Analysis of Different Random Graph Models in the Identification of Network Motifs in Complex Networks

Vom Fachbereich Informatik der Technischen Universität Kaiserslautern zur Verleihung des akademischen Grades Doktor der Naturwissenschaften (Dr. rer. nat.) genenehmigte Dissertation von

Wolfgang Eugen Schlauch

Datum der wissenschaftlichen Aussprache: 3. Dezember 2016

Dekan:	Prof. Dr. Klaus Schneider
Berichtserstatter:	Prof. Dr. Katharina Anna Zweig
	Prof. Dr. Clemence Mangien

D 386

CONTENTS

I INTRODUCTION 5

1 MOTIVATION 7
- 1.1 General Introduction to Network Analysis . . . 7
- 1.2 Which Model to Use . . . 9
- 1.3 Subgraphs . . . 9
- 1.4 The missing ingredient . . . 11

II DEFINITIONS 13

2 GRAPH MODELS 17
- 2.0.1 The Erős-Rényi model - the random graph . . . 18
- 2.1 Fixed Degree Sequence model . . . 18
  - 2.1.1 Havel-Hakimi-Algorithm . . . 19
  - 2.1.2 Configuration Model . . . 19
  - 2.1.3 Switching Algorithm . . . 21
  - 2.1.4 Sequential Importance Sampling . . . 23

3 MATHEMATICAL TOOLKIT 27
- 3.1 Statistical Analysis of Differences in Results . . . 28
- 3.2 Simple Independence Model . . . 30
- 3.3 Critique on the applicability of SIM to all graphs . . . 31
  - 3.3.1 Example 1 . . . 32
  - 3.3.2 Example 2 . . . 33

III COMPARISON OF THE DIFFERENT NULL MODELS BASED ON ARTIFICIAL AND REAL-WORLD UNDIRECTED GRAPHS 35

4 COMPARATIVE ANALYSIS BASED ON THE MODELS 39
- 4.1 Erdős-Rényi graph . . . 41
- 4.2 Forest-Fire graph . . . 43

5 SEQUENTIAL IMPORTANCE SAMPLING—WHICH PROBABILITY DISTRIBUTION TO USE? 45

6 ANALYSIS OF UNDIRECTED REAL-WORLD GRAPHS 47
- 6.1 Datasets . . . 47

7 STABLE MEASURES 51
- 7.1 Diameter . . . 52
  - 7.1.1 Approximating the Diameter . . . 52
  - 7.1.2 Model comparison . . . 53
- 7.2 Distance . . . 56
  - 7.2.1 Approximating the Distance . . . 56
  - 7.2.2 Model comparison . . . 57
- 7.3 Implications . . . 60
- 7.4 Multiple Edges and Self-Loops . . . 60

8 SENSITIVE MEASURES 67
8.1 Average Neighbor Degree . . . 67
8.1.1 Approximating the Average Neighbor Degree . . . 68
8.1.2 Model comparison . . . 68
8.2 Common Neighbors . . . 70
8.2.1 Approximating the Co-Occurrence . . . 71
8.2.2 Comparison of the models . . . 72
8.2.3 Local Co-Occurrences . . . 74
8.3 Implications . . . 74

IV COMPARISON OF THE DIFFERENT NULL MODELS BASED ON DIRECTED GRAPHS 79

9 ANALYSIS OF DIRECTED GRAPHS 83
9.1 A short history of motif analysis . . . 83
9.2 Data . . . 85
10 ON MOTIFS 87
11 DIFFERENT MODELS UNDER INVESTIGATION 91
11.1 On the Directed Configuration Model . . . 91
11.2 On the Sequential Importance Sampling . . . 99
11.3 A Faster Option to Calculate the Expected Number of Motifs . . . 105
11.4 Revisiting the Bifan . . . 120
11.4.1 Summary . . . 125
12 NODE-BASED PARTICIPATION ESTIMATION IN MOTIFS 127
12.1 Constructing Position-Based Equations . . . 128
12.2 On the Sum of Equations . . . 130
12.3 How do the Equations for Positions Fare? . . . 135
12.4 Revisiting the Bifan . . . 137
13 ON PREDICTING CO-PURCHASED ITEMS 141
13.0.1 Television Series Prediction . . . 143
13.1 The model to use and open problems . . . 152

V SUMMARY 155
14 SUMMARY AND CONCLUSIONS 157
14.1 Summary . . . 157
14.2 Conclusions to draw . . . 159
14.3 Future Work . . . 159

VI APPENDIX 169
COMPARISON WITH THE RESULTS OF ITZKOVITZ ET AL. 171
DIRECTED GRAPHS - TABLES 175
1 On the Configuration Model . . . 175
2 On the Sequential Importance Sampling Model . . . 183
3 On the Faster Option . . . 191
BIFAN EQUATION REVISITED - CONTINUED 195
PUBLICATIONS 199
1 Journal Articles . . . 199

2 Conferences 199
3 Other 199

LIST OF FIGURES

Figure 2.1 Possible two-edge swaps in an undirected graph. 21
Figure 2.2 "Unswappable" Digraph 22
Figure 2.3 Intermediate step in the graph generating process of the sequential importance sampling. 24
Figure 3.1 Kolmogorov-Smirnoff Test 29
Figure 3.2 Example 1 32
Figure 3.3 Two graphs generated from the same degree sequence with different modularity-scores. 33
Figure 4.1 Measures in unskewed graphs 42
Figure 4.2 Measures in skewed graphs 44
Figure 5.1 First comparison of the models based on assortativity 46
Figure 7.1 Example graph for which the CFG and the SIM yield many multiple edges and self-loops. 61
Figure 7.2 G(n, m) edgeloss example 63
Figure 7.3 Edgeloss in graphs with skewed degree distributions 64
Figure 7.4 Comparison of multiple edges attached to a high degree node 66
Figure 8.1 Example of a multigraph. The question is, how to calculate the average neighbor degree of v. 67
Figure 8.2 Co-occurrence vs. multigraph 71
Figure 8.3 Comparison of the number of neighbors 75
Figure 11.1 Edgeloss in directed graphs 92
Figure 11.3 Histogram of Feed-Forward Loop distribution in CFG, ECFG, and FDSM 98
Figure 11.4 Degree distributions of graphs 102
Figure 11.5 Average occurrence of edges in samples 115
Figure 11.6 Average occurrence of edges in samples, contd. 117
Figure 11.7 Histograms of the distribution of subgraphs in different graphs, compared with the estimated distribution via the SIM. 119
Figure 12.1 Positions in a Feed-Forward Loop 128
Figure 12.2 Relative error for Twopaths 133
Figure 12.3 Relative error for Feed-Forward Loops 134
Figure 12.4 Relative error based on position 136
Figure 12.5 Relative error of the Bifan 140
Figure 13.1 Correctly assessed edges 151

LIST OF TABLES

Table 6.1 Basic network statistics for the individual networks used in the article. 49
Table 7.1 Average diameter of CFG samples 54
Table 7.2 Average diameter of SIS samples 54
Table 7.3 Two-sample z-test results, the Kolmogorov-Smirnov two-sample test result and its p-value for the diameter of the graphs generated with the FDSM and the USIS. 55
Table 7.4 Average distance of CFG samples 57
Table 7.5 Two-sample z-test results, the Kolmogorov-Smirnov two-sample test result and its p-value for the average distance of the graphs generated with the FDSM and the CFG. 58
Table 7.6 Average distance of SIS samples 59
Table 7.7 Two-sample z-test results, the Kolmogorov-Smirnov two-sample test result and its p-value for the average distance of the graphs generated with the FDSM and the USIS. 59
Table 7.8 Average number of self-loops and multi-edges for samples from the CFG. The table also contains their expected value calculated with Equation (7.19), resp. Equation (7.20). 65
Table 8.1 Average neighbor degree in FDSM, CFG, and SIM 68
Table 8.2 Average neighbor degree in FDSM and SIS 69
Table 8.3 Two-sample z-score calculation in comparison with the FDSM to test whether results can be from the distribution indicated by the samples. 70
Table 8.4 Average co-occurrence in the different models 73
Table 8.5 z-score calculation with the graphs from the ECFG as samples and the number of cooccurrences in the FDSM as value to test whether it can be from the distribution indicated by the samples. 73
Table 11.1 Number of Forks found in the respective models. The standard deviation for the CFG is due to the fact that two edges between the same node do not yield a Fork. 93
Table 11.2 Number of Fans found in the respective models. 94
Table 11.3 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fork. 95
Table 11.4 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fan. 95
Table 11.5 Number of Twopaths found in the respective models. 96
Table 11.6 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Twopaths. 96
Table 11.7 Number of Feed-Forward Loops found in the respective models. . . 97
Table 11.8 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Feed-Forward Loop. 98
Table 11.9 Number of Twopaths found in the respective models. 100

Table 11.10 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Twopath. 101
Table 11.11 Percentage of nodes which violate the condition that the square of their out-, in-, or combined degree should be smaller than the sum of the respective degree sequence. 101
Table 11.12 Number of Feed-Forward Loops found in the respective models. . . 103
Table 11.13 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Feed-Forward Loop. 104
Table 11.14 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Out-Fan. 104
Table 11.15 Results of the two-sample z-score calculation between the different models without weights for the Out-fan motif. 105
Table 11.16 Number of Forks found in the respective models. 108
Table 11.17 Number of Fans found in the respective models. 110
Table 11.18 Equations developed based on the basic equation 11.1, following the approach of equation 11.3 and equation 11.9 111
Table 11.19 Number of Twopaths found in the respective models. 112
Table 11.20 Number of Feed-Forward Loops found in the respective models. . . 114
Table 11.21 z-score of Feed-Forward Loops comparing the SIM with the range of values of the FDSM, z-score of the mean of the FDSM compared to the SIM, and the two-sample z-statistic of the two models, FDSM and SIM. 114
Table 11.22 Percentage of nodes violating the condition that the square of their out-, in-, or combined degree should be smaller than the sum of the respective degree sequence. 116
Table 11.23 Number of Bifans found in the respective models. 117
Table 11.24 z-score of Bifans comparing the SIM with the range of values of the FDSM, z-score of the mean of the FDSM compared to the SIM, and the two-sample z-statistic of the two models. 120
Table 11.25 Number of Twopaths calculated with the simple equation and the slightly modified version. 121
Table 11.26 These graphs are counted additionally when the simple Bifan equation is used. Subtracting the corresponding equations yields more reasonable results. 122
Table 11.27 In the first column, the estimated number of Bifans using $\frac{\left(\langle k^2\rangle-\langle k\rangle\right)^2\left(\langle j^2\rangle-\langle j\rangle\right)^2}{4\langle k\rangle^4}$ is shown; the other columns contain the estimated number of subgraphs that do not occur in simple graphs but do contribute to the estimate in the first column. 123
Table 11.28 Change in the Bifan-equation when the principle of inclusion and exclusion is applied. 124
Table 11.29 In the first column, the estimated number of Feed-Forward Loops using $\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\langle kj\rangle\left(\langle j^2\rangle-\langle j\rangle\right)}{\langle k\rangle^3}$ is shown; the other columns contain the estimated number of subgraphs that do not occur in simple graphs but do contribute to estimate in the first column. 124

Table 12.1 Equations for the expected number of subgraphs containing node u, depending on the possible position; basic approach without consideration of the participation of the node itself in the mean. 128
Table 12.2 Equations for the expected number of subgraphs containing node u, depending on the possible position; corrected for node u by subtraction from the remaining average in the nominator. 129
Table 12.3 Relative error for the Twopath, E. coli 131
Table 12.4 Relative error for the Twopath, Ythan 131
Table 12.5 Relative error for the Feed-Forward Loop, E. coli 132
Table 12.6 Relative error for the Feed-Forward Loop, Ythan 135
Table 12.7 Equations regarding the Bifan subgraph where corrections for node u are made by subtracting its contribution from the averages. 137
Table 12.8 Relative error for the Bifan, E. coli 138
Table 12.9 Relative error for the Bifan, Ythan 138
Table 13.1 Hit-rate based on the new prediction method for the same television series as in Zweig and Kaufmann [108]. 143
Table 13.2 Average quality of recommendations based on the simple independence model for television series. Dashes in the last two columns indicate that no season of a series had all other seasons of the same series in the top 100 recommendations. 148
Table 13.3 Local PPV_k comparison 149
Table 13.4 Global PPV_k comparison 150
Table 1 Equations of Itzkovitz et al. [43] and own equations 11.18 together with the corresponding subgraph. Observe, that some do have the same equation. 172
Table 2 E. coli 173
Table 3 S. cerevisiae 173
Table 4 Little Rock 174
Table 1 Number of Double-Joins found in the respective models. 175
Table 2 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Double-Join. 175
Table 3 Number of Threecycles found in the respective models. 176
Table 4 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Threecycles. 176
Table 5 Number of Complete subgraphs found in the respective models. . . 177
Table 6 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Complete subgraph. 177
Table 7 Number of Fourcycles found in the respective models. 178
Table 8 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fourcycle. 178
Table 9 Number of Bifans found in the respective models. 179
Table 10 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Bifan. 179
Table 11 Number of Biparallel subgraphs found in the respective models. . . 180
Table 12 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Biparallel subgraph. 180

Table 13 Number of In-Fans found in the respective models. 181
Table 14 Results of the Kolmogorov-Smirnov two-sample test between the different models for the In-Fan. 181
Table 15 Number of Out-Fans found in the respective models. 182
Table 16 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Out-Fan. 182
Table 17 Number of Double-Joins found in the respective models. 183
Table 18 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Double-Join. 183
Table 19 Number of Threecycles found in the respective models. 184
Table 20 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Threecycle. 184
Table 21 Number of Complete subgraphs found in the respective models. . . 185
Table 22 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Complete subgraph. 185
Table 23 Number of Fourcycles found in the respective models. 186
Table 24 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fourcycle. 186
Table 25 Number of Bifans found in the respective models. 187
Table 26 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Bifan. 187
Table 27 Number of Biparallel subgraphs found in the respective models. . . 188
Table 28 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Biparallel subgraph. 188
Table 29 Number of In-Fans found in the respective models. 189
Table 30 Results of the Kolmogorov-Smirnov two-sample test between the different models for the In-Fan. 189
Table 31 Number of Out-Fans found in the respective models. 190
Table 32 Results of the Kolmogorov-Smirnov two-sample test between the different models for the Out-Fan. 190
Table 33 Number of Double-Joins found in the respective models. 191
Table 34 Number of Threecycles found in the respective models. 191
Table 35 Number of Complete subgraphs found in the respective models. . . 192
Table 36 Number of Fourcycles found in the respective models. 192
Table 37 Number of Bifans found in the respective models. 193
Table 38 Number of Biparallel subgraphs found in the respective models. . . 193
Table 39 Number of In-Fans found in the respective models. 194
Table 40 Number of Out-Fans found in the respective models. 194

ABSTRACT

This thesis is concerned with different null-models that are used in network analysis. Whenever it is of interest whether a real-world graph is exceptional regarding a particular measure, graphs from a null-model can be used to compare the real-world graph to. By analyzing an appropriate null-model, a researcher may find whether the results of the measure on the real-world graph is exceptional or not.

Deciding which null-model to use is hard and sometimes the difference between the null-models is not even considered. In this thesis, there are several results presented: First, based on simple global measures, undirected graphs are analyzed. The results for these measures indicates that it is not important which null-model is used, thus, the fastest algorithm of a null-model may be used. Next, local measures are investigated. The fastest algorithm proves to be the most complicated to analyze. The model includes multigraphs which do not meet the conditions of all the measures, thus, the measures themselves have to be altered to take care of multigraphs as well. After careful consideration, the conditions are met and the analysis shows, that the fastest is not always the best.

The same applies for directed graphs, as is shown in the last part. There, another more complex measure on graphs is introduced. I continue testing the applicability of several null-models; in the end, a set of equations proves to be fast and good enough as long as conditions regarding the degree sequence are met.

ACKNOWLEDGEMENT

I wanted to write a simple “Thank you”, instead of trying to list all who helped and supported me. I am sure I will forget somebody. Nevertheless, I wanted to thank explicitly my advisor Katharina Anna Zweig, since she made this possible. Emöke Ágnes Horvát, since she helped me so much with the first article. Darko Obradovic, he is the one who suggest me this position when I did not even know that there is someone working on graphs in Kaiserslautern. My parents and my sister for their constant support—I know that I do not show how much I appreciate it half as much as I do. My friends, especially but without particular order, Lisa Lenz, Jessica Weiss, Anne Bardens, Kai Wiedel, René Becker, Andreas Resch; you all showed confidence in me even when I was lacking trust in what I did. My colleagues and friends from university, Pia Achilles, Mareike Bockholt, Marsha Kleinbauer, Sude Tavassoli, Raphael Reitzig, Peter Treiber, Sebastian Wild—I will miss lunch with you, I will miss coffee with you, and I will miss seeing that I am not the only one who sometimes scratches his head and wonders what he is doing. A special thank you for Professor Nebel, not only for coffee, but also for the moral support.

In the end, I am sure that I forgot many people. I am sorry.

Part I

INTRODUCTION

1

MOTIVATION

1.1 GENERAL INTRODUCTION TO NETWORK ANALYSIS

Network analysis - about what is this person talking? That is one of the first questions that pops into someone's mind whenever they talk to somebody from the field. Even among scientist network analysis entails so many different things that it is not uncommon that two people talk about something and after some time they realize they speak about different things, even though the terminology is the same. Thus, a short explanation what I mean in this thesis is needed.

Network analysis is an ever growing field. Some of the earliest work that can be considered as (social) network analysis is from the early nineteenth century. As one of the most influential scientist in this field Jacob Levy Moreno has to be mentioned, who is considered one of the persecutors of the field. He was one of the first to use *sociograms*, i.e., depictions of social networks to display the group behavior of people [65]. A *social network* for these sociograms is a group of people, in which some of the persons interact with each other. Even earlier than Moreno, John C. Almack used questionnaires to gather information about the social relations of children at school. He did not only investigate the social connections they made. He correlated the IQs of the pairs of students. The result was that more relations existed between persons with similar IQ than dissimilar [1]. Similarly, Beth Wellman studied the social interactions of students in a high school and correlated found pairs on IQ, grades, and similar measurable scores to identify homophily between students as well [101].

These precursors of what is nowadays known as social network analysis covered only small groups, i.e., from school class-sized groups up to 150 people. The methods developed and ideas proposed by them are still applicable to a certain extent, some of the assumptions made still cannot be proven but are still considered as important [86, 33].

Scientist in graph theory noticed the trend of using sociograms, depictions as graphs. *Graph theory*, a field more or less initiated by Leonhard Euler, concerns itself with problems that can either be expressed as a graph or are based on graph structures. Examples of graph theoretic problems include the enumeration of graphs having particular properties, coloring problems (can any map in a plane be drawn using only four different colors), and path-based problems (finding the shortest path between two nodes). The two fields together are powerful, such that analysis of a sociogram, which can be expressed as a graph, based on the work of Kőnig [51], should be simple enough. Note that the term sociogram is not commonly used in graph theory; when speaking of a depiction or some relationship between a set of persons, network science usually uses the term (social) network.

Almost every first-world inhabitant will have had contact with any of the fields mentioned above. Most likely social networks touched the life in any way, since in January

2016 about 1.6 billion users were registered at Facebook[1] [23]. Other well known social networking sites are Renren (31 mio. users, China[2]) and vk.com (100 mio. users, Russland[3]), which are popular in some countries since the countries either restrict access to Facebook or their alphabet limits themselves to a somewhat closer economy. Other points of contact with graph theory are the use of route-planning software, or a decade ago the use of Google search, which was said to be based on the PageRank-algorithm. This algorithm calculated a weight for a website, based on ingoing and outgoing links, to decide which pages that match the searched term should be presented first. Many of these applications are used by the users without knowledge that algorithms from graph theory are applied to data.

Besides the enormous sizes social networks tend to have, what good does a value of $3.14 \cdot 10^{-2}$ in a measure if there is nothing to compare the result with. In *Social Network Analysis*, a subfield of network analysis that concerns itself only with social networks, one usually gathers relationships between entities, calculates some measures based on the perceived relationships, and interprets the result. This interpretation is either made with the help of an expert, for example, a sociologist, or with someone who knows the entities, for example, the head of a research and development department [20]. But there is also another way to determine whether the gathered data is exceptional with regards to the measure. Mathematics provide a solution for this, called *random graph theory*. This subfield of graph theory is based on the work of Gilbert [29], Erdős, and Rényi [24]. They gave the two most famous random graph models. In most theoretic works they are just called the random graph, or when names are used, the $\mathcal{G}(n, p)$ and the $\mathcal{G}(n, m)$ model. The random graph theory was a long time a quiet field, only in the last decades a rise of publications could be observed. With the (re-)discovery of something called "small-world"-effect or "6 degrees of separation" [61], a surge of random graph models occurred and subsequently the field was rejuvenated. These new models attempted to generate graphs that are in structure more similar to the networks observed in the real world; they attempt to model structure and behavior better, to improve results, observe anomalies, and other things. Besides the new models, many new ideas, theories and algorithms have been proposed and are in constant development.

Still, this leads to an interesting question: which graph sizes are used in graph theory and network analysis?

Physics and math tend to use comparatively small graphs while social network analysis has the option to use enormous datasets like the emails sent by a company (13k users), the players of a game (0.5bn users), or the users of Facebook (>1bn users). Most recently, scientists at Facebook discovered, that users on their social network were even closer than the famous "six degrees of separation" [93, 2]. They found that the average distance between 721 million active Facebook users (69 billion links) is only 4.74. The experiment was not run on the complete data set and Edunov et al. [23] estimate the current distance at 4.57 via approximative algorithms based on 1.6 billion users. To put the distance in a context, that is as if Tim, the neighbors son who is ten years old, writes a letter and after only four exchanges, the president of the United States of America has Tim's letter. Still, it is not known whether this is unique to Facebook or whether this is usual for a social network of this order and size. For this purpose, random graph theory can be used.

1 http://www.facebook.com
2 `http://www.renren.com`
3 `http://www.vk.com`

1.2 WHICH MODEL TO USE

Of course, random graph theory is a powerful tool to analyze whether social networks of a certain size do all show such small average distances. It can be done by generating a large enough sample of graphs, computing the average distances on these and comparing this average to the average distance of the real-world graph. But it is not entirely clear what model to use. When the field was young, only simple models existed that are not able to model necessary features of graphs such as the Facebook network. When using the $\mathcal{G}(n, m)$ model to generate a set of graphs similar in size to the Facebook graph, problems occur as soon as one analyzes them. A social network such as Facebook has completely different properties than the more simple $\mathcal{G}(n, m)$ model. The set of sampled graphs will have an average distance that most likely will not be similar to the one of Facebook, purely by these properties. Does this imply that Facebook is very special? Yes and no. Of course it is special when we consider only this model. As soon as other random graph models are considered, it may be different. There are many more random graph models that are used in network analysis and graph theory — how to chose one or a family of models such that there is a basis to compare the results of different random graph models to each other?

One of the best ideas is fixing a parameter that describes the graph the best; best, in this case, is defined by the researcher. In this thesis, a very simple but descriptive parameter of graphs is used: the degree sequence. The degree sequence of a graph describes the graph by recording how many connections a node has and noting it down in a list. In Part III nad Part IV, several models that keep this parameter constant are under investigation regarding some standard measures. Ordinarily, one would assume that models that are based on a single fixed parameter yield the same results. That has been investigated by Schlauch, Zweig, and Horvat [81]. The results presented in Part III are an extension of the work that has been done already.

Based on the result of the Facebook analysis, one may ask about the average distance per country. For the United States of America, the average distance is estimated at only 3.46[23]. That leads to next question that this thesis is concerned with: parts of larger graphs or *subgraphs*.

1.3 SUBGRAPHS

Questions in the field of graph theory are: how are subgraphs extracted, how often do subgraphs of a specific type occur, does a graph contain a certain subgraph and more. In the context of this thesis, the question of *network motif analysis* [62] is of interest. A *network motif* is a subgraph of small size, 3 to 4 nodes usually, that occurs in a graph statistical significantly more often than one would expect. Since graphs with a fixed degree sequence are of interest, this Part IV will be about which graph model are useful for network motif analysis. Do all models that are based on a fixed degree sequence yield the same average number of subgraphs? This question has not been explicitly asked before.

Milo [63] introduced the notion of network motif analysis by analyzing graphs in the context of the number of subgraphs the random graphs contain—in a way, the analysis contained the question which model is useful for the analysis, but the article was restricted to the question whether the models generate graphs uniformly at random or not. Explicitly speaking, the question was not "Do different random graph models yield the same

expected values" but "Do different random graph models generate graphs uniformly at random". In other articles, they used only one algorithm [62, 64] that showed satisfying results. This algorithm is explained more in Part II.

Now, it remains to discuss why one should care about subgraphs of such minuscule size, since looking for combinations of 3 to 4 nodes in a graph with many nodes is somewhat a strange idea. For example, a graph with 300 nodes has $\binom{300}{3} \simeq 4.5 \cdot 10^6$ possible combinations of nodes that could be checked for their connections pattern. Of course, no one will check all possible combinations but use smarter approaches. One of these approaches, while still simple, is only checking the neighborhood of a node, i.e., nodes that a node is connected to. Why should it be important if a graph contains 42 combinations instead of 10? That this is relevant can be shown with examples from biology.

In a transcriptional regulation network, the interactions of a transcription factor X regulate a second transcription factor Y, such that both jointly regulate an operon Z [84]. In terms of graph theory, node X has a directed connection to Y and another to Z, while Y has a directed connection to Z. These interactions could be considered as random. Since organisms had a long time to develop these patterns of interaction, they most likely are not random but exist for a purpose. Finding these patterns is challenging, but may help fighting diseases. Uhlmann et al. [94] employed network analytic methods in their research on breast cancer and were able to identify potential tumor suppressors. That was done via identification of co-upregulating and co-downregulating patterns in the graph consisting of proteins and miRNA (micro ribonucleic acid). In other words, two proteins had to be identified that had the same positive (up-regulating) or negative (down-regulating) influence to some miRNA. Via statistical analysis, based on random graph theory, it is thus possible to identify important patterns that may help fight serious diseases.

Subgraphs are considered as important in other fields as well. Communication networks, i.e., who talked to whom, how often, how long, at which times, can be considered as vital in some research areas. Graph theory can be applied to this information as well. Lindelauf et al. [57] used graph theory to analyze networks of a heroin distribution network and the cell-phone network of a terrorist group and showed that communication structures are not created by random but on purpose. Hindsight analysis is useful and sometimes able to discover other possible threats as well, but predicting who is dangerous is much harder. While some criminals are not the smartest [4], information on evil-doers are not as readily available as sometimes portrayed in movies. Observation of known criminals and gathering their contact networks to analyze is another useful tactic [53] that is used more frequently, even though the public rails against gathering data from public channels such as Facebook, and not so public channels like a cell phone.

It is important to note that not only the connection pattern is significant. It is also the frequency with that the pattern exists. The analysis of the frequency of a connection pattern usually consists of several steps:

1. Count how often a pattern of connections occurs in a graph;
2. Count in an ensemble of randomized graphs how often the pattern occurs;
3. Decide somehow whether the observed graph is different from the randomized graphs;

4 8 Dumb Criminals Caught Through Facebook `http://mashable.com/2012/12/12/crime-social-media/#Lm2kCY9J6aq2`

This process entails many different decisions that have to be made. The most significant choice is the ensemble of randomized graphs: How to chose the correct algorithm that generates this ensemble, or which ensemble shall be used to compare to? First models that observed graphs were compared to had either a fixed structure (lattice graphs, star graphs [57]) or were based on very simple probabilities [25, 100]. The set of possible models evolved from these simple models to be able to capture not only the size but also the behavior of large graphs. Nowadays, graph generating algorithms can model behaviors such as "the rich get richer", also called preferential attachment, certain degree requirements, communities, and other properties.

1.4 THE MISSING INGREDIENT

There are many different algorithms to generate graphs. They are put forward since researchers are missing something that they need for their research, and a new algorithm can achieve what they need. Consider the example mentioned above of the $\mathcal{G}(n, m)$ model, that generated graphs in a very simple way; researchers needed graphs that were more similar to real-world graphs in terms of the degree sequence.

What is missing from a field that is putting new ideas forth? A comparison of algorithms that claim to achieve the same, but with different effort. Comparisons regarding ease of understanding or estimated run-time are available, but a statistical evaluation of the different algorithms is not available, to the best of my knowledge. That is done in this thesis. There are several algorithms that are often used when the requirement is that the algorithms generate graphs with a given fixed degree sequence. Some researchers do not support this, their claim being that the most interesting results are achieved when the random graphs do not have the same degree sequence [39, 40]. Others support the idea of the same degree sequence [66, 108]. However, they do not necessarily share the same idea of what the algorithm should achieve [80]. In the coming pages, there will be a discussion of the algorithms that generate graphs based on a fixed degree sequence. In Part II, the algorithms are introduced as well as the statistical toolkit that is used in the analysis of the different approaches to sample graphs. Afterwards, the analysis of the algorithms is split into two parts: In Part III, algorithms used to generate undirected graphs are analyzed based on several well-known graph theoretic measures. In Part IV, directed graphs are analyzed regarding subgraph counting, a topic that has received much interest, but as good as none concerning the used graph generating algorithm. Part V will wrap the thesis up and summarize the most important findings.

Part II

DEFINITIONS

In this chapter, the necessary definitions from graph theory are introduced as well as the algorithms that are used to generate the random graphs. Furthermore, the definition of null models that this work uses is explained. Last but not least, mathematical approximations that are often used to approximate the behavior of a random graph model are introduced.

This chapter is based on work from Milo et al. [63, 64], Zweig [107, 108], Newman [66], Blitzstein and Diaconis [8], and Kim et al. [49]; these researchers developed either the algorithms that are used throughout the thesis or show clear preference of one algorithm over the others. There are more algorithms to generate graphs, but I restrict myself to only a few. The algorithms considered are considered since they are either

a) often used

b) used in different fields

c) the basis of other algorithms

Alternatively, combinations of these.

Similar arguments apply to the measures that are investigated. There is an abundance of measures that may or may not be influenced by the way graphs are generated. Again, going through all of them would take far too long, especially since the development of new measures does not stop. Thus, some measures that are expected yield the same result independent from the generating algorithm are investigated as well as measures that are expected to yield different results. Parts of this chapter were published in a paper "Different flavors of randomness" [80]; the experiments were devised after long discussion with my advisor Katharina Anna Zweig. Together with her and her former doctoral student, Emöke-Ágnes Horvát, this paper was written.

2

GRAPH MODELS

A *graph* $G = (V, E)$ is defined by a set of entities V, called *nodes*, and a set of connections between the entities E, called *edges*. Nodes are referred to as $v \in V$, the number of nodes is $|V| = n$, while edges are referred to as $e \in E = \{\{u, v\} \mid u, v \in V\}$ ($|E| = m$) for graphs where the connection is *undirected*, like friendships. If edges are directed, like followerships on Twitter, the graph is called *directed graph* or *digraph* and edges are referred to as $E = \{(u, v) \mid u, v \in V\}$. The *degree sequence* $DS = \{k_v \mid \in V, k_v \in \mathbb{N}\}$ is a simple way to write down one of the defining features of an undirected graph. For a directed graph the degree sequence is $DS = \{(k_v, j_v) \mid v \in V, k_v, j_v \in \mathbb{N}\}$ with k_v being the *out-degree* of node v and j_v being the *in-degree*. The degree sequence of a graph is the minimal information needed to construct a graph. It does not tell anything about allowed and forbidden connections. There are different classes of graphs that can all be described by the degree sequence property. A graph that does not contain *multiple edges* between two nodes, i.e., $\forall u, v \in V : \{u, v\} \leq 1$, nor *self-loops*, i.e., $\nexists e \in E : \{u, u\}$, is called a *simple* graph. Graphs that allow either of these properties are called *multi-graph*. Graphs that contain self-loops but no multiple edges are called *pseudo-graph* by some authors [36, p.10].

These classes of graphs are just the most basic. There are many more, such as *multiplex* graphs, *temporal* graphs, *weighted* graphs, and others. These are of no further interest in this thesis.

For the very basic graph classes, undirected and directed graphs, random graph theory developed many different algorithms to generate graphs. These algorithms allow to define a statistical null model $(S, \mathcal{P})$ where S is the set of possible observations and $\mathcal{P}$ is a probability distribution on S. At least, this is what one would hope to have; either the number of graphs is unknown, or the probability distribution with which graphs are returned is not known, or both. Thus, the pure statistical definition of a null model does not apply; graph theory has its own definition for a null model. A null model is a *family of graphs* with either:

a) a probability distribution on the given family of graphs

b) an algorithm that produces only graphs from the given family of graphs

To define a family of graphs, I focus on one property that all graphs of the family should share: the degree sequence DS. This family of graphs tends to be large; for example, Blitzstein and Diaconis [8] estimate the number of simple graphs that can be generated for a graph with 33 nodes and 71 edges as $(1.533 \pm 0.008) \times 10^{57}$. It would be nice if everything could be calculated and weighted according to a given probability distribution, but that is as good as impossible. Thus, different algorithms to generate graphs are used throughout the thesis.

In Chapter III, it is explained in more detail why more complex algorithms to generate graphs are necessary. The explanation of the algorithms starts out with the Erdős-Rényi

model, one of the oldest and best-understood graph generating algorithms. This is followed by a more general idea, that is then divided into several possible algorithms.

2.0.1 *The Erős-Rényi model - the random graph*

One of the best-known graph models is the Erdős-Rényi model. That is commonly called the random graph [3]. When reading papers of physicists, the mention of a random graph that is not explained refers to the Erdös-Renyi graph model. The model is based on two parameters, n and m, therefore sometimes named $\mathcal{G}(n, m)$-model. A graph with n nodes is generated, and one edge of the possible $\binom{n}{2}$ is chosen, all edges are equiprobable. This process is repeated for $\binom{n}{2}-1, \binom{n}{2}-2, \ldots \binom{n}{2}-m$ edges while all remaining possible edges are kept equiprobable. The result is one of $\binom{n}{m}$ graphs [25]. This model generates for large n the same graphs as the $\mathcal{G}(n, p)$ model given by Gilbert. The $\mathcal{G}(n, p)$ model tests for each pair of nodes if there should be an edge between them by drawing a number from $[0, 1]$ uniformly at random and comparing to the parameter p. If p is larger than the random number, an edge between the two nodes is generated. An important difference is that the $\mathcal{G}(n, m)$ model always generates a graph with m edges while the $\mathcal{G}(n, p)$ model can generate all graphs with n nodes, i.e., it can contain $m \in \left[0, \frac{n(n-1)}{2}\right]$ edges.

For large n and np being constant the degree distribution of both models follows a Poisson distribution

$$P(k_v = k) = \frac{(np)^k e^{-np}}{k!}. \tag{2.1}$$

This is a nice property, but it is not necessarily appropriate to compare a given graph with those models. When the degree distribution of the given graph is strongly skewed (see Chapter III), many measures will show very different values than the same measures taken on random samples from these models. Therefore, the real-world graph will appear extraordinary, even if the measure is not special in graphs that have the same degree sequence as the given graph. This is shown in more detail in Chapter 4.1.

That is why one cares about fixing certain parameters of a graph and tries to get a random graph model that keeps this parameter constant, but random in all other. For this work, the most simple parameter is being kept fixed, the degree sequence DS, and graphs are generated based on this sequence.

2.1 FIXED DEGREE SEQUENCE MODEL

The fixed degree sequence model is intuitively easy to understand. Given a degree sequence DS, an attempt is made to construct a graph with $|DS| = |V| = n$ nodes, each of which has a degree equal to exactly one of the values in DS, and to connect the nodes according to their given degree. Upon this, there can several more things be required, for example, a fixed clustering coefficient [38], an assortativity requirement [34], or that the graph is simple [63, 72].

Next, some algorithms are presented that can generate graphs with a fixed degree sequence. Their properties and whether the algorithm is useful for network analysis are discussed in detail.

2.1.1 *Havel-Hakimi-Algorithm*

One of the oldest algorithms to generate graphs based on a fixed degree sequence is the Havel-Hakimi-Algorithm that was developed independently by Havel [37] and Hakimi [35]. Moreover, they provided a characterization of degree sequences that are graphical, in other words, it exists at least one graph that is simple and has this degree sequence.

Definition 1 (Havel-Hakimi Theorem):
A degree sequence is graphical if and only if the sequence

$$\{k_2 - 1, k_3 - 1, \ldots, k_{k_1+1} - 1, k_{k_1+2}, \ldots, k_n\}$$

is graphical.

All degree sequences used in this thesis are graphical.
The Havel-Hakimi algorithm works as follows. Repeat until all entries of DS are zero:

1. Sort degree sequence in non-increasing order
2. Let $d = k_j$
3. Remove k_j from degree sequence
4. Subtract 1 from the first d remaining entries of the new sequence.

This algorithm has $\mathcal{O}(n \cdot \log(n))$ run-time (due to repeated sorting) and as long as the degree sequence is graphical, it provides always a simple graph. The problem with the resulting graphs is easily observable. By ordering nodes in a non-increasing order and always connecting to the k_j next nodes, the generated graph will always have many edges between high-degree nodes and much fewer edges from high degree to low degree nodes. When nodes with the same degree are indiscernible, only one graph can be generated. The preference to connect high degree nodes with each other will influence measures taken from these random graphs. Therefore, the results cannot be seen as totally random and despite being one of the first algorithms to generate graphs with a prescribed degree sequence, this algorithm has to be dismissed for analysis of real-world graphs.

2.1.2 *Configuration Model*

The configuration model was described by Bender and Canfield [5]. It was refined by Bollóbas and Thomason [10] as well as Wormald [70]. This is one of the better-studied algorithms to generate a network from a fixed degree sequence [66]. The algorithm is as follows:

1. Generate a graph $G = (V, \emptyset)$. Each node v_i has k_i stubs, k_i being the i-th value in the degree sequence.
2. Choose uniformly at random two stubs of all stubs and connect them with an edge. Repeat this process without the already chosen stubs.

This process seems to be quite simple, but a trap is hidden in this very simple description. Two stubs of the same node can be chosen, creating a self-loop, or stubs of the same pair

of nodes can be chosen several times, creating multiple edges between the two. For the created matching of stubs, the algorithm was also called matching algorithm by Alon et al. [63], but they restricted their algorithm to create only simple graphs. This variant of the configuration model forces generated graphs to be simple by including a check whether the new edge is a self-loop or a multiple edge between nodes. If so, the up to this point generated graph is discarded and the process started anew. An alternative way is creating the graph with the configuration model, including multiple edges and self-loops, but discarding everything that makes the result not simple in the end. This approach is also called *erased configuration model* [95] since edges that do not agree with the property of being simple are erased. Of course, this changes the degree distribution. The influences of these changes are discussed in Chapter 8. Bender and Canfield [5] showed for regular graphs that the probability to get a simple graph is

$$P\left(G \text{ is simple}\right) \sim \exp \frac{1-k^2}{4}$$

for $n \to \infty$ with k being the degree of the nodes in a regular graph. Blitzstein and Diaconis tested how many samples one has to draw to get a simple regular graph with degree 8. About 7 million samples were necessary [8], which is a remarkably bad result. Additionally, Janson [44, 45] showed that the probability of a graph with *any* degree sequence being simple is

$$P\left(G\left(n, (k_i)_1^n\right) \text{ is simple}\right) = \exp\left(-\sum_i \lambda_i - \sum_{i<j}\left(l_{ij} - \log\left(1+\lambda_{ij}\right)\right)\right) + o\left(1\right)$$

with

$$\lambda_i = \binom{k_i}{2}\frac{1}{n}$$
$$\lambda_{ij} = \frac{\sqrt{k_i\left(k_i-1\right)k_j\left(k_j-1\right)}}{n}.$$

This probability converges to 0 when $\frac{\sum_i k_i^2}{n} \to \infty$ but stays greater than 0 if $\sum_i k_i^2 = \mathcal{O}\left(n\right)$. Many real-world graphs have a degree distribution similar to a power-law, i.e., the second moment $\sum_i \frac{k_i^2}{n}$ is increasing, such that these graphs can be expected to have some multiple edges and/or self-loops.

The algorithm to generate directed graphs works in the same way, with the difference that one stub of the out-degree sequence and one stub of the in-degree sequence are chosen to generate edges. Besides that, the above arguments apply.

This algorithm is often used and produces results that are applied in different fields. That is most likely due to its speed, the run-time to generate a graph is in $\mathcal{O}\left(m\right)$. The problem is, that researchers rarely mention whether they accounted for the properties of the graphs this algorithm generates, i.e., whether they accounted for multi-edges or selfloops. These properties might influence the results of measures and algorithms that are intended for simple graphs severly. Other researchers prefer the erased configuration model since it generates graphs that have a degree sequence similar to the fixed degree sequence given but are thus restricted to graphs that have less edges [39, 40]. But some algorithms can generate graphs without creating self-loops or multiple edges while maintaining the degree sequence.

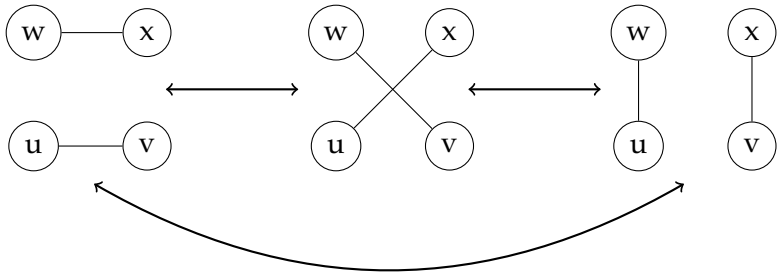


Figure 2.1: Possible two-edge swaps in an undirected graph.

2.1.3 *Switching Algorithm*

A more appropriate algorithm to generate simple graphs is the switching algorithm. This algorithm was named thus by Alon et al. [63], even though its basic idea was used already by Brualdi [13] as well as Roberts and Stone [77] without giving it a name. This algorithm is based on so-called edge swaps or matrix interchanges.

In an undirected graph the swaps depicted in Figure 2.1 are sufficient to generate all possible graphs. The problem with this approach is that the number of swaps is unknown. The basic algorithm is the following:

1. If not given an initial graph, generate a graph based on the degree sequence (with the Havel-Hakimi algorithm or any other simple graph generating algorithm)
2. Repeat for a large number of steps:
   a) Draw two edges uniformly at random.
   b) Draw a number $q \in [0, 1]$ uniformly at random. If $q \leq 0.5$ swap the edges $\{u, v\}, \{w, x\}$ to $\{u, x\}, \{w, v\}$, otherwise to $\{u, w\}, \{v, x\}$
   c) Check if the swap created a self-loop or a multiple edge; if so, undo the swap

To generate directed graphs edge swapping can be used as well with a few changes. First, it is not possible to apply all swaps from Fig. 2.1, which is evident when considering that the edges are directed. It is not possible to apply other edge swaps to $(u, v), (w, x)$ than $(u, x), (w, v)$, since it is not possible to connect out-stubs to other out-stubs. Directed graphs, in contrast to undirected graphs, can show configurations that are impossible to change by a simple two-edge swap even though it is evident that there are other realizations. An example is given by Berger and Müller-Hannemann [6] (comp. Fig. 2.2).

They consider a digraph with $3n$ nodes, $V = \{v_1, v_2, \ldots, v_{3n}\}$, where the triples $v_{3i}, v_{3i+1}, v_{3i+2}$ form three-cycles and all nodes of a single cycle are connected to any node not in this three-cycle. Any standard edge swap in a three cycle leads to a self-loop, such that this is swap is undone. Any swap of edges connecting the three-cycles either leads to the previous graph or multiple edges between two nodes, so they are undone as well. Any swap with one of the intra three-cycle edges and inter three-cycle edges leads to multiple edges. Thus, apparently only one graph exists, even though it is obvious that the direction inside a three-cycle could be swapped. A possible three-edge swap, i.e., a reorientation of a three-cycle, is introduced to solve this problem. They prove that it is possible to reach all possible realizations of a degree sequence with the additional three-edge swaps and give details on

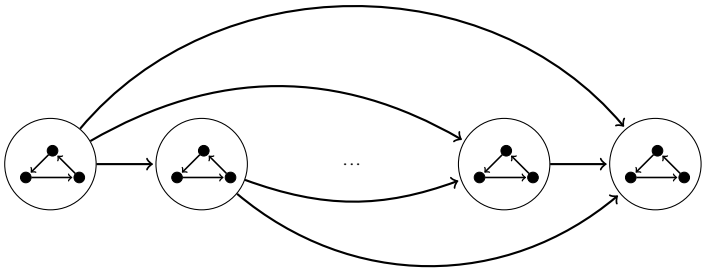


Figure 2.2: A digraph for which the standard switching algorithm (two-edge swaps) is not sufficient to find any other realization, even though there exist much more. Taken from [6].

how to recognize a degree sequence that allows using the simple two-swap algorithm (for more details, see [6]).

Of significant interest for this algorithm and its variants is the large number of steps mentioned in the description in the algorithm, written as $\mathcal{O}(c \cdot m)$. This constant was first introduced by Alon et al. [63]. They provided results of an experiment in that they measured a property of a network. The experiment was based on several 1000 samples with a small constant c; they increased c and tested whether the average of the measure on the graphs changed, i.e., they tested for when it stabilized. They found that a suitable constant would be 100, even though their experiment indicated stable values much earlier. Viger and Latapy found experimentally that for generating connected graphs $\mathcal{O}(m \cdot \log(n))$ time would be sufficient [96]. The proof is still missing. This is not constant, but provides an estimate for each graph. Ray et al. [76] found experimentally that values of $\varepsilon < 5 \cdot 10^{-3}$ are sufficient if $c = \ln\left(\frac{1}{\varepsilon}\right)$, i.e., $1 \leq c \leq 15$. For this, they took several thousand samples of very different graphs with different c and evaluated some measures on the samples. Afterward, they compared the distribution of the measures. They found, that the measures result in the same distribution for most c, but very small c. Thus, they suggest a c of 5 to 10. Since the samples are drawn from a vast set of possible graphs, the stability of the results gives confidence in this approach.

In a Bachelor thesis [79], it was tested whether this constant depends on the degree sequence. For this, a decreasing out-degree sequence was generated and kept constant, the in-degree sequence being once correlated, once un-correlated, and once anti-correlated. A "meta-graph" was generated in the following way: starting from a connected graph as a node, all possible edge swaps were denoted; if an edge swap resulted in the original graph, it was denoted as a selfloop, otherwise it was an edge to another graph (i.e. node in the meta-grpah). For all new nodes the same procedure was performed. After a time, no new graphs were found an this process terminated. Using a PageRank-like algorithm it was possible to find an approximation of c for the original graph such that a repeated evaluation of a measure with cm swaps yielded the same result as $m \log(m)$ swaps. Interestingly, even though all three bi-degree sequences yielded graphs with the same amount of edges, the constant needed to generate all possible connected graphs was different for all degree sequences. It was also much lower than 100, even though this would have been sufficient.

Another interesting result regarding the mixing time and the number of samples was discovered by Brugger et al. [14]. They tested on bipartite graphs whether it is possible to reduce the number of samples taken and whether it is possible to use only a subgraph instead of the whole graph to reach sufficient results. The tests performed were based on the Netflix-dataset and the MovieLens-dataset, i.e., movie-ratings of users. The assumption of these tests is that sequels have a higher number of common neighbors, i.e., their *co-occurrence* is higher. Brugger et al. calculate p-values denoting those pairs that have a co-occurrence as least as large as the original, i.e.,

$$p(x,y) = \sum_{i=1}^{s} \mathbb{1}_{\text{coocc}_i(x,y) \geq \text{coocc}(x,y)}.$$

Based on the p-value they calculated the *positive predictive value at k*, PPV_k, where k indicates the number of pairs of films in the ground-truth (a verified set of edges in a graph that contains also noisy data). The PPV is the fraction of correctly identified pairs from the ground truth in the set of the k highest ranked pairs of films [107]. This is similar to the *precision at k* known from information retrieval [59, Ch.8]. Their results indicate that after a few hundred generated samples the change in the PPV_k peaks and does not change significantly after this, no matter how many samples are drawn. Additionally, they were able to reduce the number of swaps to be taken from 10^9 to 10^6, which is equal to a reduction from $m \cdot \log(m)$ to $m \ln\left(\frac{1}{\varepsilon}\right)$ with $\varepsilon \sim 4.5 \cdot 10^{-5} = 10$ as shown by Ray [76].

