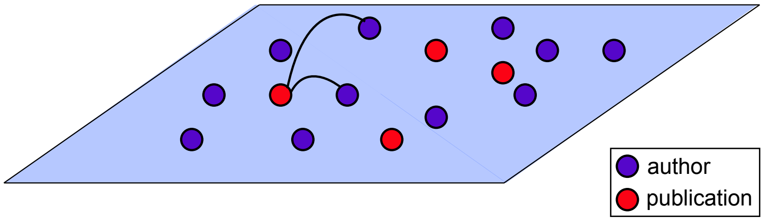


Figure A.1.8: Schematic representation of the content proximity plane. Distance among authors and publications enters the computation of the scores, eqs. (1) and (2).

case of selecting authors.

To simulate the two processes, paper production and citing papers, the number of authors N and the number of publications M has to be selected, as well as two distributions: authors per paper and citations per paper. Furthermore, to reflect the process of aging we introduce another parameter A as the maximal number of publications of an author.

Then, the workflow of the generative model is as follows: Choose the number of authors N and place them in the content proximity plane. Choose the number of publications M. For each publication choose the number of its authors, k, and place the publication into the content proximity plane, as well. Then, choose k authors from the plane according to their proximity and impact, and *publish* the paper. Finally, choose the number l of existing publications the new publication should cite and choose those publications similarly according to their proximity and impact. This process is illustrated in Figure A.1.8.

For a given publication p, we compute a score for each author in the plane. Then, we choose the k authors who should *write* p from the distribution of all author scores. The score of an author a is given by:

$$Score(a) := \alpha_1 (Rank(a) + 1)^{\beta_1} e^{-\frac{\Delta_{ap}}{2}} + (1 - \alpha_1)\frac{1}{N} \tag{A.1.1}$$

where Δ_{ap} is the Euclidian distance between a and p in the proximity plane and $Rank(a)$ is the number of citations of all publications already published by a. In other words, α_1 balances between selecting authors according to their impact (large β_1) or their proximity in the plane to p (small β_1), and between random assignment of authors to papers.

After an author a has published her/his first publication, it stays in the proximity plane for the next A publications. Afterwards, the author is marked retired and taken down from the plane and thus from the list of available authors for further publications.

In analogy to the paper production process, we select the papers each new publication should cite from the distribution of all paper scores. The score of a paper p is given by:

$$Score(p) := \alpha_2(Rank(p)+1)^{\beta_2}e^{-\frac{\Delta_{pp_{new}}}{2}} + (1-\alpha_2)\frac{1}{M} \tag{A.1.2}$$

where $\Delta_{pp_{new}}$ is the Euclidian distance between p and the new publication p_{new}, and $Rank(p)$ is the number of citations of p. Hence, α_2 balances between citing papers according to their impact (large β_2) or their proximity in the plane (small β_2), and between random citation of papers.

Once all M papers have been *published*, we extract the collaboration network by connecting any two authors that have published together and assign the citation frequencies as edge weights according to definitions 1 through 4. Hence, our model produces weighted co-authorship networks.

Although our model naturally reflects the paper production and paper citation processes, it has a rather large (and heterogeneous) parameter space. One has to choose the lifetime of the authors A, all α_1, β_1, α_2 and β_2, as well as the distribution of authors per paper and citations per paper.

In order to check, whether the empirical findings can in principle be represented in this simple model, we take the DBLP snapshot from 1990 and approximate the network using simulated annealing with respect to degree distribution, citation distribution and motif content. We take the same number of authors and papers as the original network and the empirical distribution of authors per paper. The distribution of citation per paper cannot be reconstructed from our database. Therefore, each new paper in our model cites 10 already existing papers.

The co-authorship networks generated by our model allow us to repeat the motif analysis presented in the main paper. Therefore, we perform two different evolutions based on simulated annealing. In the first case we aim at the degree distribution, the citation distribution and motif content of the real world network. The objective function is composed of the differences with respect to those three measures between the real world and the generated networks. In the second case the objective function is augmented with another term, which minimizes the difference between the ratio of the weight of motif 4 (i.e. its average citation frequency) to the box motif in the real world and the generated network. The results of both evolutions are shown in Figures A.1.9, A.1.10, A.1.11 and A.1.12.

It is easy to observe that our model not only approximates the real world network very well with respect to its topological properties, but also it is capable of reconstructing the unexpected high edge weight of the box motif.

It is a trilling question to explore and determine the size and the form of the whole solution space. Nevertheless, the preliminary results of our generative model show that the right combination of simple network processes like aging, paper production, paper citation, as well as social factors like proximity and impact, can reproduce the success of the box motif, revealed in our analysis.