This algorithm received much attention for generating graphs, statistical analysis, and tuning of the parameters that are important for its runtime. Still, there is still no proof that $c \simeq 10$ is for all graphs a large enough factor. Therefore, another algorithm that generates only simple graphs is presented.

2.1.4 *Sequential Importance Sampling*

Another algorithm to generate undirected simple graphs was developed by Blitzstein and Diaconis [8]. Later on, a very similar algorithm for directed graphs was developed by Del Genio et al. [49], who mentioned an unpublished algorithm by Blitzstein and Diaconis for the same purpose with almost the same approach.

The sequential importance sampling algorithm proceeds as follows based on a degree sequence DS.

1. Let E be an empty list of edges.
2. While DS contains non-zero elements:
   a) Choose the least i with k_i a minimal positive entry.
   b) While $k_i > 0$:
      i. Compute a candidate list

$$J = \{j \neq i \mid \{i,j\} \notin E \wedge \{k_1, k_2, \ldots, k_i - 1, \ldots, k_j - 1, \ldots, k_n\} \text{ is graphical}\}. \tag{2.2}$$

      ii. Pick $j \in J$ with probability proportional to its degree.

Figure 2.3: Intermediate step in the graph generating process of the sequential importance sampling.

iii. Add $\{i, j\}$ to E and update $\mathcal{DS}$.

3. Return E.

The test whether the remaining sequence is graphical is not used in any of the before presented algorithms but the Havel-Hakimi test for graphicality needs $\mathcal{O}(n)$ time in an elegant implementation. Additionally, there are $\sum_{i=1}^{n} k_i$ candidate lists to generate. Since it is done for each node its degree times, the overall time is $\mathcal{O}\left(n^2 \sum_{i=1}^{n} k_i\right) = \mathcal{O}\left(m \cdot n^2\right)$. This approach has some disadvantages, that are reduced by a feature concept of the algorithm. The major disadvantage is that it does not sample uniformly at random from the family of all simple graphs. As an example, consider the graph partially given in Figure 2.3.

The probability that any of the nodes in the middle is chosen is the same, the resulting graph does not look any different independent of which one is chosen first. Moreover, more than one sequence of created edges can lead to the same graph. While the overall graph is the same, the sequence constructing the graph is not, which is usually not traceable. The algorithm calculates simultaneously to choosing a neighbor the probability of choosing this node. This probability is calculated for all edges constructed and gives the overall probability of a graph and its construction sequence. Calculating the probability to get a graph helps when estimating the expected value of a measure. According to Blitzstein and Diaconis [8] the following proposition holds, where $\sigma(Y)$ is the probability that a graph Y is constructed by a given sequence, and $c(Y) = \Pi_{k=1}^{m} d_{i_k}^{i_{k-1}}!$ is assumed as the number of graphs like Y:

Propositon 1 (Blitzstein [8])**.** *Let π be a probability distribution on $\mathcal{G}_{n,d}$ and G be a random graph according to π. Let Y be a sequence of edges distributed according to σ. Then*

$$\mathbb{E}\left[\frac{\hat{\pi}}{c(Y)\,\sigma(Y)}\right] = \mathbb{E}f(G).$$

In particular, for $Y_1, \ldots, Y_N$, the output sequences of N independent runs of the algorithm,

$$\hat{\mu} = \frac{1}{N}\sigma_{i=1}^{N}\frac{\hat{\pi}(Y_i)}{c(Y_i)\,\sigma(Y_i)}\hat{f}(Y_i)$$

is an unbiased estimator of $\mathbb{E}f(G)$.

With this, it is sufficient to take fewer samples to get a good estimate of the measure f one is looking for. Computing the above equation with f being a constant function yields the approximate number of graphs with the given degree sequence, i.e.,

$$\mathbb{E}\left[\frac{1}{c(Y)\,\sigma(Y)}\right] = |\mathcal{G}_{n,d}|. \tag{2.3}$$

It is important to note that Blitzstein and Diaconis [8] also consider another probability to chose neighbors instead of the degree based choice. They suggest that a uniform choice on the set of possible neighbors may be used, but the difference between these two probabilities when choosing a neighbor has not been evaluated to the best of my knowledge.

Almost the same algorithm is used to generate directed graphs. The order of the elements in the degree sequence is a bit more complicated. For this, the definition of normal order is necessary.

Definition 2 (Normal Order [49]):
A degree sequence

$$\mathcal{D} = \{(k_1, j_1), (k_2, j_2), \ldots, (k_n, j_n)\}$$

is in normal order if and only if

$$\forall u < v, u, v \in \mathbb{N} : k_u > k_v \vee k_u = k_v \wedge j_u > j_v.$$

The degree sequence has to be in normal order before a node is chosen. The node that is chosen is the lowest-index node u with non-zero (residual) degree. This algorithm chooses uniformly at random from the set of possible neighbors, connects the two nodes with a directed edge and brings the residual degree sequence in normal order again [49]. Besides that, the algorithm has the same advantages as for undirected graphs with a worst runtime complexity of $\mathcal{O}(n^3)$.

These algorithms are well-known algorithms in the field of network analysis. The configuration model has several variants. They are much more elaborate and restrict the space of possible outcomes, not necessarily preventing multiple edges or self-loops. Moreover, sometimes very similar models have been developed, or even the same model is developed but not attributed to the configuration model (f.e. [83]).

3

MATHEMATICAL TOOLKIT

Up to this point, several null models based on algorithms have been introduced. It remains, how to compare a null model to a graph and, more importantly, how to compare null models with each other? The first part is simple.

1. Calculate the measure of interest in the real-world graph;
2. Calculate the measure of interest in a large enough number of samples drawn from an appropriate null model;
3. Calculate with an appropriate test method whether the result in the real-world graph is very different from the results of the samples.

The last step is often done with a simple z-test, even though it is most of the time unknown whether the underlying distribution is normal [63, 64]. Still, in most cases this test is enough to assess statistical significance. Alternatively, Student's t-statistic is also a measure that is useful when only a sample is taken and the mean and standard deviation are only approximated [12].

Now, how to compare two null models? As a reminder, null models are families of graphs together with a probability distribution of the occurrence of a graph. An alternative to describe a null model is by defining an algorithm that produces graphs from a set of graphs with some probability. These probability distributions are of interest. The switching algorithm uses a Monte-Carlo Markov chain that is symmetric, aperiodic, and irreducible. The states of the Markov Chain are the graphs in the family of graphs and since the properties are all accounted for [6, 88], the probability to reach any graph of the family is uniform. The configuration model is more problematic: due to the tendency to build self-loops and multiple edges [44, 45] some graphs may have higher probabilities to be generated, because they are multigraphs. Milo et al. tested once whether the modified configuration model with rejection would produce uniformly all graphs for a toy example [63]. It did not produce the graphs uniformly at random. Thus, the probability distribution of the configuration model with rejection is unknown, and the standard configuration model does not fare better. For the sequential importance sampling, it is already known that it does not produce graphs uniformly at random. Luckily, the algorithm provides information on the probability to draw the resulting graph in exactly the sequence, which allows for a simpler calculation of a weighted mean of measures taken on graphs sampled with the sequential importance sampling.

Recognizing that the null models do either not chose from the same family of graphs (multigraphs vs. simple graphs) or that they do not do so with the same probability distribution leads again to the question, how to compare different null models?

3.1 STATISTICAL ANALYSIS OF DIFFERENCES IN RESULTS

A standard way to test whether two populations (families of graphs) do share the same mean value in some measure is the two-sample z-statistic,

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where t is distributed according to Student's t-distribution, $\bar{x_1}, \bar{x_2}$ are the sample means, μ_1, μ_2 are the means, s_1, s_2 are the sample standard deviations, and n_1, n_2 are the respective sample sizes. But, in the limit of large numbers ($n_1 > 30, n_2 > 30$), it is valid to use the two-sample z-statistic

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{3.1}$$

If the population variances are not known, which they are as good as never for a family of graphs, sample variances are used. Since it is tested whether the null models coincide in a measure, the assumption is that $\mu_1 = \mu 2$, i.e. $\mu_1 - \mu_2 = 0$.

Both of these measures assume that the measure of interest is distributed according to a normal distribution $\mathcal{N}(\mu, \sigma)$. Since this is not guaranteed, the Kolmogorov-Smirnoff two-sample test is applied as well.

The exact explanation of this test can be found in e.g. [12, 81]. Here a more simple but much more illustrating example is provided. The most simple explanation is visually: The cumulative distribution function of the results of the measures in each null model is taken and the largest difference of these is the test-result D_{n_1,n_2}, where n_1 and n_2 are the respective sample sizes. In Fig. 3.1, this is visualized with some random numbers taken from two binomial distributions with the same mean but different standard deviations. This value is than compared to fixed values $c(\alpha)$; whenever $c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}$ is larger than D_{n_1,n_2}, it is assumed as plausible that the two samples are from the same distribution. Naturally, it is possible to calculate a p-value

$$p = 2\min\left\{\Pr\left(D_{n_1,n_2} \leq c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}\right), \Pr\left(D_{n_1,n_2} \geq c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}\right)\right\} \tag{3.2}$$

where $\Pr(X \leq x) = 1 - 2\sum_{i\geq 0}(-1)^{i-1}\exp -2i^{2x}$ [12]. This value allows to give an estimate of how likely it is that the two distributions are the same.

These tests suffice for statistical analysis and will be used throughout the thesis.

Another measure of interest is the so called skewness of the degree sequence. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. For a degree sequence, this implies that the degree distribution of the sequence is analyzed based on Pearson's skewness coefficient,

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathbb{E}\left[(X-\mu)^3\right]}{\left(\mathbb{E}\left[(X-\mu)^2\right]\right)^{1.5}}. \tag{3.3}$$

Distributions such as the Poisson distribution have very low skewness ($\mu^{-0.5}$), while many real-world networks show higher skewness values [81].

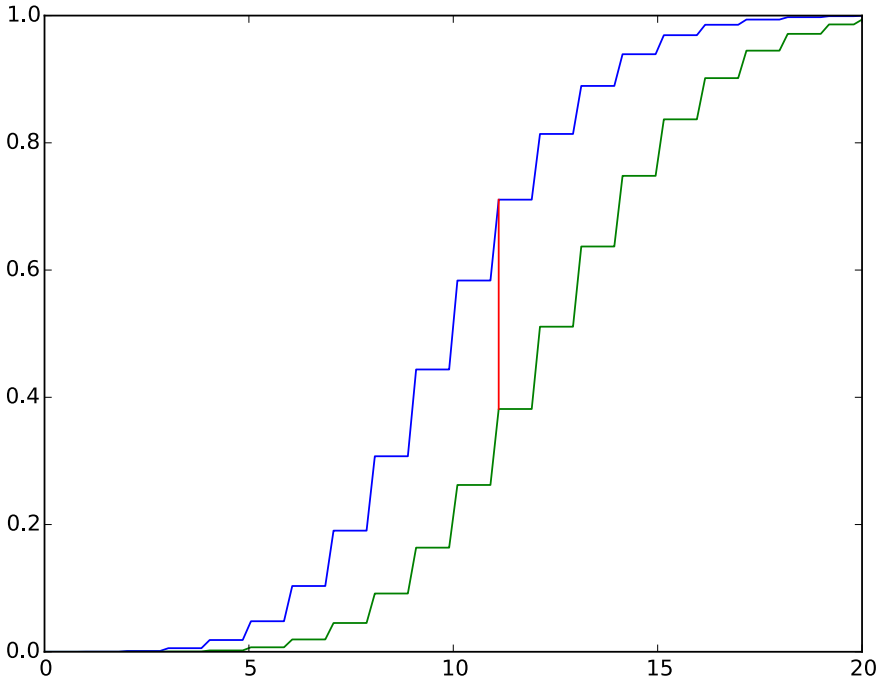


Figure 3.1: Two cumulative distribution functions of samples are computed and the maximum difference between these (marked with a red vertical line) is the result D_{n1,n_2} of the Kolmogorov-Smirnoff two-sample test.

Besides statistical testing whether another graph model is appropriate, there is one model that was not discussed above that is based on several simple assumptions. As some of the algorithms take a rather long time to calculate measures, and the sampling of random graphs is also costly, researchers always strive to improve algorithms, parallelize them or try to develop new approaches to old problems such that faster solutions to already solved problems can be applied. In Physics, it is similar, and it is not surprising that physicists developed some interesting ideas. A very prominent example that is based on very simple assumptions is explained in the following.

3.2 SIMPLE INDEPENDENCE MODEL

The *simple independence model* (SIM), as it was termed by Zweig [107], is based on a series of very simple assumptions. Consider two nodes in a graph, u, v, with degree k_u, k_v respectively. The probability that node u connects to node v in the configuration graph model is calculated as follows. There are $2m$ stubs at all, $2m-1$ subtracting one from u (this is the one looking for a a stub to connect to). Of those $2m-1$ stubs, exactly k_v belong to node v. Since there are k_u stubs coming from node u and each is equally likely to connect to node v, the probability of a connection between u and v is equal to

$$p_{u,v} = \frac{k_u k_v}{2m-1}. \tag{3.4}$$

Actually, this is the sum of the single events, connecting single stubs of u to v, therefore it is an expected value. Newman writes in [66] that for large m this term converges to a probability and is sufficient to use

$$p_{u,v} = \frac{k_u k_v}{2m}. \tag{3.5}$$

This equation can also be found in the Chung-Lu model [17].

Based on this simple expression several equations were developed to calculate measures that are quite common in graph theory and network analysis. It is often applied without considering the graphs Newman had in mind while developing this equation. Newman developed his equations with the random graph, i.e., the ER-model in mind [66, p.420]. He states that

> Although the random graph is, as we have said, not an accurate model of most real-world networks, this is, nonetheless, believed to be the basic mechanism behind the small world effect in most networks[...] [66, p.420].

While this part was focused on the diameter, this statement can be found in other places in his work as well. Still, some researchers apply the equations and state their results and do not investigate if the results are valid, i.e., if the equations are applicable to the given graph. Chung and Lu [17] state that if $\forall v \in V : k_v^2 < \sum_{v \in V} k_v$, equation 3.5 is for all node pairs between 0 and 1. Whenever this condition is not met, the results of equation 3.5 have to be considered as expected values. Still, for some measures the model is applicable as is shown in this thesis.

With the probability for an edge, some additional calculations can be made. Assume that there is an edge between nodes u and v. Since SIM assumes simple independence, there

is nothing preventing a second edge between the same nodes. The probability for a single second edge can be calculated thus as

$$\frac{k_u k_v}{2m} \frac{(k_u - 1)(k_v - 1)}{2m}. \tag{3.6}$$

Summing this over all node pairs is then equal to

$$\frac{1}{2(2m)^2} \sum_{u,v} \frac{k_u k_v}{2m} \frac{(k_u - 1)(k_v - 1)}{2m} = \frac{1}{2(2m)^2} \sum_u k_u (k_u - 1) \sum_v k_v (k_v - 1) \tag{3.7}$$

$$\overset{2m=n\langle k\rangle}{=} \frac{1}{2\langle k\rangle^2 n^2} \sum_u k_u (k_u - 1) \sum_v k_v (k_v - 1) \tag{3.8}$$

$$= \frac{1}{2\langle k\rangle^2 n^2} \sum_u \left(k_u^2 - k_u\right) \sum_v \left(k_v^2 - k_v\right) \tag{3.9}$$

$$\overset{\frac{1}{n}\sum_{v\in V} k_v^i = \langle k^i\rangle}{=} \frac{1}{2\langle k\rangle^2 n^2} n^2 \left(\langle k^2\rangle - \langle k\rangle\right)^2 \tag{3.10}$$

$$= \frac{1}{2} \left[\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right]^2. \tag{3.11}$$

Moreover, this allows also the quick estimation of how many self-loops a node might have. Considering the probability to have an edge from node u to node v, it is a simple change from v to u. Since one of the stubs is gone, the second parameter is reduced by one and results in

$$p_{u,u} = \frac{k_u (k_u - 1)}{2m}. \tag{3.12}$$

The expected number of self-loops for a graph can be calculated very easily as

$$\sum_u \frac{k_u (k_u - 1)}{2m} = \frac{\langle k^2\rangle - \langle k\rangle}{2\langle k\rangle}. \tag{3.13}$$

In social network analysis, these equations are not of much use. Social network analysis is concerned with graphs that do not permit multiple edges, as long as the underlying graph is not multiplex. Neither should a standard social network contain any self-loop since they do not provide any information in most cases. There may be exceptions where self-loops provide information, but those are rare. Nevertheless, the equations can tell the expected number of self-loops and multiple edges the configuration model develops. What the equations do not provide is information where those multiple edges and self-loops are attached.

Other equations are introduced when necessary to avoid unnecessary repetition.

3.3 CRITIQUE ON THE APPLICABILITY OF SIM TO ALL GRAPHS

While the simple independence model seems to allow for calculation of expected values, the applicability of the model should be discussed. The model and the equations used therein were developed with two premises that are either not explicitly stated or not stated at all. The model that was to be approximated was the ER-model or the $\mathcal{G}(n, m)$ model, initially. Nowadays, almost any graph considered for research purposes is said to have a

degree distribution following a power-law, being close to a power-law, or having at least a skewed degree distribution [18].

Newman used his equations on graphs with skewed degree distributions as well. For this purpose, other properties of the degree sequence should be checked before using the simple independence model, such as $\forall v \in V k_v^2 < \sum_{v \in V} k_v$ as stated by Chung and Lu [17]. Otherwise, results may be hard to interpret or wrong (such as probabilities larger than 1). But there are other caveats as well. As an example, we consider two different applications of the simple independence model that should be done only with utmost care.

3.3.1 *Example 1*

The first example is a measure called *assortativity*—the idea is to have a value between -1 and 1 that says something about the mixing structure of the graph. For example, assortativity can be calculated based on scalar values assigned to people and the scalar values associated with their neighbors [67]. When people choose other people with scalar attributes similar to their own, assortativity is high; when they choose otherwise, the assortativity is negative. We can also speak of a disassortative behavior of groups.

This is investigated in the following based on the degree.

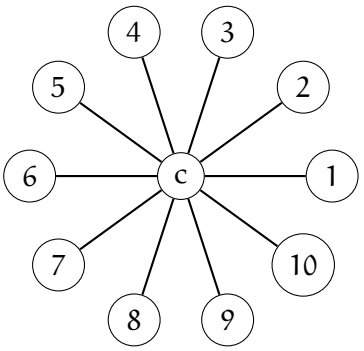
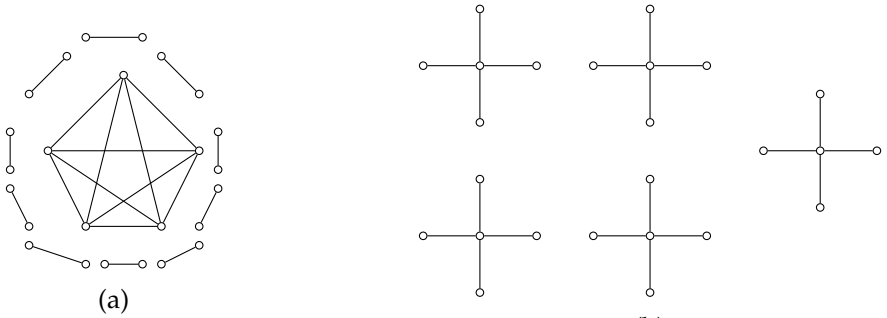


Figure 3.2: Example 1

Consider the graph given in Fig. 3.2. It is easy to see that this is the only possible realization of the graph without any self-loops or multiple edges, in other words, it is the only simple graph possible. Calculating the assortativity of a graph like this is not necessary since there are no other realizations of a degree sequence like this. Still, applying the equation for degree assortativity,

$$r = \frac{\sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) k_i k_j}{\sum_{i,j} \left(k_i \delta_{i,j} - \frac{k_i k_j}{2m} \right) k_i k_j},$$

to this graph yields a value of -1 which seems on the first look reasonable—the low degree nodes are connected to the high degree node, the graph is as disassortative as possible. The graph is compared to the "most assortative realization" possible, i.e., the divisor in the equation symbolizes a graph in the each high degree node is connected to a node of equal degree, and each low degree node is connected to a node of equal degree. For the graph in Fig. 3.2, this would imply that node c is connected to itself, and there are five two-cliques.



(a) (b)

Figure 3.3: Two graphs generated from the same degree sequence with different modularity-scores.

But that is first of all not a simple graph; second, the assumption that any node can connect to another node of the same degree is just not met in all networks. It is more likely that this assumption is not met for all nodes in any real world graph. The probability that a graph has enough nodes of each degree such that they connect only to each other is very low. That is an assumption based on recent research, that almost all graphs that are used in research either show power-law like distributions or distributions that have much less high degree than low or average degree nodes, i.e., they have skewed degree distributions [18].

3.3.2 *Example 2*

As another example, consider the modularity defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \delta\left(c_i, c_j\right),$$

where c_i, c_j are the identifiers of a dense group, called community, node k_i, k_j belong to. This is said to be a measure of the extent to which like is connected to like in a network [66]. It takes values between 1 and -1, indicating with positive values whenever more edges are between vertices of a group than expected, negative when not. It is similar to assortativity in the sense, that by dividing by the maximum of the modularity another assortativity coefficient can be calculated (for more detail see [66, p. 224f.]). Still, modularity is another measure that is often applied in network analysis. But rarely is the equation under explicit consideration whether it makes sense to use it at all.

In Fig. 3.3 two graphs are displayed that share a very simple degree sequence. There are five nodes with a degree of 4 and 20 nodes with a degree of 1. In Fig. 3.3a we have a central clique, surrounded by dyads while in Fig. 3.3b are five four-stars displayed. By the claim of Newman [66], modularity should be higher for graphs in which more similar nodes are connected, i.e., more nodes of a group. But as a calculation of the modularity shows, the complete graph in Subfig. 3.3a (all components) has a modularity of 0.725 while the star-graphs in Subfig. 3.3b (all components) have a modularity of 0.8 while a single star-graph has a modularity of 0. Of course, when each dyad is considered as its own community c_j, each star graph as its own community c_j, the results is more reasonable, yet still irritating when the five-clique in Fig 3.3a is considered.

In the paper by Newman and Girvan, modularity is explained as the difference between observed edges inside a community minus the expected number of edges within them when placed at random [68]. Implicitly, the simple independence model is used here which directly implies that self-loops, as well as multiple edges, are possible. There are attempts to calculate a more fair modularity score [15] that compensates for loops, or for loops and multiple edges, but we did see no paper but this using this approach.

Part III

COMPARISON OF THE DIFFERENT NULL MODELS BASED ON ARTIFICIAL AND REAL-WORLD UNDIRECTED GRAPHS

In the following, some research based on artificial and real-world graphs is presented. The analysis starts with artificial graphs, using a graph from the simple $\mathcal{G}(n, m)$ model, and it can be observed that null models do show very similar results for the unskewed degree distributions. But the more skewed the degree distribution is that is used for the artificial graphs, i.e., the more similar to nowadays real-world graphs the degree sequence becomes, the more the results of different null models differ.

This effect can be observed even more clearly in the analysis of the real degree sequences. The differences sometimes seem to be minuscule, but when network analysis is applied in real life, these differences can be of great significance. Consider as an example the television show "Numb3rs", in which the young Professor Charles Eppes sometimes applies network analysis to spot the culprit. In one of the episodes, it is mentioned that Professor Eppes compared the development of friendships to random Apollonian networks, which is not necessarily reasonable due to the construction process of a random Apollonian network. Something along the lines is mentioned by another character in the television show, stated as "some of the math is a little subjective". In another episode temporal network analysis is performed in which sequences of graphs are analyzed to find people who are moved out of focus of a group to find the culprit. But, as with most television shows, this does not work well, and they miss the evil-doer until the very end.

Even though this is just a television show, some parts of network analysis have found their way into reality and are applied in useful ways. Some criminals get caught since their closer social network is observed via Facebook and people upload pictures without concern. The NSA and DoD use both social network analysis, the extent of which is unknown [53]. Other papers concerned with terrorist networks or political networks have been published in the last decade (f.e. [74, 82, 97, 16]). Seldom, more in-depth analyses than the standard centralities are applied to find important persons. But centrality scores can sometimes be artifacts caused by the density or some structural property of the graph. By analyzing random graphs with the same degree sequence, it is possible to observe artifacts like this when the random graphs show the same or similar results.

Most of the work done in this part was done in collaboration with Katharina Zweig, who had the initial idea to compare measures on different null models to see whether null models yield the same range of results. This was inspired by her work on bipartite graphs in which she discovered strong discrepancies between two different null models. The paper that originated from this is "Different flavors of randomness—which graph model to use to assess statistical significance". This paper was written in collaboration with Katharina Zweig and Emöke-Ágnes Horvát, with great influence by Katharina Zweig, while Emöke-Ágnes Horvát provided helpful insights and restructured much of the text.

Synopsis First, it is important to note that different null models can result in the same average values, but that they do not have to. This effect seems to be strongly dependent on the degree sequence, but that is not provable up to now. Afterward, a discussion on how the sequential importance sampling by Blitzstein and Diaconis should choose neighbors is used to lead into a discussion of the measures that were applied in "Different flavors of randomness" [81]. While some of the results are of this paper, all discussed material that references the sequential importance sampling are new. The conclusion did not change much from "Different flavors of randomness"; some null models have side effects that influence the analysis, like multiple edges between nodes or preference of high degrees as partners when constructing a graph from scratch.

In the following we will use shorthand notation for the different graph models:

CFG Configuration model

ECFG Erased configuration model

FDSM Fixed degree sequence model with the swapping algorithm

SIS Sequential importance sampling

- DSIS with degree-based choice from the set of possible neighbors
- USIS with uniform choice from set of possible neighbors

4

COMPARATIVE ANALYSIS BASED ON THE MODELS

As an introduction to the topic, we use two artificially generated graph. One is generated with the $\mathcal{G}(n, m)$ model, the other with an algorithm provided by Leskovec et al. [55], the Forest-Fire model. While the first graph has a degree sequence that follows roughly a Poissonian distribution, the second graphs degree sequence follows a powerlaw distribution. We then generate random graphs based on the two artificial graphs, using the previousely described algorithms. The approach is to compare the artifical graph with samples drawn from a null model that generates graphs that are similar to a certain extent, i.e., they share the degree sequence but nothing else. In the following it is shown that the CFG is good to analyze graphs that have a degree sequence following a Poissonian degree distribution, but for other degree distributions, it is not the best approach.

The measures that will be investigated are the following:

DISTANCE AND DIAMETER One of the most basic questions that sparked the interest in graph theory was the "Königsberger Brücken" problem in which several parts of a city were connected by bridges over rivers. The question was "Can we devise a route that uses all of our cities seven bridges without using a bridge twice?". Nowadays, a travler in the city of Königsberg (Kaliningrad, Russia) has only five bridges and it is rather more likely that the question in his mind will be "What is the shortest distance I have to go if I want to visit all bridges in Kaliningrad?" In other words, what are the distances between the bridges? To solve this, the single-source shortest-path problem is a good analogy from graph theory.

DEFINITION 3 (SINLGE-SOURCE SHORTEST-PATH PROBLEM):
Given a source node s in a graph $G = (V, E)$ find the shortest sequence of vertices $P = (s, v_1, v_2, \ldots, v_l, t) \in V^{l+2}$ such that v_i is adjacent to v_{i+1} for $1 \leq i < l$ for all other vertices in the graph.

The single-source shortest-path problem is a subproblem of the ALL-PAIRS SHORTEST-PATH problem, a problem that searches to answer the same question for all pairs of nodes.

Calculating all distances between all pairs is a solved problem, but it is one of the old and very prominent problems. Already in 1959, Dijkstra solved it with the famous Dijkstra-algorithm and Bellman and Ford, Johnson, and Floyd-Warshall devised other solutions for weighted graphs, negatively weighted graphs and others [22, 4, 47, 27]. The problem of the algorithms is the runtime. While the single-source shortest-path problem can be solved in $\mathcal{O}(n^2)$, the all-pairs shortest-path problem still requires with these algorithms $\mathcal{O}(n^3)$ time.

Since comparing all length of shortest-paths between all pairs of nodes would take long time and would most likely yield rather inconclusive results, we are comparing the average distance.

DEFINITION 4 (AVERAGE DISTANCE):
The average distance (average path length) is one of the most robust measures of graph theory. Assume that $d(v_1, v_2) = 0$ when v_2 cannot be reached from v_1, otherwise let $d(v_1, v_2)$ denote the shortest distance between v_1 and v_2. The average distance is thus

$$\frac{1}{n(n-1)} \sum_{v_i, v_j \in V, i \neq j} d(v_i, v_j). \tag{4.1}$$

Algorithms like Dijkstra's or Bellman-Ford's calculate the shortest distance between all pairs of nodes. If there is no viable way from node u to node v in a graph, this means that they are in two different *components* and the distance between them is infinity per definition. For the algorithms used to calculate distance, this is not a problem, but for the diameter it is.

DEFINITION 5 (DIAMETER):
The diameter is the longest shortest distance between any two nodes in the graph.

Since calculating the diameter for a graph that is disconnected may end up with ambigous results, depending on the definition of the distance, for the diameter the largest connected component is considered.

DEFINITION 6 (COMPONENT):
A (connected) component is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the graph.

Most real world graphs have one so called *giant component*, i.e., a component that contains most nodes and most edges while the other components tend to be smaller such that they do not matter much. A standard approach to find connected components is breadth-first search, labeling the found nodes and continuing while there are other unmarked nodes with another label. In this work, whenever the diameter is mentioned, the diameter of the giant component is meant. Since not every algorithm to generate graphs checks whether the graph contains only one component, we check all measures only on the largest component. This implies that neither the true diameter of a graph with more than one component is recorded (∞), nor necessarily the largest diameter of all components. Consider a graph that has 98% of its nodes in one component with a diameter of 5; the rest builds a sparse second component with a diameter of 6. Still, the diameter of 5 contains more information about the speed with which messages may be passed through the graph.

A standard algorithms to calculate the diameter is the Floyd-Warshal algorithm, that solves the all-pairs-shortest-path problem first. It is then just looking up the largest value. Algorithms that are faster in praxis still have the same worst case runtime [11, 19].

AVERAGE NEIGHBOR DEGREE Another classic measure is the average neighbor degree, i.e., how many neighbors do the neighbors of a node have on average. This measure is calculated for a single node as

$$a\nu(\nu) = \frac{1}{k_\nu} \sum_{u \in N(\nu)} k_u. \tag{4.2}$$

$N(\nu)$ is the neighborhood of a node ν. This measure can be seen as an indicator of degree assortativity, i.e., when a high degree node mainly has high degree neighbors, the average neighbor degree should be high for the node itself. The global average neighbor degree, i.e., the average of the average neighbor degree, is calculated as

$$AV(G) = \frac{1}{n} \sum_{\nu \in V} a\nu(\nu) = \sum_{\nu \in V} \frac{1}{k_\nu} \sum_{u \in N(\nu)} k_u. \tag{4.3}$$

According to Feld [26], the mean among friends is empirical at least as great as the mean among individuals, i.e., friends seem always to have more friends than an individual.

AVERAGE CLUSTERING COEFFICIENT The clustering coefficient is divided into two different versions: a global and a local one. While the global clustering coefficient gives an indication of the tendency of the nodes to build triangles, i.e., three nodes connected to each other, the local clustering gives an indication of the embeddedness of a single node. Embeddedness means, how well is the neighborhood of a node connected, or how close is the neighborhood of a node to being a clique. The local measure is easy to calculate via

$$C_i = \frac{2\,|\{e_{j,k} \mid \nu_j, \nu_k \in N(\nu_i), e_{j,k} \in E\}|}{k_i(k_i - 1)}. \tag{4.4}$$

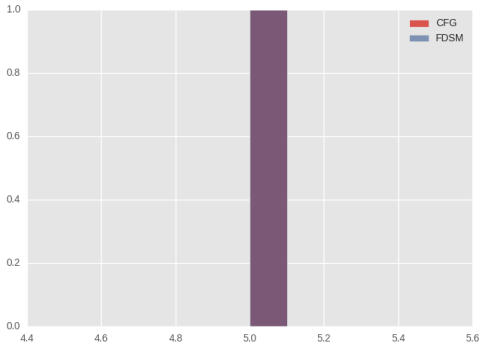
This measure is the number of triangles that ν_i is a part of divided by the number of possible triangles. We average this over all nodes in the largest component.

4.1 ERDŐS-RÉNYI GRAPH

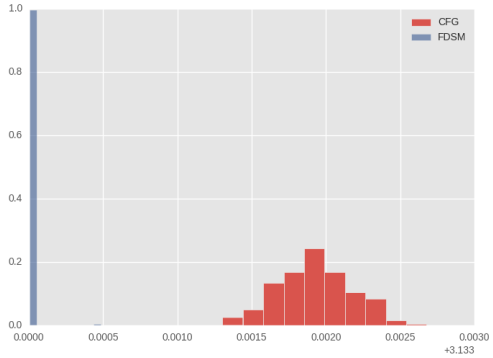
The first test will show that there are graphs that follow a degree distribution for which it is not important which null model is chosen. The graph that is used is generated with the Erdős-Rényi graph model. The Erdős-Rényi graph model, $\mathcal{G}(n, m)$ is similar to Gilbert's $\mathcal{G}(n, p)$ model, especially in the context of the number of edges attached to a node, i.e., the degree sequence. In the latter, two nodes build an edge with probability p, which implies that there are on average $\binom{n}{2}p$ edges. Graphs from the $\mathcal{G}(n, m)$ model are generated by choosing pairs of nodes at random and generating edges until m edges exist.

The graph generated with the $\mathcal{G}(n, m)$ model used has 7012 nodes and 78512 edges. It consists of one component, has an average degree of 22.41, shows neither assortative nor dis-assortative behavior, and the degree sequence follows a Poissonian degree distribution, as expected.

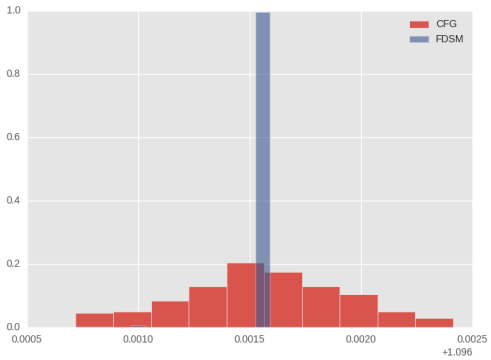
In Fig. 4.1 all results are shown. First, observe that the diameter is, in this case, a stable measure and it does not change based on the null model graphs are generated with. The average distance between nodes varies only to a very small extent (comp. Fig. 4.1b). The



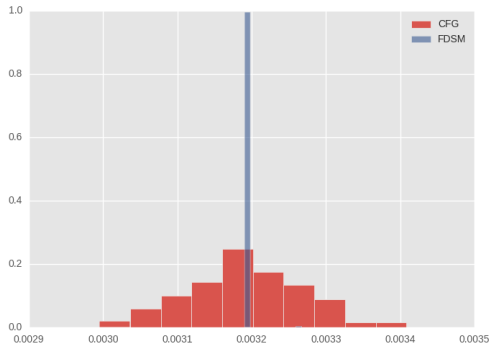
(a) Diameter around 5



(b) Average Distance FDSM exact 3.13, CFG around this value.



(c) Average Neighbor Degree FDSM exact 1.096, FDSM around this value.



(d) Clustering Coefficient FDSM exact 0.032, CFG around this value.

Figure 4.1: Histograms of measures on several hundred graphs generated with the two different based on a graph from the $\mathcal{G}(n, m)$ model.

results are very close, the only difference being that the average distance graphs generated with FDSM is always the same, while the average distance in graphs generated with CFG varies to a certain but negligible degree. The other measures, i.e., the average neighbor degree and the average clustering coefficient, do also show little variation in both null models.

Based on this example, it is not necessary to use the FDSM due to its run-time complexity. With a run-time in $\mathcal{O}(m \log m)$, the FDSM takes much longer than the CFG, that has a run-time in $\mathcal{O}(m)$. The question that remains is whether this is true for all graphs, i.e., is the CFG always the algorithm to choose or does it show problems when applied to graphs with a more skewed degree distribution.

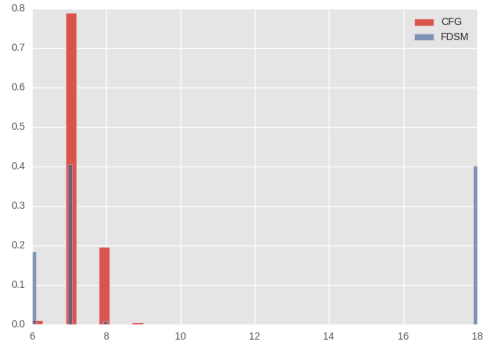
4.2 FOREST-FIRE GRAPH

The Forest-Fire graph algorithm, developed by Leskovec et. al. [55], works as follows: To some not further defined base graph a new node arrives. This new node chooses an ambassador it connects to from the existing graph. With probability p neighbors of this ambassador are added to the neighbors of the new node (i.e., they are "burnt"). This process is repeated recursively, until "the fire dies", i.e., no new nodes are added to the neighborhood. Already burned nodes cannot be burned by a node again.

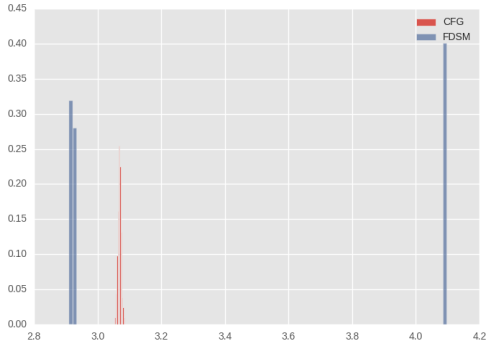
The graph generated with the Forest-Fire algorithm has 7012 nodes and 78572 edges, so it is very similar in this regard to the graph from the $\mathcal{G}(n, m)$ model from the former section. It consists of one component, has an average degree of 22.41, has a slightly dis-assortative tendency (-0.07), and the degree sequence follows a power law distribution with $\gamma = 1.79$. Recall, that the CFG connects stubs of nodes uniformly at random. The higher the degree of a pair of nodes, the more likely it is that they will have at least one, possibly several edges between them. Thus, this power law distribution will change the analysis based on the CFG, since there will be multiple edges and self-loop edges attached to the high degree nodes (more on this see Section 7.3) (and not contribute to the analysis).

Several points can be observed. First, observe that the graphs generated with the FDSM have in each figure in Fig. 4.2 an accumulation point that makes about 40% of all results. The results of the measures at this point are very close to the measures taken on the original graph, which indicates that the results of the graphs generated with the FDSM did change the graph only marginally or that the edge-swaps lead to the original artificial Forest-Fire graph. Second, observe that the results do only overlap in Fig. 4.2a, i.e., only the diameter is similar in graphs sampled with the two different models. For all other measures, there is a rather stark contrast between the graphs generated with the CFG from the graphs generated with the FDSM. Since this experiment is based on an artificial graph, this is not explained in more detail, but it is important to stress that the models do not necessarily yield the same results.

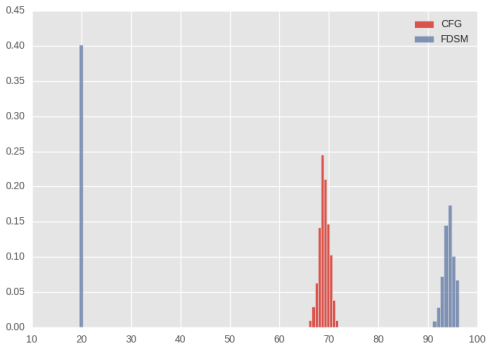
In the following section the SIS is explored more in depth, since the description of the algorithm given by Blitzstein and Diaconis [8] as well as the description by DelGenio et al. [21] does not investigate a rather important detail. This detail is studied more in the following.



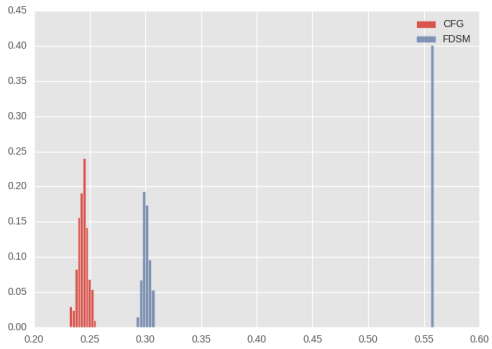
(a) Diameter



(b) Average Distance



(c) Average Neighbor Degree



(d) Clustering Coefficient

Figure 4.2: Histograms of measures based on several hundred of graphs generated with the two different models based on a graph from the Forest-Fire model.

5

SEQUENTIAL IMPORTANCE SAMPLING—WHICH PROBABILITY DISTRIBUTION TO USE?

As described in Section 2.1.4, SIS chooses the next neighbor based on the remaining degrees of the nodes in the set of possible neighbors. Blitzstein and Diaconis [8] suggested additionally using a uniform distribution, a discussion that was not introduced by Del Genio et al. [21]. Still, an analysis of the results that the different distributions yield was not performed to the best of my knowledge. Therefore, an investigation of the results based on the two different distributions is performed throughout the thesis. Here, a short preview based on the degree-assortativity is given.

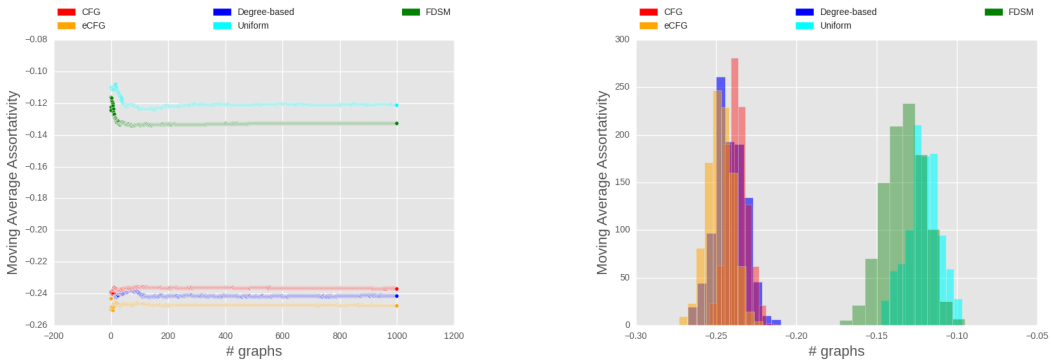
Even though the degree-assortativity coefficient was described as a somewhat strange measure due to the comparison to some assumed perfect graph in Section 3.3, it indicates to a certain degree the connectivity between nodes. Since this section is only about the probability distributions that is used to choose the neighbor of a node, it should be sufficient to compare the degree-assortativity coefficient of graphs generated with CFG, ECFG, and the FDSM with graph generated with the SIS using different probability distributions. The comparison is done once with the moving average

$$\mu_{av}(j) = \frac{1}{j}\sum_{i=1}^{j} \text{assortativity}(G_i), \tag{5.1}$$

where G_i is the i-th generated graph based on some null model. After a certain point there will be not much change in this average, indicating that most graphs do show similar assortativity values. Since the graphs generated with the FDSM are considered as ground truth, the closer the moving average of the degree-assortativity of any model is to the one of the FDSM, it can be assumed as more likely that other analyses do show similar results. As a second comparison, histograms of the degree-assortativity are used that will show the distribution of the degree-assortativity more clearly. Histogramms will show as well that the change in assortativity is rather small.

For this analysis a real-world graph is used that encodes protein-protein interactions; this graph has 418 nodes and 519 edges.

In Fig. 5.1a the moving average of the assortativity of graphs generated with SIS with uniform choosing of neighbors (USIS), the degree-based choosing of neighbors (DSIS), CFG, ECFG, and FDSM are compared. The graphs generated with DSIS do show a dis-assortative behavior very similar to graphs generated with the CFG or the ECFG. Graphs generated with a uniform probability when choosing a neighbor, USIS, are close to the graphs generated with the FDSM, such that the uniform sampling seems to be preferable. By comparing the distribution of the assortativity, see Fig. 5.1b, it can be seen that the FDSM produces a broader range of results than USIS, but they are very close to each other.



(a) Moving average of assortativity

(b) Histogram of assortativity

Figure 5.1: 5.1a the moving average of the degree-assortativity coefficient is plotted for graphs generated with five different algorithms; 5.1b histograms of the degree-assortativity coefficient for graphs generated with the different algorithms are shown. Overall, the ECFG, CFG, and DSIS show similar results, while the FDSM and USIS build another cluster of graphs similar regarding their assortativity.

Still, the number of samples is small (1000) compared to the full space of possible graphs such that it is hard to decide which probability distribution to support (recall that a smaller graph had about 10^{55} realizations). Since we consider as ground truth the FDSM, the uniform probability distribution to choose neighbors seems to produce graph more similar to that found in the FDSM. Still, testing both probability distributions with other measures will give insight which of the two probability distributions yields more likely results.

The graphs based on the non-uniform probability distribution do show a lower assortativity, which indicates that more connections are between (high, low)-degree pairs than otherwise. This means that measures based on graphs sampled with this choice of probability distribution will show different results from graphs build with a uniform distribution. For example, the average neighbor degree of low degree nodes in graphs generated with DSIS will be high; in graphs generated with USIS the average neighbor degree should be more evenly distributed.

This analysis is based on one example. Still, from the probability distribution to choose a neighbor together with the algorithmic description of how a graph is constructed, it is likely that there will be more edges between low degree and high degree nodes when a non-uniform probability distribution is chosen. This assumption is based on the description, that the node with the lowest degree starts searching neighbors, combined with choosing neighbors based on their degree. Therefore, based on the chosen distribution, the algorithm constructs either graphs that resemble more graphs built with CFG, which tends to connect any node more likely to high degree nodes, or graphs that resemble graphs drawn from FDSM.

6

ANALYSIS OF UNDIRECTED REAL-WORLD GRAPHS

Up to now, it was shown that different null models can influence the result of analysis. The graphs used for the comparison of the CFG and the FDSM were artificial. On graphs with unskewed degree-sequences, the faster CFG *seems* to be applicable without causing problems (for more details see [80]). However, on graphs with more skewed degree-sequences, the results are quite different from each other such that it is reasonable to test whether this is true for other graphs as well. Therefore, it is interesting to see how the null models perform in the analysis of real-world graphs. Before we start with the analysis of graphs, we describe the data used throughout this chapter.

6.1 DATASETS

There is an abundance of existing datasets and the possibility to extend these is almost limitless. The only real limits are the regulations of websites, social networks, games, and the capabilities of the computer one has to work with. For example, the complete social network of Facebook would be a great dataset to access, and it might even be possible to crawl it. Jernigan and Mistree crawled Facebook to gather all members of the MIT-subgraph [46]. They did this automatically with a self-written crawler "Arachne". Even though they did nothing to prevent this crawling from being detected they were not stopped, but Facebook stated that they had a monitoring system to detect "misuse". It is not sure if this activity was observed and classified as being "not misuse" or if it just was not detected. Edunov et al. calculated, based on some approximative measures, average distances for 1.6 billion users of Facebook [23]. However, they work for Facebook, so the problem of acquiring data is smaller.

The datasets used are from different fields of science.

A the network of email contacts provided in the Enron corpus (Email-Enron) [50, 56], four Facebook networks between students of Georgetown, Princeton, Oklahoma, and Caltech from Traud et al. [92], a Facebook excerpt by McAuley and Leskovec [60] ($\text{Facebook}_{\text{ML}}$), a Facebook-like social network compiled by Opsahl [73] are examples of **social** networks;

B networks between arXiv-authors of the Condensed Matter, High Energy Physics - Phenomenology (High Energy Physics), General Relativity, Astro-Physics, and the High Energy Physics - Theory (High Energy Physics Theory) category are examples of **collabortation-networks**[1];

1 http://www.snap.edu/data/index.html/canets

C *Saccharomyces cerevisiae* (S. cerevisiae) transcriptional regulation network, *Caenorhabditis elegans*[2] (C. elegans) neural network, the human-disease interaction network used by Goh et al. [30] (Diseases), and *Escherichia coli* transcriptional regulation network (E. coli)[3] are examples of **biological** networks. While transcriptional regulation networks are in general directed, since this analysis is only interested in the effect of graph generation algorithms on the measures, we discard direction between nodes.

Some general information about the graphs can be found in Table 6.1. Additionally, results of the measures applied in the following are displayed as well.

2 http://www-personal.umich.edu/~mejn/netdata/
3 http://www.weizman.ac.il/mcb/UriAlon/

Graph	n	m	skewness	diameter	avg. distance	avg. neighbor degree	co-occurrence
Email-Enron	36 692	183 831	16.30	13	4.03	236.43	25 566 893
Georgetown	9414	425 638	2.30	11	2.76	143.08	67 751 053
Princeton	6596	293 320	1.38	9	2.68	143.51	46 139 701
Oklahoma	17 425	892 528	3.49	9	2.77	185.99	194 235 901
Caltech	769	16 656	1.32	6	2.34	73.98	1 231 412
$\text{Facebook}_{\text{ML}}$	4039	88 234	4.52	8	3.69	105.55	9 314 849
Facebook-like	1899	13 838	4.25	8	3.06	69.79	755 882
Condensed Matter	23 133	93 497	5.76	15	5.35	17.32	1 972 996
High Energy Physics	12 008	118 521	5.02	13	4.67	42.07	15 280 441
General Relativity	5242	14 496	3.83	17	6.05	9.69	230 017
Astro Physics	18 772	198 110	3.85	14	4.19	44.16	12 755 612
High Energy Physics Theory	9877	25 998	3.02	18	5.95	10.11	300 167
S. cerevisiae	685	1051	6.36	15	5.23	21.57	13 017
C. elegans	297	2148	4.06	5	2.46	32.00	53 804
Diseases	867	1527	4.52	15	6.49	6.03	10 063
E. coli	418	519	9.35	13	4.82	14.72	5290

Table 6.1: Basic network statistics for the individual networks used in the article.

7

STABLE MEASURES

Up to now, the results of the analysis indicate that there are two different classes of results: one class contains algorithms and models that tend to connect nodes more likely to high degree nodes, such as the CFG, ECFG, and the DSIS; the other class of results contains algorithms based on a uniform choice, either of neighbors (USIS) or of edges in the rewiring process of edge-swap algorithms (FDSM). By the high degree of some nodes, they are longer available to connect to (USIS) or edges connected to them are more often chosen (FDSM), but the results of the previous sections on artificial graphs showed already differences between uniform and degree based choices. Since the FDSM has a better worst-case runtime than the USIS ($\mathcal{O}(m \log(m))$ vs. $\mathcal{O}(n^3)$), this is the preferred algorithm to generate graphs from the family of graphs with a fixed degree sequence. But the examples above were artificial and do not necessarily reflect real-world graphs. Therefore, an analysis similar to the one performed with the $\mathcal{G}(n, m)$ model and the Forest-Fire model is done with real world graphs more in depth. From each graph, between 150 and 1000 examples were generated with the respective model.

Recall, that the samples based on the graph from the $\mathcal{G}(n, m)$ model were quite close to each other with respect to the diameter and the average distance between all nodes. This is true for different types of random graph models, including the $\mathcal{G}(n, m)$ model [17], the Barabási-Albert graph model [9], and for graph models with a degree sequence following a powerlaw with $2 < \beta < 3$ with $\forall v \in v : k_v^2 \leq \sum_{u \in V} k_u$ [17]. For all of these models, the diameter of graphs generated with one of these models is very similar. The same is not necessarily true for real-world graphs, to which nodes may be added that connect to other nodes that have only few connections. These will initially increase the diameter, but when they become more integrated to the graph, the diameter will shrink again.

Thus, randomly generated versions of a real-world graph intuitively should have a lower diameter. This effect is similar to the one observed by Watts and Strogatz [100] in their seminal work that analyzed the diameter and the clustering coefficient in graphs that were regular and clearly structured and were rewired with increasing probability. They discovered that the higher the probability of rewiring an edge is, the smaller becomes the diameter. Similar effects are to be expected for real-world graphs and generated graphs based on the fixed degree sequence of the real-world graphs. Nodes that are more on the "outskirts" in the real-world graph, nodes with few neighbors and of low importance, may in a generated graph be connected to more central nodes and therefore, the diameter may shrink.

The same may happen in the CFG with the difference, that the process that generates a graph tends to connect high degree nodes to each other more than high and low degree nodes. Thus, remaining groups of low degree nodes may be left that can either be on the

outskirts of the giant component or they can build small, separate components. For the graphs under investigation this happened, but the giant component consisted, on average, out of 82% of the nodes of a graph (E. coli) up to 100% of the graph ($\text{Facebook}_{\text{ML}}$). It may happen, that the diameter of the small component is larger than the diameter of the giant component, but since the giant component is most of the time more important for analysis than the small components, analysis of small components is omitted in this thesis.

Considering the CFG, one has to always to remember that multiple edges between nodes or self-loops may be created. For the following discussion of the diameter and the average distance, this is of no importance, since the graphs are all considered as unweighted. Two edges between two nodes do not reduce or increase the number of steps a traveler on the graph can take, thus, the results of the CFG and ECFG coincide. Therefore, results of the ECFG are omitted.

7.1 DIAMETER

For the discussion of the diameter, in all generated graphs the diameter is measured. The results are then averaged for each graph generating null model, and the standard deviation is calculated. The model that does not generate graphs, SIM, has several options to approximate the expected diameter as shown in the following.

7.1.1 *Approximating the Diameter*

To approximate the diameter, one can assume to start at any node in the graph and follow from this node all possible paths to gather all distances and to calculate based on this the diameter. When there are no nodes, one has to make assumptions on the number of neighbors of each hypothetical node. The most simple approximation is to assume that each node has the same number of neighbors. In fact, this assumption reduces the model to a tree. Still, this simple assumption allows calculating an approximation of the diameter. The degree that is used to approximate the degree of a node is the average degree of a node in the graph before shown in equation 8.1. As a reminder,

$$AV(G) = \frac{1}{n}\sum_{v\in V} av(v) = \sum_{v\in V}\frac{1}{k_v}\sum_{u\in N(v)} k_u = \langle k\rangle.$$

For equations of the SIM, the notation of Newman is used, i.e., averages over node degrees are displayed as $\langle k^i\rangle = \frac{1}{n}\sum_{v\in V} k_v^i$.

Under the assumption that every node has on average this number of neighbors, one can sum over the neighbors at distance d and do so until we reach the number of nodes in the graph, i.e.,

$$n = \sum_{i=1}^{l}\langle k\rangle. \tag{7.1}$$

Reformulating this from a sum to an expression that is easily solvable for the expected diameter m is possible, using an index shift we get

$$n = \sum_{i=1}^{l} \langle k \rangle^i \tag{7.2}$$

$$= \sum_{i=0}^{l-1} \langle k \rangle^{i+1} \tag{7.3}$$

$$= \langle k \rangle \sum_{i=0}^{l-1} \langle k \rangle^i \tag{7.4}$$

$$= \langle k \rangle \frac{\langle k \rangle^{(l-1)+1} - 1}{\langle k \rangle - 1}. \tag{7.5}$$

Solving this equation for m yields

$$l = \frac{\log\left(\frac{n(\langle k \rangle - 1)}{\langle k \rangle} + 1\right)}{\log\left(\langle k \rangle\right)} \tag{7.6}$$

An easier approximation is the maximization of the inner term.

$$n = \langle k \rangle^l. \tag{7.7}$$

Solving this for m yields

$$l = \frac{\log(n)}{\log\left(\langle k \rangle\right)}. \tag{7.8}$$

The results of both approximations are shown in this order in the following as SIM_1 and SIM_2.

7.1.2 *Model comparison*

In Table 7.1 the average diameter and the standard deviation of the graphs generated with the FDSM and the CFG are shown. As expected, the results are very close to each other. This result shows, that even though graphs generated with the CFG do have multiple edges and do sometimes connect nodes to themselves instead of connecting to other nodes, the giant components of the graphs are similar in diameter to the graphs generated with the FDSM. On the other hand, the equations of SIM severely underestimate the diameter in all cases.

While the directly solvable equation for the diameter estimate, equation 7.8, is very close to the more involved equation 7.6, the result is always too low to be even considered as the diameter for the given graphs. The approach to estimate the diameter via equations does not work. The assumptions that have been made, i.e., each node having the same number of neighbors, all nodes are equal to a certain extent, are implausible for real-world graphs. The equations may work for graphs following a Poissonian distribution or trees, but for graphs that follow a skewed degree distribution, they do not work in general.

In Table 7.2 the average diameter and the standard deviation of graphs generated with the FDSM, the USIS, and the DSIS are shown. Remarkable are the results of graphs generated with DSIS. They are much lower than the results of the other algorithms but still closer to the

Graph	FDSM	CFG	SIM_1	SIM_2
Email-Enron	8.46 ± 1.10	8.49 ± 1.17	4.51	4.56
Georgetown	5.07 ± 0.26	5.07 ± 0.26	2.03	2.03
Princeton	5.06 ± 0.24	5.08 ± 0.27	1.96	1.96
Oklahoma	5.10 ± 0.30	5.09 ± 0.29	2.11	2.11
Caltech	4.85 ± 0.38	4.94 ± 0.34	1.76	1.76
$Facebook_{ML}$	5.10 ± 0.30	5.11 ± 0.31	2.19	2.20
Facebook-like	6.46 ± 0.54	6.75 ± 0.55	2.79	2.82
Condensed Matter	9.79 ± 0.77	9.78 ± 0.79	4.75	4.81
High Energy Physics	6.92 ± 0.50	6.93 ± 0.43	3.13	3.15
General Relativity	9.99 ± 0.77	10.10 ± 0.67	4.89	5.01
Astro Physics	7.07 ± 0.41	6.97 ± 0.39	3.21	3.23
High Energy Physics Theory	10.85 ± 0.79	10.85 ± 0.75	5.41	5.54
S. cerevisiae	10.70 ± 1.18	10.95 ± 1.11	5.47	5.82
C. elegans	5.04 ± 0.20	5.13 ± 0.34	2.11	2.13
Diseases	11.13 ± 0.92	11.27 ± 0.97	5.11	5.37
E. coli	10.57 ± 1.19	11.09 ± 1.20	6.07	6.64

Table 7.1: Average and standard deviation of the diameter in the ground truth model FDSM, the CFG, and the two equations of the SIM.

Graph	FDSM	USIS	DSIS
Email-Enron	8.46 ± 1.10	9.43 ± 0.57	5.01 ± 0.07
Georgetown	5.07 ± 0.26	5.47 ± 0.50	4.00 ± 0.00
Princeton	5.06 ± 0.24	5.30 ± 0.46	4.00 ± 0.00
Oklahoma	5.10 ± 0.30	5.71 ± 0.45	4.00 ± 0.00
Caltech	4.85 ± 0.38	4.99 ± 0.23	4.00 ± 0.00
$Facebook_{ML}$	5.10 ± 0.30	5.41 ± 0.49	4.00 ± 0.00
Facebook-like	6.46 ± 0.54	6.93 ± 0.48	4.34 ± 0.47
Condensed Matter	9.79 ± 0.77	10.43 ± 0.58	7.05 ± 0.21
High Energy Physics	6.92 ± 0.50	7.62 ± 0.55	4.33 ± 0.47
General Relativity	9.99 ± 0.77	11.03 ± 0.78	7.21 ± 0.40
Astro Physics	7.07 ± 0.41	7.73 ± 0.51	5.00 ± 0.00
High Energy Physics Theory	10.85 ± 0.79	12.20 ± 0.80	8.52 ± 0.51
S. cerevisiae	10.70 ± 1.18	11.07 ± 1.11	9.84 ± 0.95
C. elegans	5.04 ± 0.20	5.00 ± 0.00	4.00 ± 0.00
Diseases	11.13 ± 0.92	11.61 ± 0.73	9.82 ± 0.64
E. coli	10.57 ± 1.19	10.60 ± 1.13	11.74 ± 1.43

Table 7.2: Average and standard deviation of the diameter in the ground truth model FDSM, the USIS, and the DSIS.

average diameter of the FDSM than the results of the equations 7.6 resp. 7.8. The algorithm starts with low degree nodes and connects them according to a probability distribution based on the remaining degree of potential neighbors. Thus, in the beginning, it is more likely that low degree nodes connect to high degree nodes. This is similar to the CFG that connects any node to high degree nodes more likely. The difference is, that the DSIS starts with low degree nodes and connects them with higher probability to high degree nodes; on the other hand, the CFG tends to connect high degree nodes together. Thus, it is plausible that some star-like structures exist that are connected via higher degree nodes, and the average distance is small.

On the other hand, graphs generated with USIS have an average diameter that is close to the average diameter of graphs generated with the FDSM. When applying a two-sample z-test to the averages and standard deviations of FDSM and USIS with the assumption that there is no difference between the models, the results indicate that the difference between them does not only exist, but is in most cases rather large. The two-sample z-test assumes that both groups of results follow a normal distribution. That is not guaranteed for the diameter of graphs (or any graph-measure in particular) since the family of graphs is too large to know the exact distribution. Therefore, the Kolmogorov-Smirnov two-sample test is applied, that compares the cumulative distributions of the frequency of the results directly. The results of both tests are in Table 7.3.

Graph	$\|z\|$	D	p
Email-Enron	11.14	0.49	$2.14 \cdot 10^{-21}$
Georgetown	9.19	0.40	$5.38 \cdot 10^{-10}$
Princeton	5.97	0.24	$7.17 \cdot 10^{-4}$
Oklahoma	13.44	0.60	$1.33 \cdot 10^{-21}$
Caltech	3.23	0.13	$2.32 \cdot 10^{-1}$
$\text{Facebook}_{\text{ML}}$	6.75	0.31	$3.57 \cdot 10^{-6}$
Facebook-like	7.37	0.41	$2.14 \cdot 10^{-10}$
Condensed Matter	11.50	0.38	$2.66 \cdot 10^{-17}$
High Energy Physics	10.98	0.52	$2.73 \cdot 10^{-16}$
General Relativity	10.96	0.56	$7.53 \cdot 10^{-19}$
Astro Physics	12.06	0.58	$3.35 \cdot 10^{-20}$
High Energy Physics Theory	13.87	0.68	$1.12 \cdot 10^{-27}$
S. cerevisiae	3.18	0.17	$7.48 \cdot 10^{-3}$
C. elegans	2.04	0.04	$10.00 \cdot 10^{-1}$
Diseases	4.49	0.24	$7.17 \cdot 10^{-4}$
E. coli	0.36	0.03	$9.39 \cdot 10^{-1}$

Table 7.3: Two-sample z-test results, the Kolmogorov-Smirnov two-sample test result and its p-value for the diameter of the graphs generated with the FDSM and the USIS.

Only for three of the graphs under investigation, the difference between the graphs generated with the FDSM and USIS are acceptable as very close to each other (E. coli, C. elegans, Caltech); for all other graphs, the two algorithms give very different results. While the z-

scores show all but the E. coli graph and the C. elegans graph as rather implausible ($z < 3$), the Kolmogorov-Smirnoff test says differently. Recall, that

$$D < c(\alpha) \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

indicates, that it is possible that the two distributions are the same; since this is harder to interpret than a p-value, here also these are shown. A p-value below a certain threshold, usually 0.05, indicates that the distributions are different. Thus, the Caltech graphs with a p-value of 0.23 yield also diameters that are from the same distribution, i.e., the algorithms could be used interchangeably. Nevertheless, the rest of the tests show different results. Therefore, neither probability distribution to chose a neighbor of the SIS gives results that are reasonably close to the results of the FDSM, which are considered as ground truth.

7.2 DISTANCE

As before, in all generated graphs distances are measured in the giant component and then averaged per node and then per graph. For the SIM, the equations are developed in the following.

7.2.1 *Approximating the Distance*

Again, the equation that is used to approximate the expected average distance for graphs has to be constructed based on the SIM. For this, Newman defines the average number of neighbors at distance i as

$$c_i = \left(\frac{\langle k^2 \rangle}{\langle k \rangle} \right)^{i-1} \langle k \rangle. \tag{7.9}$$

With equation 7.9 it is possible to derive an equation such as

$$n = 1 + \sum_{i=1}^{d} c_i = 1 + \langle k \rangle \sum_{i=1}^{d} \left(\frac{\langle k^2 \rangle}{\langle k \rangle} \right)^{i-1}, \tag{7.10}$$

which equates the number of nodes in a graph with the number of neighbors of a single node at any distance plus this node. As before, c_i is an approximation, therefore, equation 7.10 is also an approximation. With an index shift we derive

$$1 + \langle k \rangle \sum_{i=1}^{d} \left(\frac{\langle k^2 \rangle}{\langle k \rangle} \right)^{i-1} = 1 + \langle k \rangle \sum_{i=0}^{d-1} \left(\frac{\langle k^2 \rangle}{\langle k \rangle} \right)^{i} \tag{7.11}$$

$$= 1 + \langle k \rangle \frac{\left(\frac{\langle k^2 \rangle}{\langle k \rangle} \right)^{(d-1)+1} - 1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}. \tag{7.12}$$

This equation can be solved for d, such that

$$d = \frac{\log\left(\frac{n-1}{\langle k \rangle} \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right) + 1 \right)}{\log\left(\frac{\langle k^2 \rangle}{\langle k \rangle} \right)}. \tag{7.13}$$

Alternatively, it is possible to calculate the maximum distance in which a number of nodes as large as n is,

$$n = c_d = \left(\frac{\langle k^2 \rangle}{\langle k \rangle}\right)^{d-1} \langle k \rangle. \tag{7.14}$$

Solving this equation for d yields

$$d = \frac{\log\left(\frac{n}{\langle k \rangle}\right)}{\log\left(\frac{\langle k^2 \rangle}{\langle k \rangle}\right)} + 1. \tag{7.15}$$

How these equations perform, after seeing the former set of equations failed, is shown in the following

7.2.2 *Model comparison*

Graph	FDSM	CFG	SIM_1	SIM_2
Email-Enron	1.59 ± 1.59	1.62 ± 1.62	2.66	2.66
Georgetown	2.29 ± 0.01	2.32 ± 0.01	1.91	1.92
Princeton	2.23 ± 0.02	2.24 ± 0.02	1.85	1.85
Oklahoma	2.35 ± 0.01	2.36 ± 0.01	1.95	1.95
Caltech	2.05 ± 0.06	2.10 ± 0.05	1.66	1.67
$Facebook_{ML}$	2.44 ± 0.03	2.39 ± 0.03	1.97	1.97
Facebook-like	2.60 ± 0.08	2.65 ± 0.07	2.21	2.21
Condensed Matter	4.07 ± 0.03	4.03 ± 0.03	3.56	3.57
High Energy Physics	2.73 ± 0.03	2.77 ± 0.03	2.32	2.32
General Relativity	3.88 ± 0.08	3.91 ± 0.08	3.40	3.43
Astro Physics	3.05 ± 0.02	3.06 ± 0.02	2.62	2.62
High Energy Physics Theory	4.55 ± 0.09	4.56 ± 0.09	3.95	3.98
S. cerevisiae	3.25 ± 0.65	3.43 ± 0.73	3.05	3.08
C. elegans	2.29 ± 0.10	2.41 ± 0.12	1.91	1.93
Diseases	4.32 ± 0.38	4.42 ± 0.33	3.65	3.72
E. coli	2.83 ± 0.43	2.78 ± 0.45	3.08	3.12

Table 7.4: Average and standard deviation of the average distance in the ground truth model FDSM, the CFG, and the two equations of the SIM.

The averages of the average distance of graphs generated with FDSM, the CFG, and the corresponding equations of the SIM are shown in Table 7.4. As with the diameter, the equations to estimate the average distance deliver results that misestimate the expected value, but for the S. cerevisiae, C. elegans, Diseases, and the E. coli graph. For these four, the z-score indicates that the equations could replace the sampling approach.

The differences between the CFG and the FDSM are small. In fact, they tend to be about the size of the standard deviation. When applying the two-sample z-score, the result for almost all graphs indicates that the distribution is relatively close, i.e., they can be considered as distributions with the same mean and standard deviation, which implies that the graphs

Graph	z	D	p
Email-Enron	0.01	0.21	$2.34 \cdot 10^{-4}$
Georgetown	1.23	0.64	$6.66 \cdot 10^{-19}$
Princeton	0.37	0.27	$1.03 \cdot 10^{-3}$
Oklahoma	0.63	0.36	$2.85 \cdot 10^{-6}$
Caltech	0.60	0.46	$5.70 \cdot 10^{-10}$
$\text{Facebook}_{\text{ML}}$	1.07	0.55	$4.52 \cdot 10^{-14}$
Facebook-like	0.45	0.35	$5.96 \cdot 10^{-6}$
Condensed Matter	0.93	0.55	$6.63 \cdot 10^{-27}$
High Energy Physics	0.95	0.52	$1.27 \cdot 10^{-12}$
General Relativity	0.31	0.24	$5.04 \cdot 10^{-3}$
Astro Physics	0.27	0.24	$5.04 \cdot 10^{-3}$
High Energy Physics Theory	0.04	0.10	$6.77 \cdot 10^{-1}$
S. cerevisiae	0.18	0.21	$2.05 \cdot 10^{-2}$
C. elegans	0.79	0.54	$1.40 \cdot 10^{-13}$
Diseases	0.20	0.21	$2.05 \cdot 10^{-2}$
E. coli	0.08	0.12	$3.67 \cdot 10^{-1}$

Table 7.5: Two-sample z-test results, the Kolmogorov-Smirnov two-sample test result and its p-value for the average distance of the graphs generated with the FDSM and the CFG.

should be similar. When applying a Kolmogorov-Smirnoff two-sample test, the results look quite different. Almost all D values are larger than the critical value $c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}$, and the p-values indicate that that only the E. coli and the High Energy Physics Theory graphs have samples with similar distances (see Table 7.5).

The differences between the USIS, DSIS, and the FDSM do not merit analysis with the two-sample z-score or the Kolmogorov-Smirnov two-sample test. The standard deviations are very small, such that the mean value of the average distance should coincide to be from the same distribution, which it does not. Therefore, the z-score will not show them as very similar; the same applies for the Kolmogorov-Smirnoff two-sample test. Recall that the maximum difference between the two cumulative distributions is the value D that is compared to $c(\alpha)$ with some modifier specific to the sample sizes. With such small standard deviations, there is a large gap between the plots of the two cumulative distribution functions such that D will be large and the p-value small (see Table 7.7). As can be observed in Table 7.7, only half of the rows have a z-score entry small enough such that one might consider the distribution of the average distance from the USIS similar to that in the FDSM. Based on this, also the graphs should not be too similar. The large D-values, all almost reaching the maximum value of 1, together with the low p-values, strengthen this assumption.

Interestingly, the difference between the two versions of the SIS is rather small in some cases (e.g. Email-Enron, Condensed Matter), in some cases, it is large (e.g. E. coli, Oklahoma), such that a test between these can only be considered as inconclusive.

Graph	FDSM	USIS	DSIS
Email-Enron	1.59 ± 1.59	3.68 ± 0.00	3.65 ± 0.00
Georgetown	2.29 ± 0.01	2.48 ± 0.00	2.35 ± 0.00
Princeton	2.23 ± 0.02	2.44 ± 0.00	2.30 ± 0.00
Oklahoma	2.35 ± 0.01	2.58 ± 0.00	2.42 ± 0.00
Caltech	2.05 ± 0.06	2.25 ± 0.00	2.12 ± 0.00
$\text{Facebook}_{\text{ML}}$	2.44 ± 0.03	2.60 ± 0.00	2.43 ± 0.00
Facebook-like	2.60 ± 0.08	3.05 ± 0.01	2.88 ± 0.01
Condensed Matter	4.07 ± 0.03	4.36 ± 0.00	4.32 ± 0.00
High Energy Physics	2.73 ± 0.03	3.21 ± 0.00	3.18 ± 0.01
General Relativity	3.88 ± 0.08	4.33 ± 0.01	4.25 ± 0.00
Astro Physics	3.05 ± 0.02	3.40 ± 0.00	3.33 ± 0.00
High Energy Physics Theory	4.55 ± 0.09	4.86 ± 0.01	4.86 ± 0.00
S. cerevisiae	3.25 ± 0.65	4.07 ± 0.04	4.22 ± 0.03
C. elegans	2.29 ± 0.10	2.39 ± 0.01	2.27 ± 0.01
Diseases	4.32 ± 0.38	4.69 ± 0.05	4.74 ± 0.02
E. coli	2.83 ± 0.43	4.05 ± 0.09	4.63 ± 0.10

Table 7.6: Average and standard deviation of the average distance in the ground truth model FDSM, the USIS, and the DSIS.

Graph	$\lvert z \rvert$	D	p
Email-Enron	1.31	1.00	$1.42 \cdot 10^{-89}$
Georgetown	12.91	1.00	$3.07 \cdot 10^{-60}$
Princeton	12.29	1.00	$3.07 \cdot 10^{-60}$
Oklahoma	16.69	1.00	$1.99 \cdot 10^{-52}$
Caltech	3.07	1.00	$3.07 \cdot 10^{-60}$
$\text{Facebook}_{\text{ML}}$	4.47	1.00	$3.07 \cdot 10^{-60}$
Facebook-like	5.53	1.00	$1.57 \cdot 10^{-60}$
Condensed Matter	9.65	1.00	$1.42 \cdot 10^{-89}$
High Energy Physics	18.02	1.00	$3.07 \cdot 10^{-60}$
General Relativity	5.85	1.00	$3.07 \cdot 10^{-60}$
Astro Physics	18.38	1.00	$3.07 \cdot 10^{-60}$
High Energy Physics Theory	3.49	1.00	$3.07 \cdot 10^{-60}$
S. cerevisiae	1.26	0.90	$7.10 \cdot 10^{-49}$
C. elegans	0.95	0.83	$1.24 \cdot 10^{-41}$
Diseases	0.96	0.84	$1.25 \cdot 10^{-42}$
E. coli	2.74	0.98	$1.07 \cdot 10^{-92}$

Table 7.7: Two-sample z-test results, the Kolmogorov-Smirnov two-sample test result and its p-value for the average distance of the graphs generated with the FDSM and the USIS.

7.3 IMPLICATIONS

Diameter and distance are basic measures. Still, they contain much information about a graph and can be of vital importance, depending on the desired outcome. If one wants to calculate centralities, many of these include calculating either directly or indirectly distances as well. The betweenness centrality is the number of shortest path between all pairs of nodes that go through a node divided by the number of all shortest path between all pairs of nodes, the closeness centrality of a node is defined as the inverse of the mean of the geodesic distances, flow-analyzing algorithms most surely need connections and are related to distances as well [85, 91]. The diameter is used not as often in algorithms, but is still an important measure.

The analyses based on real-world graphs confirm what was shown for artificial graphs. The average distances and the average diameter are for most graph models very similar, for the ECFG and the CFG they are even the same. Most surprising in this part of research is that the results of the models containing only simple graphs and the results of the model containing multi-graphs are close together. In fact, the two-sample z-score yields for some measures small enough values, such that one may believe it not important which model is used to sample graphs from. Only an analysis based on the Kolmogorov-Smirnoff two-sample test shows that the distribution of some measure is not distributed the same.

Judging from the results of the analysis so far, it appears as if it is not important which model is used to compare a real-world graph to when global measures are used. The results of the FDSM and the CFG are always close to each other. The values of the SIS are not as satisfactory—while the diameter measured in the USIS is similar to the diameter measured in the FDSM, the average distance of both variants of the SIS do appear to be very different from the other models. Since the worst-case run-time of the SIS is the worst of the compared algorithms, it is not likely to be used in an in-depth analysis. Still, this behavior was not observed beforehand. It may be interesting to analyze this as well. Differently, the SIM underestimates both measures severely, since the approximations assume that nodes are, on average, the same.

Thus, the CFG and FDSM both can be used. Since the CFG is the faster algorithm of the two, the CFG would be the algorithm of choice. At least this result is what one would have to believe if research would stop here. In the following, one interesting problem regarding the CFG is investigated.

7.4 MULTIPLE EDGES AND SELF-LOOPS

One has to be aware that the CFG does generate multiple edges and self-loops not uniformly at random. The higher the degree of a node, the more likely it is that a self-loop is attached to it. The same is valid for high degree pairs of nodes; it is more likely that two nodes of high degree build more multiple edges between them than other pairs. It can even occur that it is more likely that more than two edges between a pair of high degree nodes exist than between high and low degree nodes. A first analysis of this was given by Schlauch et al. [80].

In Figure 7.1 are two nodes, both with degree $n-2$, and $n-2$ nodes, each with degree 2. There are several things to be said about this graph. First of all, there are only very few realizations when one is restricted to simple graphs. Second, a graph generated with CFG

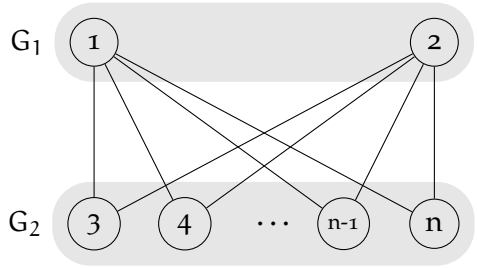


Figure 7.1: Example graph for which the CFG and the SIM yield many multiple edges and self-loops.

is likely to have multiple edges between the high degree nodes and self-loops attached to the high degree nodes.

Following the argumentation of Newman [66, p.441f], the expected number of multiple edges can be calculated. This is based on the simple independence assumption; the expected number of edges between two nodes u, v can be calculated with the following equation

$$\frac{k_u k_v}{2m} \frac{(k_u - 1)(k_v - 1)}{2m}. \tag{7.16}$$

This equation, when $\forall v \in V : k_v^2 < \sum_{v \in V} k_v$, is the probability of a first edge times the probability of a second edge. Since real-worlds graphs do not necessarily follow this restriction, this is the expected number of a first edge times the likelihood that a second edge exists.

Summing this over all pairs of nodes yields

$$\frac{1}{2(2m)^2} \sum_{u,v} \frac{k_u k_v}{2m} \frac{(k_u - 1)(k_v - 1)}{2m} = \frac{1}{2(2m)^2} \sum_u k_u (k_u - 1) \sum_v k_v (k_v - 1) \tag{7.17}$$

$$= \frac{1}{2\langle k \rangle^2 n^2} \sum_u k_u (k_u - 1) \sum_v k_v (k_v - 1) \tag{7.18}$$

$$= \frac{1}{2} \left[\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right]^2. \tag{7.19}$$

To calculate the number of self-loops, a similar approach is taken. The "probability" that a node u has a self-loop is given by

$$p_{u,u} = \frac{k_u (k_u - 1)}{2m}. \tag{7.20}$$

Thus, the expected number of self-loops can be calculated via summing over all nodes

$$\sum_u \frac{k_u (k_u - 1)}{2m} = \frac{\langle k^2 \rangle - \langle k \rangle}{2\langle k \rangle}. \tag{7.21}$$

Furthermore, it is possible to estimate exactly what is contributed by nodes of which degree. For the example in Fig. 7.1 this is done explicitly in the following.

$$\mathit{selfloops} = \frac{\langle k^2 \rangle - \langle k \rangle}{2\langle k \rangle} = \frac{n-2}{4} \tag{7.22}$$

$$\mathit{multiple\ edges} = \frac{1}{2} \left[\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right]^2 = \frac{(n-2)^2}{8} \tag{7.23}$$

$$\text{selfloops}_{k=2} = (n-2)\,\frac{2}{4m} = \frac{2\,(n-2)}{4\,(2\,(n-2))} = \frac{1}{4} \tag{7.24}$$

$$\text{selfloops}_{k=n-2} = 2\frac{(n-2)\,(n-3)}{4m} = \frac{(n-2)\,(n-3)}{2\,(2\,(n-2))} = \frac{n-3}{4} \tag{7.25}$$

$$\begin{aligned}\text{multiple edges}_{k=2} &= (n-2)\,\frac{1}{2}\frac{2}{(2m)^2}\sum_i \left(k_i^2 - k_i\right) \\ &= \frac{n-2}{2\,(2\,(n-2))}\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} = \frac{n-2}{8}\end{aligned} \tag{7.26}$$

$$\begin{aligned}\text{multiple edges}_{k=n-2} &= 2\frac{(n-2)\,(n-3)}{(2m)^2}\sum_i \left(k_i^2 - k_i\right) \\ &= \frac{(n-2)\,(n-3)}{2\,(2\,(n-2))}\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} = \frac{n^2-5n+6}{8}\end{aligned} \tag{7.27}$$

It can be observed, that nodes of degree 2 would contribute together only $\frac{1}{4}$ of all self-loops (each node contributes thus $\frac{1}{4(n-2)}$), while the two nodes of degree $n-2$ would contribute $\frac{n-3}{4}$ to the estimated number of self-loops, most of the possible self-loops would therefore be contributed by the two high degree nodes. The same behavior can be observed for multiple edges; here, the two large nodes would contribute again much more than the rest of the graph's nodes.

This behavior was shown by Schlauch et al. [81]. It was most likely not discovered before because the CFG and the SIM were developed by physicists that were concerned with graphs that had a Poissonian distribution. Therefore, the first investigation is shown in the following.

To investigate this behavior, for a graph generated with the CFG it is measured how many edges each node would loose if it would have been created with the ECFG, i.e., we measure the change in the degree of each node. This can be written as

$$f(k) = \frac{\sum_{v\in V, k_v=k} k_{obs}}{k\,|\{v \mid k_v = k\}|}. \tag{7.28}$$

Equation 7.28 adds the remaining degree of all nodes of desired degree k up and divides by k times the number of nodes that are supposed to have degree k. This equation results in values between 0, i.e., none of the nodes with wanted degree k has any edge, and 1, which implies that all nodes realized all edges. In Fig 7.2 the average, minimal, and maximal edgeloss are denoted as measured with equation 7.28 on graphs from the $\mathcal{G}(n,m)$-model. From this plot, it is clearer why the CFG or the SIM have such a high appeal to anybody who deals with Poissonian degree distributions. The average edgeloss is for each degree always below 1%, the maximum edge loss occurs on the low degree nodes but still only reaches $\sim 7\%$. An average edgeloss less than 1% indicates that almost each node realized its wanted degree, thus almost all edges were realized, thus the loss of edges is rare. Thus, the CFG is a perfect candidate to use to analyze graphs that have a degree sequence that follows a Poissonian distribution. The CFG is fast and almost exact for these graphs with slight losses that tend to occur more often on high degree nodes (see the averages), but these variations are hardly worth the trouble of using a more elaborate model.

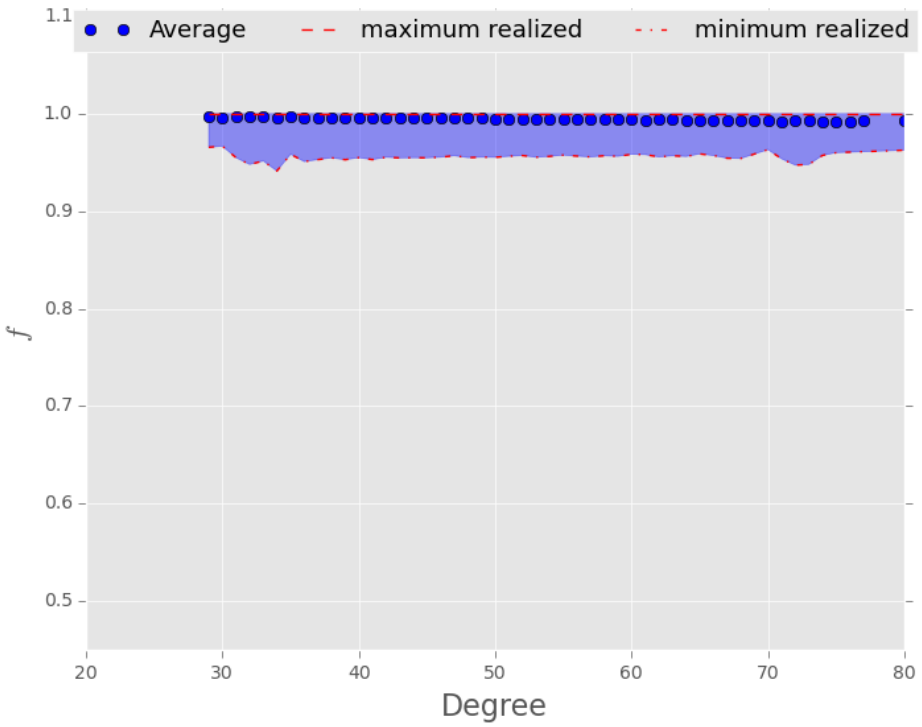
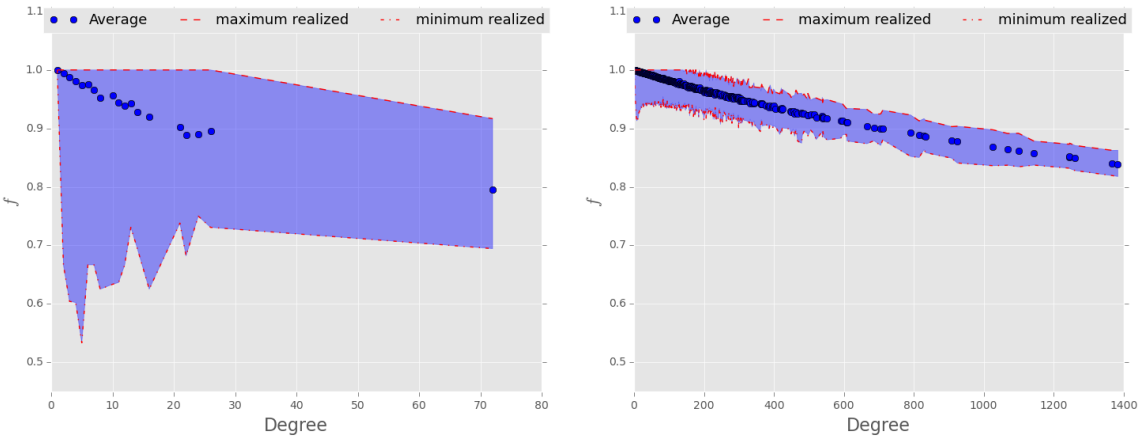


Figure 7.2: Results of measuring edgeloss per degree, measured on 200 graphs with Poissonian degree distribution, $\mathcal{G}(5000, 127754)$. The red striped line is the maximum edgeloss; the red dotted line is the minimum edgeloss, and the dots are the average edgeloss.



(a) E. coli edgeloss

(b) Email Enron edgeloss

Figure 7.3: Results of measuring edgeloss per degree, measured on 200 graphs. The red striped line is the maximum edgeloss; the red thin dotted line is the minimum edgeloss and the dots are the average edgeloss. In Fig 7.3a the graphs generated are based on the E. coli graph (skewness 9.34), in Fig. 7.3b the graphs generated are based on the Email-Enron graph (skewness 16.3).

Poissonian degree distributions are unskewed. What happens when the underlying probability distribution of a graph is more skewed? This is shown in the following on two different graphs.

In Fig. 7.3a the edgeloss in graphs generated with the CFG with the degree sequence of the E. coli graph is plotted. The underlying distribution of this degree sequence can be considered as a power law distribtuion $k^{-\gamma}$ with $\gamma = 3.04$. It is easy to note that the edgeloss is very different from the edgeloss shown in Fig 7.2. The highest degree node, a node with degree 72, has an average edgeloss of about 20% and in none of the generated graphs its degree was fully realized. The other nodes do also lose edges, and it is to observe that the low degree nodes do compensate better on average for occurring edgeloss. Compensate in the sense, that even though a node of degree 2 may lose one or both edges, the sum of the remaining degrees of nodes with wanted degree 2 tends to be closer to the sum of the wanted degrees, since there are more nodes of this degree than of the high degree nodes.

In Fig 7.3b, the underlying degree distribution of the Email-Enron graph is a power law distribution with exponent $\gamma = 2.1$, and the maximum degree is significantly higher (1383) than in the E. coli graph. The edgeloss is much more severe in total, even though the edgeloss for the maximum degree node is on average only $\sim 15\%$. Almost all nodes with degrees over 100 did not realize their full degree in any of the samples drawn with the CFG.

In Table 7.8, the expected edgeloss by self-loops and multiple edges (equation 7.19, resp. 7.20) is shown together with the observed number of self-loops and multiple edges in all graphs. The results of the equations are very good estimates of the number of "bad" edges in the samples.

To show that multiple edges occur more often between high degree node-pairs instead of high-low degree node pairs, we used a reformulation of equation 7.19. Instead of summing

Network	average # self-loops		average # multiple edges	
	measured	expected	measured	expected
Email-Enron	68	70	3944	4830
Georgetown	80	80	6274	6320
Princeton	78	79	6111	6162
Oklahoma	109	109	11 611	11 772
Caltech	38	37	1279	1332
$\text{Facebook}_{\text{ML}}$	53	53	2608	2756
Facebook-like	28	27	678	729
Condensed Matter	11	11	122	110
High Energy Physics	65	64	3875	4096
General Relativity	8	8	71	56
Astro Physics	32	32	1053	1024
High Energy Physics Theory	6	6	40	30
S .cerevisiae	6	6	38	36
C. elegans	13	13	144	156
Diseases	3	3	13	9
E. coli	5	5	22	25

Table 7.8: Average number of self-loops and multi-edges for samples from the CFG. The table also contains their expected value calculated with Equation (7.19), resp. Equation (7.20).

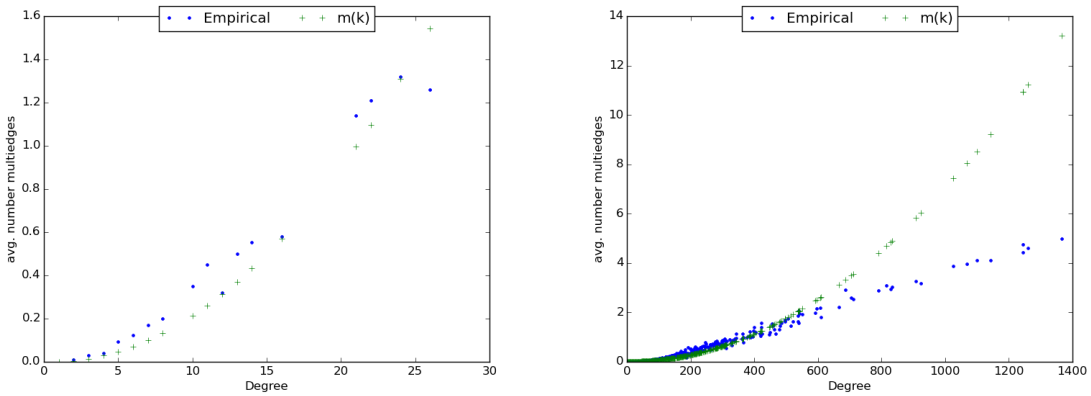
over all nodes, we compute for the highest degree node the number of possible multiple edges, i.e.

$$\begin{aligned} m(k) &= \frac{1}{2}\frac{k_{max}\cdot k}{2m}\frac{(k_{max}-1)(k-1)}{2m} \\ &= \frac{\left(k_{max}^2-k_{max}\right)}{2(4m^2)}\left(k^2-k\right) \end{aligned} \tag{7.29}$$

This changes the equation such that plugging in valid values of the degree sequences directly provides information on how many multiple edges between the maximum degree node and all other nodes are to be expected. Observing the same in samples drawn with the CFG, it is then possible to plot them against each other. This is done for the E. coli and the Email Enron graphs.

In Fig. 7.4a the observed number of multiple edges for the E. coli degree sequence is plotted against the calculated number of expected multiple edges via $m(k)$. Equation 7.29 is a good estimate for the number of multiple edges generated by the CFG based on this degree sequence. Moreover, it is obvious that the higher the degree, the higher is the number of multiple edges built to nodes of that degree.

For the Email-Enron graphs degree sequence the results are similar. The higher the node degree, the more multiple edges are to be expected. The observations confirm this for this graph as well, but only to a certain extent. The higher the degree of the node, i.e., the higher k is chosen, the larger the difference between the observed and the expected number of multiple edges becomes. This is most likely reasoned by the quadratic influence of k



(a) E. coli

(b) Email-Enron

Figure 7.4: Comparative plots of the number of multiple edges of the node with maximum degree and other nodes between observed values and expected values by $m(k)$. For Fig. 7.4a, the degree sequence of the E. coli graph was used to generate sample graphs with CFG, for Fig 7.4b the Email Enron graph was used.

(as well as k_{max}) that results in an overestimate of the importance of high degree nodes. Therefore, the estimation of equation 7.19 might be a good approximation in sum, but for the single nodes it will be strongly influenced by the node's degree. This can actually be observed for the low degree nodes as well, even if not as strongly (since it is averaged for the sampled graphs). For low degree nodes the plot shows that the generated graphs do actually show slightly more multiple edges than the equations expected.

These observations were not made before and are very important. Considering other measures that are calculated on graphs from the CFG, the number of multiple edges between high degree nodes can influence the result. In the coming section it is shown that this influences analysis in a significant way on real-world graphs.

8

SENSITIVE MEASURES

Before, it was shown that the analysis of real-world graphs is not much influenced by the algorithm graphs are generated with or which graph family they belong to. Now, a closer look at a local level, i.e., measures that are taken based on single nodes, may reveal a different picture. The first example is the average neighbor degree.

8.1 AVERAGE NEIGHBOR DEGREE

The average neighbor degree is a good example that is easy to calculate on graphs. The average neighbor degree of a node v is defined as the sum of its neighbors $N(v)$ degrees divided by its own degree, i.e.,

$$AV(v) = \frac{1}{k_v} \sum_{u \in N(v)} k_u. \tag{8.1}$$

This, summed over all nodes of a graph, divided by the number of nodes in the graph is then the average of the average, or the global measure of the average neighbor degree.

$$AV(G) = \sum_{v \in V} AV(v) = \sum_{v \in V} \frac{1}{k_v} \sum_{u \in N(v)} k_u \tag{8.2}$$

This definition is fine for simple graphs. The definition of this measures for multigraphs is, to the best of my knowledge, not clear.

In Fig. 8.1, the problem that can occur in multigraphs is shown. Node v has a degree of 2, but only one neighbor u. In a simple graph, it would have two different neighbors and the average neighbor degree of v would be simple to calculate. In a multigraph, there are at least two different options. First, the neighbors are considered as a set, i.e., each neighbor appears at most once. This implies that the example in Fig. 8.1 yields either $\frac{5}{2}$ (i.e., neighbor degree divided by the actual degree of v) or $\frac{5}{1}$ (i.e., erasing multiple edges as

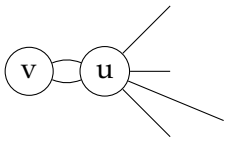


Figure 8.1: Example of a multigraph. The question is, how to calculate the average neighbor degree of v.

in the ECFG). Alternatively, each edge's endpoint can be considered as a separate neighbor. For the example in Fig. 8.1 this implies that $AV(v) = \frac{5+5}{2}$. We will use two of the proposed methods to measure the average neighbor degree. Once, to be consistent with the approach from before, we will use the approach to erase multiple edges and self-loops and calculate for the resulting graphs from the ECFG the average neighbor degree. Furthermore, the last presented method, which considers each endpoint of a node as a new neighbor, is used.

8.1.1 *Approximating the Average Neighbor Degree*

The SIM has capabilities to estimate the average number of neighbors as well. Newman provides for this purpose an equation [66, p.447]

$$AV_{SIM}(G) = \sum_k k \frac{kp_k}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle}. \tag{8.3}$$

p_k is the total fraction of nodes with degree k in a network, the total number of nodes with degree k is np_k. Given this fraction, the product np_k is the number of nodes with degree k. The probability that any edge connects to a particular node with degree k is then $\frac{knp_k}{2m} = \frac{kp_k}{\langle k \rangle}$.