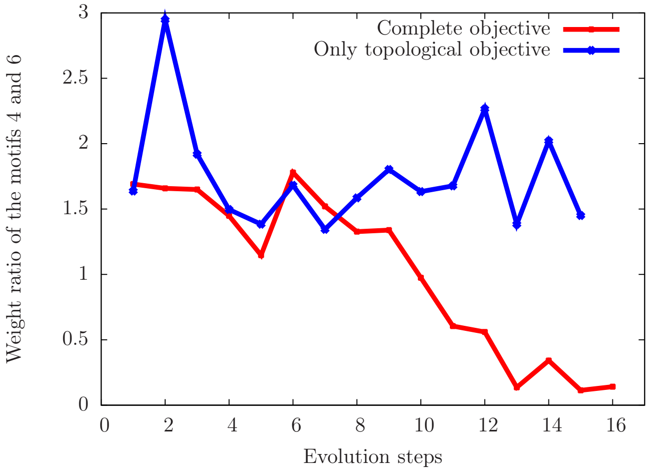


Figure A.1.9: Approximating the DBLP snapshot from 1990. Once with respect to degree distribution, citation distribution and motif content only, and once augmented with the ratio in weight of motif 4 to motif 6.

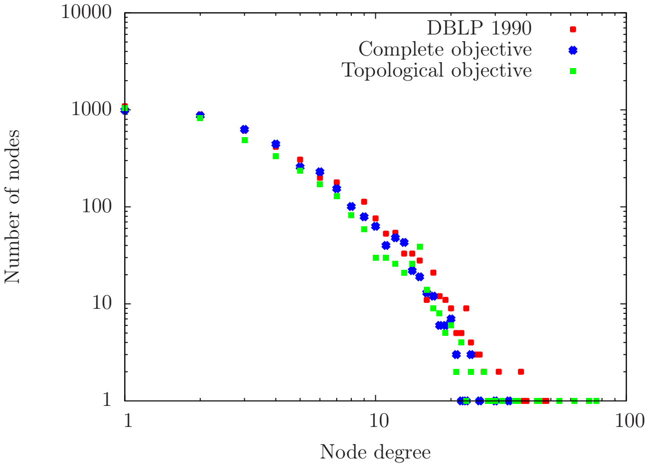


Figure A.1.10: Approximating the degree distribution of the DBLP snapshot from 1990. Once with respect to topological properties only and once augmented with the ratio in weight of motif 4 to motif 6.

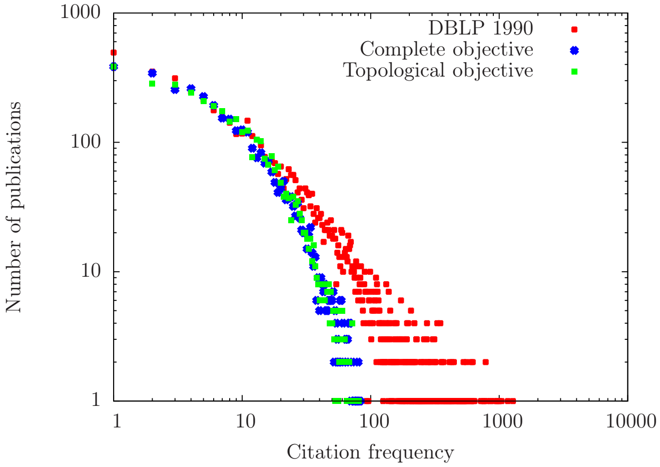


Figure A.1.11: Approximating the ciation distribution of the DBLP snapshot from 1990. Once with respect to topological properties only and once augmented with the ratio in weight of motif 4 to motif 6.

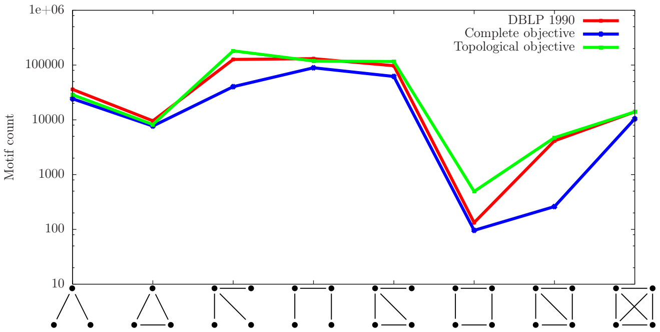


Figure A.1.12: Approximating the motif content of the DBLP snapshot from 1990. Once with respect to topological properties only and once augmented with the ratio in weight of motif 4 to motif 6.

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig und ohne fremde Hilfe verfasst habe.
Ich habe keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt.
Die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen sind als solche kenntlich gemacht worden.
Ich habe mich bisher nicht um den Doktorgrad beworben.

Bremen, den 22. Juli 2011

Christoph Fretter