8.1.2 *Model comparison*

Graph	FDSM	CFG	ECFG	SIM
Email-Enron	156.10 ± 0.77	140.08 ± 0.79	128.61 ± 0.82	140.08
Georgetown	161.41 ± 0.29	160.19 ± 0.27	155.86 ± 0.30	160.18
Princeton	159.38 ± 0.32	158.29 ± 0.32	153.25 ± 0.29	158.30
Oklahoma	221.39 ± 0.59	218.60 ± 0.35	212.41 ± 0.36	218.62
Caltech	77.26 ± 0.55	74.91 ± 0.54	66.55 ± 0.45	74.93
$\text{Facebook}_{\text{ML}}$	111.57 ± 0.51	106.44 ± 0.51	99.00 ± 0.47	106.57
Facebook-like	60.56 ± 0.67	55.68 ± 0.76	49.93 ± 0.62	55.62
Condensed Matter	22.17 ± 0.09	22.10 ± 0.08	22.02 ± 0.09	22.10
High Energy Physics	137.73 ± 0.55	129.91 ± 0.64	121.36 ± 0.60	129.93
General Relativity	17.00 ± 0.13	16.85 ± 0.12	16.67 ± 0.15	16.87
Astro Physics	65.87 ± 0.19	65.37 ± 0.20	64.59 ± 0.20	65.39
High Energy Physics Theory	12.56 ± 0.06	12.54 ± 0.06	12.50 ± 0.06	12.55
S. cerevisiae	14.77 ± 0.35	13.42 ± 0.36	12.19 ± 0.52	13.39
C. elegans	27.88 ± 0.62	26.04 ± 0.46	22.36 ± 0.50	26.05
Diseases	7.71 ± 0.15	7.58 ± 0.15	7.42 ± 0.15	7.59
E. coli	13.08 ± 0.41	11.14 ± 0.61	9.69 ± 0.76	11.19

Table 8.1: Average and standard deviation of the average neighbor degree in the ground truth model FDSM, the CFG, the ECFG, and the corresponding equations of the SIM.

In Table 8.1 the means and standard deviations of the FDSM, the variants of the CFG and the estimated average neighbor degree with the corresponding equation 8.3 of the SIM are

shown. For almost all graphs it is evident that the results do not coincide. This implies that the samples these measures were taken on are not from the same distribution. The ECFG underestimates for most graphs the mean of the average neighbor degree, which is not surprising since edges are lost due to the construction process. The CFG and the SIM do, in all cases, show values that are so close together that analysis with the CFG would only be necessary if one is interested in the difference of a real-world graph and the result of simulations, i.e., the z-score. When it is just to show that the structure of the graph is different from random, the quick calculation with equation 8.3 would be enough. But for almost all graphs the difference between the results of the FDSM and the SIM (and therefore the CFG) are rather large. This difference can have an enormous influence on the outcome of any analysis. Consider the following: The real-world Oklahoma graph has an average neighbor degree of 185.99, the real-world $\text{Facebook}_{\text{ML}}$ graph has an average neighbor degree of 105.55. Now, considering just the averages of the CFG and the result of this measure in the real-world graph would indicate the Oklahoma graph as exceptional while the $\text{Facebook}_{\text{ML}}$ graph is not very special. When computing z-scores, the Oklahoma graph is still exceptional independent of the set the random graphs are from. However, the $\text{Facebook}_{\text{ML}}$ graph shows that it makes a difference. When computing the z-score based on the sampling process with the CFG, a z-score of 1.74 is the result. This score indicates that the real-world graph could be a result of the generation process using the CFG. On the other hand, computing the z-score with the FDSM, the z-score is 11.8. This score indicates that a graph like the real-world graph is very unlikely, even though not impossible, to be a result of the generating process of the FDSM. The real-world graph of the $\text{Facebook}_{\text{ML}}$ would thus be considered as not very special when using the CFG, but very special when using the FDSM.

Graph	FDSM	USIS	DSIS
Email-Enron	156.10 ± 0.77	151.15 ± 0.92	420.15 ± 0.78
Georgetown	161.41 ± 0.29	161.11 ± 0.32	243.04 ± 0.58
Princeton	159.38 ± 0.32	159.30 ± 0.28	226.70 ± 0.25
Oklahoma	221.39 ± 0.59	220.62 ± 0.39	377.39 ± 1.14
Caltech	77.26 ± 0.55	76.99 ± 0.52	104.68 ± 0.34
$\text{Facebook}_{\text{ML}}$	111.57 ± 0.51	110.06 ± 0.55	186.42 ± 0.97
Facebook-like	60.56 ± 0.67	59.25 ± 0.74	119.33 ± 0.58
Condensed Matter	22.17 ± 0.09	22.09 ± 0.09	36.11 ± 0.10
High Energy Physics	137.73 ± 0.55	135.71 ± 0.68	292.94 ± 0.25
General Relativity	17.00 ± 0.13	16.97 ± 0.13	27.92 ± 0.12
Astro Physics	65.87 ± 0.19	65.70 ± 0.21	127.45 ± 0.25
High Energy Physics Theory	12.56 ± 0.06	12.54 ± 0.06	18.75 ± 0.06
S. cerevisiae	14.77 ± 0.35	14.38 ± 0.42	22.44 ± 0.40
C. elegans	27.88 ± 0.62	27.37 ± 0.42	35.20 ± 0.46
Diseases	7.71 ± 0.15	7.65 ± 0.16	10.42 ± 0.15
E. coli	13.08 ± 0.41	12.72 ± 0.44	17.88 ± 0.43

Table 8.2: Average and standard deviation of the average neighbor degree in the ground truth model FDSM, the USIS, and the DSIS

Considering the results in Table 8.2, the USIS is close to the results of the FDSM in almost all cases. The results of the FDSM and the USIS are close enough in almost all cases such that one set of graphs could have been generated using the other algorithm (see Table 8.3).

Graph	z_{CFG}	z_{ECFG}	z_{USIS}	z_{DSIS}
Email-Enron	14.51	24.43	4.14	240.84
Georgetown	3.09	13.42	0.69	125.69
S. cerevisiae	2.68	4.13	0.71	14.32
Diseases	0.59	1.33	0.27	12.59

Table 8.3: Two-sample z-score calculation in comparison with the FDSM to test whether results can be from the distribution indicated by the samples.

On the other hand, the graphs generated with DSIS are completely off. Even small graphs such as E. coli show large differences in the average result, resulting in enormous z-statistics. Considering that the results of the CFG and ECFG were deemed unreliable, these results are even worse. The graph generating process causes the high average neighbor degree. Recall that the algorithm starts with assigning neighbors to low-degree nodes. For DSIS, this decision is degree based, i.e., connections to high degree nodes are more likely. Starting with connecting low degree nodes to high degree nodes with higher probability influences the result of the average neighbor calculation quite strongly, as can be observed from the z-scores in Table 8.3.

8.2 COMMON NEIGHBORS

The common neighbor degree is the next measure to investigate. A global measure for this is, for simple graphs, simple to evaluate. By counting for each pair of nodes the number of common neighbors,

$$\operatorname{cooc}(u, v) = |N(u) \cap N(v)|, \tag{8.4}$$

and adding these numbers up, it is easy to observe that

$$\sum_{u,v \in V, u \neq v} \operatorname{cooc}(u, v) = \sum_{v \in V} \binom{k_v}{2}. \tag{8.5}$$

As explanation, a node v has exactly as many as $\binom{k_v}{2}$ pairs of neighbors, i.e., there are $\binom{k_v}{2}$ pairs of neighbors. The right-hand side of equation 8.5 is much simpler to evaluate than the left-hand side. Moreover, as a global measure in simple fixed degree sequence graphs it is a constant. Thus, only the FDSM is displayed for algorithms that do not generate multigraphs and keep the degree sequence fixed[1].

For the CFG, again there is the problem of multiple edges between nodes. If u and v both have the neighbor w, and assume there are 5 edges between v and w and 2 edges between u and w, how often is w a common neighbor of the other two nodes? w is either once a common neighbor, as it would be in a graph sampled with ECFG, it could also be

1 Of course it was tested whether the algorithms of the SIS both produce the same result as the FDSM. They do, as they are supposed to.

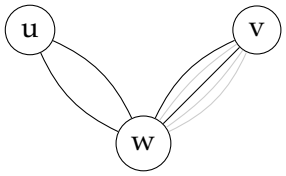


Figure 8.2: Calculating the co-occurrence in a multigraph is not defined.

two times considered as a neighbor since there can two different two-paths be constructed (comp. Fig. 8.2).

Another option is the maximum possible combinations of the edges, which is for this example 10. For the co-occurrence, there is no definition of how to count this in a multigraph. Additionally, self-loops should be accounted for. Since this would yield again several possible options on how to account for the co-occurrence in a multigraph, the CFG is dropped for this analysis.

8.2.1 *Approximating the Co-Occurrence*

Recall, that the expected number of multiple edges between a pair u, v of nodes could be calculated as

$$\frac{k_u k_v}{2m} \frac{(k_u - 1)(k_v - 1)}{2m}.$$

Using a similar approach, it is possible to calculate the expected number of edges from two nodes u, v to a third node l as

$$\frac{k_u k_l}{2m} \frac{k_v (k_l - 1)}{2m}. \tag{8.6}$$

Approximating the number of common neighbors of u, v is possible by summing over the nodes l

$$\sum_{l \in V} \frac{k_u k_l}{2m} \frac{k_v (k_l - 1)}{2m} = \frac{k_u k_v}{(2m)^2} \sum_{l \in V} k_l^2 - k_l \tag{8.7}$$

$$= \frac{k_u k_v}{(2m)^2} n \left(\langle k^2 \rangle - \langle k \rangle \right) \tag{8.8}$$

$$= \frac{k_u k_v}{2m} \frac{\left(\langle k^2 \rangle - \langle k \rangle \right)}{\langle k \rangle}. \tag{8.9}$$

This equation can also be found in Newman [66, p.441]. It yields the approximation of the number of common neighbors of two nodes; summing over the remaining variables, u, v, yields

$$\text{cooc}_{\text{SIM}_2}(G) = \sum_{u,v \in V} \frac{k_u k_v}{2m} \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)}{\langle k \rangle} \tag{8.10}$$

$$= \sum_{u \in V} n\langle k \rangle \frac{k_u}{2m} \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)}{\langle k \rangle} \tag{8.11}$$

$$= \sum_{u \in V} k_u \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)}{\langle k \rangle} \tag{8.12}$$

$$= n\langle k \rangle \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)}{\langle k \rangle} \tag{8.13}$$

$$= n\left(\langle k^2 \rangle - \langle k \rangle\right). \tag{8.14}$$

Another approximation, for bipartite graphs, is given by Zweig and Kaufmann [108] as

$$\text{cooc}_{\text{SIM}_1}(G) = \sum_{u,v \in L, u \neq v} \frac{k_u k_v}{n}, \tag{8.15}$$

where $L \cup R = V, L \cap R = \emptyset$. For this equation, Zweig and Kaufmann showed that it misestimates the total co-occurrence for a certain class of graphs. This class of graphs consists out of bipartite graphs, i.e., two disjoint sets of nodes that are connected by a number of edges. Additionally, the degree sequences for both groups of nodes are $L = R = \{1, 2, \ldots, n\}$. The difference for this specific class of graphs is in $\Omega\left(n^3\right)$ [108].

Nevertheless, both equations, the FDSM, and the ECFG are compared to see whether any equation of the SIM could be used instead of sampling from the ECFG and whether there is a (large) gap between the equations and the FDSM.

8.2.2 *Comparison of the models*

The ECFG obviously fails to deliver results that are even close to the result in the other models (see Table 8.4). To show this more pronounced, we calculated the z-score of the real-world graph in comparison to the samples from the ECFG. Remember, that when a fixed degree sequence model is used, the z-score would be 0 (actually not computable, since the standard deviation would be 0 as well). Considering the results in Table 8.5, it is obvious that the global co-occurrence is affected by the edgeloss rather severely.

As is shown in Table 8.5, the real-world graph's global co-occurrence (and therefore the global co-occurrence of all graphs with this exact degree sequence) would be recognized as something very, very special. That is unreasonable. Thus, the ECFG is not necessarily a good model for comparative analysis. On the other hand, Van Hoorn claimed that these changes would make the analysis more interesting/useful [39, 40].

For equation 8.15, the results are bad as well. The global co-occurrence is underestimated constantly. Since Zweig and Kaufmann [108] showed that equation 8.15 yields misestimates for a very particular case, this confirms that it also misestimates on more general degree sequences.

Graph	FDSM	ECFG	SIM_1	SIM_2
Email-Enron	$2.56 \cdot 10^7$	$2.18 \cdot 10^7 \pm 6.59 \cdot 10^4$	$1.84 \cdot 10^6$	$2.56 \cdot 10^7$
Georgetown	$6.78 \cdot 10^7$	$6.47 \cdot 10^7 \pm 3.97 \cdot 10^4$	$3.85 \cdot 10^7$	$6.77 \cdot 10^7$
Princeton	$4.61 \cdot 10^7$	$4.36 \cdot 10^7 \pm 3.16 \cdot 10^4$	$2.61 \cdot 10^7$	$4.61 \cdot 10^7$
Oklahoma	$1.94 \cdot 10^8$	$1.85 \cdot 10^8 \pm 9.41 \cdot 10^4$	$9.14 \cdot 10^7$	$1.94 \cdot 10^8$
Caltech	$1.23 \cdot 10^6$	$9.97 \cdot 10^5 \pm 5.05 \cdot 10^3$	$7.20 \cdot 10^5$	$1.23 \cdot 10^6$
$\text{Facebook}_{\text{ML}}$	$9.31 \cdot 10^6$	$8.26 \cdot 10^6 \pm 1.84 \cdot 10^4$	$3.85 \cdot 10^6$	$9.31 \cdot 10^6$
Facebook-like	$7.56 \cdot 10^5$	$6.21 \cdot 10^5 \pm 3.88 \cdot 10^3$	$2.01 \cdot 10^5$	$7.54 \cdot 10^5$
Condensed Matter	$1.97 \cdot 10^6$	$1.96 \cdot 10^6 \pm 1.61 \cdot 10^3$	$7.56 \cdot 10^5$	$1.97 \cdot 10^6$
High Energy Physics	$1.53 \cdot 10^7$	$1.34 \cdot 10^7 \pm 2.71 \cdot 10^4$	$2.34 \cdot 10^6$	$1.53 \cdot 10^7$
General Relativity	$2.30 \cdot 10^5$	$2.25 \cdot 10^5 \pm 6.20 \cdot 10^2$	$8.01 \cdot 10^4$	$2.30 \cdot 10^5$
Astro Physics	$1.28 \cdot 10^7$	$1.25 \cdot 10^7 \pm 8.72 \cdot 10^3$	$4.18 \cdot 10^6$	$1.28 \cdot 10^7$
High Energy Physics Theory	$3.00 \cdot 10^5$	$2.98 \cdot 10^5 \pm 3.18 \cdot 10^2$	$1.37 \cdot 10^5$	$3.00 \cdot 10^5$
S. cerevisiae	$1.30 \cdot 10^4$	$1.07 \cdot 10^4 \pm 3.50 \cdot 10^2$	$3.20 \cdot 10^3$	$1.29 \cdot 10^4$
C. elegans	$5.38 \cdot 10^4$	$4.17 \cdot 10^4 \pm 8.20 \cdot 10^2$	$3.09 \cdot 10^4$	$5.35 \cdot 10^4$
Diseases	$1.01 \cdot 10^4$	$9.63 \cdot 10^3 \pm 1.35 \cdot 10^2$	$5.37 \cdot 10^3$	$1.00 \cdot 10^4$
E. coli	$5.29 \cdot 10^3$	$3.99 \cdot 10^3 \pm 2.53 \cdot 10^2$	$1.27 \cdot 10^3$	$5.23 \cdot 10^3$

Table 8.4: Average and standard deviation of the cooccurrences in the ground truth model FDSM, the CFG, the ECFG, and the corresponding equations of the SIM.

Graph	z-score
Email-Enron	57.39
Georgetown	75.96
Princeton	80.67
Oklahoma	95.79
Caltech	46.43
$\text{Facebook}_{\text{ML}}$	57.45
Facebook-like	34.68
Condensed Matter	8.58
High Energy Physics	68.04
General Relativity	7.45
Astro Physics	30.72
High Energy Physics Theory	5.31
S. cerevisiae	6.58
C. elegans	14.72
Diseases	3.22
E. coli	5.13

Table 8.5: z-score calculation with the graphs from the ECFG as samples and the number of cooccurrences in the FDSM as value to test whether it can be from the distribution indicated by the samples.

On the other hand, based on the results in Table 8.4, the difference between the FDSM and SIM_2 are negligible. Simulation is not necessary to get the co-occurrence and calculating the global co-occurrence with the SIM is possible. That this is not the case is shown in the following. First, consider again equation 8.14. There are several averages that are removed during the reformulation. But averages are over *all* nodes, i.e., in equation 8.12 the sum considers implicitly pairs u, u as well. Since this was not considered before, we take a look at the local, node-based co-occurrence estimates.

8.2.3 *Local Co-Occurrences*

It is possible to calculate for a fixed node degree the number of expected common neighbors based on equation 8.9. By fixing one node, the equation depends only on a single variable, i.e.,

$$\text{cooc}_{\text{fixed}}(w) = k_{\text{fixed}} \frac{k_w}{2m} \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \tag{8.16}$$

$$= c k_w \tag{8.17}$$

Plotting the results of equation 8.17 against the average results of the simulations shows the problem that occurs when using the equations. When setting $k_{\text{fixed}} = max_{u \in V} k_u$, the calculation overestimates the co-occurrence for all nodes by far. When the fixed degree is high, as in Fig. 8.3, equation 8.17 overestimates the number of common neighbors. More importantly, we omitted the highest degree node, since this node occurs only once in this graph; calculating the co-occurrence for a node with itself is a strange idea and should yield zero. The SIM would not care and yield some value following the linear scaling. Thus, for high degree nodes such as the node with the highest degree, the SIM overestimates locally. For other fixed degrees, as for a node with an average or below average degree, the results of the equation are closer to the values that occur on average in the sample graphs. For the minimum degree, the approximation underestimates many values; only the high degree nodes are still overestimated. Since there are more nodes of low degree than of high degree in real-world graphs, this overall balances out.

8.3 IMPLICATIONS

From the standpoint of network analysis, seeing that the CFG was used in a research paper should be a warning to the reader. If the authors did not check that they used graphs that have an unskewed degree distribution, did not clarify if they used the erased configuration model or the multigraph generating model, or if the authors do not mention at all how they handle multiple edges or self-loops, the results of the paper have to be checked. Not only has to be checked whether equations and measures applied in the analysis handle multiple edges at all, but it has to be checked as well whether the conclusions are reasonable compared to the model. When a graph has an underlying degree-distribution that is unskewed, it is not as bad to use the ECFG and claim that the CFG is used as was shown when the edgeloss in a graph with Poissonian degree distribution was measured. When the degree distribution is strongly skewed, this does carry weight, and it has to be carefully examined whether the results are reasonable in any way.

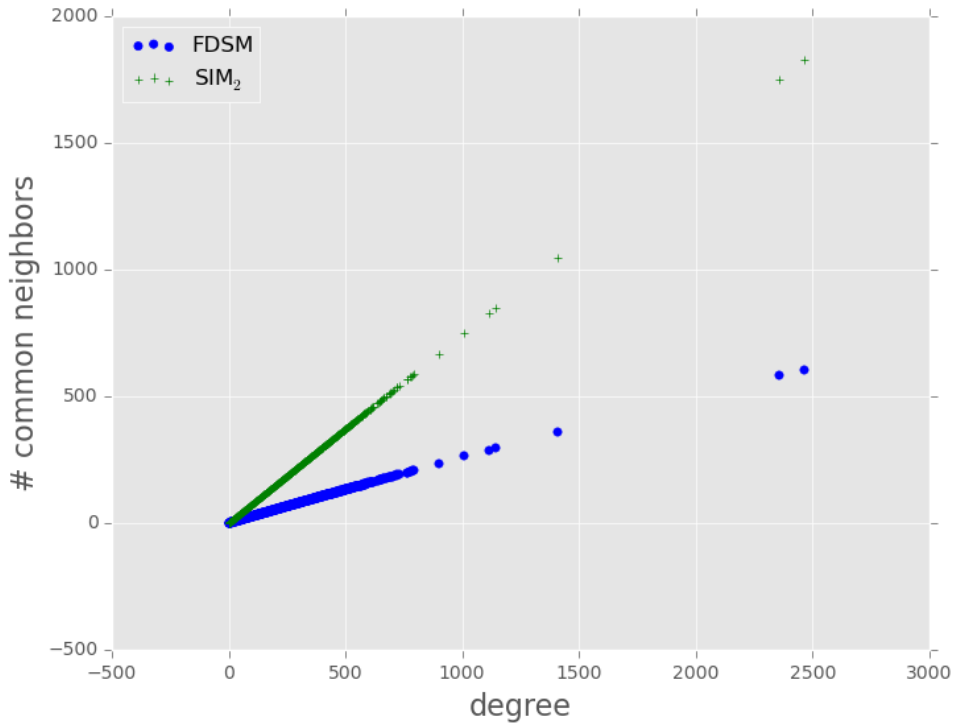


Figure 8.3: Shown are the average number of common neighbors a node of any degree has with nodes of the maximum degree for the Oklahoma graph. This is measured in the generated graphs of the FDSM (blue dots) and the estimated number of co-occurrences (green crosses).

There are also arguments to use the ECFG. The CFG is one of the simplest random graph models for constructing graphs. Still, usually simple graphs are used for analysis and therefore, the ECFG is primarily of interest [39]. That a fraction of edges is lost in total, is most of the time not considered or at least not regarded as having much influence. Comparing the number of lost edges to the existing number of edges, it is a rather low number for most graphs. Another work suggests that the ECFG would even be the null model to choose when testing for statistical significance, since the constraint of a fixed degree sequence would constrain the set of possible graphs to much [40]. Even though this argument sounds reasonable, the ECFG will only loose or maintain the degree sequence. To see whether graphs that follow the same degree distribution show different results than one real-world graph, it would be more reasonable to use the exponential random graph model, that generates not the specific degree sequence but uses the same distribution [28]. The ECFG may or may not use the same distribution, but high degree nodes will always lose edges as was shown above.

The equations of the SIM did show results that were, overall, bad. We checked whether bad results might have been caused by too many nodes with $k_u^2 > \sum_{v \in V} k_v$; this was not the case, at most 1.3 percent of the nodes had this indicator that the equations should not be used. Still, the global approximations of the diameter and the distance were not very good; this is caused by the assumption that nodes have about the same degree—an assumption that can be made for a graph with a Poissonian degree distribution, but not for graphs with more skewed degree distributions. The same applies for the average neighbor degree. The global co-occurrence is surprisingly well estimated by equation 8.14. Still, when calculating the co-occurrences that a fixed degree node has with any other node, it is evident that there are many misestimates.

For the SIS, there are two different versions of the algorithm: one, that uses a uniform distribution to choose neighbors from the set of possible neighbors. The other uses a probability distribution based on the remaining degree of possible neighbors. The former shows for almost all measures taken very similar results to the FDSM, the model considered as ground truth. The latter does sample graphs for that the measures are very different from the results found in graphs from the FDSM. The reason for this is the construction process of the SIS. When considering the DSIS, the degree based algorithm, many of the low degree nodes will be connected to high degree nodes. This may reduce the diameter, but since high degree nodes, hubs, do not connect as well as in other models, the distance is higher. The average neighbor degree is higher for the same reason. On the other hand, the USIS randomizes graphs in a way such that the results are much closer to the ground truth model, FDSM. Still, the average distances are larger than found in the graphs from the FDSM. Since the run-time estimate of the SIS is worse than the estimate for the FDSM, I suggest the following process for analysis of undirected graphs:

1. Check which distribution the real-world graph's degree sequence follows
   a) If the degree sequence is unskewed, e.g. the Poissonian distribution, one may use the CFG or ECFG
   b) If the degree sequence is strongly skewed, use the FDSM
2. Generate a number of samples with the chosen model
3. Evaluate the measures on the samples

4. Evaluate the measure on the real-world graph
5. Test with a t-test or, if enough samples have been taken, with a z-test, whether the result of the measure on the real-world graph may be a result of coincidence, i.e., if the calculated score is small, the measure is not outstanding in the graph and should not be treated as such

Regarding the SIM, for undirected graphs, it seems to be not applicable to data that is skewed. Nevertheless, it can (and only should) be used as an approximation to see whether the real-world graph differs strongly from the estimate of the equations of the SIM. Whenever there is a "large" difference, a more detailed analysis should be started. Large is a very problematic term since sometimes large is 0.3 (diameter) while the same value can also be very small (average neighbor degree). Thus, analysis should always be done with the utmost care.

In the following, we change the graphs of interest and go from undirected to directed graphs, using this opportunity to change the analyzed measure as well.

Part IV

COMPARISON OF THE DIFFERENT NULL MODELS BASED ON DIRECTED GRAPHS

In the last chapter, we observed that there are differences when using algorithms to generate graphs. The configuration model uses a large family of graphs and does not generate graphs uniformly at random. The sequential importance sampling uses a smaller family of graphs since multigraphs are excluded. The use of so-called simple graphs makes analysis straightforward, but this algorithm still does not generate graphs uniformly at random, independent of the probability distribution which is used to create edges between nodes and their possible neighbors. One probability distribution favors connections to high degree nodes; the other does not favor any node. However, the algorithm starts always with the lowest degree node. Therefore, some edges may not be so random. Even if the SIS would be able to generate graphs uniformly at random, it would still be slower than other algorithms. The sequential importance sampling algorithm also gives a weighting factor back such that analysis may take into account how likely it is to end up with a graph. The fixed degree sequence model is the only one that does everything needed. It samples from the same family as the sequential importance sampling algorithm uniformly at random. Even though it is faster than the sequential importance sampling algorithm, it is still not fast enough to apply to Big Data. The simple independence model would be fast enough, but the results on undirected graphs are not good enough when they follow skewed degree distributions.

The question that we are discussing in this chapter is based on the results of the former chapter. When the models show such different results on undirected graphs, how do they apply to directed graphs? Since the measures from the undirected graphs section are very simple, now some more interesting problems are to be investigated.

In the following, network motif analysis is performed in the steps outlined below.

- Define a subgraph of interest;
- Count how often this subgraph occurs in a real-world graph;
- Generate random graphs and count the occurrence of the subgraph in the generated graphs;
- Check whether the subgraph occurs statistically significantly often in the real graph.

When the subgraph does occur statistically significantly often, then the subgraph is called a *network motif*, following Milo et al. [63].

SYNOPSIS First, a short history of network motif analysis is given to inform the reader about the coming research. Afterward, the subgraphs of interest, the motifs, are introduced. The null models are tested for their quality regarding network motif analysis. While the SIS does not need any new definition, the CFG proves again problematic due to its generation process and the resulting multiple edges. For the SIM a set of equations is developed and applied. The results show great promise, but sometimes the estimates are wrong. An investigation yields several possible ways to remedy this problem to a certain extent. Additionally, the participation of a node in a subgraph is investigated based on the SIM.

9

ANALYSIS OF DIRECTED GRAPHS

Instead of repeating the steps of the analysis from the last chapter for directed graphs, distances, diameter and so on, another topic is covered. Here, subgraph counting on directed graphs is of interest. In subgraph counting one is either searching for specific connection patterns of a fixed number of nodes in a graph and compares the result obtained in the real-world graph to the results obtained when counting subgraphs in a set of sampled graphs with the same degree sequence. Alternatively, all patterns of up to certain size are counted in the real-world graph as well as in the sampled graphs. The comparison itself is usually performed with the z-score. If the z-score is larger than 3, the subgraph in question can be considered as a *network motif*. Kashtan et al. [48] suggested that the z-score should be at least 5 to call a subgraph a *network motif*.

9.1 A SHORT HISTORY OF MOTIF ANALYSIS

The first paper that performed *network motifs analysis* as such is from Milo et al. [62]. In this article, the researchers checked for two things. First of all, they tested three different algorithms to generate graphs whether they would produce graphs uniformly at random. The models under investigation were the following: the configuration model with resampling when multiple edges were created; the switching algorithm, that is equal to the FDSM; a new algorithm called "go with the winners", which samples several graphs simultaneously and discards any graph that contains multiple edges or self-loops. For one part of their research they use the "go with the winners" as ground truth, since it samples all possible simple graphs, even though it is very slow. They did this with a very simple toy-graph that consists out of 12 nodes and 20 edges for that only 91 configurations exist. They compared the "go with the winners" algorithms results to the other two. The result was that the FDSM is a better fit regarding uniformity. The other test they performed was for the mixing time of the FDSM. They tested how many edge swaps had to be carried out to reach a stable average number of occurrences of a certain subgraph, i.e., when does a larger number of sampling not change the average number of subgraphs found in a graph. From this experiment, they concluded that $100 \cdot |E|$ would be enough, even though the plot in the paper indicates that a much lower number would be sufficient. In other works by Milo et al. [63, 64], they perform similar research, analyzing how well this form of analysis applies to other graphs and if other network motifs can be found as easily.

The most simple algorithm to find all subgraphs that follow a predetermined pattern is to check for each node its neighborhood and the connections within. If the pattern occurs, a counter is increased. Usually, the patterns investigated consist out of 3 to 4 nodes, such that the simple way takes a long time. Considering the amount of swaps that is used and

the number of samples that has to be drawn from the family of graphs, this process is very time-consuming.

Several improvements regarding motif analysis have been made. Instead of scanning for a fixed subgraph, nowadays algorithms scan a graph for all possible subgraphs of a given size and count their occurrence. The most well-known algorithm is by Wernicke [103]. The algorithm is called ESU with the variant RANDESU. The former starts with single nodes, extends their neighborhood for nodes that have a larger index than the node itself and that are not connected to a node that is already in the set of neighbors. This procedure builds the so-called ESU-tree. According to Wernicke, all size-k subgraphs are output exactly once. The variant RANDESU does what the name implies; it performs the ESU algorithm with a random element added, i.e., a probability is introduced; with this probability, a node is added to the set. Otherwise, it is skipped. This speeds up the calculation. An additional interesting factor is a change in the resampling process. The resampling process does not resample the complete graph but decides based on some given probabilities whether the ESU-tree is traversed further or if some levels are skipped. For more details see [103]. An even faster approach is provided by Itzkovitz et al. [43]. Based on the degree sequence and some additional information, i.e., in-, out-, and mutual degree, they approximated the number of subgraphs in a given graph. This idea is quite similar to the approach explained later, see Section 12.1, but we were not aware of this work until after the development of our approach. Additionally, our approach needs less information[1].

Another work regarding this topic is by Birmele [7], who considered subgraphs of motifs, called subpatterns, as a possible indicator of the presence of a motif. If the subpatterns are overrepresented, the graph should contain certain motifs as well. It is important to note that there is a difference between locally overrepresented and globally overrepresented subpatterns. The models that are used for comparison do not necessarily maintain the degree sequence, which might skew the results.

To the best of my knowledge, despite all these efforts, it was never tested whether it is even possible to use the multigraph generating version of the configuration model to get and estimate of the number of subgraphs in a graph and compare the result with the real-world graph. Moreover, it was not investigated whether the directed SIS yields the same results as the FDSM. This task was most likely not done, since the SIS was developed after much of the analysis regarding network motifs was done. Still, the possibility that the results are similar to those of the FDSM exists since both models draw from the same family of graphs, i.e., all simple directed graphs are considered by both algorithms. It is important to mention, that even though the SIS does not draw uniformly at random from the family of simple graphs, that standard edge-swap algorithms do also sometimes not reach every possible graph. For this, Berger and Müller-Hannemann suggest a modified algorithm that swaps three edges at once, when certain patterns of connection are discovered [6]. It is not known if that is considered in the research of Milo et al. On the other hand, SIS can generate all possible graphs since it generates in each try a new graph from scratch. The only restriction that has to be mentioned at this point is the probability distribution the choice of a new neighbor is based on.

Therefore, the following chapter is divided into several parts. First, the subgraphs that are considered are introduced. Second, a comparison of the multigraph generating configuration model, including a thorough description on how to count motifs in multigraphs,

1 For a comparison to their approach, see Appendix VI

with the baseline model, FDSM, is performed. In this comparison the multigraphs generated by the CFG as well as the simplified graphs of the ECFG are considered. Third, a comparison of the FDSM and SIS with two different probability distributions for choosing a new neighbor is performed. This is done to see whether the SIS could be used for the analysis of subgraphs and which probability distribution is a better fit to the results of the FDSM. Fourth, a set of equations is developed and compared to the baseline model. These equations are also used to calculate the participation of single nodes in subgraphs.

Despite using directed graphs in this chapter, the abbreviations stay the same as for undirected graphs, i.e., the abbreviations used are FDSM for the fixed degree sequence model, CFG for the configuration model, ECFG for the erased configuration model, USIS for uniformity-based sequential importance sampling, DSIS for degree-based sequential importance sampling, and finally SIM for the sequential importance sampling.

9.2 DATA

For this chapter, directed graphs are required. Beforehand, graphs were used as undirected graphs. Therefore, the following graphs have been selected:

FOOD WEB Examples of predatory graphs, i.e., who hunts whom, are given as Silwood graph, Ythan Estuary graph, Little Rock graph, St. Marks Island graph, St. Martin Seagrass graph, and Grassland graph [2]

ELECTRICAL CIRCUITS Two small electrial circuits, s208 and s420, with gates as nodes and the flow of electrical signals between them as edges, as used by Milo et al.[64]

COMPUTER NETWORK Snapshots of the peer-to-peer file sharing network Gnutella from 08.09.2002 [3] and 09.09.2002 [4]

BIOLOGICAL Beforehand, we used the transcription regulation graphs as undirected graph; now we include the direction for the E. coli graph and the S. Cerevisiae graph

No numbers for the occurrence of subgraphs are presented for the real-world graphs since they do not get compared to the random graph models.

2 http://pil.phys.uniroma1.it/~gcalda/cosinsite/extra/data/foodwebs/
3 http://snap.stanford.edu/data/p2p-Gnutella08.html
4 http://snap.stanford.edu/data/p2p-Gnutella09.html

10

ON MOTIFS

Usually, motif analysis considers only motifs of size 3 or 4. For näive implementation, this is very understandable; one possibility is calculating powers of the adjacency matrix which is in $\mathcal{O}\left(n^3\right)$. Problematic is in cases such as this the discovery of more complex subgraphs. Another option is to check all k-tuples of nodes, which is not much better. Therefore, initial investigation of motifs considered only three-node and four-node subgraphs to be motifs. Newer algorithms, like ESU by Wernicke [102, 103] can find motifs with more nodes, but research regarding motifs with more nodes is challenging to find if it exists at all. A reimplementation by one of the developers of the Stanford Network Analysis Package, short SNAP[1], implemented the algorithms with specific optimizations for three and four nodes as fixed parameters. The optimizations speed the calculation up but restrict the researcher to a specific number of nodes in subgraphs.

When all combinations of three nodes are taken into account, one speaks of the triad census [99]. With three nodes, fifteen different patterns of connections can be constructed. One of them is a special case, the not connecting pattern, i.e., three nodes without any connections. Only a subset of these is considered here, but it would be enough to consider even less to be able to decide whether different algorithms to sample graphs do give similar results regarding subgraph counting. Additionally, several four node subgraphs are considered.

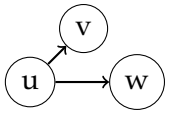
When graph sampling is performed to assess statistical significance of a feature of a graph, one can argue whether a graph should be restricted to use a specific degree sequence or whether different sequences yield more interesting results [40]. Since the aim of this chapter is to find network motifs, it is more reasonable to use the same degree sequence. Otherwise, a real-world graph might seem overwhelmingly rich in a certain subgraph, even though other degree sequences are not able to generate this particular subgraph. Still, the ECFG only approximates the degree sequence due to edgeloss, therefore, also these cases are covered. As will be shown, the results of the ECFG are always much lower than the results of any other algorithm.

In the following the subgraphs that are under consideration are described in more detail; there are three and four node subgraphs, most of them occurred in motif-based research already.

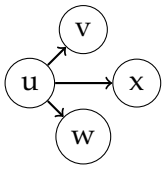
FORK: Forks are one of the subgraphs that are used in this work as a baseline. Whenever a model is not able to yield the same amount of this subgraph for any graph generated, either the degree sequence has been changed, or the model is just not able to capture even this very simple pattern. This subgraph consists of three nodes, one of which

1 http://snap.stanford.edu

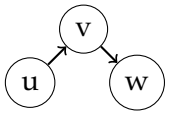
is connected to the other two (Fig 10.1a). The number of Forks attached to a node is equal to the number of pairs of neighbors a node has.



(a) Fork



(b) Fan



(c) Two-path

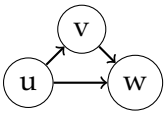
FAN Just as the Fork, this can be used as a baseline comparison of the different graph models. It is easy to see that this is just a way to describe that a node has different triples of successors, i.e., that the edges $(u, v), (u, w), (u, x)$ exist (see Fig. 10.1b).

TWOPATH: Twopaths (Fig. 10.1c) are one of the most common subgraphs and not considered as a motif in many fields. Two-paths consists out of three nodes $\{u, v, w\}$ and two edges, $(u, v), (v, w)$. Still, this pattern is of interest, for example, in graphs of predators (who eats whom) or in social sciences. Often it is heard that "the enemy of my enemy is my friend" or "a friend of a friend is a friend of mine" and for some of these, there is even evidence [89, 90]. Based on this, a researcher may look at the connection patterns in graphs and predict whether new connections may come into existence between persons (e.g. [78, 58, 42]).

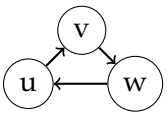
FEED-FORWARD LOOP: One of the most popular, i.e., most researched motifs, is the Feed-Forward Loop (see Fig. 10.2a). It is, more or less, a combination of the Twopath and the Fork, i.e., the edges $(u, v), (v, w)$ are to be found as well as an additional edge (u, w). Depending on the field of study the researcher is from, there can be different types of meaning to each edge. Be it either social sciences as mentioned for Twopaths, where "a friend of my friend is a friend of mine" can be tested, or biology, where edges can be either be marked as excitatory or inhibitory. Considering that there are different types of edges is usually not done in motif analysis since it requires more computation time to do so.

THREECYCLE This motif can also be interpreted as three Twopaths, one being $(u, v), (v, w)$, another $(v, w), (w, u)$, and $(w, u), (u, v)$ (see Fig. 10.2b).

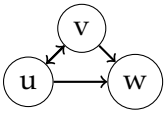
DOUBLEJOIN This is another good example of overlapping motifs that is considered as a motif in its own right. Whenever a Doublejoin is found, the Two-path counter increases by two and the Feed-Forward Loop counter by one. The two Twopaths consists out of the edges $(u, v), (v, w)$ and $(v, u), (u, w)$ (see Fig. 10.2c).



(a) Feed-Forward Loop

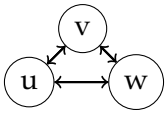


(b) Threecycle

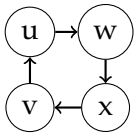


(c) Double-Join

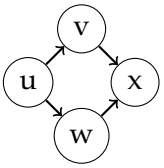
COMPLETE In social science it could be an indicator of strong friendships since each of the nodes is in a reciprocated relationship with all other nodes (see Fig. 10.3a). Depending on the context of research it occurs in, it can either indicate strong friendships, good collaborators, but it can also indicate strong enmity. It is one of the rather rare motifs to be found in a graph.



(a) Complete



(b) Fourcycle



(c) Biparallel

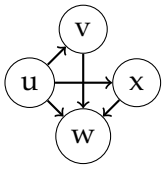
FOURCYCLE This subgraph can be found in electronic circuits [64], as in regulating circuits or similar. Based on the theory of Simmel [86] and Granovetter [33], it is rather unlikely that such a subgraph occurs in social networks. Node u is connected to nodes v and w, and according to the theory it is rather unlikely that if the connection is strong, that v and w are not connected.

BIPARALLEL This motif was also found in electronic circuits and synaptic connections. The fact that there is branching with $(u, v), (u, w)$ and then a merge $(v, x), (w, x)$ (cf. Fig.10.3c) is especially interesting for synapses, considering that evolution is the process of maximizing potential while using minimal resources.

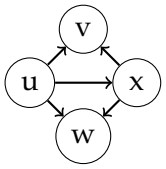
IN-FAN This can be considered as a mixture of Fan and Biparallel, or as an extension to both. Additionally to the edges in the Fan, an edge (v, w) and an edge (x, w) are added (see Fig. 10.4a).

OUT-FAN Similar to the In-Fan, in this subgraph one of the targets of the Fan subgraph connects to the other two (see Fig. 10.4b).

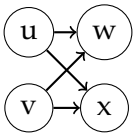
BIFAN This is a special case of overlapping motifs since two Forks are overlapping such that the targets of the edges are the same, but the sources are different (see Fig.10.4c). This subgraph was reported to be found in electronic circuits, in gene regulation graphs, and in synaptic connection graphs.



(a) In-Fan



(b) Out-Fan



(c) Bifan

Most of these subgraphs have been looked at before [64], the most popular being the Feed-Forward Loop, which has been analyzed in several papers [62, 78, 105, 106]. Two baseline checks have been added to the set of previously existing motifs, such that there is some other simple way to check whether the results of different graph generating algorithms should be or can be used in motif analysis.

In the next chapter, different models are compared, starting with the two version of the CFG with the baseline model, the FDSM.

11

DIFFERENT MODELS UNDER INVESTIGATION

For undirected graphs, the analysis was straightforward - all models were compared based on a single measure to the FDSM. Here, a different approach is taken; instead of comparing a single subgraphs occurrence in the different models, we compare the occurrence of all subgraphs considered at once. This investigation is done as follows: first, the results of the two versions of the CFG are compared to the results of the FDSM. Second, the results of the two variants of the SIS are compared to the ones of the FDSM. Third, a set of equations based on the SIM is developed and tested against the FDSM.

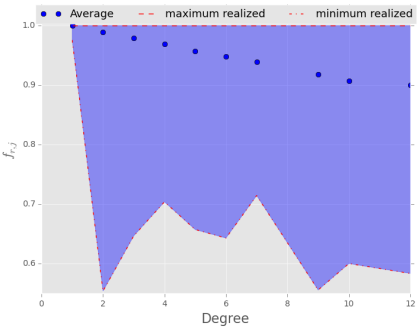
11.1 ON THE DIRECTED CONFIGURATION MODEL

The CFG for directed graphs shows the same problem as it does for undirected graphs. It generates multiple edges between nodes, the difference being that it now tends to generate more edges between high out-degree and high in-degree nodes. Similar to Fig. 7.3, which showed the edgeloss in samples of undirected graphs per degree once for the E. coli graph and once for the Email-Enron graph, the same is possible for directed graphs and done in the following.

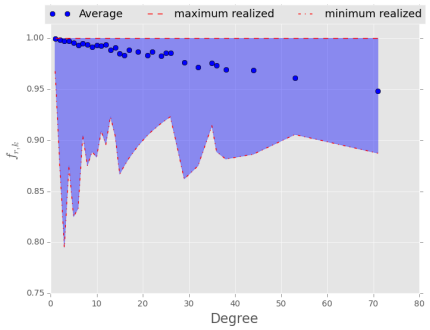
For Fig. 11.1, equation 7.28 has been adapted to measure once for in-degree and once for out-degree the loss of edges attached to each node of fixed degree k. Most interestingly, for the in-degree sequence of the S. cerevisiae graph, the loss is rather mild, but the out-degree sequence has a rather severe loss for high-degree nodes; in the artificially generated example, losses in both sequences are about the same and rather severe for high degree nodes. The example graph was generated with the Forest-Fire model; it has 4038 nodes, 87121 edges and its degree distribution follows a power law with an exponent of $\gamma \sim 1.6$; this example graph has only one weakly connected component.

One of the unanswered (and before not asked) questions is whether the edgeloss of the ECFG influences the results of motif analysis severely. Another issue is how motif counting in multigraphs should be defined and whether this yields reasonable results in comparison to the slower sampling with the FDSM. Note that we consider a subgraph as one of the above when the number of nodes participating is as described. For example, the Twopath has to have three nodes which are connected $(u, v), (v, w)$; it is not allowed that self-loops, i.e., $(u, u), (u, v)$, or only two edges, $(u, v), (v, u)$, are considered as Twopath in this work.

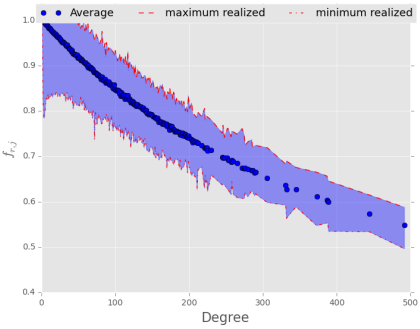
The biggest problem with the last question is the following. In Fig. 11.2a, most algorithms will find exactly one Feed-Forward Loop and no more. This is, for many purposes, correct. But when trying to find *all* possible motifs, it is not defined how this is done. One has to consider what is possible to count in a multigraph; the lowest possible number of subgraphs is found in the ECFG; edges are deleted such that the results of an analysis



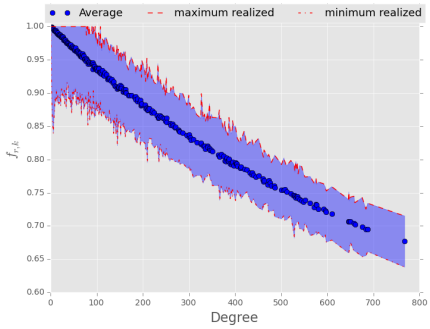
(a) S. cerevisiae in-degree loss



(b) S. cerevisiae out-degree loss

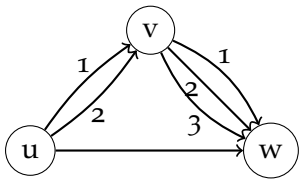


(c) Artificial graph in-degree loss

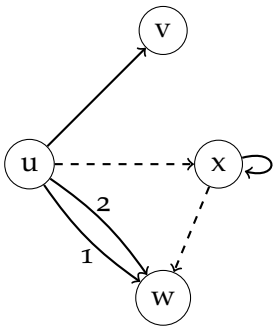


(d) Artificial graph out-degree loss

Figure 11.1: Edgeloss measured on the graph of S. cerevisiae (11.1a, 11.1b) and an artificial example (11.1c, 11.1d); in Fig. 11.1a, resp. 11.1c the loss of in-degree is shown, in Fig. 11.1b, resp. 11.1d, the loss of out-degree is shown.



(a) How to count Feed-Forward Loops in multigraphs?



(b) The graph contains two Forks, $(u, w_1), (u, v)$ and $(u, w_2), (u, v)$; a simple graph would have the dashed edges instead of the self-loop and the multiple edge and thus three Forks.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$8.20 \cdot 10^2 \pm 0.00$	$7.99 \cdot 10^2 \pm 4.49$	$6.30 \cdot 10^2 \pm 3.29 \cdot 10^1$
Silwood	$3.67 \cdot 10^3 \pm 0.00$	$3.63 \cdot 10^3 \pm 5.48$	$2.89 \cdot 10^3 \pm 1.02 \cdot 10^2$
St. Martin Island	$7.64 \cdot 10^2 \pm 0.00$	$7.34 \cdot 10^2 \pm 5.51$	$5.50 \cdot 10^2 \pm 3.03 \cdot 10^1$
Ythan Estuary	$4.47 \cdot 10^3 \pm 0.00$	$4.40 \cdot 10^3 \pm 8.38$	$3.27 \cdot 10^3 \pm 1.08 \cdot 10^2$
Little Rock	$2.55 \cdot 10^4 \pm 0.00$	$2.51 \cdot 10^4 \pm 2.00 \cdot 10^1$	$1.68 \cdot 10^4 \pm 2.78 \cdot 10^2$
Grassland	$2.38 \cdot 10^2 \pm 0.00$	$2.32 \cdot 10^2 \pm 2.36$	$2.08 \cdot 10^2 \pm 1.14 \cdot 10^1$
s208	$1.64 \cdot 10^2 \pm 0.00$	$1.63 \cdot 10^2 \pm 8.98 \cdot 10^{-1}$	$1.58 \cdot 10^2 \pm 5.40$
s420	$3.80 \cdot 10^2 \pm 0.00$	$3.79 \cdot 10^2 \pm 9.59 \cdot 10^{-1}$	$3.72 \cdot 10^2 \pm 7.10$
Gnutella 08.08.2002	$9.35 \cdot 10^4 \pm 0.00$	$9.21 \cdot 10^5 \pm 1.67 \cdot 10^3$	$9.27 \cdot 10^4 \pm 9.91 \cdot 10^1$
Gnutella 09.08.2002	$1.22 \cdot 10^5 \pm 0.00$	$1.36 \cdot 10^6 \pm 2.30 \cdot 10^3$	$1.21 \cdot 10^5 \pm 1.37 \cdot 10^2$
E .coli	$4.82 \cdot 10^3 \pm 0.00$	$4.81 \cdot 10^3 \pm 2.87$	$4.41 \cdot 10^3 \pm 1.27 \cdot 10^2$
S. cerevisiae	$1.18 \cdot 10^4 \pm 0.00$	$1.18 \cdot 10^4 \pm 4.38$	$1.12 \cdot 10^4 \pm 1.62 \cdot 10^2$

Table 11.1: Number of Forks found in the respective models. The standard deviation for the CFG is due to the fact that two edges between the same node do not yield a Fork.

based on the ECFG can have a broad range of results due to different structures of the underlying multigraph generated with the CFG. This option is shown in the results as ECFG. Another option would be to count all possible subgraphs, i.e., all permutations of edges. In Fig. 11.2a this implies that there are the following Feed-Forward Loops: $(u,v)_1$, $(v,w)_1$, (u,w); $(u,v)_1$, $(v,w)_2$, (u,w); and $(u,v)_1$, $(v,w)_3$, (u,w), moreover $(u,v)_2$, $(v,w)_1$, (u,w); $(u,v)_2$, $(v,w)_2$, (u,w); and $(u,v)_2$, $(v,w)_3$, (u,w). This option does count the maximum possible number of subgraphs. In the results, it is shown as CFG. While the first option will most likely yield lower results than the FDSM due to lost edges which do contribute to the number of subgraphs in the graphs generated with the FDSM, the second option can be expected to yield higher results. Since we consider the results of the FDSM as gold standard due to the research by Milo et al. [63, 64], the comparison with the FDSM is necessary.

First, a comparison of the baseline subgraphs is performed. Counting these two, the Fork and the Fan, should always result in the same number of subgraphs when a graph with a fixed degree sequence is generated. In Table 11.1 the results of counting the Fork (Fig. 10.1a) are shown, in Table 11.2 the results of counting the Fan (Fig. 10.1b) are shown. The first observation is that neither the CFG nor the ECFG do yield the same number of subgraphs as the FDSM. This is a bad sign for the CFG, since it uses a fixed degree sequence but it fails to confirm the assumption of the baseline. In Fig. 11.2b we show explicitly what can happen when using the CFG. The FDSM allows only one graph to be generated and results always in the same number of Forks.. The CFG may generate multiple edges and thus forfeit possible Forks. The ECFG samples less subgraphs due to edgeloss, as expected. However, there are several ways to calculate whether two models yield reasonably similar results. The most famous in regards to network motif counting is the z-score or to be more exact the standard score calculation. Since there are two groups of samples, the two-sample z-statistic [54] with the hypothesis that the means are equal ($\mu_1 - \mu_2 = 0$) $\frac{\bar{\mu_1} - \bar{\mu_2}}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ should be used, where $\bar{\mu_i}$ is the observed average number of subgraphs for samples i. If the score is small, it is likely that the models are equal. However, the test, as well as the two-sample z-score, assumes that the distribution the samples are from is a normal distribution. Therefore, using the

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$2.26 \cdot 10^3 \pm 0.00$	$2.09 \cdot 10^3 \pm 3.97 \cdot 10^1$	$1.45 \cdot 10^3 \pm 1.46 \cdot 10^2$
Silwood	$3.01 \cdot 10^4 \pm 0.00$	$2.92 \cdot 10^4 \pm 1.37 \cdot 10^2$	$2.04 \cdot 10^4 \pm 1.25 \cdot 10^3$
St. Martin Island	$1.76 \cdot 10^3 \pm 0.00$	$1.56 \cdot 10^3 \pm 3.85 \cdot 10^1$	$1.04 \cdot 10^3 \pm 1.03 \cdot 10^2$
Ythan Estuary	$3.13 \cdot 10^4 \pm 0.00$	$2.99 \cdot 10^4 \pm 2.00 \cdot 10^2$	$1.82 \cdot 10^4 \pm 1.20 \cdot 10^3$
Little Rock	$2.18 \cdot 10^5 \pm 0.00$	$2.07 \cdot 10^5 \pm 5.93 \cdot 10^2$	$1.05 \cdot 10^5 \pm 3.40 \cdot 10^3$
Grassland	$3.87 \cdot 10^2 \pm 0.00$	$3.61 \cdot 10^2 \pm 1.32 \cdot 10^1$	$3.04 \cdot 10^2 \pm 3.55 \cdot 10^1$
s208	$2.22 \cdot 10^2 \pm 0.00$	$2.19 \cdot 10^2 \pm 4.30$	$2.10 \cdot 10^2 \pm 1.66 \cdot 10^1$
s420	$6.32 \cdot 10^2 \pm 0.00$	$6.27 \cdot 10^2 \pm 6.09$	$6.06 \cdot 10^2 \pm 3.57 \cdot 10^1$
Gnutella 08.08.2002	$3.09 \cdot 10^5 \pm 0.00$	$3.09 \cdot 10^5 \pm 8.07 \cdot 10^1$	$3.03 \cdot 10^5 \pm 1.62 \cdot 10^3$
Gnutella 09.08.2002	$4.57 \cdot 10^5 \pm 0.00$	$4.57 \cdot 10^5 \pm 9.25 \cdot 10^1$	$4.49 \cdot 10^5 \pm 2.92 \cdot 10^3$
E .coli	$7.13 \cdot 10^4 \pm 0.00$	$7.09 \cdot 10^4 \pm 1.55 \cdot 10^2$	$5.94 \cdot 10^4 \pm 4.15 \cdot 10^3$
S. cerevisiae	$1.50 \cdot 10^5 \pm 0.00$	$1.49 \cdot 10^5 \pm 1.86 \cdot 10^2$	$1.34 \cdot 10^5 \pm 4.39 \cdot 10^3$

Table 11.2: Number of Fans found in the respective models.

Kolmogorov-Smirnov two-sample test may give more information since it compares the distributions of the subgraphs. Therefore, in Table 11.3, resp. 11.4, the results D of this test, the maximal difference between the cumulative distribution functions, are listed for the two motifs Fork and Fan. Recall that a low D is necessary to judge the distributions as similar. For $\alpha = 0.05$ a D-value below 0.136 would be low enough to judge the distribution of the number of subgraphs as indistinguishable. It is to observe that all D-values are close to 1 or 1 such that it is unlikely that the graphs are from the same family of graphs. Moreover, even the CFG and ECFG do show results regarding the Kolmogorov-Smirnov two-sample test close to 1, indicating that even the models which are sometimes used interchangeably do not yield the same results regarding the number of subgraphs. P-values are omitted, since they would add no more information but that it is rather unlikely that the subgraph distributions are the same ($p < 10^{-5}$).

Of course, these two subgraphs are very basic, and the total number of these subgraphs is supposed to be fixed. The ECFG may lose possible interaction partners while the CFG may build multiple edges and therefore produce entirely different results. The results of the analyses of the Fork and the Fan show that using the CFG or ECFG may yield results which can easily lead to misinterpretation of the graph the samples are compared to. A too high/low number of subgraphs and a large standard deviation may show the real-world graphs number of motifs as extraordinarily small/large. On the other hand, the FDSM may indicate the number of subgraphs in the real-world graph as ordinary, as it is the case with the number of Forks or Fans.

That multiple edges, self-loops or the deletion of edges can influence the number of other subgraphs is true for other subgraphs as well. Another graph structure which is simple to count is the Twopath subgraph. Even though the D-values are lower in general (see Table 11.6), they are still larger than 0.2 to be considered as to be drawn from the same distribution, such that the graphs are as well not likely to be from the same distributions.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	1.00	1.00	1.00
Silwood	1.00	1.00	1.00
St. Martin Island	1.00	1.00	1.00
Ythan Estuary	1.00	1.00	1.00
Little Rock	1.00	1.00	1.00
Grassland	1.00	1.00	$9.45 \cdot 10^{-1}$
s208	$5.45 \cdot 10^{-1}$	$7.60 \cdot 10^{-1}$	$6.20 \cdot 10^{-1}$
s420	$6.50 \cdot 10^{-1}$	$8.50 \cdot 10^{-1}$	$7.10 \cdot 10^{-1}$
Gnutella 08.08.2002	1.00	1.00	1.00
Gnutella 09.08.2002	1.00	1.00	1.00
E .coli	1.00	1.00	1.00
S. cerevisiae	1.00	1.00	1.00

Table 11.3: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fork.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	1.00	1.00	1.00
Silwood	1.00	1.00	1.00
St. Martin Island	1.00	1.00	1.00
Ythan Estuary	1.00	1.00	1.00
Little Rock	1.00	1.00	1.00
Grassland	1.00	1.00	$7.85 \cdot 10^{-1}$
s208	$4.95 \cdot 10^{-1}$	$6.90 \cdot 10^{-1}$	$3.55 \cdot 10^{-1}$
s420	$6.25 \cdot 10^{-1}$	$7.85 \cdot 10^{-1}$	$4.20 \cdot 10^{-1}$
Gnutella 08.08.2002	1.00	1.00	1.00
Gnutella 09.08.2002	1.00	1.00	1.00
E .coli	1.00	1.00	$9.95 \cdot 10^{-1}$
S. cerevisiae	1.00	1.00	1.00

Table 11.4: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fan.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$8.99 \cdot 10^2 \pm 4.47$	$8.73 \cdot 10^2 \pm 1.54 \cdot 10^1$	$7.20 \cdot 10^2 \pm 3.28 \cdot 10^1$
Silwood	$9.16 \cdot 10^2 \pm 2.92$	$9.01 \cdot 10^2 \pm 1.02 \cdot 10^1$	$7.00 \cdot 10^2 \pm 4.61 \cdot 10^1$
St. Martin Island	$8.57 \cdot 10^2 \pm 4.33$	$8.22 \cdot 10^2 \pm 2.13 \cdot 10^1$	$6.40 \cdot 10^2 \pm 3.47 \cdot 10^1$
Ythan Estuary	$3.48 \cdot 10^3 \pm 6.18$	$3.37 \cdot 10^3 \pm 5.66 \cdot 10^1$	$2.51 \cdot 10^3 \pm 9.36 \cdot 10^1$
Little Rock	$2.85 \cdot 10^4 \pm 3.42 \cdot 10^1$	$2.83 \cdot 10^4 \pm 2.92 \cdot 10^2$	$2.14 \cdot 10^4 \pm 3.24 \cdot 10^2$
Grassland	$1.42 \cdot 10^2 \pm 1.45$	$1.41 \cdot 10^2 \pm 1.92$	$1.32 \cdot 10^2 \pm 5.35$
s208	$2.68 \cdot 10^2 \pm 1.93$	$2.65 \cdot 10^2 \pm 2.83$	$2.62 \cdot 10^2 \pm 4.58$
s420	$5.86 \cdot 10^2 \pm 2.11$	$5.83 \cdot 10^2 \pm 3.02$	$5.78 \cdot 10^2 \pm 4.88$
Gnutella 08.08.2002	$9.42 \cdot 10^4 \pm 6.14$	$9.32 \cdot 10^4 \pm 9.28 \cdot 10^1$	$9.32 \cdot 10^4 \pm 1.57 \cdot 10^2$
Gnutella 09.08.2002	$1.09 \cdot 10^5 \pm 5.39$	$1.08 \cdot 10^5 \pm 1.31 \cdot 10^2$	$1.08 \cdot 10^5 \pm 1.46 \cdot 10^2$
E .coli	$2.02 \cdot 10^2 \pm 5.72 \cdot 10^{-1}$	$1.64 \cdot 10^2 \pm 2.94 \cdot 10^1$	$1.92 \cdot 10^2 \pm 9.28$
S. cerevisiae	$3.27 \cdot 10^2 \pm 4.15 \cdot 10^{-1}$	$3.26 \cdot 10^2 \pm 1.51$	$3.16 \cdot 10^2 \pm 1.01 \cdot 10^1$

Table 11.5: Number of Twopaths found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$8.70 \cdot 10^{-1}$	1.00	1.00
Silwood	$7.60 \cdot 10^{-1}$	1.00	1.00
St. Martin Island	$9.05 \cdot 10^{-1}$	1.00	1.00
Ythan Estuary	$9.75 \cdot 10^{-1}$	1.00	1.00
Little Rock	$5.80 \cdot 10^{-1}$	1.00	1.00
Grassland	$3.95 \cdot 10^{-1}$	$9 \cdot 10^{-1}$	$7.80 \cdot 10^{-1}$
s208	$4.70 \cdot 10^{-1}$	$6.80 \cdot 10^{-1}$	$4 \cdot 10^{-1}$
s420	$5.05 \cdot 10^{-1}$	$7.35 \cdot 10^{-1}$	$4.25 \cdot 10^{-1}$
Gnutella 08.08.2002	1.00	1.00	$2.26 \cdot 10^{-1}$
Gnutella 09.08.2002	1.00	1.00	$4.20 \cdot 10^{-1}$
E .coli	$6.95 \cdot 10^{-1}$	$7.75 \cdot 10^{-1}$	$6.15 \cdot 10^{-1}$
S. cerevisiae	$2.65 \cdot 10^{-1}$	$9 \cdot 10^{-1}$	$6.60 \cdot 10^{-1}$

Table 11.6: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Twopaths.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$1.84 \cdot 10^2 \pm 1.06 \cdot 10^1$	$2.29 \cdot 10^2 \pm 3.16 \cdot 10^1$	$1.20 \cdot 10^2 \pm 1.27 \cdot 10^1$
Silwood	$1.71 \cdot 10^2 \pm 1.60 \cdot 10^1$	$2.32 \cdot 10^2 \pm 4.23 \cdot 10^1$	$1.01 \cdot 10^2 \pm 1.39 \cdot 10^1$
St. Martin Island	$2.50 \cdot 10^2 \pm 1.29 \cdot 10^1$	$2.94 \cdot 10^2 \pm 3.75 \cdot 10^1$	$1.38 \cdot 10^2 \pm 1.63 \cdot 10^1$
Ythan Estuary	$6.68 \cdot 10^2 \pm 3.04 \cdot 10^1$	$1.03 \cdot 10^3 \pm 1.57 \cdot 10^2$	$3.81 \cdot 10^2 \pm 2.92 \cdot 10^1$
Little Rock	$1.06 \cdot 10^4 \pm 1.45 \cdot 10^2$	$1.09 \cdot 10^4 \pm 3.14 \cdot 10^2$	$5.47 \cdot 10^3 \pm 1.67 \cdot 10^2$
Grassland	$1.21 \cdot 10^1 \pm 2.90$	$1.74 \cdot 10^1 \pm 6.56$	9.25 ± 2.83
s208	2.34 ± 1.38	6.96 ± 3.98	2.55 ± 1.56
s420	2.81 ± 1.68	7.55 ± 4.39	2.68 ± 1.61
Gnutella 08.08.2002	$6.22 \cdot 10^2 \pm 3.03 \cdot 10^1$	$7.86 \cdot 10^2 \pm 4.95 \cdot 10^1$	$6.00 \cdot 10^2 \pm 2.95 \cdot 10^1$
Gnutella 09.08.2002	$5.41 \cdot 10^2 \pm 2.54 \cdot 10^1$	$6.20 \cdot 10^2 \pm 6.74 \cdot 10^1$	$5.27 \cdot 10^2 \pm 2.42 \cdot 10^1$
E .coli	7.94 ± 3.39	$1.06 \cdot 10^1 \pm 7.65$	6.19 ± 2.67
S. cerevisiae	$1.18 \cdot 10^1 \pm 3.71$	$1.55 \cdot 10^1 \pm 8.41$	$1.06 \cdot 10^1 \pm 3.86$

Table 11.7: Number of Feed-Forward Loops found in the respective models.

The standard procedure for comparing two means is a variant of the z-score, the two-sample z-statistic [54], as mentioned before

$$z = \frac{(\bar{\mu_1} - \bar{\mu_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}}.$$

Remember that $\bar{\mu_i}$ is the observed average number of subgraphs while μ_i is the expected number of subgraphs. This value is not known, but, since it is a test for equality, one may assume $\mu_1 = \mu_2$, thus, the difference yields 0. For the values in Table 11.5, the comparison between the FDSM and the CFG yields values larger than 5 (for appropriate μ_i, σ_i see Table 11.5). Still, since the general distribution of the number of motifs is unknown, it is not necessarily valid to apply any variant of the z-score. Moreover, when we test each generated graph from one of the sets to the other set, the following happens: first, assume that we are given a graph from the FDSM to be tested whether it belongs to the samples from the CFG. Since the CFG produces a broad range of outcomes concerning the number of subgraphs, the standard deviation is "large". Thus, since we divide by the standard deviation, the resulting z-score will almost always be small. The other way around, this does not hold. The standard deviation of the FDSM is smaller than the standard deviation of the CFG and many graphs have a large z-score.

One of the more complex but popular motifs is the Feed-Forward Loop (see Fig. 10.2a). In Table 11.7 the observed means and standard deviations are denoted for this subgraph. For this subgraph, the differences are more pronounced. The standard deviation in the CFG is much larger than in the other null models and the mean number of occurrences for this subgraph is always higher. The effect can be observed more clearly with the D-values of the Kolmogorov-Smirnov two-sample tests, which are always large (see Table 11.8). In Fig. 11.3, the distributions of the Feed-Forward Loop subgraphs in the different graphs is visualized. Even though the distributions look like normal distributions, it is not guaranteed that the subgraphs are distributed according to a normal distribution. The sets of samples are small compared to the family of simple graphs. Even though the histograms overlap in many places, the CFGS histograms are much flatter and wider than the histograms of the other models. Moreover, the mean of the CFG is higher than the means of the other models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$7.65 \cdot 10^{-1}$	1.00	1.00
Silwood	$7.25 \cdot 10^{-1}$	$9.75 \cdot 10^{-1}$	$9.95 \cdot 10^{-1}$
St. Martin Island	$6.50 \cdot 10^{-1}$	1.00	1.00
Ythan Estuary	$9.85 \cdot 10^{-1}$	1.00	1.00
Little Rock	$5.40 \cdot 10^{-1}$	1.00	1.00
Grassland	$4.70 \cdot 10^{-1}$	$3.95 \cdot 10^{-1}$	$6.75 \cdot 10^{-1}$
s208	$6.40 \cdot 10^{-1}$	$7.50 \cdot 10^{-2}$	$5.70 \cdot 10^{-1}$
s420	$5.93 \cdot 10^{-1}$	$3.25 \cdot 10^{-2}$	$6.25 \cdot 10^{-1}$
Gnutella 08.08.2002	$9.92 \cdot 10^{-1}$	$3.04 \cdot 10^{-1}$	1.00
Gnutella 09.08.2002	$6.05 \cdot 10^{-1}$	$2.65 \cdot 10^{-1}$	$6.85 \cdot 10^{-1}$
E .coli	$1.80 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$3.45 \cdot 10^{-1}$
S. cerevisiae	$2.35 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$	$2.65 \cdot 10^{-1}$

Table 11.8: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Feed-Forward Loop.

Figure 11.3: Histograms and kernel density estimates of the density of subgraphs in the different graphs. This visual clue hints that the CFG overestimates while the ECFG understimates the number of motifs.

The histograms of the ECFG are more similar to the histograms of the FDSM, but the mean is usually lower. Therefore, when the ECFG is used, it can happen that the results are close to the results the FDSM would yield (row 1, column 1 in Fig. 11.3) but on the other hand, this is not guaranteed (row 2, columns 4 and row 3, column 3 in Fig. 11.3).

The other motifs do show very similar results in the sense of the Kolmogorov-Smirnoff two-sample test, see Appendix 1. The versions of the CFG in comparison to the FDSM are impractical in several ways. First, the CFG samples not uniformly at random from the family of graphs [63, 8]. Second, the number of subgraphs is easily overestimated when counting with the maximum approach. Third, when using the ECFG edges vanish non-uniformly (see Section 11.1). Fourth, when using the ECFG the number of motifs can be underestimated which is due to the lost edges. Van Hoorn and Litvak [39] suggest the ECFG as the best model in their research since they want to compare to random and slightly different graphs which follow the same distribution. For research that investigates the expected number of a certain subgraph, changed degree sequences can skew the results. Therefore, we suggest not to use any model of the CFG family for the network motif discovery process.

The first point was found by Milo et al. [62] and Newman [66], but in their work no algorithm besides the edge swap algorithm, the FDSM, was under investigation that sampled uniformly at random without resampling all graphs from a family of graphs. Therefore, in the next section the SIS is under investigation.

11.2 ON THE SEQUENTIAL IMPORTANCE SAMPLING

The SIS has the advantage of generating only simple graphs in comparison to the CFG. The disadvantage compared to the FDSM is that the edges are generated according to a probability distribution which is either uniform, degree based, or any other probability distribution. SIS and its method to choose neighbors of a node have been discussed only shortly by Blitzstein and Diaconis [8] as well as DelGenio et. al. [21] for undirected graphs; for the directed case there was no discussion which probability distribution should be chosen but only the uniform distribution was proposed [49]. Again, this implies that a node chooses uniformly at random from the set of allowed nodes, i.e., the set of nodes which a node is not connected to already and which does not prohibit the generation of the graph. The probability distribution which is used to choose neighbors is investigated in this chapter as well. The distributions investigated are the uniform choice between possible neighbors, as suggested by DelGenio et al. [49], and the distribution based on the remaining degree, as suggested by Blitzstein and Diaconis [8]. USIS is the abbreviation for the SIS with the uniform choice of nodes, DSIS is the abbreviation for SIS for which the choice of a new neighbor is based on the remaining degree.

As for the CFG, the first subgraphs under investigation are the most simple subgraphs, the Fork, and the Fan. These subgraphs should coincide with the number of subgraphs in the FDSM, regardless of the probability distribution used.

Since both, USIS and DSIS generate simple graphs with a fixed degree sequence, they do sample the simple subgraphs very well and without a fault. This behavior is expected when the models generate only simple graphs since the number of Forks (Fans) for a node in a simple graph is equal to the number of pairs (triples) of successors a node has. Thus, the sum over the nodes is a constant for simple graphs. Thus, these results are non-exceptional,

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$8.99 \cdot 10^2 \pm 4.47$	$9.00 \cdot 10^2 \pm 4.25$	$9.01 \cdot 10^2 \pm 4.09$
Silwood	$9.16 \cdot 10^2 \pm 2.92$	$9.17 \cdot 10^2 \pm 3.03$	$9.16 \cdot 10^2 \pm 2.73$
St. Martin Island	$8.57 \cdot 10^2 \pm 4.33$	$8.55 \cdot 10^2 \pm 4.49$	$8.58 \cdot 10^2 \pm 5.65$
Ythan Estuary	$3.48 \cdot 10^3 \pm 6.18$	$3.49 \cdot 10^3 \pm 6.02$	$3.49 \cdot 10^3 \pm 6.73$
Little Rock	$2.85 \cdot 10^4 \pm 3.42 \cdot 10^1$	$2.85 \cdot 10^4 \pm 1.49 \cdot 10^1$	$2.85 \cdot 10^4 \pm 1.36 \cdot 10^1$
Grassland	$1.42 \cdot 10^2 \pm 1.45$	$1.42 \cdot 10^2 \pm 9.40 \cdot 10^{-1}$	$1.42 \cdot 10^2 \pm 1.22$
s208	$2.68 \cdot 10^2 \pm 1.93$	$2.68 \cdot 10^2 \pm 1.92$	$2.67 \cdot 10^2 \pm 2.07$
s420	$5.86 \cdot 10^2 \pm 2.11$	$5.86 \cdot 10^2 \pm 2.12$	$5.86 \cdot 10^2 \pm 2.08$
Gnutella 08.08.2002	$9.42 \cdot 10^4 \pm 6.14$	$9.42 \cdot 10^4 \pm 6.68$	$9.42 \cdot 10^4 \pm 6.63$
Gnutella 09.08.2002	$1.09 \cdot 10^5 \pm 5.39$	$1.09 \cdot 10^5 \pm 5.87$	$1.09 \cdot 10^5 \pm 5.39$
E .coli	$2.02 \cdot 10^2 \pm 5.72 \cdot 10^{-1}$	$2.02 \cdot 10^2 \pm 5.43 \cdot 10^{-1}$	$2.02 \cdot 10^2 \pm 7.26 \cdot 10^{-1}$
S. cerevisiae	$3.27 \cdot 10^2 \pm 4.15 \cdot 10^{-1}$	$3.27 \cdot 10^2 \pm 4.36 \cdot 10^{-1}$	$3.27 \cdot 10^2 \pm 4.75 \cdot 10^{-1}$

Table 11.9: Number of Twopaths found in the respective models.

and the results regarding the other subgraphs are more interesting than for the most simple ones.

Regardless of the way to measure the difference, counting the Twopath motifs in the graphs drawn with the three models gives values which are quite similar. As a reminder, the two-sample z-score as applied in the former section gave z-scores which were quite large, but for some graphs not large enough to deny the possibility that the samples had the same subgraph distribution. The Kolmogorov-Smirnov two-sample tests were more explicit and showed that the distributions were indeed not the same. This indicates that the graphs were also very different and from different families of graphs. Again, D-values larger than 0.14 are indicators for different distributions of the subgraph counts.

The results in Table 11.10 show that it is more likely that the graphs are drawn with the FDSM and any SIS are from the same graph family. Exceptional are the graphs of Little Rock, Ythan Estuary, Silwood, and St. Martin Island, for which the D-value indicates that it is rather implausible that the number of subgraphs is from the same distribution independent of the probability to choose a neighbors. Furthermore, large D-values can be seen as a pointer that not only the distribution of the number of subgraphs is different, but also that the graphs are not from the same distribution. For the Twopath, it seems that the USIS is the more likely to yield graphs which are similar regarding subgraphs to the FDSM. Still, there are cases for which the DSIS has a smaller D-value; despite the cases mentioned above, the D-value is small enough in both cases such that for the Twopath uniform choosing of neighbors seems to be the better alternative.

However, since the FDSM is considered as the gold standard for motif analysis, it has to be investigated why the SIS seems to be a viable alternative for some graphs but with some it is not usable. We conducted several comparison based on the degree sequences of these and the other graphs (skewness, correlation between in- and out-degree, the exponent of the power law if the graph followed a power law, etc.), but none yielded a clear indicator. The set of graphs used contains graphs with much more skewed, but also not as skewed degree sequences as these four graphs. The Pearson correlation coefficient r between in- and out-degree sequences is also not a sufficient indicator; one of the defective graphs has r close to 0 (Silwood), for another r is close to 0.15 (Ythan Estuary), while the other two have

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$9 \cdot 10^{-2}$	$2.05 \cdot 10^{-1}$	$1.80 \cdot 10^{-1}$
Silwood	$3.20 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$	$2.40 \cdot 10^{-1}$
St. Martin Island	$2.25 \cdot 10^{-1}$	$1.45 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$
Ythan Estuary	$1.80 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Little Rock	$3.30 \cdot 10^{-1}$	$3.30 \cdot 10^{-1}$	$9 \cdot 10^{-2}$
Grassland	$1.20 \cdot 10^{-1}$	$2.05 \cdot 10^{-1}$	$1 \cdot 10^{-1}$
s208	$1.40 \cdot 10^{-1}$	$2.20 \cdot 10^{-1}$	$3.60 \cdot 10^{-1}$
s420	$4 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	$1.20 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.30 \cdot 10^{-1}$	$1.32 \cdot 10^{-1}$	$6 \cdot 10^{-2}$
Gnutella 09.08.2002	$1.25 \cdot 10^{-1}$	$7 \cdot 10^{-2}$	$1.20 \cdot 10^{-1}$
E .coli	$1 \cdot 10^{-2}$	$1.10 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
S. cerevisiae	$5 \cdot 10^{-3}$	$1.50 \cdot 10^{-2}$	$1 \cdot 10^{-2}$

Table 11.10: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Twopath.

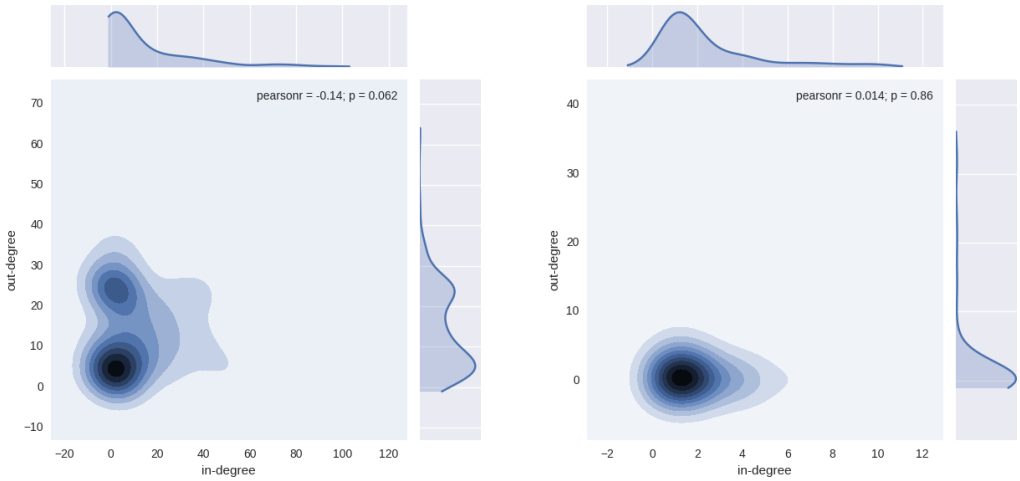
a negative r value. Still, Little Rock has a higher r than other graphs (s208 and s420 are lower), such that this cannot be the only reason. A further check on the combined degree distributions also gave no result, since the shape of the combined distribution of in- and out-degree of these three graphs resembles the combined distributions of the s208 and s420 graphs, which do yield rather low values (comp. Fig. 11.4).

As a last option, we checked whether $\forall v \in V : k_v^2 < \sum_{v \in V} k_v \wedge j_v^2 < \sum_{v \in V} j_v \wedge (k_v + j_v)^2 < \sum_{v in V} (k_v + j_v)$ [17] and denoted the percentage of nodes which violated this condition.

Graph	k	j	k+j
St Marks Seagrass	0.00	0.00	0.00
Silwood	5.84	0.00	3.25
St. Martin Island	0.00	6.67	4.44
Ythan Estuary	4.44	0.74	2.96
Little Rock	0.55	6.01	4.37
Grassland	0.00	1.14	1.14
s208	0.00	0.00	0.00
s420	0.00	0.00	0.00
Gnutella 08.09.2002	0.00	0.00	0.00
Gnutella 09.09.2002	0.00	0.00	0.00
E. coli	0.72	0.00	0.24
S. cerevisiae	1.02	0.00	0.29

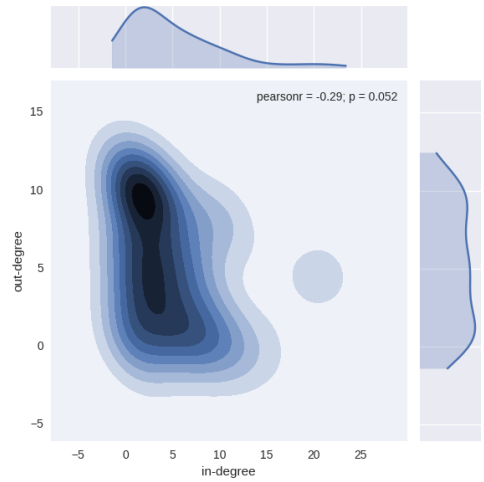
Table 11.11: Percentage of nodes which violate the condition that the square of their out-, in-, or combined degree should be smaller than the sum of the respective degree sequence.

This is the only experiment with promising results. All graphs for which the simple independence model was off have large values in at least one of the measures. The Grassland

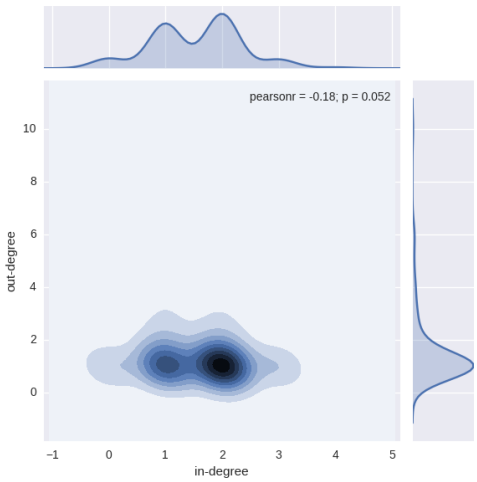


(a) Little Rock

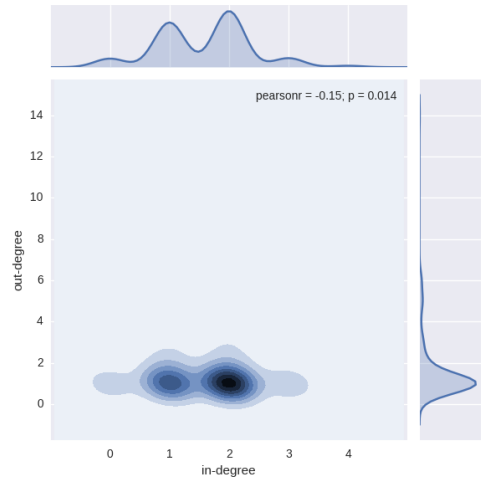
(b) Silwood



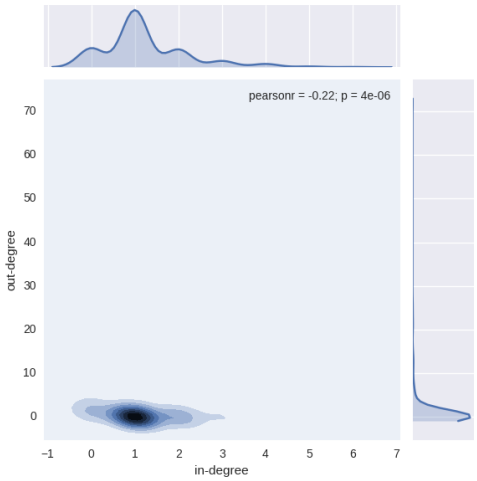
(c) St. Martin Island



(d) s208



(e) s420



(f) E. Coli

Figure 11.4: The in-, out-, and joint degree-distributions of the 11.4a Little Rock graph, 11.4b the Silwood graph, 11.4c the St. Martin Island, 11.4d the s208 graph, 11.4e the s420 graph, and 11.4f the E. coli graph.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$1.84 \cdot 10^2 \pm 1.06 \cdot 10^1$	$1.81 \cdot 10^2 \pm 1.07 \cdot 10^1$	$1.81 \cdot 10^2 \pm 1.15 \cdot 10^1$
Silwood	$1.71 \cdot 10^2 \pm 1.60 \cdot 10^1$	$1.66 \cdot 10^2 \pm 1.43 \cdot 10^1$	$1.68 \cdot 10^2 \pm 1.98 \cdot 10^1$
St. Martin Island	$2.50 \cdot 10^2 \pm 1.29 \cdot 10^1$	$2.29 \cdot 10^2 \pm 1.27 \cdot 10^1$	$2.30 \cdot 10^2 \pm 1.44 \cdot 10^1$
Ythan Estuary	$6.68 \cdot 10^2 \pm 3.04 \cdot 10^1$	$6.33 \cdot 10^2 \pm 3.46 \cdot 10^1$	$6.29 \cdot 10^2 \pm 3.17 \cdot 10^1$
Little Rock	$1.06 \cdot 10^4 \pm 1.45 \cdot 10^2$	$9.72 \cdot 10^3 \pm 1.44 \cdot 10^2$	$9.79 \cdot 10^3 \pm 1.64 \cdot 10^2$
Grassland	$1.21 \cdot 10^1 \pm 2.90$	$1.20 \cdot 10^1 \pm 2.31$	$1.20 \cdot 10^1 \pm 3.48$
s208	2.34 ± 1.38	1.93 ± 1.35	2.48 ± 1.84
s420	2.81 ± 1.68	2.79 ± 1.23	3.56 ± 1.58
Gnutella 08.08.2002	$6.22 \cdot 10^2 \pm 3.03 \cdot 10^1$	$6.17 \cdot 10^2 \pm 2.55 \cdot 10^1$	$6.22 \cdot 10^2 \pm 2.62 \cdot 10^1$
Gnutella 09.08.2002	$5.41 \cdot 10^2 \pm 2.54 \cdot 10^1$	$5.34 \cdot 10^2 \pm 2.54 \cdot 10^1$	$5.44 \cdot 10^2 \pm 2.68 \cdot 10^1$
E .coli	7.94 ± 3.39	7.50 ± 3.15	9.53 ± 4.10
S. cerevisiae	$1.18 \cdot 10^1 \pm 3.71$	$1.17 \cdot 10^1 \pm 3.05$	$1.08 \cdot 10^1 \pm 3.93$

Table 11.12: Number of Feed-Forward Loops found in the respective models.

graph has the lowest score of these "bad" graphs, thus for some motifs the sampling with the importance sampling algorithm may have similar results to the results of the FDSM. The scores of the E. coli graph and the S. cerevisiae graph are both larger than 0 but still small and as the analysis shows, they did not pose a problem.

For the Feed-Forward Loop, this is even more pronounced (see Table 11.12). The D-values are rather high for all graphs which belong to the FoodWeb-category in the columns that compare with the FDSM. The other categories are still in a good range. The Gnutella graphs have the highest value of the other categories, but the results indicate that they are still acceptable as being from the same distribution. The choice of neighbor also influences the result for this subgraph. Thus, the algorithms are not interchangeable.

For all other subgraphs, this behavior is the same - the FoodWeb-category graphs yield high D-values when comparing the FDSM and any SIS. For the remaining graphs, the comparison via the Kolmogorov-Smirnov two-sample test yields different results depending on the model used (see Appendix 2). While the number of subgraphs of graphs drawn with USIS does yield values that are either below or only slightly above the threshold which is used to decide whether the samples are from the same distribution, the DSIS shows different results. Often, when the comparison of FDSM and USIS is close to the threshold, the comparison of FDSM and DSIS yields a D-value which is above the threshold (for example, see Tables 28, 32).

When the two-sample z-statistic is used for the calculation whether the samples are from the same distribution, the results are different to some extent. While the z-statistic is often $|2| \leq z < |3.5|$ for both models compared to the FDSM, the result does not necessarily coincide with the result of the Kolmogorov-Smirnov two-sample test. For example, considering the D-scores (see Table 32) and the z-scores (see Table 11.15) of the St. Marks Seagrass graph concerning the Out-Fan, the results of the first test imply that DSIS is a better fit while the z-scores suggest the opposite. Considering the Silwood graph, it is the other way around, while for the two electronic circuits s208 and s420 the results both point to the USIS as a better fit.

It is important to keep in mind that the general form of the distribution of the number of subgraphs in the family of graphs is unknown, and it remains unknown whether the

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$2.40 \cdot 10^{-1}$	$2.10 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$
Silwood	$2.90 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$	$1.80 \cdot 10^{-1}$
St. Martin Island	$6.60 \cdot 10^{-1}$	$6.75 \cdot 10^{-1}$	$2.10 \cdot 10^{-1}$
Ythan Estuary	$5.25 \cdot 10^{-1}$	$6.05 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Little Rock	1.00	1.00	$3.10 \cdot 10^{-1}$
Grassland	$9.50 \cdot 10^{-2}$	$2.10 \cdot 10^{-1}$	$2.70 \cdot 10^{-1}$
s208	$1.55 \cdot 10^{-1}$	$6.50 \cdot 10^{-2}$	$1.20 \cdot 10^{-1}$
s420	$1 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$2.70 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.63 \cdot 10^{-1}$	$9.73 \cdot 10^{-2}$	$1.10 \cdot 10^{-1}$
Gnutella 09.08.2002	$1.75 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$1.70 \cdot 10^{-1}$
E .coli	$1.30 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$
S. cerevisiae	$1.20 \cdot 10^{-1}$	$1.40 \cdot 10^{-1}$	$2.30 \cdot 10^{-1}$

Table 11.13: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Feed-Forward Loop.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$1.65 \cdot 10^{-1}$	$1.40 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
Silwood	$2.25 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$	$3.30 \cdot 10^{-1}$
St. Martin Island	$4.05 \cdot 10^{-1}$	$4.30 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
Ythan Estuary	$1.35 \cdot 10^{-1}$	$3.35 \cdot 10^{-1}$	$4.20 \cdot 10^{-1}$
Little Rock	$8.80 \cdot 10^{-1}$	$8 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
Grassland	$1.55 \cdot 10^{-1}$	$1.05 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
s208	$6.25 \cdot 10^{-2}$	$1.75 \cdot 10^{-2}$	$8 \cdot 10^{-2}$
s420	$2 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
Gnutella 08.08.2002	$1.98 \cdot 10^{-1}$	$1.51 \cdot 10^{-1}$	$6 \cdot 10^{-2}$
Gnutella 09.08.2002	$6 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$9 \cdot 10^{-2}$
E .coli	$9 \cdot 10^{-2}$	$2.25 \cdot 10^{-1}$	$2.40 \cdot 10^{-1}$
S. cerevisiae	$1.30 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$	$1.70 \cdot 10^{-1}$

Table 11.14: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Out-Fan.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	2.58	1.94	$5.00 \cdot 10^{-1}$
Silwood	1.14	$1.30 \cdot 10^{-1}$	$5.49 \cdot 10^{-1}$
St. Martin Island	6.27	7.36	$7.72 \cdot 10^{-1}$
Ythan Estuary	$2.47 \cdot 10^{-1}$	3.35	2.70
Little Rock	$2.21 \cdot 10^{1}$	$2.14 \cdot 10^{1}$	1.61
Grassland	3.19	1.93	4.35
s208	1.70	2.55	1.84
s420	$8.55 \cdot 10^{-1}$	2.24	1.35
Gnutella 08.08.2002	2.76	3.08	$1.94 \cdot 10^{-1}$
Gnutella 09.08.2002	$9.97 \cdot 10^{-3}$	1.08	$9.58 \cdot 10^{-1}$
E .coli	$1.08 \cdot 10^{-1}$	2.74	2.33
S. cerevisiae	$4.59 \cdot 10^{-1}$	1.63	1.99

Table 11.15: Results of the two-sample z-score calculation between the different models without weights for the Out-fan motif.

z-scores can be applied or not; most likely, it also will not be known due to the size of the family of graphs. Even for small, undirected graphs (Chesapeake Bay, 33 nodes, 71 edges), the estimate for the number of graphs which can be generated is $(1.533 \pm 0.008) \cdot 10^{57}$ [8]. For larger graphs, the number will only become even larger. For directed graphs, the estimation of the number of graphs is not a direct byproduct of the sampling process, but the number of graphs in the family of graphs will not be smaller. The general form of the distribution is not known for a simple reason. It may be that the FDSM with $\mathcal{O}(m^2)$ edge-swaps traverses some parts of the underlying Markov-Chain, but depending on the number of samples, structure of the graph which is analyzed, and maybe even more features, it is not guaranteed that the complete scope of results is sufficiently covered. Ray et al. [76] showed that when using $\mathcal{O}(cm)$ edge-swaps there is no difference between $c \sim 5$ and $c = 100$ for some measures. It has to be pointed out that this can be important, and the number of samples has to be considered carefully.

Still, the uniform choice of a neighbor yields results closer to that of the FDSM, even though it may take longer than the degree based choice [8]. Blitzstein and Diaconis [8] as well as Kim et al. [21] gave worst-case runtime estimates of their algorithm which are in $\mathcal{O}(n^3)$, but claim that the algorithm is usually much faster. On the other hand, the FDSM needs $\mathcal{O}(m \log(m))$ to $\mathcal{O}(m^2)$ swaps (plus additional time for checks)

In fact, all sampling approaches are slow because they have to sample and then count. Therefore, in the next section, a set of equations is developed and tested against the FDSM to see whether the results of the equations are good enough to estimate the expected number of motifs.

11.3 A FASTER OPTION TO CALCULATE THE EXPECTED NUMBER OF MOTIFS

Equations for estimating measures can be developed using the basic equation of the SIM. While the basic equation 3.5 for undirected graphs has a dividend of $2m$, the equation for

directed graphs has to be changed. First, the equation has to include the difference between in- and out-degree; second, the dividend changes to m, such that

$$p_{u,v} = \frac{k_u j_v}{m} \tag{11.1}$$

is the probability to connect node u to node v. Of course, this is only an approximation of a probability, and it is possible that $p_{u,v}$ is larger than 1, as seen in the undirected graphs. In this chapter, I explore the possibility to use equation 11.1 to calculate the expected number of subgraphs.

The two basic subgraphs, Fork and Fan, are the first two subgraphs for which equations are developed. Considering the structure of the Fork, two edges go from one node to two other nodes, i.e., edges (u, v) and (u, w). Since this subgraph has two nodes which are isomorphous, i.e., they both have in- and out-degree $(0, 1)$, the result is divided by the factorial of the number of isomorphous nodes, 2, to prevent double counting.

$$\frac{1}{2} \sum_{u,v,w} \frac{k_u j_v}{m} \frac{(k_u - 1)\, j_w}{m} = \frac{1}{2m^2} \sum_{u,v,w} \left(k_u^2 - k_u\right) j_v j_w \tag{11.2}$$

$$\overset{\langle j\rangle = \frac{\sum_{w\in V} j_w}{n}}{=} \frac{n\langle j\rangle}{2m^2} \sum_{u,v} \left(k_u^2 - k_u\right) j_v$$

$$\overset{\langle k\rangle = \frac{n}{m}}{=} \frac{n\langle j\rangle^2}{2m\langle k\rangle} \sum_{u} \left(k_u^2 - k_u\right)$$

$$= \frac{n\left(\langle k^2\rangle - \langle k\rangle\right)\langle j\rangle^2}{2\langle k\rangle^2}$$

$$\overset{\langle j\rangle = \langle k\rangle}{=} \frac{n\left(\langle k^2\rangle - \langle k\rangle\right)}{2} \tag{11.3}$$

This equation coincides with the straightforward approach; to get the number of Forks, it is possible to count for each node how many pairs of successors it has.

$$\sum_{v\in V} \binom{k_v}{2} = \sum_{v\in V} \frac{k_v^2 - k_v}{2} \tag{11.4}$$

$$= \frac{n\left(\langle k^2\rangle - \langle k\rangle\right)}{2} \tag{11.5}$$

For the Fan, it would be how many triples of successors a node has.

$$\sum_{v\in V} \binom{k_v}{3} = \sum_{v\in V} \frac{k_v^3 - 3k_v^2 + 2k_v}{6} \tag{11.6}$$

$$= \frac{n\left(\langle k^3\rangle - 3\langle k^2\rangle + 2\langle k\rangle\right)}{6} \tag{11.7}$$

For the Fan, there are three nodes which are isomorphous, i.e. $(0, 1)$. Therefore, the equation is divided by $3! = 6$.

$$\frac{1}{6}\sum_{u,v,w,x}\frac{k_u j_v}{m}\frac{(k_u-1)\,j_w}{m}\frac{(k_u-2)\,j_x}{m} = \frac{1}{6m^3}\sum_{u,v,w,x}\left(k_u^3-3k_u^2+2k_u\right)j_u j_w j_x \tag{11.8}$$

$$= \frac{n\langle j\rangle}{6m^3}\sum_{u,v,w}\left(k_u^3-3k_u^2+2k_u\right)j_u j_v$$

$$= \frac{n\langle j\rangle^2}{6m^2\langle k\rangle}\sum_{u,v}\left(k_u^3-3k_u^2+2k_u\right)j_u$$

$$= \frac{n\langle j\rangle^3}{6m\langle k\rangle^2}\sum_{u,v}\left(k_u^3-3k_u^2+2k_u\right)$$

$$= \frac{n\left(\langle k^3\rangle-3\langle k^2\rangle+2\langle k\rangle\right)}{6} \tag{11.9}$$

A more interesting question is the calculation of the variance and thus the standard deviation of the number of subgraphs. In theory, it is straightforward. The simple independence assumption that the equations are based on allows calculating the variance of the subgraph occurrence without problems. Now, consider the following. A graph which has many low degree nodes and two hubs, i.e., high degree nodes. These high degree nodes have a higher probability to be connected. Thus, the likelihood that they are part of more than one subgraph is increased as well. Therefore, the occurrence of a subgraph would be judged as a dependent event, i.e., it is not possible to calculate the variance of a subgraph under the assumption of independence, but the equation for correlated variables has to be used.

First, the variance of an edge has to be calculated. This is done under the assumption that edges either are between two nodes or they are not. There are m experiments done to create edges (provided that a graph with m edges is to be created). Thus, the variance of an edge is

$$\mathbb{V}(e) = m\frac{k_u j_v}{m^2}\left(1-\frac{k_u j_v}{m^2}\right) \tag{11.10}$$

$$= \frac{k_u j_v}{m}\left(1-\frac{k_u j_v}{m^2}\right) \tag{11.11}$$

$$= \frac{k_u j_v}{m}-\frac{k_u^2 j_v^2}{m^3}. \tag{11.12}$$

For this equation it was used that the expected number of edges between two nodes is given as $\frac{k_u j_v}{m}$ and there are m edges, thus the probability of an edge is assumed to be $\frac{k_u j_v}{m^2}$. An approach to calculate the variance of not necessarily independent variables was given by Goodman [31, 32]. For two variables Goodman states the variance of not necessarily independent variables is

$$\begin{aligned}\mathbb{V}(xy) = {}& \mu_x^2\mathbb{V}(y)+\mu_y^2\mathbb{V}(x)+2\mu_x\mu_y\mathrm{Cov}(x-\mu_x, y-\mu_y)\\ &+2\mu_x\mathrm{Cov}\left(x-\mu_x, y^2-2y\mu_y+\mu_y^2\right)+2\mu_y\mathrm{Cov}\left(x^2-2x\mu_x+\mu_x^2\right),\\ &+\mathrm{Cov}\left(x^2-2x\mu_x+\mu_x^2, y^2-2y\mu_y+\mu_y^2\right)-(\mathrm{Cov}(x-\mu_x, y-\mu_y))^2\end{aligned} \tag{11.13}$$

Graph	FDSM	SIM
St. Marks Seagrass	$8.20 \cdot 10^2 \pm 0.00$	$8.20 \cdot 10^2 \pm 2.07 \cdot 10^1$
Silwood	$3.67 \cdot 10^3 \pm 0.00$	$3.67 \cdot 10^3 \pm 4.86 \cdot 10^1$
St. Martin Island	$7.64 \cdot 10^2 \pm 0.00$	$7.64 \cdot 10^2 \pm 2.29 \cdot 10^1$
Ythan Estuary	$4.47 \cdot 10^3 \pm 0.00$	$4.47 \cdot 10^3 \pm 6.04 \cdot 10^1$
Little Rock	$2.55 \cdot 10^4 \pm 0.00$	$2.55 \cdot 10^4 \pm 1.74 \cdot 10^2$
Grassland	$2.38 \cdot 10^2 \pm 0.00$	$2.38 \cdot 10^2 \pm 8.28$
s208	$1.64 \cdot 10^2 \pm 0.00$	$1.64 \cdot 10^2 \pm 2.93$
s420	$3.80 \cdot 10^2 \pm 0.00$	$3.80 \cdot 10^2 \pm 3.07$
Gnutella 08.08.2002	$9.35 \cdot 10^4 \pm 0.00$	$9.35 \cdot 10^4 \pm 1.48 \cdot 10^1$
Gnutella 09.08.2002	$1.22 \cdot 10^5 \pm 0.00$	$1.22 \cdot 10^5 \pm 1.44 \cdot 10^1$
E .coli	$4.82 \cdot 10^3 \pm 0.00$	$4.82 \cdot 10^3 \pm 3.92 \cdot 10^1$
S. cerevisiae	$1.18 \cdot 10^4 \pm 0.00$	$1.18 \cdot 10^4 \pm 4.01 \cdot 10^1$

Table 11.16: Number of Forks found in the respective models.

while the equation for three and more variables is much more complicated. An alternative form of the equation is given by Goodman [32] as

$$\mathbb{V}(xy) = \mu_x^2 \mu_y^2 \left\{ \mathbb{E}\left\{ [\delta_x + \delta_y + \delta_x \delta_y]^2 \right\} - A^2 \right\}, \tag{11.14}$$

where $\delta_i = \frac{p(i) - \mathbb{E}[i]}{\mathbb{E}[i]}$ and $A = B - 1 = \frac{M}{\Pi_i \mathbb{E}[i]} - 1 = \frac{\mathbb{E}[\Pi_i p(i)]}{\Pi_i \mathbb{E}[i]} - 1$. Observe, that $\delta_x = \delta_y$ since

$$\frac{\frac{k_u j_v}{m^2} - \frac{k_u j_v}{m}}{\frac{k_u j_v}{m}} = m \frac{k_u j_v - m k_u j_v}{m^2 k_u j_v} \tag{11.15}$$

$$= \frac{1}{m} - 1. \tag{11.16}$$

Using equations 11.14 and 11.16 to get the variance of the Fork subgraph yields the following equation

$$\begin{aligned} \sum_{u,v,w \in V} \left(\frac{k_u j_v}{m} \frac{(k_u - 1) j_w}{m} \right)^2 \Bigg\{ \mathbb{E}\left\{ \left[\left(\frac{1}{m} - 1 \right)^4 + 4 \left(\frac{1}{m} - 1 \right)^3 + 4 \left(\frac{1}{m} - 1 \right)^2 \right] \right\} \\ - \left(\frac{m^2 \left(\langle k^2 \rangle - \langle k \rangle \right)}{\langle k \rangle^2 \left(k_u^2 - k_u \right)^2 j_v^2 j_w^2} - \frac{m \left(k_u^2 - k_u \right)^2}{\langle k \rangle \left(k_u^2 - k_u \right) j_v j_w} + 1 \right) \Bigg\} \end{aligned} \tag{11.17}$$

$$\begin{aligned} = \frac{\left(\langle k^4 \rangle - 2\langle k^3 \rangle + \langle k^2 \rangle \right) \langle j^2 \rangle^2}{m \langle k \rangle^3} \left(\left(\frac{1}{m} - 1 \right)^4 + 4 \left(\frac{1}{m} - 1 \right)^3 + 4 \left(\frac{1}{m} - 1 \right)^2 - 1 \right) \\ - \left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle^2} \right)^2 \left(\frac{1}{m^2} - 2 \right) \end{aligned} \tag{11.18}$$

Evaluating this equation and taking the square root yields rather large values, see Table 11.16.

While the number of subgraphs itself is estimated well, the standard deviation is off. Note that the number of Forks (and Fans) in a family of graphs with fixed degree sequence

is constant, and the standard deviation of samples is, therefore, zero. However, the modeling approach with the SIM results in values larger than zero, which is obviously not good for these subgraphs.

For the Fan, the standard deviation is harder to calculate. According to Goodman [32], a general form of the variance is

$$\mathbb{V}\left(\Pi_i x_i\right) = \Pi_i \mu_i^2 \left\{ \left[\sum_i \delta_i + \sum_{i,j,i<j} \delta_i \delta_j + \cdots + \Pi_i \delta_i \right]^2 - A^2 \right\}. \tag{11.19}$$

Calculating the variance of the Fan subgraph based on equation 11.19 yields for most graphs **negative** variances. Therefore, no standard deviation can be calculated. This is true for more complex subgraphs as well. To decide whether a subgraph is a network motif, some measure of variation is needed. Thus, as an approximation of the variance and the standard deviation, I decided to use another equation provided by Goodman [31, 32]. The general equation to calculate the variance of the product of **independent** variables is given as

$$\mathbb{V}\left(\Pi_i x_i\right) = \Pi_i \mu_i^2 \left(\sum_i \frac{\mathbb{V}\left(x_i\right)}{\mu_i^2} + \sum_{i,j,i<j} \frac{\mathbb{V}\left(x_i\right)}{\mu_i^2} \frac{\mathbb{V}\left(x_j\right)}{\mu_j^2} + \cdots + \Pi_i \frac{\mathbb{V}\left(x_i\right)}{\mu_i^2} \right). \tag{11.20}$$

By using equation 11.20 we lose some information, i.e., it is possible that all of a subgraph use one same edge—but the equation will not acknowledge this and thus misestimate the variance.

Still, some estimate is better than no estimate. As δ_i before, the square of the coefficient of variation simplifies to a certain degree.

$$\frac{\mathbb{V}\left(x_i\right)}{\mu_i^2} = \frac{m \frac{k_u j_v}{m^2} \left(1 - \frac{k_u j_v}{m^2}\right)}{\left(\frac{k_u j_v}{m}\right)^2} \tag{11.21}$$

$$= \frac{m}{k_u j_v} - \frac{1}{m} \tag{11.22}$$

Graph	FDSM	SIM
St. Marks Seagrass	$2.26 \cdot 10^3 \pm 0.00$	$2.26 \cdot 10^3 \pm 1.38 \cdot 10^2$
Silwood	$3.01 \cdot 10^4 \pm 0.00$	$3.01 \cdot 10^4 \pm 5.15 \cdot 10^2$
St. Martin Island	$1.76 \cdot 10^3 \pm 0.00$	$1.76 \cdot 10^3 \pm 1.31 \cdot 10^2$
Ythan Estuary	$3.13 \cdot 10^4 \pm 0.00$	$3.13 \cdot 10^4 \pm 6.29 \cdot 10^2$
Little Rock	$2.18 \cdot 10^5 \pm 0.00$	$2.18 \cdot 10^5 \pm 2.02 \cdot 10^3$
Grassland	$3.87 \cdot 10^2 \pm 0.00$	$3.87 \cdot 10^2 \pm 4.67 \cdot 10^1$
s208	$2.22 \cdot 10^2 \pm 0.00$	$2.22 \cdot 10^2 \pm 2.08 \cdot 10^1$
s420	$6.32 \cdot 10^2 \pm 0.00$	$6.32 \cdot 10^2 \pm 3.05 \cdot 10^1$
Gnutella 08.08.2002	$3.09 \cdot 10^5 \pm 0.00$	$3.09 \cdot 10^5 \pm 5.13 \cdot 10^2$
Gnutella 09.08.2002	$4.57 \cdot 10^5 \pm 0.00$	$4.57 \cdot 10^5 \pm 6.21 \cdot 10^2$
E .coli	$7.13 \cdot 10^4 \pm 0.00$	$7.13 \cdot 10^4 \pm 6.65 \cdot 10^2$
S. cerevisiae	$1.50 \cdot 10^5 \pm 0.00$	$1.50 \cdot 10^5 \pm 6.91 \cdot 10^2$

Table 11.17: Number of Fans found in the respective models.

From equations 11.20, 11.22 we derive for the Fan

$$\begin{aligned}
&\sum_{u,v,w} \left(\frac{k_u j_v}{m} \frac{(k_u-1)\, j_w}{m} \frac{(k_u-2)\, j_x}{m}\right)^2 \left(\left(\frac{m}{k_u j_v} - \frac{1}{m}\right) + \left(\frac{m}{(k_u-1)\, j_w} - \frac{1}{m}\right)\right. \\
&\quad + \left(\frac{m}{(k_u-2)\, j_x} - \frac{1}{m}\right) + \left(\frac{m}{k_u j_v} - \frac{1}{m}\right)\left(\frac{m}{(k_u-1)\, j_w} - \frac{1}{m}\right) \\
&\quad + \left(\frac{m}{k_u j_v} - \frac{1}{m}\right)\left(\frac{m}{(k_u-2)\, j_x} - \frac{1}{m}\right) \\
&\quad + \left(\frac{m}{(k_u-1)\, j_w} - \frac{1}{m}\right)\left(\frac{m}{(k_u-2)\, j_x} - \frac{1}{m}\right) \\
&\quad \left. + \left(\frac{m}{k_u j_v} - \frac{1}{m}\right)\left(\frac{m}{(k_u-1)\, j_w} - \frac{1}{m}\right)\left(\frac{m}{(k_u-2)\, j_x} - \frac{1}{m}\right)\right)
\end{aligned} \tag{11.23}$$

$$\begin{aligned}
&= \left(1 - \frac{1}{m}\right)^2 \frac{\langle j^2\rangle}{m\langle k^3\rangle}\left(3\langle k^5\rangle - 15\langle k^4\rangle + 26\langle k^3\rangle - 18\langle k^2\rangle + 4\langle k\rangle\right) \\
&\quad + \left(1 - \frac{1}{m}\right)\frac{j^2}{\langle k\rangle^2}\left(3\langle k^4\rangle - 12\langle k^3\rangle + 15\langle k^2\rangle - 6\langle k\rangle\right) \\
&\quad - \left(3 - \frac{3}{m} + \frac{1}{m^2}\right)\frac{\langle k^6\rangle - 6\langle k^5\rangle + 13\langle k^4\rangle - 12\langle k^3\rangle + 4\langle k^2\rangle}{m^3\langle k\rangle^4} \\
&\quad + n\left(\langle k^3\rangle - 3\langle k^2\rangle + 2\langle k\rangle\right)
\end{aligned} \tag{11.24}$$

The intermediate steps are left out since they only involve multiplication, taking averages, and rearranging the equation.

Observe in Table 11.17, that this approximation seems to be not very good—when using graph generating models, this amount of discrepancy between the standard deviation in the FDSM and the model under investigation would already be enough to discard the model. Still, the expected number of subgraphs is calculated well. Too see whether the expected number of subgraphs is approximated as well as for Forks and Fans, other subgraphs are considered as well.

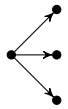
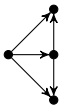
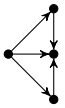
Motif		Equation
Fork		$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)n}{2}$
Fan		$\frac{\left(\langle k^3\rangle-3\langle k^2\rangle+2\langle k\rangle\right)n}{6}$
Twopath		$n\cdot\langle kj\rangle$ (revised: $n\langle kj\rangle-\frac{\langle kj\rangle^2}{\langle k\rangle^2}$)
Double-Join		$\frac{\left(\langle j^2\rangle-\langle j\rangle\right)\left(\langle k^2j\rangle-\langle kj\rangle\right)^2}{2m\langle k\rangle^3}$
Feed-Forward Loop		$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle j^2\rangle-\langle j\rangle\right)\langle kj\rangle}{\langle k\rangle^3}$
Threecycle		$\frac{\langle kj\rangle^3}{3\langle k\rangle^3}$
Fourcycle		$\frac{\langle kj\rangle^4}{4\langle k\rangle^4}$
Complete		$\frac{\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)^3}{6m^3\langle k\rangle^3}$
Out-Fan		$\frac{\left(\langle j^2\rangle-\langle j\rangle\right)^2\left(\langle k^2j\rangle-\langle kj\rangle\right)\left(\langle k^3\rangle-3\langle k^2\rangle+2\langle k\rangle\right)}{2m\langle k\rangle^4}$
In-Fan		$\frac{\left(\langle k^3\rangle-3\langle k^2\rangle+2\langle k\rangle\right)\left(\langle j^3\rangle-3\langle j^2\rangle+2\langle j\rangle\right)\langle kj\rangle^2}{2m\langle k\rangle^4}$
Biparallel		$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle j^2\rangle-\langle j\rangle\right)\langle kj\rangle^2}{2\langle k\rangle^4}$

Table 11.18: Equations developed based on the basic equation 11.1, following the approach of equation 11.3 and equation 11.9

For all subgraphs under investigation, the equations are given in Table 11.18. Some are quite simple, some more complex, reflecting the structure of the underlying subgraph-structure. Still, deriving the equations and calculating them is much faster than using simulations to find the expected occurrences of subgraphs. The equations for the basis-check perform well enough, such that the advanced motifs are under investigation in the following; the standard deviation, on the other hand, is much harder to calculate and up to now the results have been not satisfactory.

The next subgraph under investigation is the Twopath. The results are shown in Table 11.19. The means shown in Table 11.19 for the SIM are good approximations of the result of the simulations using the FDSM. Still, a hand-waving comparison is not accurate. Thus, some measure of quality is needed. Even though for comparison the z-score was avoided when possible, since the real distribution of the number of subgraphs is not known, for single-value tests it is the most prominent way to test whether a value can be from a given dataset. Thus, the test is whether the result of the equation of the SIM could be a result of the FDSM. The z-scores for the Twopath are within the bounds by Kashtan et

Graph	FDSM	SIM
St. Marks Seagrass	$8.99 \cdot 10^2 \pm 4.47$	$9.16 \cdot 10^2 \pm 3.67 \cdot 10^1$
Silwood	$9.16 \cdot 10^2 \pm 2.92$	$9.22 \cdot 10^2 \pm 3.79 \cdot 10^1$
St. Martin Island	$8.57 \cdot 10^2 \pm 4.33$	$8.72 \cdot 10^2 \pm 3.74 \cdot 10^1$
Ythan Estuary	$3.48 \cdot 10^3 \pm 6.18$	$3.52 \cdot 10^3 \pm 7.93 \cdot 10^1$
Little Rock	$2.85 \cdot 10^4 \pm 3.42 \cdot 10^1$	$2.86 \cdot 10^4 \pm 2.15 \cdot 10^2$
Grassland	$1.42 \cdot 10^2 \pm 1.45$	$1.43 \cdot 10^2 \pm 1.30 \cdot 10^1$
s208	$2.68 \cdot 10^2 \pm 1.93$	$2.70 \cdot 10^2 \pm 1.68 \cdot 10^1$
s420	$5.86 \cdot 10^2 \pm 2.11$	$5.88 \cdot 10^2 \pm 2.46 \cdot 10^1$
Gnutella 08.08.2002	$9.42 \cdot 10^4 \pm 6.14$	$9.42 \cdot 10^4 \pm 3.10 \cdot 10^2$
Gnutella 09.08.2002	$1.09 \cdot 10^5 \pm 5.39$	$1.09 \cdot 10^5 \pm 3.32 \cdot 10^2$
E .coli	$2.02 \cdot 10^2 \pm 5.72 \cdot 10^{-1}$	$2.02 \cdot 10^2 \pm 1.50 \cdot 10^1$
S. cerevisiae	$3.27 \cdot 10^2 \pm 4.15 \cdot 10^{-1}$	$3.27 \cdot 10^2 \pm 1.88 \cdot 10^1$

Table 11.19: Number of Twopaths found in the respective models.

al. [48], i.e., the z-scores of the Twopath subgraph are smaller than 5. Still, for the graphs from the biology category, the results are rather high. Considering the equation given in Table 11.18, this can be alleviated. The base equation is

$$\mathrm{Twopath} = \sum_{u,v,w \in V, u \neq v \neq w} \frac{k_u j_v}{m} \frac{k_v j_w}{m} \tag{11.25}$$

However, by taking the average of the degrees, the nodes are not guaranteed to be disjunct. This implies, that instead of only Twopaths, also reciprocal edges, $(u, v), (v, u)$, are included in this equation as well. Thus, subtracting the average number of reciprocal edges should give results closer to the values of the FDSM.

$$\mathrm{Reciprocity} = \frac{1}{2} \sum_{u,v \in V} \frac{k_u j_v}{m} \frac{k_v j_u}{m} \tag{11.26}$$

The factor $\frac{1}{2}$ is due to the degree structure of the subgraph; both nodes have the same in and out-degree. While this equation would be necessary to estimate the number of reciprocal subgraphs, in the correction of equation 11.25 it does not appear. The reason for this is the straightforward application of the idea of the Twopath: equation 11.25 overestimates for combinations such as $(u, v), (v, u)$. Since there is no factor in this equation, none is used for the correction in the revised equation. Of course, some errors are still made when applying these equations, i.e., self-loops are still included by combinations such as $(u, u), (u, v)$. To avoid counting combinations which are not permitted, the averages have to be considered more carefully. Building averages allows for combinations such as the reciprocal, but also other combinations in which u is the only node considered. Thus, subtracting nodes which

are still to be summed over from the averages should yield better results. For the Twopath, this implies the following

$$\sum_{u,v,w\in V,u\neq v\neq w} \frac{k_u j_v}{m}\frac{k_v j_w}{m} = \sum_{u,v\in V,u\neq v} \frac{k_u j_v k_v}{m^2}\left(n\langle j\rangle - \frac{j_u}{n} - \frac{j_v}{n}\right) \tag{11.27}$$

$$= \sum_{u,v\in V,u\neq v}\left(\frac{k_u k_v j_v}{m^2}n\langle j\rangle - \frac{k_u j_u k_v j_v}{m^2 n} - \frac{k_u k_v j_v^2}{m^2 n}\right) \tag{11.28}$$

$$= \sum_{u\in V}\frac{k_u}{m^2}n\langle j\rangle\left(n\langle kj\rangle - \frac{k_u j_u}{n}\right) - \sum_{u\in V}\frac{k_u j_u}{m^2 n}\left(n\langle kj\rangle - \frac{k_u j_u}{n}\right)$$

$$- \sum_{u\in V}\frac{k_u}{m^2 n}\left(n\langle kj^2\rangle - \frac{k_u j_u^2}{n}\right) \tag{11.29}$$

$$= n^3\frac{\langle k\rangle\langle kj\rangle\langle j\rangle}{m^2} - n\frac{\langle j\rangle\langle k^2 j\rangle}{m^2} - n\frac{\langle kj\rangle^2}{m^2} - n\frac{\langle kj^2\rangle\langle k\rangle}{m^2} + 2n\frac{\langle k^2 j^2\rangle}{m^2} \tag{11.30}$$

$$= n\langle kj\rangle - \frac{\langle k^2 j\rangle}{m} - \frac{\langle kj\rangle^2}{m\langle k\rangle} - \frac{\langle kj^2\rangle}{m} + 2\frac{\langle k^2 j^2\rangle}{m^2 n} \tag{11.31}$$

This equation was tested, but the results coincide for most tested subgraphs with the results of the equations given in Table 11.18. Thus, I will use the basic equations for the subgraphs.

On the other hand, the more complex subgraphs are much better estimated by the corresponding equation. Moreover, considering the standard deviation, there is an even simpler approximation than the assumption of independence of the edges which participate in a subgraph. For the Twopath, we investigated the equation and its results closer. There is one dominant term in equation 11.31, that avoids tuples such as (u, u, u): the third term, that is $\frac{2}{n}$ times the expected number of reciprocal subgraphs. Thus, this approach is promising and will be considered for other subgraphs as well.

Observe the same happening in equation 11.24, which has as a last term the equation given in Table 11.18 without a factor of $\frac{1}{6}$. This term is also the dominant, largest term of equation 11.24. The rest of the equation is neglibgble in contrast. Considering that for more complicated subgraphs the equations becoming more complex and much longer, but that they still contain the equation for the estimated number of subgraphs, it is easier to use it to approximate the standard deviation. Thus, from now on the standard deviation is approximated by the square root of the estimated number of motifs. It is important to note that this is not exact, but for the purpose of this research, it is good enough. We stress, that when the equations are used in applied research, using this approximation may result in erroneous rejections or acceptance as a network motif.

As before, the next subgraph under investigation is the Feed-Forward Loop. The results of the equation are shown in Table 11.20. Overall, the results are quite good for most graphs. This is reflected in the statistics. The z-statistic regarding the occurrence of the Feed-Forward Loop in comparison to the FDSM is in almost all graphs below 1 (see Table 11.21), such that it seems plausible to use equations instead of time-consuming simulations. Still, the Ythan Estuary graph and the Little Rock graph are examples which show that the equations are not precise estimates. As can be seen in Appendix 3 in Tables 35 to 40, the average number of subgraphs of these two graphs are hard to estimate with the equations. In particular, the high estimate for Complete subgraphs in the Ythan Estuary graph (cf. Table 35) shows that the equations are not the best solution. Since the equa-

Graph	FDSM	SIM
St. Marks Seagrass	$1.84 \cdot 10^2 \pm 1.06 \cdot 10^1$	$1.74 \cdot 10^2 \pm 1.95 \cdot 10^1$
Silwood	$1.71 \cdot 10^2 \pm 1.60 \cdot 10^1$	$1.72 \cdot 10^2 \pm 2.17 \cdot 10^1$
St. Martin Island	$2.50 \cdot 10^2 \pm 1.29 \cdot 10^1$	$2.35 \cdot 10^2 \pm 2.56 \cdot 10^1$
Ythan Estuary	$6.68 \cdot 10^2 \pm 3.04 \cdot 10^1$	$8.31 \cdot 10^2 \pm 6.34 \cdot 10^1$
Little Rock	$1.06 \cdot 10^4 \pm 1.45 \cdot 10^2$	$1.03 \cdot 10^4 \pm 1.83 \cdot 10^2$
Grassland	$1.21 \cdot 10^1 \pm 2.90$	$1.23 \cdot 10^1 \pm 5.33$
s208	2.34 ± 1.38	2.33 ± 1.63
s420	2.81 ± 1.68	2.72 ± 1.71
Gnutella 08.08.2002	$6.22 \cdot 10^2 \pm 3.03 \cdot 10^1$	$6.22 \cdot 10^2 \pm 2.58 \cdot 10^1$
Gnutella 09.08.2002	$5.41 \cdot 10^2 \pm 2.54 \cdot 10^1$	$5.44 \cdot 10^2 \pm 2.40 \cdot 10^1$
E .coli	7.94 ± 3.39	7.49 ± 3.39
S. cerevisiae	$1.18 \cdot 10^1 \pm 3.71$	$1.16 \cdot 10^1 \pm 3.93$

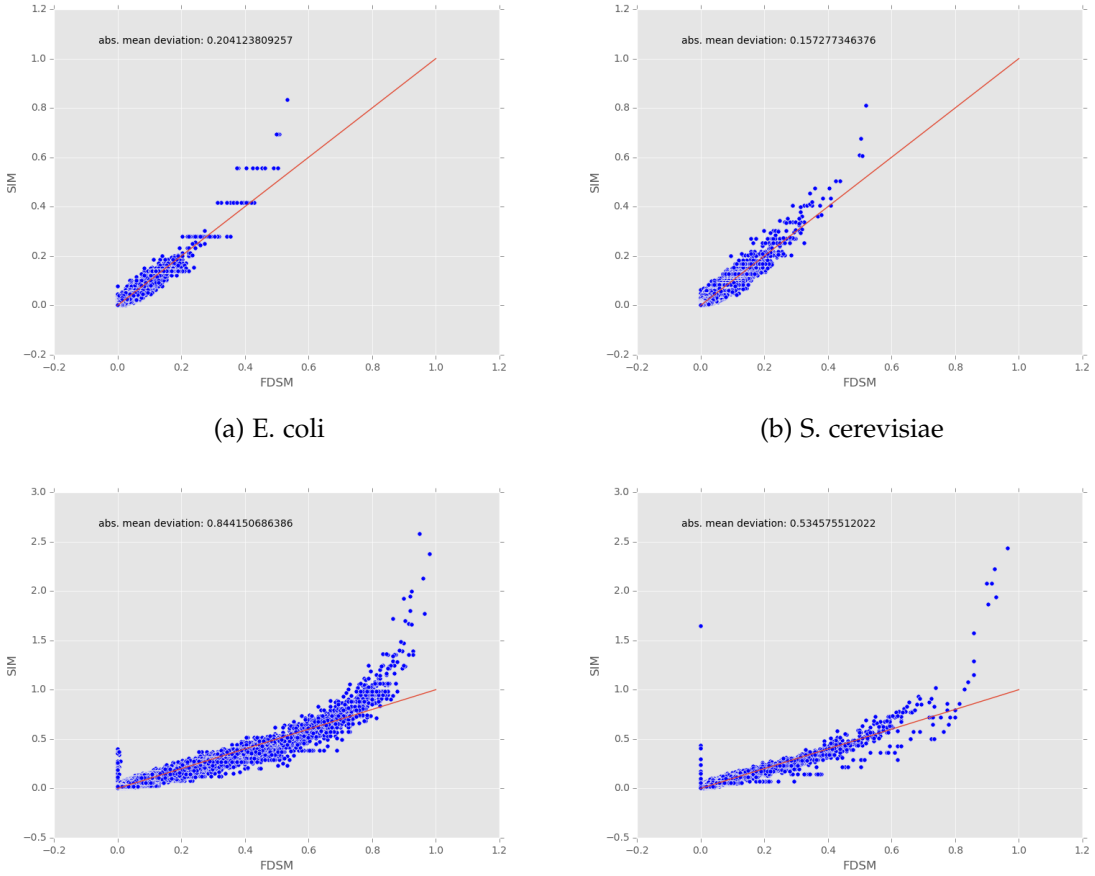
Table 11.20: Number of Feed-Forward Loops found in the respective models.

Graph	$\text{z-score}_{\text{FDSM}}$	$\text{z-score}_{\text{SIM}}$	two-sample z-score
St. Marks Seagrass	0.93	0.75	0.58
Silwood	0.09	0.12	0.07
St. Martin Island	1.19	1.00	0.76
Ythan Estuary	5.37	5.66	3.90
Little Rock	1.93	2.76	1.58
Grassland	0.06	0.05	0.04
s208	0.00	0.00	0.00
s420	0.06	0.06	0.04
Gnutella 08.08.2002	0.01	0.01	0.01
Gnutella 09.08.2002	0.10	0.11	0.08
E .coli	0.13	0.16	0.10
S. cerevisiae	0.05	0.05	0.04

Table 11.21: z-score of Feed-Forward Loops comparing the SIM with the range of values of the FDSM, z-score of the mean of the FDSM compared to the SIM, and the two-sample z-statistic of the two models, FDSM and SIM.

tions are only considered as an approximation, it can be good enough, but they have to be handled with care when the result is interpreted.

A possible way to test whether the SIM can be used to approximate the number of subgraphs is shown in the following. Comparing the average adjacency matrix, i.e., the average of the adjacency matrices of the sampled graphs to the expected probability of connections with the SIM, yields an indicator of whether the results of the equations may be reasonable. The comparison is edgewise, and one of the easiest ways is a visualization of this. In Fig. 11.5 for four graphs this is tested. The plots are akin to P-P plots



(a) E. coli

(b) S. cerevisiae

(c) Little Rock

(d) Ythan Estuary

Figure 11.5: Depicted is the average occurrence of an edge in the FDSM versus the probability of an edge calculated in the SIM. Observe that the 11.5a E. coli and 11.5b S. cerevisiae are almost perfectly aligned with only small deviations, while the 11.5c Little Rock and 11.5d Ythan Estuary show large deviations for some edges.

or Q-Q plots, with the difference that they have on the x-axis the SIM probabilities, on the y-axis the FDSM probabilities, instead of quantiles or the cumulative distribution function. The perfect plot would be a diagonal, or clustered closely around the diagonal. The closer the estimates are around the identity, the better are the results that the equations yield. In fact, the s208, the s420, and the Gnutella graphs are even more tightly clus-

tered around the identity line than the presented graphs (see Fig. 11.6). Another possible estimator error is the percentage of nodes v violating any of following conditions: $k_v^2 < \sum_{u \in V} k_u \wedge j_v^2 < \sum_{u \in V} j_u \wedge (k_v + j_v)^2 < \sum_{u \in V} (k_u + j_u)$. The results are shown in Table 11.22.

Graph	k	j	k+j
St Marks Seagrass	0.00	0.00	0.00
Silwood	5.84	0.00	3.25
St. Martin Island	0.00	6.67	4.44
Ythan Estuary	4.44	0.74	2.96
Little Rock	0.55	6.01	4.37
Grassland	0.00	1.14	1.14
s208	0.00	0.00	0.00
s420	0.00	0.00	0.00
Gnutella 08.09.2002	0.00	0.00	0.00
Gnutella 09.09.2002	0.00	0.00	0.00
E. coli	0.72	0.00	0.24
S. cerevisiae	1.02	0.00	0.29

Table 11.22: Percentage of nodes violating the condition that the square of their out-, in-, or combined degree should be smaller than the sum of the respective degree sequence.

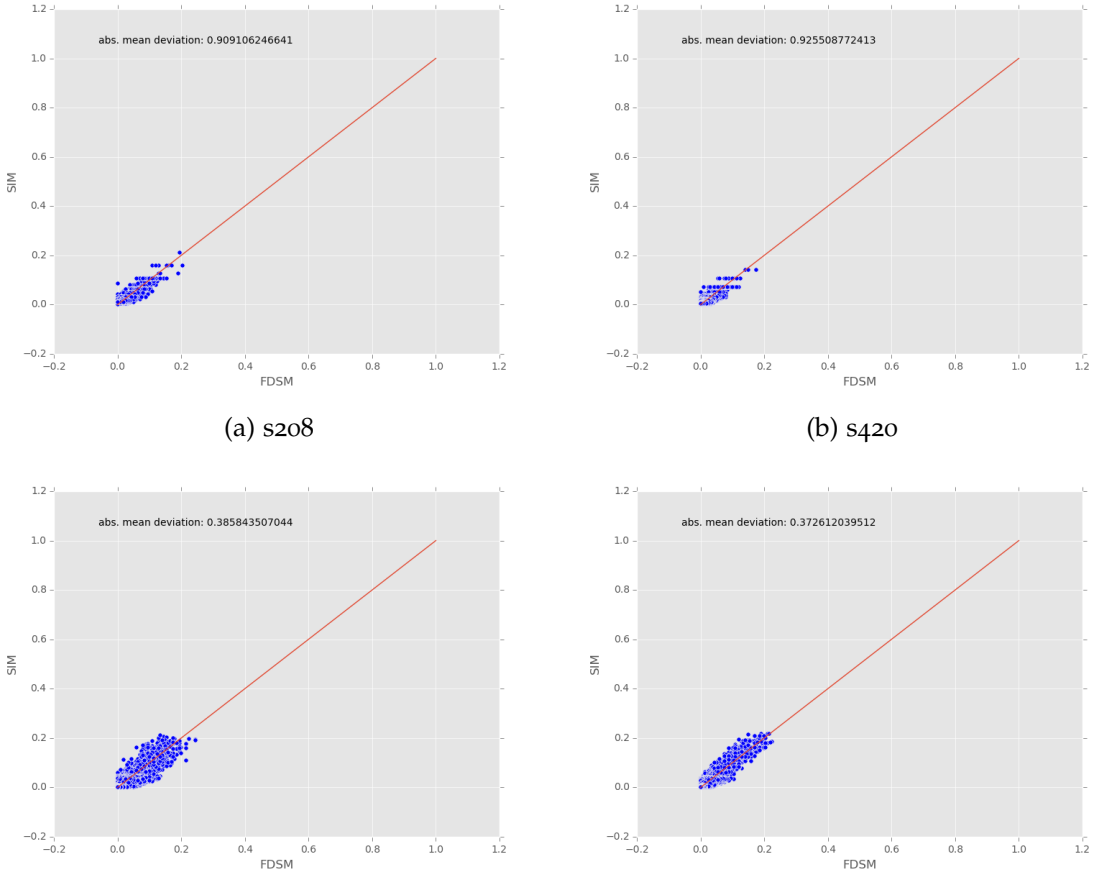
As stated before, the Little Rock graph, the Ythan Estuary graph, the Silwood graph, the St. Martin Island graph, and the Grassland graph do all show values larger than 1% for this measure—and considering the results show in Appendix 3, for these graphs the results are often not as good as for the other graphs.

Additionally, in the plot is the absolute mean error of the expected probability of the samples from FDSM and the probabilities of an edge calculated with the SIM. Graphs which are very close to the optimal line show lower values. To make this point even more clear, histograms of the observed distributions in graphs generated with the FDSM are shown together with the estimated distributions of the SIM in Figs. 11.7.

Still, the equations are good estimators for the number of subgraphs even for graphs like Little Rock. The equations of the SIM show surprisingly good results. However, for the Bifan the results are remarkably bad.

Observe in Table 37, that for almost all graphs the equation overestimates the number of occurrences of this subgraph. In Table 11.24, the z-score of the calculated expected value versus the samples of the FDSM, the average number of occurrences in the FDSM versus the calculated values of the SIM, and the two-sample z-statistic are shown for the Bifan subgraph. While for about half the graphs the results are good, i.e., $|z| < 1$, for the other half the results are worse by far; even for small graphs as the E. coli graph, the z-statistic is rather large and the result of the equation is rather unlikely to be from the same distribution as the results of the sampling (and vice versa).

For the Twopath, the equation does count some subgraphs which were not intended, i.e., instead of keeping $u \neq v, u \neq w, v \neq w$, the averages allowed implicitly for combinations like (u, v, u) or, even worse, (u, u, u). Subtracting the estimated number of reciprocal sub-



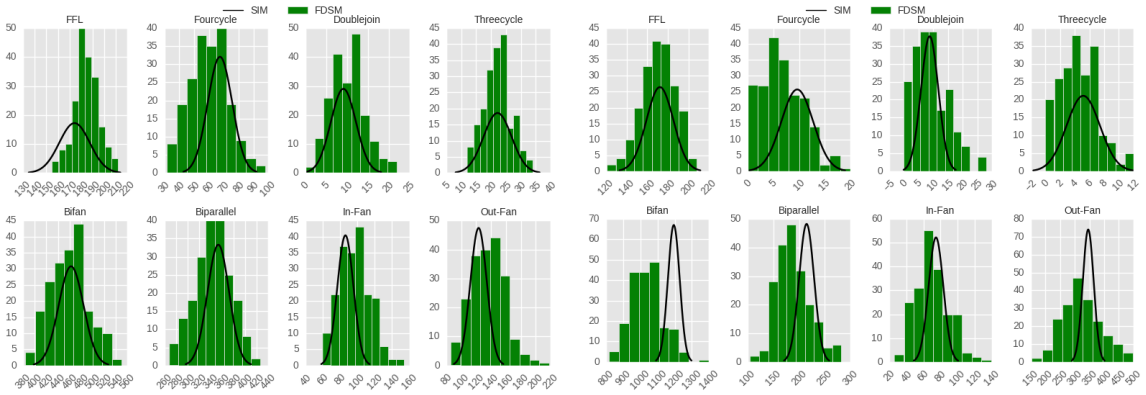
(a) s208 (b) s420

(c) Gnutella 08.09.2002 (d) Gnutella 09.09.2002

Figure 11.6: Depicted is the average occurrence of an edge in the FDSM versus the probability of an edge calculated in the SIM. All graphs show a good alignment.

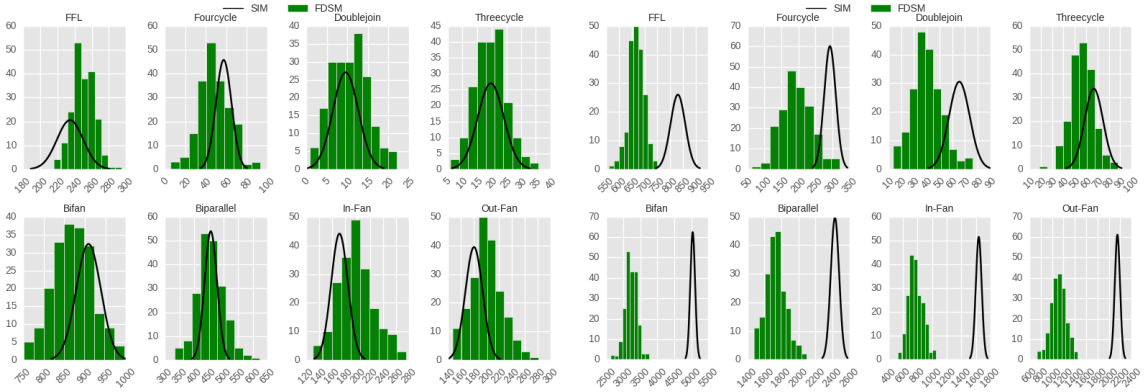
Graph	FDSM	SIM
St. Marks Seagrass	$4.60 \cdot 10^2 \pm 3.21 \cdot 10^1$	$4.63 \cdot 10^2 \pm 2.15 \cdot 10^1$
Silwood	$1.03 \cdot 10^3 \pm 9.35 \cdot 10^1$	$1.19 \cdot 10^3 \pm 3.46 \cdot 10^1$
St. Martin Island	$8.71 \cdot 10^2 \pm 4.78 \cdot 10^1$	$9.09 \cdot 10^2 \pm 3.01 \cdot 10^1$
Ythan Estuary	$3.22 \cdot 10^3 \pm 1.77 \cdot 10^2$	$5.04 \cdot 10^3 \pm 7.10 \cdot 10^1$
Little Rock	$1.67 \cdot 10^5 \pm 2.70 \cdot 10^3$	$2.01 \cdot 10^5 \pm 4.48 \cdot 10^2$
Grassland	$1.43 \cdot 10^1 \pm 6.22$	$3.46 \cdot 10^1 \pm 5.88$
s208	$6.48 \cdot 10^{-1} \pm 8.96 \cdot 10^{-1}$	$6.68 \cdot 10^{-1} \pm 8.17 \cdot 10^{-1}$
s420	$7.73 \cdot 10^{-1} \pm 8.78 \cdot 10^{-1}$	$8.49 \cdot 10^{-1} \pm 9.21 \cdot 10^{-1}$
Gnutella 08.08.2002	$4.61 \cdot 10^3 \pm 1.36 \cdot 10^2$	$4.70 \cdot 10^3 \pm 6.86 \cdot 10^1$
Gnutella 09.08.2002	$4.14 \cdot 10^3 \pm 1.20 \cdot 10^2$	$4.22 \cdot 10^3 \pm 6.50 \cdot 10^1$
E .coli	$6.48 \cdot 10^1 \pm 1.32 \cdot 10^1$	$9.26 \cdot 10^1 \pm 9.63$
S. cerevisiae	$3.06 \cdot 10^2 \pm 3.50 \cdot 10^1$	$3.49 \cdot 10^2 \pm 1.87 \cdot 10^1$

Table 11.23: Number of Bifans found in the respective models.



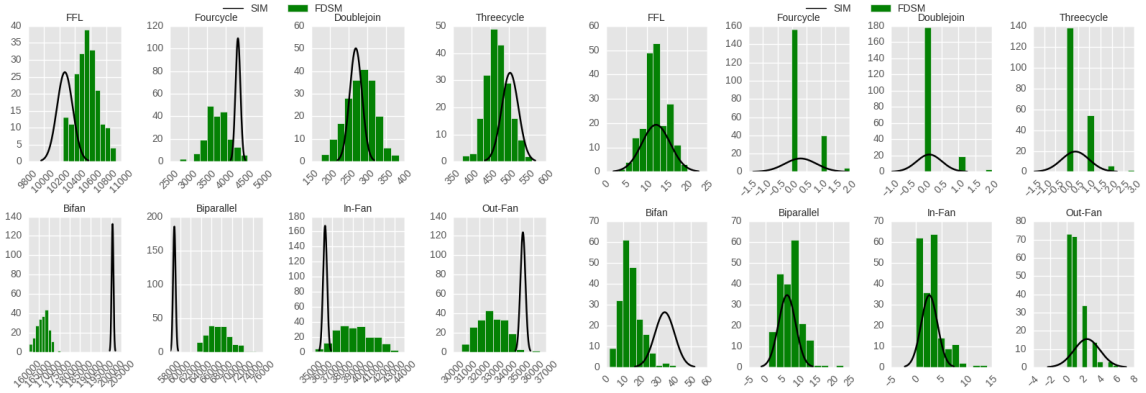
(a) St. Marks Seagrass

(b) Silwood



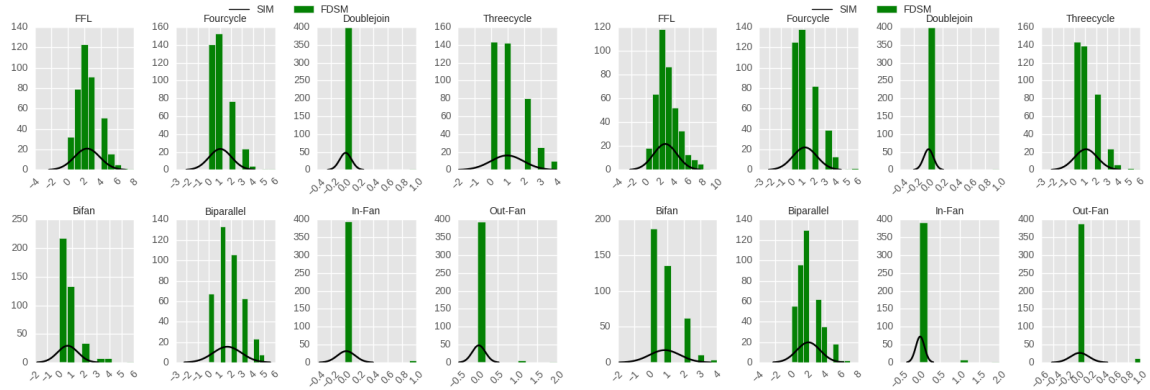
(c) St. Martin Island

(d) Ythan Estuary

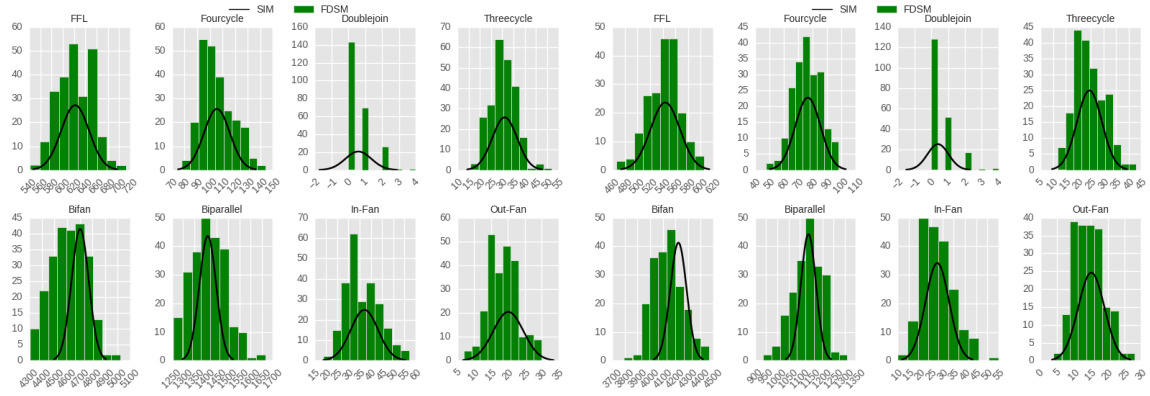


(e) Little Rock

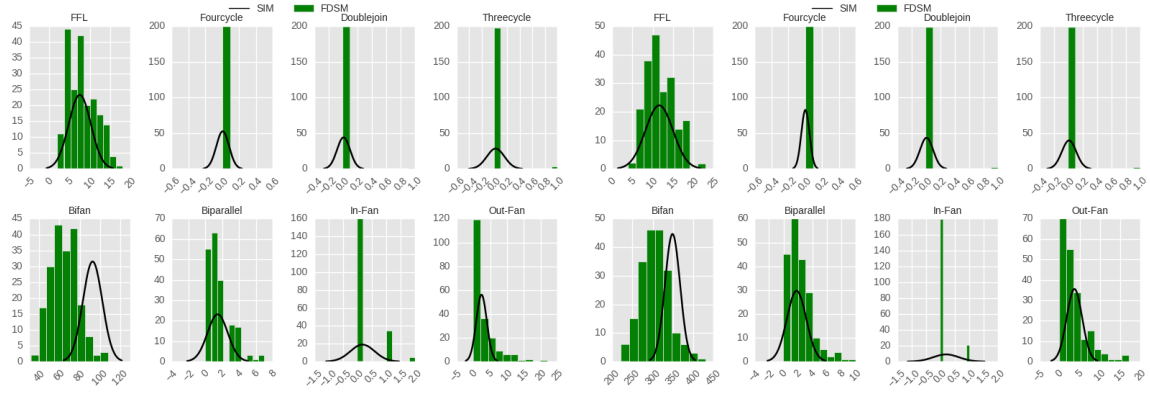
(f) Grassland



(g) s208 (h) s420



(i) Gnutella 08.08.2002 (j) Gnutella 09.08.2002



(k) E. coli (l) S. cerevisiae

Figure 11.7: Histograms of the distribution of subgraphs in different graphs, compared with the estimated distribution via the SIM.

Graph	z-score$_{\text{FDSM}}$	z-score$_{\text{SIM}}$	two-sample z-score
St. Marks Seagrass	0.09	0.13	0.07
Silwood	1.71	4.62	1.60
St. Martin Island	0.79	1.25	0.67
Ythan Estuary	10.31	25.64	9.57
Little Rock	12.41	74.87	12.24
Grassland	3.28	3.46	2.38
s208	0.02	0.02	0.02
s420	0.09	0.08	0.06
Gnutella 08.08.2002	0.66	1.31	0.59
Gnutella 09.08.2002	0.72	1.33	0.63
E .coli	2.11	2.89	1.70
S. cerevisiae	1.22	2.28	1.08

Table 11.24: z-score of Bifans comparing the SIM with the range of values of the FDSM, z-score of the mean of the FDSM compared to the SIM, and the two-sample z-statistic of the two models.

graphs alleviated this problem already. The remaining parts of the equations were close to 0 (see Table 11.25 for change in means).

For the sake of completeness, all overcounted combinations would have to be calculated as well. In applied research, this would be a necessity. Otherwise, it may happen that results are accepted/rejected falsely.

This is shown in more detail on the example of the Bifan. The estimates of the Bifan with the simple equation are, in some cases, far off (see Table 37). Thus, the correction is done in the following.

11.4 REVISITING THE BIFAN

For the Bifan, additional constellations have to be subtracted to get more reasonable results. There are several way to derive the equations that have to be subtracted. First, one can think of re-using a node to have a three-node structure such as (u, v, w, w), (u, v, x, x), (u, v, v, x) and so forth and draw all possible graphs which come into existence first for three, then for two, then for one node. Explicitly speaking, this is calculating all possible graphs which contribute to the overestimate in the equation. Constructing for each possible visualization the corresponding equation is straightforward as is shown in Table 11.26. The only thing that has to be considered is the principle of inclusion and exclusion; when subtracting the equations for three-node sub-subgraphs from the equation for a four-node subgraph, the two-node sub-subgraphs are already subtracted twice such that they have to be added while the one-node sub-subgraph has to be subtracted again. The second option is to calculate the equation directly.

Graph	FDSM	SIM_{old}	SIM_{mod}
St. Marks Seagrass	$8.99 \cdot 10^2$	$9.16 \cdot 10^2$	$9.00 \cdot 10^2$
Silwood	$9.16 \cdot 10^2$	$9.22 \cdot 10^2$	$9.16 \cdot 10^2$
St. Martin Island	$8.57 \cdot 10^2$	$8.72 \cdot 10^2$	$8.57 \cdot 10^2$
Ythan Estuary	$3.48 \cdot 10^3$	$3.52 \cdot 10^3$	$3.48 \cdot 10^3$
Little Rock	$2.85 \cdot 10^4$	$2.86 \cdot 10^4$	$2.85 \cdot 10^4$
Grassland	$1.42 \cdot 10^2$	$1.43 \cdot 10^2$	$1.42 \cdot 10^2$
s208	$2.68 \cdot 10^2$	$2.70 \cdot 10^2$	$2.68 \cdot 10^2$
s420	$5.86 \cdot 10^2$	$5.88 \cdot 10^2$	$5.86 \cdot 10^2$
Gnutella 08.08.2002	$9.42 \cdot 10^4$	$9.42 \cdot 10^4$	$9.42 \cdot 10^4$
Gnutella 09.08.2002	$1.09 \cdot 10^5$	$1.09 \cdot 10^5$	$1.09 \cdot 10^5$
E .coli	$2.02 \cdot 10^2$	$2.02 \cdot 10^2$	$2.02 \cdot 10^2$
S. cerevisiae	$3.27 \cdot 10^2$	$3.27 \cdot 10^2$	$3.27 \cdot 10^2$

Table 11.25: Number of Twopaths calculated with the simple equation and the slightly modified version.

$$\begin{aligned}
&\frac{1}{m^4} \sum_{u,v,w,x \in V} \left(k_u^2 - k_u\right)\left(k_v^2 - k_v\right)\left(j_w^2 - j_w\right)\left(j_x^2 - j_x\right) \\
&\qquad = \frac{1}{m^4} \sum_{u,v,w \in V} \left(k_u^2 - k_u\right)\left(k_v^2 - k_v\right)\left(j_w^2 - j_w\right) \\
&\left(n\left(\langle j^2\rangle - \langle j\rangle\right) - \left(\frac{j_u^2}{n} - \frac{j_u}{n}\right) - \left(\frac{j_v^2}{n} - \frac{j_v}{n}\right) - \left(\frac{j_w^2}{n} - \frac{j_w}{n}\right)\right) \qquad (11.32)
\end{aligned}$$

This equation is way more complicated to handle, and the construction of the equation, i.e., introduction of new terms, occurs always as soon as new averages are introduced. However, overall, it results in the same equation as the first option. The complete development of the equation is not shown. Thus, the first option is much simpler.

The equations in Table 11.26 result in the values given in Table 11.27. This results in the improved scores in Table 11.28. In the last column in Table 11.28, the z-statistic is calculated between the corrected Bifan-equation result and the results from the samples from the FDSM are shown. For almost all graphs, the z-score improved, but the St. Martin Island graph and the St. Marks Seagrass graph (FDSM: 871, resp. 460; SIM_{old}: 908.92, resp. 462.81; $SIM_{c}orrected$: 748.34, resp. 397.20). The Gnutella graphs have a bit higher z-statistic, but they are still very low. Thus, the modification of the simple equation could prove useful. The results of the simple equation for the Ythan Estuary graph and the Little Rock graph were not even close to the real result. With the modified equation, the z-statistic is still large, even though the results are closer to the averages from the samples in the FDSM (see Table 37).

This begs the question why the correction factors and the correct way to derive them are introduced this late and not earlier. The answer to this problem is simple. When applying the SIM, the result is often thought of as an approximation to reality [66, p.448]. Thus, the results of the equations in comparison to the averages from samples was already satisfying for most subgraphs under investigation. For the Bifan, I showed that these modifiers do

Node combination	Depiction	Equation
(u, v, w, u)		$\frac{\left(\langle k^2j^2\rangle-\langle kj^2\rangle-\langle k^2j\rangle+\langle kj\rangle\right)\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle j^2\rangle-\langle j\rangle\right)}{4m\langle k\rangle^3}$
(u, v, w, w)		$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)^2\left(\langle j^4\rangle-2\langle j^3\rangle+\langle j^2\rangle\right)}{4m\langle k\rangle^3}$
(u, w, w, x)		$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle j^2\rangle-\langle j\rangle\right)\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)}{4m\langle k\rangle^3}$
(u, u, w, x)		$\frac{\left(\langle j^2\rangle-\langle j\rangle\right)^2\left(\langle k^4\rangle-2\langle k^3\rangle+\langle k^2\rangle\right)}{4m\langle k\rangle^3}$
(u, u, u, w)		$\frac{\left(\langle j^2\rangle-\langle j\rangle\right)\left(\langle k^4j^2\rangle-2\langle k^3j^2\rangle+\langle k^2j^2\rangle-\langle k^4j\rangle+2\langle k^3j\rangle-\langle k^2j\rangle\right)}{4m^2\langle k\rangle^2}$
(u, u, w, w)		$\frac{\left(\langle k^4\rangle-2\langle k^3\rangle+2\langle k^2\rangle\right)\left(\langle j^4\rangle-2\langle j^3\rangle+\langle j^2\rangle\right)}{4m^2\langle k\rangle^2}$
(u, w, u, w)		$\frac{\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)^2}{4m^2\langle k\rangle^2}$
(u, w, w, w)		$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle k^2j^4\rangle-2\langle k^2j^3\rangle+\langle k^2j^2\rangle-\langle kj^4\rangle+2\langle kj^3\rangle-\langle kj^2\rangle\right)}{4m^2\langle k\rangle^2}$
(u, u, u, u)		$\frac{\langle k^4j^4\rangle-2\langle k^4j^3\rangle+\langle k^4j^2\rangle-2\langle k^3j^4\rangle+4\langle k^3j^3\rangle-2\langle k^3j^2\rangle+\langle k^2j^4\rangle-2\langle k^2j^3\rangle+\langle k^2j^2\rangle}{4m^3\langle k\rangle}$

Table 11.26: These graphs are counted additionally when the simple Bifan equation is used. Subtracting the corresponding equations yields more reasonable results.

influence the result. Moreover, during research regarding network motifs, I tended to think more in terms of simple digraphs and not, as shown in Table 11.26, in digraphs containing self-loops or multiple edges. Of course, when the SIM is applied, it should be obvious that those are implicitly contained in the equations as shown throughout this work. However, results which are very close to the results found with the FDSM can occlude simple facts very easily. Worse, when calculating the sub-subgraphs of the Feed-Forward Loop, the estimated values are low (see Table 11.29). Since the Feed-Forward Loop is the first subgraph which is more complicated but still simple enough to be thoroughly investigated, it was not obvious that these sub-subgraphs are of importance.

The error by over-counting combinations such as (u, v, u) and similar is small, for almost all subgraphs considered. Thus, as a first estimate, it seems applicable to use the simple equations as in Table 11.18. When this estimate is significantly different from the real-world graph's number of subgraphs, there are two choices: either using the still fast and exact equation to calculate the estimated number of subgraphs more accurately or sampling graphs from the appropriate null model, the FDSM. Using the more accurate and more complicated equations as well as the correct equations for the standard deviation may already be enough. For example, the expected number of Twopaths was estimated very well by the exact equation which takes care of overestimates; the standard deviation calculated using the equation for independent variables (equation 11.20) is also more exact than taking the square root of the expected number of subgraphs. Thus, this can give more insight to whether the subgraph under consideration is, in fact, a network motif.

Graph	Bifan									
St. Marks Seagrass	$4.63 \cdot 10^{2}$	6.28	$2.53 \cdot 10^{1}$	6.28	$3.01 \cdot 10^{1}$	$2.64 \cdot 10^{-1}$	1.64	$8.53 \cdot 10^{-2}$	$3.17 \cdot 10^{-1}$	$9.63 \cdot 10^{-3}$
Silwood	$1.19 \cdot 10^{3}$	9.07	$4.75 \cdot 10^{1}$	9.07	$1.20 \cdot 10^{2}$	$4.17 \cdot 10^{-1}$	4.78	$6.88 \cdot 10^{-2}$	$4.11 \cdot 10^{-1}$	$1.79 \cdot 10^{-2}$
St. Martin Island	$9.09 \cdot 10^{2}$	9.95	$1.03 \cdot 10^{2}$	9.95	$4.43 \cdot 10^{1}$	$2.57 \cdot 10^{-1}$	5.01	$1.09 \cdot 10^{-1}$	1.07	$1.75 \cdot 10^{-2}$
Ythan Estuary	$5.04 \cdot 10^{3}$	$1.16 \cdot 10^{2}$	$5.81 \cdot 10^{2}$	$1.16 \cdot 10^{2}$	$3.29 \cdot 10^{2}$	6.17	$3.80 \cdot 10^{1}$	2.66	$2.89 \cdot 10^{1}$	1.62
Little Rock	$2.01 \cdot 10^{5}$	$4.80 \cdot 10^{2}$	$8.16 \cdot 10^{3}$	$4.80 \cdot 10^{2}$	$3.51 \cdot 10^{3}$	3.19	$1.43 \cdot 10^{2}$	1.15	$1.07 \cdot 10^{1}$	$4.40 \cdot 10^{-2}$
Grassland	$3.46 \cdot 10^{1}$	$2.32 \cdot 10^{-2}$	$1.41 \cdot 10^{1}$	$2.32 \cdot 10^{-2}$	3.38	$3.42 \cdot 10^{-3}$	1.38	$1.55 \cdot 10^{-5}$	$10.00 \cdot 10^{-5}$	$1.47 \cdot 10^{-5}$
s208	$6.68 \cdot 10^{-1}$	$2.61 \cdot 10^{-3}$	$1.46 \cdot 10^{-2}$	$2.61 \cdot 10^{-3}$	$7.72 \cdot 10^{-2}$	$7.95 \cdot 10^{-5}$	$1.69 \cdot 10^{-3}$	$1.02 \cdot 10^{-5}$	$1.22 \cdot 10^{-4}$	$4.28 \cdot 10^{-6}$
s420	$8.49 \cdot 10^{-1}$	$1.56 \cdot 10^{-3}$	$9.05 \cdot 10^{-3}$	$1.56 \cdot 10^{-3}$	$6.54 \cdot 10^{-2}$	$3.19 \cdot 10^{-5}$	$6.97 \cdot 10^{-4}$	$2.88 \cdot 10^{-6}$	$2.61 \cdot 10^{-5}$	$6.39 \cdot 10^{-7}$
Gnutella 08.09.2002	$4.70 \cdot 10^{3}$	1.68	$5.89 \cdot 10^{1}$	1.68	4.88	$8.18 \cdot 10^{-4}$	$6.12 \cdot 10^{-2}$	$6.00 \cdot 10^{-4}$	$2.36 \cdot 10^{-2}$	$1.06 \cdot 10^{-5}$
Gnutella 09.09.2002	$4.22 \cdot 10^{3}$	$9.63 \cdot 10^{-1}$	$4.98 \cdot 10^{1}$	$9.63 \cdot 10^{-1}$	4.93	$4.18 \cdot 10^{-4}$	$5.81 \cdot 10^{-2}$	$2.19 \cdot 10^{-4}$	$1.27 \cdot 10^{-2}$	$4.51 \cdot 10^{-6}$
E. coli	$9.26 \cdot 10^{1}$	$2.98 \cdot 10^{-2}$	1.65	$2.98 \cdot 10^{-2}$	$2.76 \cdot 10^{1}$	$2.66 \cdot 10^{-4}$	$4.91 \cdot 10^{-1}$	$9.59 \cdot 10^{-6}$	$5.11 \cdot 10^{-4}$	$4.49 \cdot 10^{-6}$
S. cerevisiae	$3.49 \cdot 10^{2}$	$4.31 \cdot 10^{-2}$	5.75	$4.31 \cdot 10^{-2}$	$2.90 \cdot 10^{1}$	$3.36 \cdot 10^{-4}$	$4.77 \cdot 10^{-1}$	$5.32 \cdot 10^{-6}$	$3.56 \cdot 10^{-4}$	$2.54 \cdot 10^{-6}$

Table 11.27: In the first column, the estimated number of Bifans using $\frac{\left(\langle k^2\rangle - \langle k\rangle\right)^2 \left(\langle j^2\rangle - \langle j\rangle\right)^2}{4\langle k\rangle^4}$ is shown; the other columns contain the estimated number of subgraphs that do not occur in simple graphs but do contribute to the estimate in the first column.

Graph	Old	Corrected	z_{FDSM}
St. Marks Seagrass	462.81	397.20	1.95
Silwood	1194.62	1014.56	0.22
St. Martin Island	908.92	748.34	2.57
Ythan Estuary	5039.89	3972.14	4.26
Little Rock	200 577.01	188 102.33	7.79
Grassland	34.62	18.43	0.67
s208	0.67	0.57	0.08
s420	0.85	0.77	0.14
Gnutella 08.09.2002	4703.21	4636.11	0.17
Gnutella 09.09.2002	4224.32	4167.74	0.25
E. coli	92.64	63.85	0.07
S. cerevisiae	348.98	314.65	0.24

Table 11.28: Change in the Bifan-equation when the principle of inclusion and exclusion is applied.

Graph	FFL				
St. Marks Seagrass	$1.74 \cdot 10^{2}$	$4.83 \cdot 10^{-2}$	$1.33 \cdot 10^{-1}$	$1.30 \cdot 10^{-1}$	$3.29 \cdot 10^{-1}$
Silwood	$1.72 \cdot 10^{2}$	$8.49 \cdot 10^{-3}$	$5.86 \cdot 10^{-2}$	$3.53 \cdot 10^{-2}$	$4.17 \cdot 10^{-1}$
St. Martin Island	$2.35 \cdot 10^{2}$	$5.71 \cdot 10^{-2}$	$1.50 \cdot 10^{-1}$	$3.16 \cdot 10^{-1}$	$4.66 \cdot 10^{-1}$
Ythan Estuary	$8.31 \cdot 10^{2}$	$1.41 \cdot 10^{-1}$	$2.28 \cdot 10^{-1}$	$6.28 \cdot 10^{-1}$	8.82
Little Rock	$1.03 \cdot 10^{4}$	$1.34 \cdot 10^{-1}$	$3.36 \cdot 10^{-1}$	1.01	1.21
Grassland	$1.23 \cdot 10^{1}$	$9.35 \cdot 10^{-5}$	$6.21 \cdot 10^{-3}$	$1.35 \cdot 10^{-4}$	$2.03 \cdot 10^{-3}$
s208	2.33	$7.47 \cdot 10^{-5}$	$2.93 \cdot 10^{-4}$	$2.52 \cdot 10^{-4}$	$8.32 \cdot 10^{-4}$
s420	2.72	$1.98 \cdot 10^{-5}$	$8.76 \cdot 10^{-5}$	$6.28 \cdot 10^{-5}$	$2.23 \cdot 10^{-4}$
Gnutella 08.09.2002	$6.22 \cdot 10^{2}$	$3.52 \cdot 10^{-5}$	$5.12 \cdot 10^{-5}$	$5.14 \cdot 10^{-4}$	$2.90 \cdot 10^{-3}$
Gnutella 09.09.2002	$5.44 \cdot 10^{2}$	$1.53 \cdot 10^{-5}$	$3.12 \cdot 10^{-5}$	$2.50 \cdot 10^{-4}$	$1.43 \cdot 10^{-3}$
E. coli	7.49	$5.77 \cdot 10^{-6}$	$1.62 \cdot 10^{-4}$	$1.31 \cdot 10^{-4}$	$7.93 \cdot 10^{-4}$
S. cerevisiae	$1.16 \cdot 10^{1}$	$2.09 \cdot 10^{-6}$	$1.15 \cdot 10^{-4}$	$5.74 \cdot 10^{-5}$	$4.99 \cdot 10^{-4}$

Table 11.29: In the first column, the estimated number of Feed-Forward Loops using $\frac{\left(\langle k^2\rangle - \langle k\rangle\right)\langle kj\rangle\left(\langle j^2\rangle - \langle j\rangle\right)}{\langle k\rangle^3}$ is shown; the other columns contain the estimated number of subgraphs that do not occur in simple graphs but do contribute to estimate in the first column.

Even though the approach with the equations is elegant, I recommend in the end using the sampling approach whenever the equations tell the researcher that the subgraph is a network motif. Of course, it is also possible to only apply the equations to approximate the number of subgraphs. However, when decisions are made based on the results of these approximations, the decisions may be wrong.

11.4.1 *Summary*

It was tested whether the CFG can be used reliably in the analysis of digraphs to estimate the number of subgraphs of a particular type. The CFG produces multiple edges and self-loops, including them in the process of subgraph counting shows that the CFG does not give reliable results. When self-loops and multiple edges are dismissed, i.e., the ECFG is used, the results of the subgraph counting algorithm are much lower than the ones from the samples from the FDSM. The SIS is different than in the chapter before; while the uniform choice of neighbors from the set of possible neighbors yields for most graphs results closer to the results of the FDSM, the DSIS is not far off and in some cases even better. The algorithm of the SIS has a worst-case run-time of $\mathcal{O}\left(n^3\right)$; the swapping algorithm which is used for the FDSM uses $\mathcal{O}\left(m \log\left(m\right)\right)$ swaps with some additional checks and the swaps themselves; this will result in at most $\mathcal{O}\left(n\right)$ such that, the runtime of the swapping algorithm is in $\mathcal{O}\left(nm \log\left(m\right)\right)$. Considering that researchers in social network analysis assume that in real-world graphs the order and the size of a graph are asymptotically equal, $n \approx m$ [71], the swapping algorithm is to be considered faster.

Still, the swapping algorithm and the analysis of the number of subgraphs is time-intensive work. Being able to give equations that allow for quick calculation of the expected number of subgraphs is much easier. As shown, most of the equations work well for most graphs—even without the correction. As a quick estimate, the equations are good enough in most cases. Whenever the number of subgraphs in the real-world graph differs more than three standard deviations estimated with the SIM, a more in-depth analysis with the FDSM to check whether the real-world graph is truly different from generated graphs has to be applied. This additional analysis is needed to confirm that the subgraph is a network motif. Otherwise, when the number of subgraphs in the real-world graph and the estimated number of subgraphs by the equation coincide or are close to each other, it can be considered as optional.

As the example of the Bifan subgraph shows, sometimes it is useful to consider also the subgraphs which are implicitly included in the equations, even though they do not follow the restriction of a simple graph. Calculating the expected value for them can be done explicitly as shown in Tables 11.26, 11.27.

In the next part, the equations are investigated more in-depth. For undirected graphs, the global co-occurrence measure of the SIM was quite good, the results are very close to that of the FDSM. However, when investigating the co-occurrence on a node base, i.e., how often does a specific node co-occur with others, the results were worse. One would expect that this effect translates to the subgraph counting as well. This is tested in the following.

12

NODE-BASED PARTICIPATION ESTIMATION IN MOTIFS

There has not been much research concerning local network motifs; neither is much done on the participation of nodes in subgraphs. Sometimes, local network motifs are called node-based triad patterns [104]. For this thesis, a local network motif is defined as a subgraph a node is a part of more often than expected. While this is akin to the standard definition of a network motif (a subgraph occurring more often than expected), standard research concerns itself with the whole graph. In this part of the research, we are interested in a single node. Research that was done based on local network motifs is sparse. Koschützki et al. [52] developed a motif-based centrality measure based on real-world graphs without sampling. The centrality measure is straightforward; it works by counting in how many subgraphs of a type a node participates. The more subgraphs it participates in, the higher the centrality. In a recent paper Wang et al. [98] developed a new importance measure for nodes which uses the participation of nodes in subgraphs as part of the measure and builds upon the work of Koschützki et al. [52] but uses the sampling approach with the FDSM. In 2015, Winkler and Reichardt [104] published their work on finding the node specific triad-census, i.e., they compare the number of subgraphs a node participates in with the expected number of subgraphs. The expected number of subgraphs a node participates in is the mean of several hundred samples which have been drawn with the FDSM.

Developing equations to derive the estimated number of subgraphs in a graph worked well, but up to now this is a global measure. The SIM tends to be good globally and bad for local measures, since it misestimates node-behavior based on the degree of nodes as seen for the co-occurrence in Section 8.2. In this chapter, I will present results on whether SIM can be applied in the context of local network motifs. This is done with three subgraphs from the former section, i.e., the Twopath, the Feed-Forward Loop, and the Bifan. Afterward, I will show how the idea performs in comparison to sophisticated and established methods in the field of data science.

The analysis whether it can be applied to count how often a node participates in a subgraph is structured as follows:

1. Generate several hundred random graphs according to the degree sequence of a graph from the FDSM.
2. Measure for the Twopath and the Feed-Forward Loop how often a node participates;
3. Calculate mean and standard deviation of the participation of a node in these motifs;
4. Develop equations based on the simple independence assumption;

The second point entails that not only the involvement of a node in a subgraph is measured, but also in which place the node is. This *position-restricted expected participation* can yield

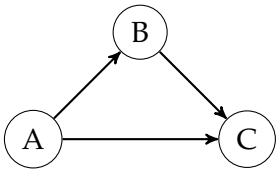


Figure 12.1: In a Feed-Forward Loop there are three different positions a node can take and each position may have a significance to it.

$\text{Twopath}_{\text{Node}}$	Equation	FFL_{Node}	Equation
	$k_u \frac{\langle kj \rangle}{\langle k \rangle}$		$\frac{k_u^2 - k_u}{m} \frac{\langle kj \rangle \left(\langle j^2 \rangle - \langle j \rangle \right)}{\langle k \rangle^2}$
	$k_u j_u$		$\frac{k_u j_u}{m} \frac{\left(\langle k^2 \rangle - \langle k \rangle \right) \left(\langle j^2 \rangle - \langle j \rangle \right)}{\langle k \rangle^2}$
	$j_u \frac{\langle kj \rangle}{\langle k \rangle}$		$\frac{j_u^2 - j_u}{m} \frac{\langle kj \rangle \left(\langle k^2 \rangle - \langle k \rangle \right)}{\langle k \rangle^2}$

Table 12.1: Equations for the expected number of subgraphs containing node u, depending on the possible position; basic approach without consideration of the participation of the node itself in the mean.

information on the function a node has in a real-world graph. As is shown in Fig. 12.1, in a Feed-Forward Loop there are three positions a node can maintain. Depending on the field of research, a node's position in a Feed-Forward Loop may entail different meanings. Note that this does not change the algorithm to count the participation of nodes in subgraphs.

12.1 CONSTRUCTING POSITION-BASED EQUATIONS

The equations need only small changes. Before, it was a sum over all possible tuples of nodes. When calculating the position-restricted expected participation of a node in a subgraph, there should be exactly as many equations needed as there are different positions in a subgraph. For two test-subgraphs, Twopath and Feed-Forward Loop, this implies that there are three different equations. For example, the expected number of Feed-Forward Loops a node might participate in as node A (see Fig. 12.1) can be given as

$$\sum_{B,C \in V} \frac{k_A j_B}{m} \frac{k_B j_C}{m} \frac{(k_A - 1)(j_C - 1)}{m} = \frac{1}{m^3} \left(k_A^2 - k_A \right) \sum_{B,C \in V} k_B j_B \left(j_C^2 - j_C \right) \tag{12.1}$$

$$= \frac{k_A^2 - k_A}{m} \frac{\langle kj \rangle \left(\langle j^2 \rangle - \langle j \rangle \right)}{\langle k \rangle^2}. \tag{12.2}$$

This equation can be calculated for every node in a graph. Equations for the other positions can be found following the same principle (results in Table 12.1).

The equations yield expected values for the participation of a node in a particular position in a subgraph. When the equations for a subgraph are summed over all possible positions, the result is the expected number of subgraphs of this type a node participates in. The sum over the equations over all nodes yields the equations given in Table 11.18, i.e., $n\langle kj \rangle$ for the Twopath and $\frac{\langle kj \rangle \left(\langle k^2 \rangle - \langle k \rangle \right) \left(\langle j^2 \rangle - \langle j \rangle \right)}{\langle k \rangle^3}$ for the Feed-Forward Loop.

$\text{Twopath}_{\text{Node}}$	Equation	FFL_{Node}	Equation
	$k_u \frac{\langle kj\rangle - \frac{k_u j_u}{n}}{\langle k\rangle}$		$\frac{(k_u^2 - k_u)}{m} \frac{\left(\langle j^2\rangle - \frac{j_u^2}{n} - \langle j\rangle + \frac{j_u}{n}\right)\left(\langle kj\rangle - \frac{k_u j_u}{n}\right)}{\langle k\rangle^2}$
	$k_u j_u$		$\frac{k_u j_u}{m} \frac{\left(\langle k^2\rangle - \frac{k_u^2}{n} - \langle k\rangle + \frac{k_u}{n}\right)\left(\langle j^2\rangle - \frac{j_u^2}{n} - \langle j\rangle + \frac{j_u}{n}\right)}{\langle k\rangle^2}$
	$j_u \frac{\langle kj\rangle - \frac{k_u j_u}{n}}{\langle k\rangle}$		$\frac{j_u^2 - j_u}{m} \frac{\left(\langle k^2\rangle - \frac{k_u^2}{n} - \langle k\rangle + \frac{k_u}{n}\right)\left(\langle kj\rangle - \frac{k_u j_u}{n}\right)}{\langle k\rangle^2}$

Table 12.2: Equations for the expected number of subgraphs containing node u, depending on the possible position; corrected for node u by subtraction from the remaining average in the nominator.

However, based on the knowledge from the analysis of the Bifan subgraph, it is known that the simple equations yield too high values. Since the sum of the positions yields the same as the equations in Table 11.18, it is evident that tuples such as (u, v, u) are considered. Therefore, a simple modification can be made. Averages which remain in the nominator are computed over all nodes, i.e., node u is considered as often as averages are taken, additionally to the explicit position. For example in the Twopath, node u contributes to the first equation k_u, but is also contained in each averaging of the degrees. Implictly, combinations as such stand for combinations $(u, u, u,)$. Thus, subtracting the nodes contribution from the averages alleviates the problem to a certain degree. Table 12.2 displays the modified equations. The problem with this is the following.

$$\sum_{v,w} \frac{k_u j_v}{m} \frac{k_v j_w}{m} = k_u \frac{\langle k\rangle\langle kj\rangle}{\langle k\rangle^2} = k_u \frac{\langle kj\rangle}{\langle k\rangle} \tag{12.3}$$

In equation 12.3, the equation for the first position in the Twopath without correction, one can see that there is another average, $\langle k\rangle$, in the nominator, which is not considered in Table 12.2. This is due to the reduction of the fraction. The correct equation for the first position in the Twopath under careful consideration of all nodes is therefore

$$\sum_{v,w\in V} \frac{k_u j_v}{m} \frac{k_v j_w}{m} = \frac{1}{m^2} \sum_{v\in V} k_u j_v k_v \left(n\langle j\rangle - \frac{j_u}{n} - \frac{j_v}{n}\right) \tag{12.4}$$

$$= \frac{1}{m^2}\left(k_u n\langle j\rangle\left(n\langle kj\rangle - \frac{k_u j_u}{n}\right) - \frac{k_u j_u}{n}\left(n\langle kj\rangle - \frac{k_u j_u}{n}\right) - \frac{k_u}{n}\left(n\langle kj^2\rangle - \frac{k_u j_u^2}{n}\right)\right)$$

$$= \frac{1}{m^2}\left(k_u n^2\langle j\rangle\langle kj\rangle - k_u^2 j_u\langle j\rangle - k_u j_u\langle kj\rangle + 2\frac{k_u^2 j_u^2}{n^2} - k_u\langle kj^2\rangle\right) \tag{12.5}$$

Usually, the computation of the equations is over **all** triples of nodes, i.e., triples such as (u, u, u) are included and vanish in the averages. By subtracting them explicitly, as done in equation 12.5, the equation becomes more complex, the subtractions create new terms. Instead of the simple equation displayed in Table 12.2, we are now faced with a longer

equation that supposedly yields the expected number of first positions a node u has in Twopaths. For the second and third position, the equations are equally complicated.

$$\mathtt{second} = \frac{1}{m^2}\left(n^2\langle k\rangle^2 k_u j_u - \langle j\rangle k_u^2 j_u - \langle kj\rangle k_u j_u + 2\frac{k_u^2 j_u^2}{n^2} + \langle k\rangle k_u j_u^2\right) \tag{12.6}$$

$$\mathtt{third} = \frac{1}{m^2}\left(n^2\langle k\rangle\langle kj\rangle j_u - \langle kj\rangle k_u j_u - \langle k^2 j\rangle j_u + 2\frac{k_u^2 j_u^2}{n^2} - \langle k\rangle k_u j_u^2\right) \tag{12.7}$$

The average over u is then

$$n\langle kj\rangle - \frac{\langle kj\rangle^2}{m\langle k\rangle} - \frac{\langle k^2 j\rangle}{m} - \frac{\langle kj^2\rangle}{m} + 2\frac{\langle k^2 j^2\rangle}{nm^2}, \tag{12.8}$$

which is far more complicated than the simple equation that is used to estimate the expected number of Twopaths with a high precision (see Table 11.19). Moreover, this equation follows the principle of inclusion and exclusion outlined previously (see Section 11.4; the initial equation $n\langle kj\rangle$ for the three nodes, equations for two nodes are subtracted, and the equation for a single node is added.

The more complicated a subgraph is, the more complex the equations will be; for a (relatively) simple subgraph as the Feed-Forward Loop, the equation to be in first position (position A in Fig. 12.1) would already be

$$\begin{aligned}\frac{1}{m^3}\Bigg(&n^2\langle kj\rangle\left(\langle j^2\rangle - \langle j\rangle\right)\left(k_u^2 - k_u\right) - \langle kj\rangle\left(k_u^2 j_u^2 - k_u j_u^2 - k_u^2 j_u + k_u j_u\right)\\ &- \left(\langle j^2\rangle - \langle j\rangle\right)\left(k_u^3 j_u - k_u^2 j_u\right) - \left(\langle kj^3\rangle - \langle kj^2\rangle\right)\left(k_u^2 - k_u\right)\\ &+ \frac{2}{n^2}\left(k_u^3 j_u^3 - k_u^2 j_u^3 - k_u^3 j_u^2 + k_u^2 j_u^2\right)\Bigg)\end{aligned}$$

Again, summing each part of the equation over all nodes u yields the same results as presented in Table 11.29, i.e, the number of subgraphs which are implicitly assumed by the equation to exist but are not allowed in a simple graph. In Table 11.29, it was shown that it is not necessarily useful to take care of these "forbidden" sub-subgraphs. On the other hand, for the Bifan calculating the number of these sub-subgraphs and subtracting them from the expected number of Bifans in a graph yielded much better results. Whether or not this detailed research is necessary cannot be decided in a thesis. A good recommendation is using the simple modification first (see Table 12.2 and whenever a node shows a different behavior than expected, using more complicated equations such as equation 12.8 to verify their odd behavior. Only when the estimated value of participations of a node differs from the number participations in the real-world graph, then a more involved research which specializes on this node should be performed.

12.2 ON THE SUM OF EQUATIONS

Despite the efforts to show that it is possible to estimate the number of subgraphs more precisely than with the set of equations in Table 11.18, the following investigation tests

whether the simple modifications are already enough (see Table 12.2)[1]. The following discussion is based on only two of the twelve graphs from above, the E. coli graph and the Ythan Estuary graph, i.e., one graph for which the equations fit sampling results very well and one for which the equations do not fit at all.

Equation	≤ 0.1	≤ 0.5	≤ 1	$= 0$
Sum_{unmod}	0.59	0.94	0.98	0.00
Sum_{mod}	0.59	0.94	0.98	0.00
A_{unmod}	0.83	0.89	1.00	0.75
A_{mod}	0.83	0.89	1.00	0.75
B_{unmod}	0.98	0.98	1.00	0.94
B_{mod}	0.98	0.98	1.00	0.94
C_{unmod}	0.62	0.89	0.99	0.18
C_{mod}	0.63	0.89	0.99	0.18

Table 12.3: Relative error $\frac{observed-expected}{observed}$ is measured for the total number of participations and each possible position in the Twopath subgraph. Sum refers to the expected participation in any of the three places. Columns 2-4 show the percentage of nodes with a smaller relative error than shown in the head row. The last column keeps track of the percentage of how many nodes did never show up in any Twopath in graphs generated with the FDSM.

Equation	≤ 0.1	≤ 0.5	≤ 1	$= 0$
Sum_{unmod}	0.67	1.00	1.00	0.00
Sum_{mod}	0.68	1.00	1.00	0.00
$Start_{unmod}$	0.56	0.94	0.98	0.39
$Start_{mod}$	0.56	0.95	0.98	0.39
$Middle_{unmod}$	0.93	0.96	0.99	0.39
$Middle_{mod}$	0.93	0.96	0.99	0.39
$Sink_{unmod}$	0.74	0.93	0.97	0.01
$Sink_{mod}$	0.78	0.93	0.97	0.01

Table 12.4: Relative error $\frac{observed-expected}{observed}$ is measured for the total number of participations and each possible position in the Twopath subgraph. Sum refers to the expected participation in any of the three places. Columns 2-4 show the percentage of nodes with a smaller relative error than shown in the head row. The last column keeps track of how many nodes did never show up in any Twopath in graphs generated with the FDSM.

For this, the relative error, $\frac{\mu_{FDSM}-\mu_{SIM}}{\mu_{FDSM}}$, for each node of a graph is calculated. In Table 12.3, resp. 12.4, the percentages lower than the fixed thresholds in the top row are shown.

1 The more accurate and more complex set of equations would also have been possible to test, but when the equation from Table 12.2 suffice, this test is unnecessary.

Considering the Twopath subgraph approximations of the E. coli graph, the results are apparently not very good. With a relative error larger than 0.1 for 41% of the nodes, the equations cannot be reasonable.

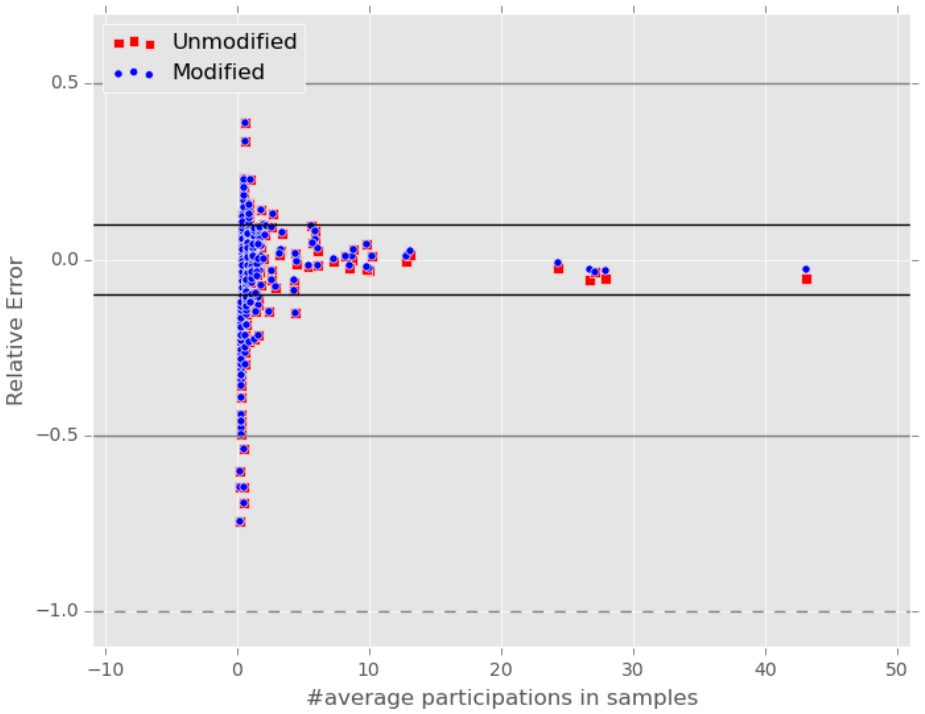
As can be observed in Fig. 12.2 the estimates of the lower degree nodes are all very low. For a randomly chosen node of this degree range, the observed average number of Twopath subgraphs it participates in is 0.11; the estimated participations by the equation is 0.076. The relative error for such a low number of found and estimated participations would be 0.31. Thus, yes, the estimates are not very good, objectively. In Fig. 12.2, nodes are plotted according to their expected number of participations in a Twopath in samples from the FDSM versus the relative error, using once the sum of the unmodified equations (Table 12.1, red squares) and once the sum of the modified equations (see Table 12.2, blue dots).

It is obvious, that the E. coli graph is estimated well for nodes which participate more often than once or twice in a Twopath subgraph (relative error lower than 0.1, Fig 12.2a). For the Ythan the result is similar; only very few nodes of interest have a relative error large enough to be of interest, as long as they participate often enough in Twopath subgraphs. For the unmodified equation, the results are worse for all nodes of interest.

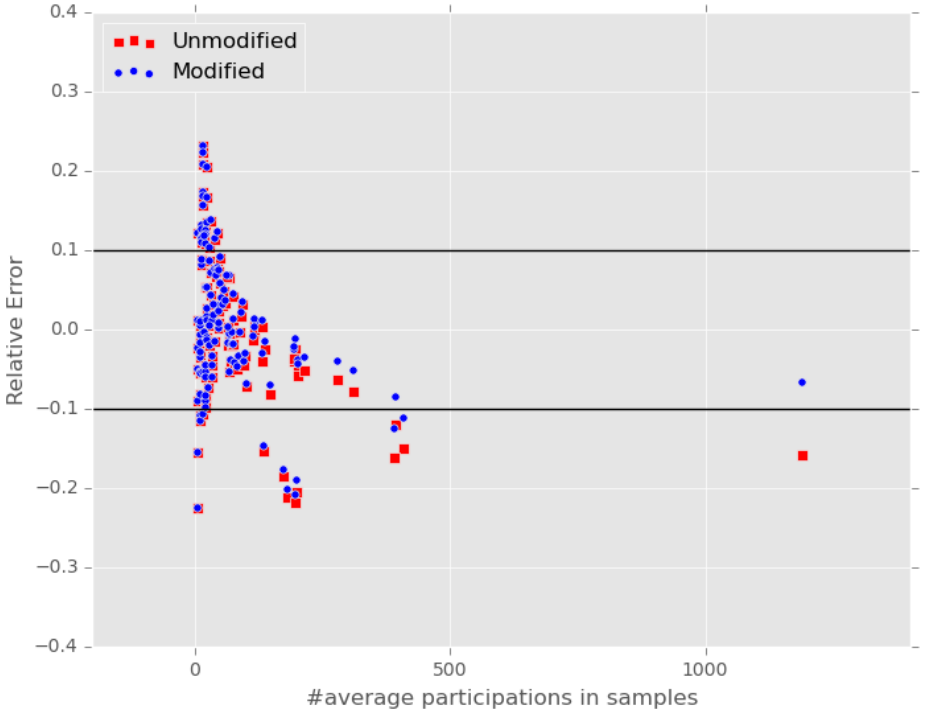
Equation	≤ 0.1	≤ 0.5	≤ 1	$= 0$
$\mathrm{Sum}_{\mathrm{unmod}}$	0.73	0.93	0.98	0.60
$\mathrm{Sum}_{\mathrm{mod}}$	0.73	0.93	0.98	0.60
$\mathrm{A}_{\mathrm{unmod}}$	0.92	0.97	1.00	0.88
$\mathrm{A}_{\mathrm{mod}}$	0.92	0.97	1.00	0.88
$\mathrm{B}_{\mathrm{unmod}}$	0.95	0.98	1.00	0.93
$\mathrm{B}_{\mathrm{mod}}$	0.95	0.98	1.00	0.93
$\mathrm{C}_{\mathrm{unmod}}$	0.80	0.93	0.99	0.73
$\mathrm{C}_{\mathrm{mod}}$	0.80	0.93	0.99	0.73

Table 12.5: Relative error $\frac{\mathrm{observed}-\mathrm{expected}}{\mathrm{observed}}$ is measured for the total number of participations and each possible position in the Feed-Forward Loop subgraph. Sum refers to the expected participation in any of the three places. Columns 2-4 show the percentage of nodes with a smaller relative error than shown in the head row. The last column keeps track of how many nodes did never show up in any Twopath in graphs generated with the FDSM.

For the Feed-Forward Loop, the results are similar. The relative error seems to be high for both graphs; for the Ythan Estuary graph, half of the results of the equations are off by more than a relative error of 0.1 while most are at least below 0.5 (see Table 12.6). However, a comparison with Fig. 12.3b shows, that the interesting nodes which often participate in Feed-Forward Loops and are misestimated are not as many as the "uninteresting" nodes, i.e., nodes which participate as good as never. It is important to note that the modified equation estimates the participation in Feed-Forward Loops much better than the unmodified one. A prominent example is the highest degree node which has a relative error of over $|1|$ with the unmodified equation while the modified equation yields a score of about $|0.4|$. The E. coli graph, on the other hand, shows even more clearly that the misestimated nodes are the unimportant ones.

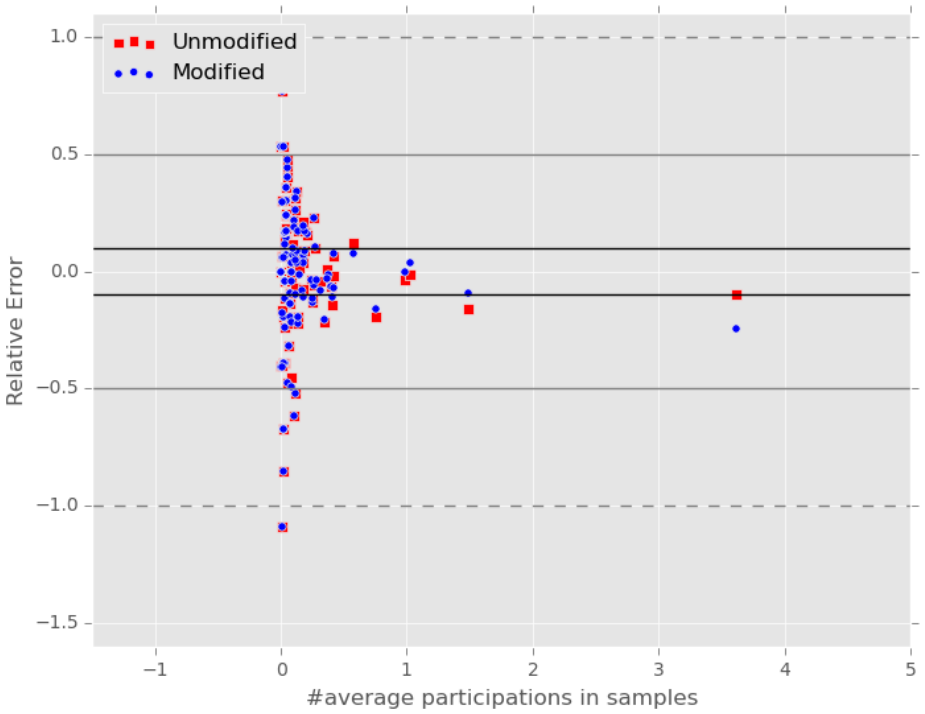


(a) E. coli with 1000 samples

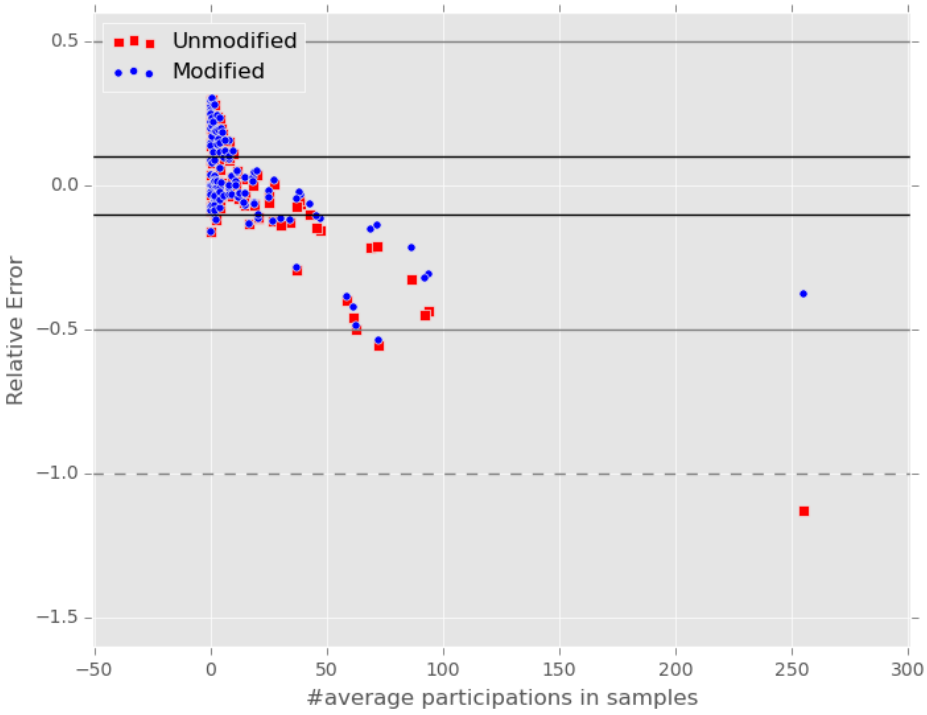


(b) Ythan Estuary with 350 samples

Figure 12.2: Shown are relative errors considering the Twopath subgraph in the E. coli graph and the Ythan Estuary graph.



(a) E. coli with 1000 samples



(b) Ythan Estuary with 350 samples

Figure 12.3: Shown are relative errors considering the Feed-Forward Loop subgraph in the E. coli graph and the Ythan Estuary graph.

Equation	≤ 0.1	≤ 0.5	≤ 1	$= 0$
Sum_{unmod}	0.50	0.98	0.99	0.05
Sum_{mod}	0.50	0.99	1.00	0.05
A_{unmod}	0.64	0.89	0.97	0.53
A_{mod}	0.64	0.91	0.97	0.53
B_{unmod}	0.55	0.96	0.99	0.39
B_{mod}	0.58	0.96	0.99	0.39
C_{unmod}	0.79	0.96	0.99	0.38
C_{mod}	0.81	0.96	1.00	0.38

Table 12.6: Relative error $\frac{observed-expected}{observed}$ is measured for the total number of participations and each possible position in the Feed-Forward Loop subgraph. Sum refers to the expected participation in any of the three places. Columns 2-4 show the percentage of nodes with a smaller relative error than shown in the head row. The last column keeps track of how many nodes did never show up in any Twopath in graphs generated with the FDSM.

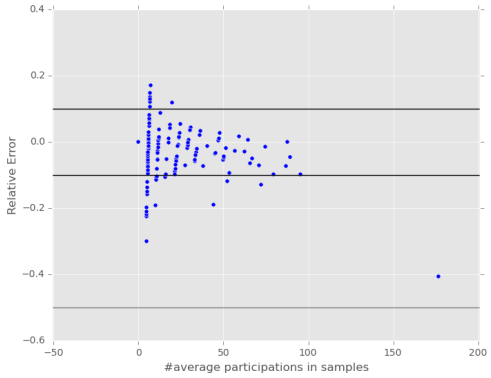
12.3 HOW DO THE EQUATIONS FOR POSITIONS FARE?

Up to this point, the approach to use equations to estimate the number of subgraphs a node participates in has had some serious drawbacks. Even though it is much faster than the standard approach via simulation, it is off with a relative error larger than 0.1 too often. Not mentioned was that at least half of the nodes was estimated very well, and almost all nodes were estimated better than half-off. Most of the errors were in the lower degree region, and the high-degree nodes cause most of the global misestimates. However, how do the equations perform concerning the node-specific position-specific participation in a subgraph? This is evaluated on the Ythan Estuary graph.

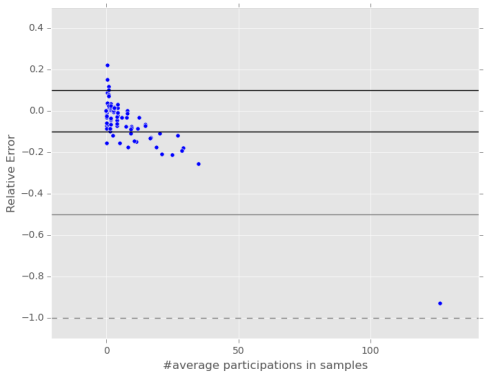
The participation of a node in the middle position is estimated for the Twopath subgraph well, for the Feed-Forward Loop many nodes are estimated well (see Figs. 12.4c, 12.4d). Those that are not estimated well are most of the time low to average degree nodes. Calculating the average error, $\frac{\mu_{FDSM}-\mu_{SIM}}{\mu_{FDSM}}$, takes into account how often a node participated in a subgraph; it may have happened that some of them never participated in any subgraph of interest, even though they could. Thus, the error may be unreasonably large for some of the nodes. Especially the second position is interesting. Here, the error is mostly based on nodes that could be in a Twopath but are instead in a reciprocal subgraph.

The Feed-Forward Loop (Fig. 12.4b) is estimated much better by the equations than for the Twopath subgraph (Fig. 12.4a) for all nodes but one. This is due to the fact that many nodes of these graph have a low out-degree (E. coli ~ 85%, Ythan Estuary ~ 53%), such that $(k_u^2 - k_u) = 0$, while $k_u \geq 0$. Only the highest degree node is misestimated for both (in fact all) positions. The end position is similar in all accounts.

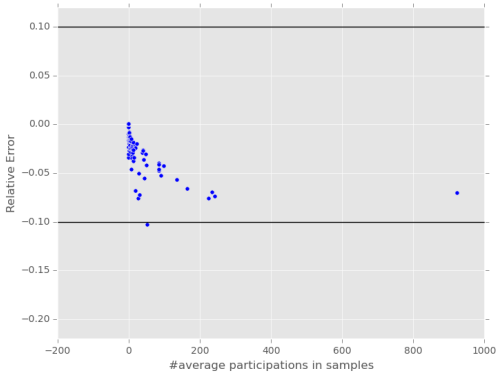
For the E. coli graph, figures are omitted, but the results are quite similar. In the low degree region, there is not much difference independent of the position. Towards the higher degrees, the difference between the average of the results of the FDSM and the equation become more pronounced; still, many of the nodes which have a relative error larger than 0.1 are of a lower degree.



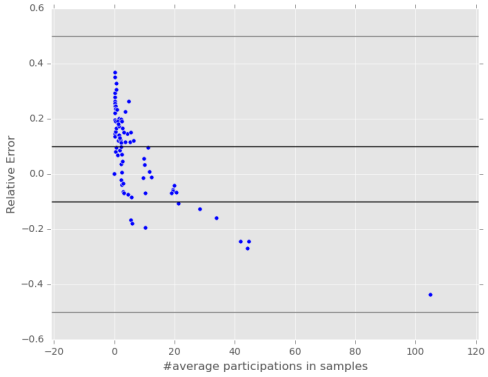
(a) Twopath first position



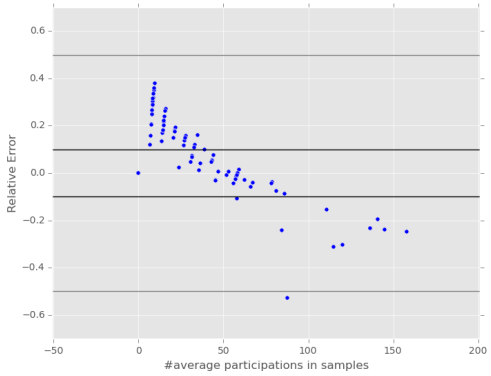
(b) Feed-Forward Loop first position



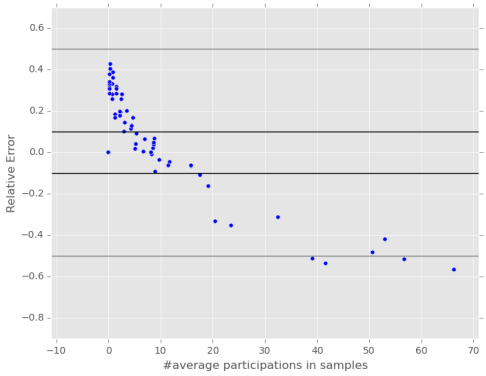
(c) Twopath middle position



(d) Feed-Forward Loop middle position



(e) Twopath end position



(f) Feed-Forward Loop end position

Figure 12.4: Shown are relative errors considering the possible positions in 12.4a, 12.4c, 12.4e the Twopath subgraph and 12.4b, 12.4d, 12.4f the Feed-Forward Loop subgraph, based on the Ythan Estuary graph (350 samples).

The equations perform on average good enough to be reasonable for a first estimate. If there is a significant discrepancy to be observed between the results of the equations and the real-world graph, a second, more in-depth analysis with the corrected equations would be in order (cf. equation 12.5). If there is still a significant discrepancy, one can either conclude that the graph and particular nodes are special; or, which is more reasonable and more thorough, validate the claim with the FDSM. Seeing that the equations misestimate the higher-degree nodes, an interesting question which came up beforehand may be answered now. For the Bifan, the global estimate by the simple equation was off most of the times, sometimes even very far off (see Table 37, worst example Ythan Estuary with a discrepancy of ~ 35000). How do corrected node-based equations fare for the Bifan and were the errors caused by only a few nodes?

12.4 REVISITING THE BIFAN

Estimating how often a node is in a certain position in the Bifan subgraph is about as complicated as for the other subgraphs. One node is being kept fixed while the others are averaged. To calculate for this node corrected equations like equation 12.5 requires more effort. The resulting equations contain 24 summands; some of these summands occur more than once. This approach implies that remaining nodes are excluded from the moments, thus

$$\begin{aligned}
&\frac{1}{4m^4} \sum_{u,v,w,x \in V} \left(k_u^2 - k_u\right)\left(k_v^2 - k_v\right)\left(j_w^2 - j_w\right)\left(j_x^2 - j_x\right) \\
&\quad = \frac{1}{4m^4} \sum_{u,v,w \in V} \left(k_u^2 - k_u\right)\left(k_v^2 - k_v\right)\left(j_w^2 - j_w\right) \\
&\left(n\left(\langle j^2\rangle - \langle j\rangle\right) - \left(\frac{j_u^2 - j_u}{n}\right) - \left(\frac{j_v^2 - j_v}{n}\right) - \left(\frac{j_w^2 - j_w}{n}\right)\right)
\end{aligned} \tag{12.9}$$

This is only the start of the calculation, the rest is in Appendix 3.

Position	Equation
	$\left(k_u^2 - k_u\right) \frac{\left(\langle k^2\rangle - \frac{k_u^2}{n} - \langle k\rangle + \frac{k_u}{n}\right)\left(\langle j^2\rangle - \frac{j_u^2}{n} - \langle j\rangle + \frac{j_u}{n}\right)^2}{4m\langle k\rangle^3}$
	$\left(j_u^2 - j_u\right) \frac{\left(\langle k^2\rangle - \frac{k_u^2}{n} - \langle k\rangle + \frac{k_u}{n}\right)^2\left(\langle j^2\rangle - \frac{j_u^2}{n} - \langle j\rangle + \frac{j_u}{n}\right)}{4m\langle k\rangle^3}$

Table 12.7: Equations regarding the Bifan subgraph where corrections for node u are made by subtracting its contribution from the averages.

We continue with the simple correction of excluding the node under consideration from each moment (cf. Table 12.7). The factor in front of equation 12.9 is $\frac{1}{4}$, but in Table 12.7 it is only $\frac{1}{2}$; this is due to the place the node participates in, since a node u can be in the upper and the lower position. The two equations for this are the same and thus the factor is $2\frac{1}{4} = \frac{1}{2}$. The results of the equations in Table 12.7 are displayed in Table 12.8, respectively Table 12.9. As before, only the fractions of nodes with an error below 0.1, 0.5, respectively

Equation	≤ 0.1	≤ 0.5	≤ 1	$= 0$
Sum_{unmod}	0.66	0.88	0.96	0.61
Sum_{mod}	0.67	0.89	0.96	0.61
A_{unmod}	0.92	0.97	1.00	0.86
A_{mod}	0.92	0.97	1.00	0.86
B_{unmod}	0.74	0.91	1.00	0.73
B_{mod}	0.74	0.92	1.00	0.73

Table 12.8: Relative error $\frac{observed-expected}{observed}$ is measured for the total number of participations and each possible position in the Bifan subgraph. Sum refers to the expected participation in any of the four places. Columns 2-4 show the percentage of nodes with a smaller relative error than shown in the head row. The last column keeps track of how many nodes did never show up in any Twopath in graphs generated with the FDSM.

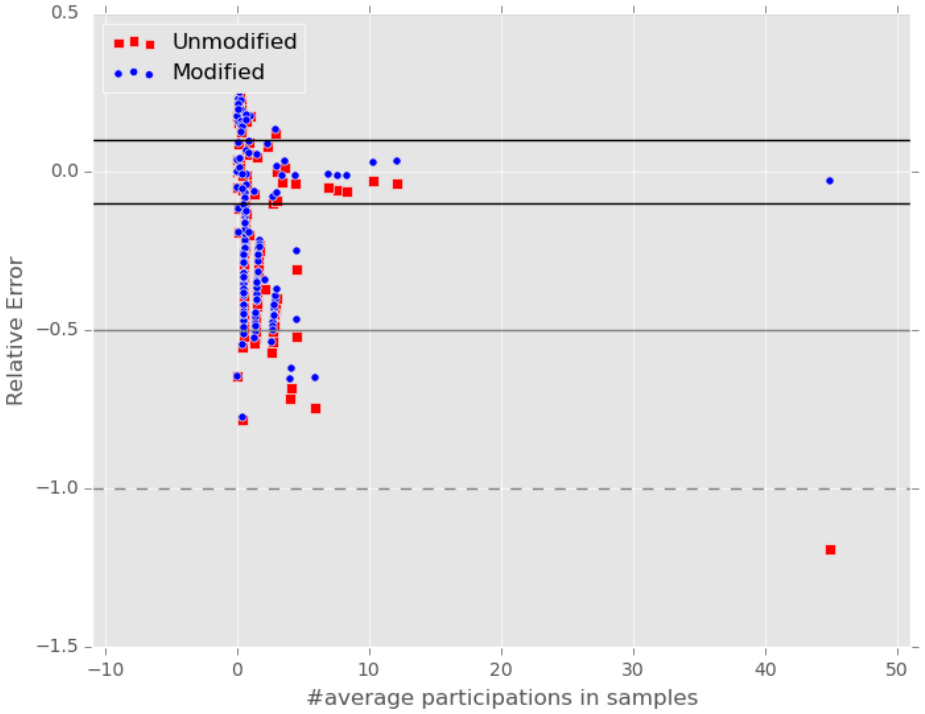
Equation	≤ 0.1	≤ 0.5	≤ 1	$= 0$
Sum_{unmod}	0.56	0.92	0.99	0.16
Sum_{mod}	0.59	0.94	1.00	0.16
$Start_{unmod}$	0.63	0.87	0.96	0.53
$Start_{mod}$	0.67	0.88	0.96	0.53
$Sink_{unmod}$	0.74	0.95	0.99	0.38
$Sink_{mod}$	0.76	0.96	1.00	0.38

Table 12.9: Relative error $\frac{observed-expected}{observed}$ is measured for the total number of participations and each possible position in the Bifan subgraph. Sum refers to the expected participation in any of the four places. Columns 2-4 show the percentage of nodes with a smaller relative error than shown in the head row. The last column keeps track of how many nodes did never show up in any Twopath in graphs generated with the FDSM.

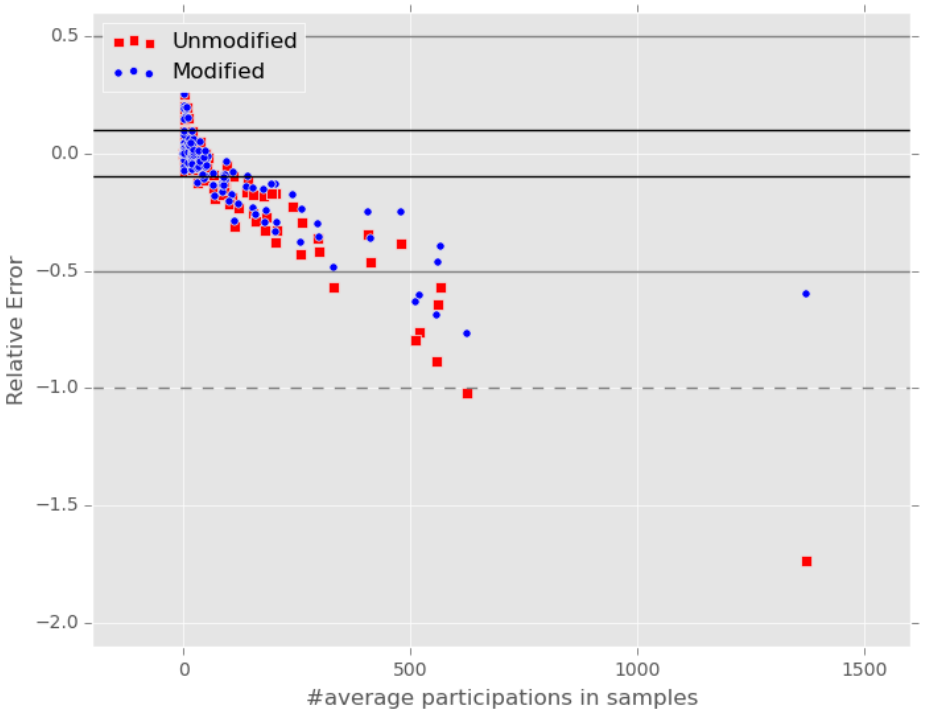
1, are noted down as well as the fraction of nodes which have no Bifan subgraph attached in the FDSM. The results are about as "good" as for the Feed-Forward Loop.

In Fig. 12.5, the relative errors concerning the sum of the equations against the expectations from the samples from the FDSM are shown. As before, low degree nodes with degree $k_u < 2 \vee j_u < 2$ all have a small value in the analysis with the FDSM and the equation results are equally small. The relative errors are small, seldom larger than 0.1. The higher the degree, the more severe the errors become. This may be caused by the quadratic influence of the node itself, as well as the other squared factors. For the E. coli graph (cf. Fig 12.5a), the modified equations have in general a lower relative error; still, there are more nodes with a higher relative error than 0.5 than for any other subgraph and much more with a higher relative error than 0.1. Most curious is the highest degree node that is completely misestimated by the simple equation but estimated very well by the modified equations. The results of the Ythan Estuary graph are even worse; almost all nodes of higher degree are misestimated by the equations. The relative error of these nodes is almost always larger than $|0.1|$, not a few of them even worse than a relative error larger than $|0.5|$. The modified equations fare better than the simple equations, again, but still, they are not good. If one summed the results of both equations in Table 12.7 for all nodes, the result would be higher than the averages in the FDSM, even though they would be lower than the initial equation of $\frac{\left(\langle k^2\rangle - \langle k\rangle\right)^2 \left(\langle j^2\rangle - \langle j\rangle\right)^2}{4\langle k\rangle^4}$.

This result confirms the approach suggested at the end of the former section. The equations can be used to approach the expected number of participations of a node in a subgraph. When the results of the equations differ much from the results of the analysis of the real-world graph, the second test with a more sophisticated set of equations can be useful. The last step that should not be omitted is analyzing the nodes that show considerable differences in a more thorough analysis with the FDSM.



(a) E. coli with 1000 samples



(b) Ythan Estuary with 350 samples

Figure 12.5: Shown are relative errors considering the Bifan subgraph in the E. coli graph and the Ythan Estuary graph.

13

ON PREDICTING CO-PURCHASED ITEMS

The SIM has been weighed, measured, and found wanting considering the prediction of co-purchased items. This investigation was done in the context of the Netflix Dataset [107], [108] as well as a sample of MovieLens data [87]. These datasets comprise the ratings of several thousand users to a large number of movies and television series. Zweig [107], Zweig and Kaufmann [108], and Spitz et al. [87] tested several different approaches to assess the similarity between nodes, including the FDSM, a model based on the idea of the SIM, and a set of equations to estimate the similarity of nodes. Most importantly for this thesis is the so-called *leverage*, i.e.,

$$\mathit{leverage}(u, v) = \frac{1}{|V|}\left(\mathrm{cooc}(u, v) - \frac{k_u k_v}{|V|}\right). \tag{13.1}$$

This equation can also be found in [107, 108] and stems from Piatetsky-Shapiro [75]. It is used to measure the difference between the expected and observed support of u and v. In this case, the observed support is the co-occurrence in a real world graph; the expected support is the expected co-occurrence. Again, the co-occurrence of two nodes is the number of nodes they share as neighbors, i.e.,

$$\mathrm{cooc}(u, v) = |N(u) \cap N(v)|, \tag{13.2}$$

that uses $N(u)$, the set of neighbors of a node. Equation 13.1 uses as expected number of co-occurrences in the simple independence model the equation

$$\mathbb{E}[\mathrm{cooc}_{\mathrm{SIM}}(u, v)] = \frac{j_u j_v}{n}.$$

This equation is considered as the expected co-occurrence of two nodes in a bipartite graph [108], such as the data sets in this chapter comprise. A graph can be called bipartite when the set of nodes can be divided into two disjoint sets such that each edge in the graph is between these sets. In movie-user data sets such as these, this is easily possible. Based on similarity measures, recommendations on "what to watch next" can be made. Thus, in the following we will often refer to the result of an algorithm as a suggestion, a recommendation, or a prediction. Prediction is a somewhat misleading term, but still applicable. Based on what users watched and rated positive, developing a system which predicts what a users is likely to watch next, is in a sense very similar to making recommendations to a user.

Most interestingly, the SIM and equation 13.1 performed in all articles remarkably bad. For example, when looking for recommendations based on "The X-Files: Season 1", sensible top ten recommendations what to watch next should include all other seasons of "The

X-Files". Instead, the SIM produced recommendations such as "Pirates of the Caribbean I", "The Matrix", "Lord of the Rings: The Fellowship of the Ring". These results are based on the equation that takes only into account the degree of the movie, and these movies are easily some of the most famous that the databases had at that time.

Now, considering the definition of the Fork, i.e., a node and one of its pairs of successors, it is possible to define the subgraph which describes a single common neighbor. It consists out of three nodes $\{u, v, w\}$ and the edges $\{(u, w), (v, w)\}$. Considering the bipartite graph as a directed graph allows constructing an equation that yields the expected number of these subgraphs for all pairs of nodes.

$$\mathbb{E}\left(\mathrm{cooc}_{\mathrm{SIM}}(u, v)\right) = \sum_{w \in V} \frac{k_u j_w}{m} \frac{k_v (j_w - 1)}{m} \tag{13.3}$$

$$= \frac{1}{m^2} \sum_{w \in V} k_u k_v \left(j_w^2 - j_w\right) \tag{13.4}$$

$$= \frac{k_u k_v}{m^2} \left(n \left(\langle j^2 \rangle - \langle j \rangle\right) - \left(\frac{j_u^2 - j_u}{n}\right) - \left(\frac{j_v^2 - j_v}{n}\right)\right) \tag{13.5}$$

$$= \frac{k_u k_v}{m} \frac{\langle j^2 \rangle - \langle j \rangle}{\langle j \rangle}. \tag{13.6}$$

In equation 13.4, one might consider excluding k_u, k_v from the sum directly, since the sum is only over the node w and not all triples of nodes. Still, as mentioned when reconsidering the Bifan, the averages $\langle j \rangle$ over all nodes $w \in V$ do include nodes u, v as well. Thus, in the next step in equation 13.5, these nodes are still excluded from the averages. Since we are considering a bipartite graph as a directed graph, a node has either an out-degree of zero and an in-degree larger than zero or vice versa. Thus, in the last step we discard the subtracted elements (they are zero) and end up with equation 13.6. Note that in a standard directed graph, i.e., one that is not based on disjoint sets of nodes, discarding the subtraction may yield results as shown for the Bifan, i.e., results which are far off from the expected values which the FDSM yields.

We compare the recommendations the equation yields to the results in the papers by Zweig, Zweig and Kaufmann, and Spitz et al. Their ground truth uses a simple and intuitive idea. Television series are usually separated in different seasons and any reasonable prediction should point to other seasons first, to similar shows next, and after this to other series. For movies, the same theory applies, only that seasons are called prequel or sequel. Since computing the expected co-occurrence using the FDSM is expensive in resources as well as in time, using equations that are much faster to calculate and much less expensive has advantages, as long as the results of the equations are either close enough to the results of the FDSM or good enough for an expert in the field. Since we assume the sampling approach with the FDSM as a baseline model, we assume that having results as the FDSM should be enough.

We start out with a comparison of the results of Zweig [107] and the results which equation 13.6 yields on the given data sets.

13.0.1 *Television Series Prediction*

In the Netflix dataset are many television series included. From these, a subset has been selected by Zweig and Kaufmann [108] to be displayed in more detail; these include the set of "Star Trek: The Next Generation", "Star Trek: Voyager", "Star Trek: Deep Space Nine", "Buffy the Vampire Slayer", "Sex and the City", "Stargate SG-1", and "Friends". For these, it is important that the sequels and prequels should be the recommended shows to watch. When talking about recommending the next shows to watch after considering a single show, it is a *local prediction*, since predictions are based on the observed and expected co-occurrences of this series alone. A short recapitulation of the results of Zweig and Kaufmann shows that the FDSM is highly superior to the SIMS equation used in their work. For example, for almost each season of "Star Trek: The Next Generation" as input, the algorithm using the SIM had a hit rate of 0, i.e., it suggested none of the other seasons but movies like "Independence Day" and "The Matrix". For seven seasons, Zweig and Kaufmann had only the top five suggestions included—the algorithm using the SIM gave only 20% reasonable results, while the sampling approach using the FDSM yielded only seasons of "Star Trek: The Next Generation", i.e., a perfect result.

While it would be simple to present the predictions of the new equation 13.6 with the changed equation in tables, showing what had been presented, it seems more enlightening to present it in a different way. Each series spans a row; each column presents a different season. In the respective field is the hit-rate, i.e., what percentage of the suggested k were part of the series, where k is the number of seasons. Shown in Table 13.1 are the results which equation 13.6 yields, i.e., the new approach to calculate the co-occurrence using the bipartite graph as a directed graph.

Serie	Season								
	1	2	3	4	5	6	7	8	9
Star Trek: The Next Generation	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
Star Trek: Deep Space Nine	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
Star Trek: Voyager	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
Buffy the Vampire Slayer	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
Friends	0.5	0.5	0.5	0.5	0.63	0.63	0.63	0.63	0.5
Sex and the City	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
Stargate SG-1	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	
The X-Files	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 13.1: Hit-rate based on the new prediction method for the same television series as in Zweig and Kaufmann [108].

The series which did not achieve the 100% hit-rate[1] are "Friends" and "Stargate SG-1"; for "Friends", the suggested other series were "Best of Friends", which is reasonable but not a direct pre- or sequel of "Friends". For "Stargate SG-1", the erroneous prediction was "Stargate: Atlantis Season 1", which is closely related to the original series "Stargate SG-1", but not a direct sequel as well. For "Stargate SG-1", the next recommendation would have been the missing season. For "Friends", the first recommendation that is outside the

1 Marketing term to express the relationship to selling sequels/prequels; in mathematical terms it is the percentage of correctly predicted (pre-/)sequels.

"Friends"-universe, i.e., neither "Friends" nor "Best of Friends", comes in more than 50% of the seasons after the "Friends"-universe. In other words, "Friends" and "Best of Friends" are closely related, such that they are all recommended before anything else in more than 50% of all seasons of "Friends". The other seasons recommend all the same series. Since the data sets do not include the names of all series, shows, and movies but only the ones for which the ground truth contains entries, it is unknown how closely related these series are.

To compare the results of the new equation 13.6 to the old equation 13.1, consider "Star Trek: The Next Generation". The results of the new equation show a tremendous improvement to before. While the old approach, using $\frac{k_u k_v}{n}$ as the expected co-occurrence of the SIM, yielded only a meager 20% hit-rate, the new approach delivers a 100% hit-rate with a considerable speed-up towards the sampling approach.

Results are shown in Table 13.2; this table shows for each television series in the Netflix-Series data set the prediction quality of the new SIM equation 13.6 with simple to measure scores:

n The number of seasons a series contains

pbr The average percentage of series for which the highest-ranked recommendation is another part of the same series; 100% is optimal

pra the average percentage of series for which all seasons were listed in the top 100 recommendations; 100% is optimal

$\overline{\texttt{first}}$ the average rank of the first listed season from the same series, calculated only with series that had 100% in pra; 1.0 is optimal

$\overline{\texttt{last}}$ the average rank of the last listed season from the same series, calculated only with series that had 100% in pra; $n - 1$ is optimal

This idea is taken from Zweig [107]; no direct comparison between all series is made since not all series we have are in the listing of Zweig (especially the two season series are omitted there).

Title of series	n	pbr	pra	$\overline{\texttt{first}}$	$\overline{\texttt{last}}$
3rd Rock from the Sun	2	0.0	50.0	13.0	13.0
7th Heaven	2	100.0	100.0	1.0	1.0
Alf	2	0.0	50.0	2.0	2.0
American Chopper	2	100.0	100.0	1.0	1.0
Arrested Development	2	100.0	100.0	1.0	1.0
Bewitched	2	0.0	0.0	-	-
CSI: Miami	2	0.0	100.0	5.0	5.0
Chappelle's Show	2	100.0	100.0	1.0	1.0
Charlie's Angels	2	0.0	100.0	67.0	67.0
Charmed	2	100.0	100.0	1.0	1.0
Crank Yankers 1: Uncensore	2	0.0	100.0	2.5	2.5
Da Ali G Show	2	100.0	100.0	1.0	1.0

Dallas	2	50.0	100.0	18.5	18.5
Danger Mouses 1 and	2	0.0	50.0	87.0	87.0
Dark Angel	2	100.0	100.0	1.0	1.0
Dead Like Me	2	100.0	100.0	1.0	1.0
Doogie Howser, M.D.	2	50.0	100.0	2.5	2.5
Ellen	2	50.0	100.0	2.5	2.5
Family Business	2	100.0	100.0	1.0	1.0
Family Guy: Vol. 1s 1-	2	100.0	100.0	1.0	1.0
Green Acres	2	100.0	100.0	1.0	1.0
Have Gun Will Travel	2	50.0	100.0	2.5	2.5
Highway to Heaven	2	0.0	50.0	44.0	44.0
Hogan's Heroes	2	50.0	100.0	30.0	30.0
Home Improvement	2	100.0	100.0	1.0	1.0
Home Movies	2	100.0	100.0	1.0	1.0
Jem and The Holograms 3: Part	2	0.0	0.0	-	-
Knight Rider	2	50.0	100.0	3.0	3.0
Las Vegas	2	100.0	100.0	1.0	1.0
MTV: Punk'd	2	100.0	100.0	1.0	1.0
Mad About You	2	100.0	100.0	1.0	1.0
Magnum P.I.	2	100.0	100.0	1.0	1.0
Mail Call: The Best of Season	2	50.0	100.0	1.5	1.5
Mary Tyler Moore	2	50.0	100.0	6.0	6.0
Michael Moore's The Awful Truth	2	100.0	100.0	1.0	1.0
Mutant X	2	0.0	50.0	42.0	42.0
NYPD Blue	2	100.0	100.0	1.0	1.0
Nero Wolfe	2	50.0	100.0	8.0	8.0
Newlyweds: Nick and Jessica	2	100.0	100.0	1.0	1.0
NipandTuck	2	100.0	100.0	1.0	1.0
Once and Again	2	50.0	100.0	3.5	3.5
One Tree Hill	2	50.0	100.0	9.5	9.5
Penn and Teller: Bullsh*t!	2	100.0	100.0	1.0	1.0
Popular	2	100.0	100.0	1.0	1.0
Project Greenlight	2	100.0	100.0	1.0	1.0
Punky Brewster	2	0.0	100.0	3.0	3.0
Reno 911	2	100.0	100.0	1.0	1.0
Samurai Jack	2	50.0	100.0	8.5	8.5
Sledge Hammer!	2	50.0	100.0	1.5	1.5
Sliders 1 and	2	50.0	100.0	7.0	7.0
Tales from the Crypt	2	50.0	100.0	27.0	27.0
Teen Titans	2	50.0	50.0	1.0	1.0
That '70s Show	2	100.0	100.0	1.0	1.0
The Bob Newhart Show	2	0.0	50.0	5.0	5.0
The Challenge of the Superfriends	2	0.0	50.0	15.0	15.0
The Fresh Prince of Bel Air	2	0.0	50.0	2.0	2.0
The Golden Girls	2	100.0	100.0	1.0	1.0
The L Word	2	100.0	100.0	1.0	1.0

The Lost World	2	0.0	0.0	-	-
The Monkees	2	0.0	100.0	5.0	5.0
The New Avengers	2	50.0	100.0	2.5	2.5
The O.C.	2	100.0	100.0	1.0	1.0
The Outer Limits: The Original Series	2	100.0	100.0	1.0	1.0
The Pretender	2	50.0	100.0	7.5	7.5
The Simple Life	2	100.0	100.0	1.0	1.0
The Twilight Zone	2	0.0	50.0	14.0	14.0
The Waltons	2	0.0	50.0	33.0	33.0
The Wire	2	100.0	100.0	1.0	1.0
Touched by an Angel	2	0.0	100.0	18.0	18.0
Trailer Park Boys	2	0.0	0.0	-	-
Viva La Bam	2	100.0	100.0	1.0	1.0
Wildboyz	2	50.0	100.0	1.5	1.5
Wiseguy 1: Part	2	100.0	100.0	1.0	1.0
World Poker Tour	2	50.0	100.0	5.5	5.5
X-Men: Evolution	2	100.0	100.0	1.0	1.0
21 Jump Street	3	100.0	100.0	1.0	4.67
2	3	100.0	100.0	1.0	2.0
Beast Wars Transformers	3	100.0	100.0	1.0	2.0
Boy Meets World	3	100.0	100.0	1.0	26.33
Cold Feet	3	66.67	100.0	10.33	38.0
Columbo	3	100.0	100.0	1.0	3.33
ER	3	100.0	100.0	1.0	11.33
G.I. Joe 1: Part	3	100.0	100.0	1.0	3.33
Gilligan's Island	3	33.33	66.67	2.0	5.0
Kung Fu	3	66.67	50.0	-	-
La Femme Nikita	3	100.0	100.0	1.0	2.0
Land of the Lost	3	100.0	83.33	1.0	42.5
MacGyver	3	100.0	100.0	1.0	6.67
Millennium	3	100.0	100.0	1.0	15.67
Monk	3	100.0	100.0	1.0	2.0
Northern Exposure	3	100.0	100.0	1.0	2.0
Quantum Leap	3	100.0	100.0	1.0	2.0
Ren and Stimpy 3 and a Half-is	3	66.67	100.0	5.33	27.33
Rocky and Bullwinkle and Friends	3	66.67	33.33	-	-
Roswell	3	100.0	100.0	1.0	3.67
Russell Simmons Presents Def Poetry	3	33.33	33.33	-	-
Sealab 2021	3	66.67	100.0	1.67	12.67
Seinfeld	3	100.0	100.0	1.0	2.0
Silk Stalkings	3	66.67	66.67	1.0	2.0
SpongeBob SquarePants	3	100.0	100.0	1.0	36.67
Star Trek: Enterprise	3	66.67	100.0	3.67	22.33
Starsky and Hutch	3	100.0	100.0	1.0	16.0
Strangers with Candy	3	100.0	100.0	1.0	2.0
Survivor 1: Borne	3	100.0	100.0	1.0	26.33

Taxi	3	100.0	100.0	1.0	7.0
The Andy Griffith Show	3	66.67	100.0	1.33	8.33
The Brady Bunch	3	33.33	50.0	-	-
The Dead Zone	3	100.0	100.0	1.0	2.33
The Flintstones	3	66.67	100.0	1.33	2.67
The Greatest American Hero	3	100.0	100.0	1.0	61.33
The Jamie Kennedy Experiment	3	100.0	100.0	1.0	7.67
The Kids in the Hall	3	100.0	66.67	1.0	3.0
The Osbournes	3	100.0	100.0	1.0	6.33
The Shield	3	100.0	100.0	1.0	2.0
Tour of Duty	3	66.67	100.0	5.33	17.0
What's Happening!!	3	0.0	50.0	-	-
Wonder Woman	3	100.0	66.67	1.0	2.0
Alias	4	100.0	100.0	1.0	3.75
All in the Family	4	100.0	100.0	1.0	10.25
CSI	4	100.0	100.0	1.0	3.0
Curb Your Enthusiasm	4	100.0	100.0	1.0	3.25
Degrassi Junior High	4	75.0	66.67	1.0	22.0
Everybody Loves Raymond	4	100.0	100.0	1.0	29.5
Farscape	4	100.0	100.0	1.0	3.0
Felicity	4	100.0	100.0	1.0	3.0
Gilmore Girls	4	100.0	100.0	1.0	3.0
In Living Color	4	100.0	75.0	1.0	3.0
Jeeves and Wooster	4	100.0	100.0	1.0	3.0
King of the Hill	4	100.0	100.0	1.0	16.25
Lost in Space	4	75.0	75.0	1.0	14.0
Married... with Children	4	75.0	75.0	1.0	3.0
Mr. Show	4	100.0	100.0	1.0	3.0
Profiler	4	100.0	100.0	1.0	5.0
Queer as Folk	4	100.0	100.0	1.0	3.0
Six Feet Under	4	100.0	100.0	1.0	3.0
Smallville	4	100.0	100.0	1.0	3.0
Soap	4	100.0	75.0	1.0	4.0
The Best of Friends	4	75.0	100.0	1.5	5.75
The Dukes of Hazzard	4	75.0	91.67	1.33	20.0
The Jeffersons	4	75.0	50.0	-	-
The King of Queens	4	100.0	100.0	1.0	7.75
The Man Show 1: Vol.	4	75.0	83.33	3.0	65.5
The West Wing	4	100.0	100.0	1.0	3.0
Three's Company	4	100.0	100.0	1.0	7.75
Transformers	4	75.0	100.0	1.25	3.75
Will and Grace	4	100.0	100.0	1.0	19.75
Andromeda	5	80.0	80.0	1.0	26.0
Angel	5	100.0	100.0	1.0	10.2
Babylon	5	100.0	100.0	1.0	4.4
Coupling	5	100.0	95.0	1.0	33.25

Dawson's Creek	5	100.0	100.0	1.0	5.0
Good Times	5	100.0	85.0	1.0	40.5
I Love Lucy	5	100.0	100.0	1.0	6.4
Law and Order	5	40.0	100.0	1.8	9.4
Oz	5	100.0	100.0	1.0	4.2
Saved by the Bell	5	100.0	100.0	1.0	37.6
The Dick Van Dyke Show	5	100.0	100.0	1.0	4.0
The Sopranos	5	100.0	100.0	1.0	4.0
Upstairs, Downstairs	5	100.0	100.0	1.0	4.0
Cheers	6	100.0	100.0	1.0	34.83
Combat! Season 1: Campaign	6	83.33	96.67	1.0	26.2
Dr. Quinn, Medicine Woman	6	100.0	100.0	1.0	12.33
Hercules: The Legendary Journeys	6	66.67	93.33	2.25	41.75
Highlander	6	100.0	100.0	1.0	7.67
Homicide: Life on the Street	6	100.0	100.0	1.0	8.33
Sanford and Son	6	100.0	83.33	1.0	82.0
South Park	6	100.0	100.0	1.0	32.33
The Simpsons	6	100.0	100.0	1.0	8.83
Xena: Warrior Princess	6	100.0	100.0	1.0	6.17
A Touch of Frost	7	57.14	50.0	-	-
Buffy the Vampire Slayer	7	100.0	100.0	1.0	6.0
Frasier	7	100.0	100.0	1.0	44.0
Sex and the City	7	100.0	100.0	1.0	6.0
Star Trek: Deep Space Nine	7	100.0	100.0	1.0	6.0
Star Trek: The Next Generation	7	100.0	100.0	1.0	6.0
Star Trek: Voyager	7	100.0	100.0	1.0	6.0
MASH	8	100.0	100.0	1.0	7.0
Stargate SG-1	8	100.0	100.0	1.0	30.13
Friends	9	100.0	100.0	1.0	12.78
Little House on the Prairie	9	88.89	79.17	-	-
The X-Files	9	100.0	100.0	1.0	8.78

Table 13.2: Average quality of recommendations based on the simple independence model for television series. Dashes in the last two columns indicate that no season of a series had all other seasons of the same series in the top 100 recommendations.

Observe, that many series have a `pbr` of 100%, i.e., the best recommendation is another season of this series. The `pra` is usually quite high as well, i.e., the first 100 recommendations contain often the other seasons of a series. The last two parameters that are measured in this list are not always perfect, but for many of the two seasons series the result is astonishingly good, since two seasons are not much and are likely to be mixed up with other similar series (e.g. "Family Guy" and "Simpsons", which are both by Matt Groening). Other examples with more seasons, such as "Hercules: The Legendary Journeys" do show estimates for the $\overline{\texttt{last}}$ parameter which are not as good as expected; this series is especially strange, since "Xena: Warrior Princess" is much better estimated. These two

series are closely related and have several overlaps. A closer look at the data reveals that the later seasons of "Hercules: The Legendary Journeys" have the complete series in their top 100 recommendations, while early seasons do not.

Spitz et al. [87] compared different prediction algorithms with each other. Here, the most interesting for the comparison are their SIM approach, since they use the old equation 13.1, and the $z*$-rating as the best result in almost all cases. The $z*$-rating uses the standard FDSM approach, orders results by a p-value defined by the authors and decides ties between two movies via the z-score (see [87] for more information).

Spitz et al. had a ground truth of sequels of films and television series and compared these with the results the different algorithms yielded. One of the authors provided their data set and ground truth such that a comparison between the adapted equation and their result is possible. Even though in their work they used other similarity measures as well, I will focus here on the old SIM equation 13.1, the FDSM with the $z*$-approach, and the new equation for the SIM, equation 13.6.

The provided ground truth contains 15 James Bond movies. For local predictions, the graph spanned by these movies is considered as a directed graph. When making local forecasts, it is reasonable to use directed graphs since not every movie has the same predictions. Within the ground truth, the James Bond movies build a clique with 210 edges, i.e., each movie is connected with any other. The $z*$-method was capable of recovering 201 of the 210 edges [87], which is a very good result. On the same data set, the algorithm using the new equation 13.6 yields an even better result with 203 of 210 possible edges, while the old equation 13.1 was not tested. Thus, testing whether the adapted equation yields for all datasets results which are as good seems promising.

Computing the leverage with the adapted equation yields quick results. Even in the most naive implementation, calculating the expected co-occurrence takes only n multiplications for a single movie. The sampling approach has to take some samples, and each sample requires $\mathcal{O}(m \log(m))$ swaps. Thus, a small loss in quality would be acceptable for speed. The ground truth consists out of titles, and their k follow-ups. Thus, a comparison of the top k recommendations based on the leverage to the ground truth for all series and movies is performed. The average of the local PPV_k is shown in Table 13.3.

Data set	$z*$ [87]	SIM [87]	$\text{SIM}_{\text{adapted}}$
Netflix Series	0.8694	0.6851	0.7136
Netflix Movies	0.57774	0.28954	0.3456
MovieLens	0.5122	0.2346	0.3461

Table 13.3: Comparison of the results of Spitz et al. and the algorithm using the adapted equation of the average local PPV_k.

The local predictions did in all cases improve on average to the old way to calculate the expected co-occurrence. Still, the $z*$ approach which uses sampling in the FDSM is better in all cases. As was observed on the "James Bond"-movie series, it may be for large movie series or TV series that the fast calculation with the SIM may yield results which are at least as good as the ones from the sampling approach.

Similar results are found when the list of the globally top recommendations is under investigation. This list consists out of the top y predictions of all predictions. The test is how many of the ground truth edges are within the top y predictions.

Data set	$z*$ [87]	second best [87]	$\text{SIM}_{adapted}$
Netflix Series	0.72	0.65	0.43
Netflix Movies	0.39	0.21	0.22
MovieLens	0.30	0.22	0.21

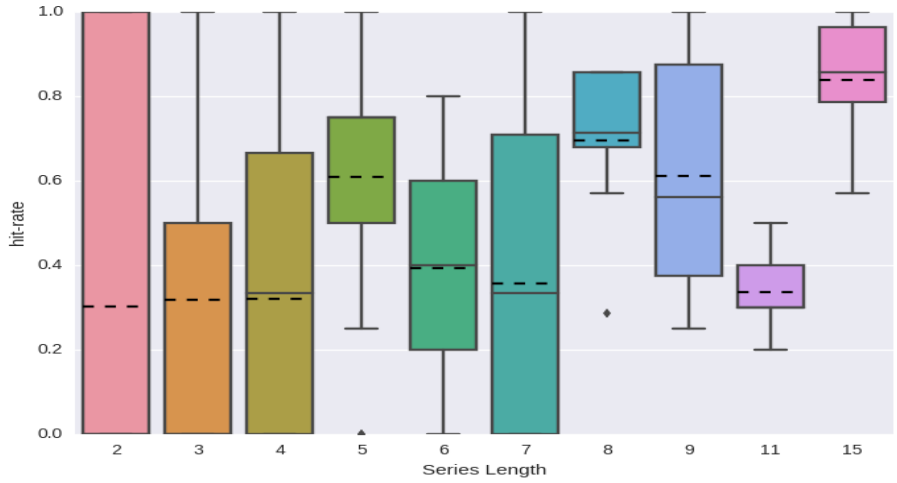
Table 13.4: Comparison of the results of Spitz et al. and the algorithm using equation 13.6 of the global PPV_k.

The Netflix Series data set is not very well predicted on a global level; the adapted equation does not even perform better than the second best method from Spitz et al. For the other data sets, Netflix Movies and MovieLens, the new equation 13.6 performs about as good as the second-best method, but still not as good as $z*$. For the "James Bond" data set, the local predictions of the adapted equation yielded more ground truth edges than the $z*$ approach (203 instead of 201).

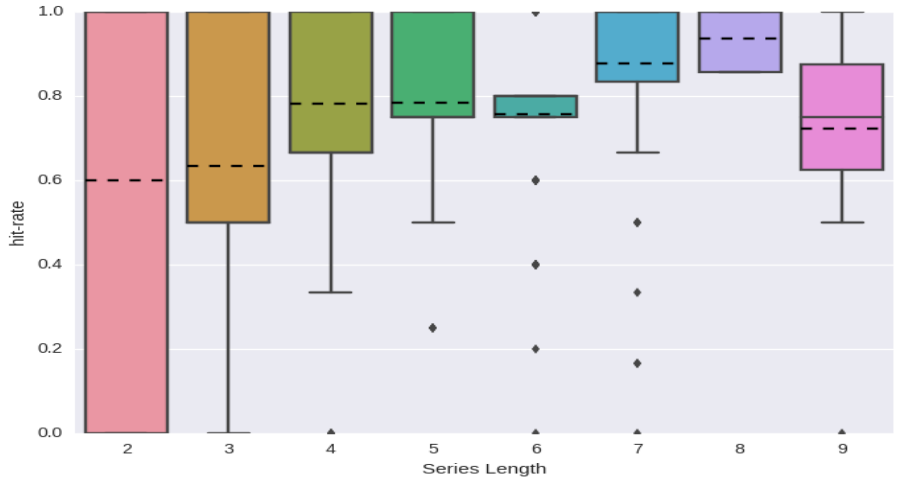
For global predictions, this is true as well. For global predictions, the graph is considered as undirected. It has 15 nodes, one for each movie, and in the ground truth 105 edges, one between each pair of movies. Spitz et al. showed that the $z*$ approach to this problem yields 55 out of 105 possible edges when using the 904 globally most similar movie pairs. The adapted equation yields only 51 edges. When using a larger number of most similar movies, i.e., using the length of the ground truth, the approach using the adapted equation yields 72 edges. As a side note, one of the James Bond movies was found neither in the 904 most similar movies nor the longer list. This movie is "The Living Daylights", in the 904 pairs long list is also "License to Kill" missing, i.e., these movies are never recommended by any of the other movies in the top list - the two movies in which James Bond is portrayed by Timothy Dalton, who played a darker, more gritty Bond than the actors before him. The actor may be the reason it is not as similar to the other movies from the series.

That the "James Bond" movies are predicted quite well could be a hint to a relation between "number of elements of a series" to movies (series) watched and rated by the same person. This assumption seems to be somewhat plausible. When a viewer watches consistently all episodes of a TV series (or movie series), it is more likely that they all have a rating and that they do not get confused with other series. If there is only one sequel to a movie or only a second season to a TV show, it is much harder to find evidence that they belong together.

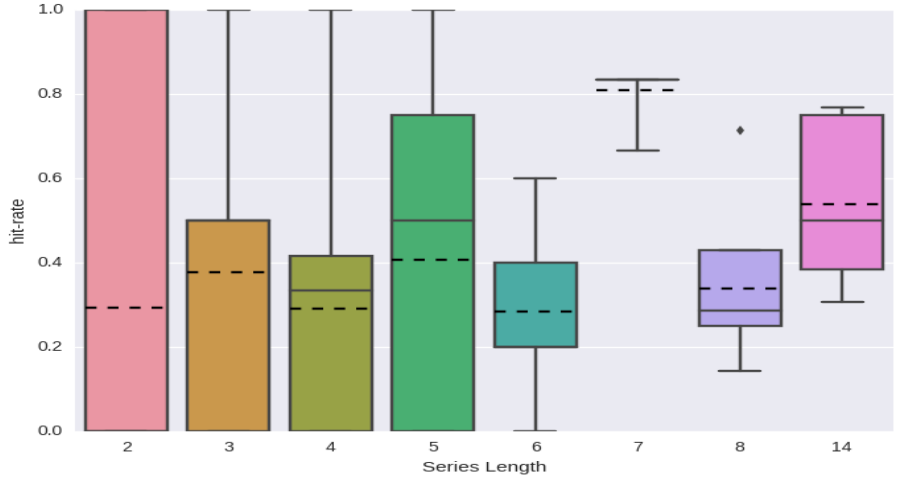
As can be seen in Fig. 13.1, TV series (Fig 13.1b) or movie series (Fig. 13.1a and Fig. 13.1c) with more elements are predicted better than the ones with fewer parts. There are two cases for which this assumption seems to be completely wrong; thus, they need investigation: The Netflix Movie data set for the case of 11 pre-/sequels and the MovieLens data set with eight pre-/sequels. A quick look at the ground truth shows that these movies are 11 Godzilla-based and a set of 8 Muppet/Sesame Street films. The latter collection was released in a time frame of 20 years; this may be the reason edges could not be assessed correctly. The former collection was released over a time span of 36 years. It contains only Japanese productions; Netflix officially started streaming in Japan in 2015 and the



(a) Netflix Movies



(b) Netflix Series



(c) MovieLens

Figure 13.1: Overview of hit-rate of correctly assessed edges; the more elements a series has, the more likely edges appear to be assessed correctly.

dataset was released in 2006, special effects improved rather much, Japanese movies are not necessarily the primary interest of the standard viewer of Netflix, the producers, cast, and story changed—all of these and more may be reasons why the Godzilla collection is not recognized very well. It may also be related to the construction of the data set—only ratings with a grade of 4 and higher are included. In IMDB[2] the movies had a rating of about 6 out of 10, Rotten Tomatoes [3] did not have a rating for all of them but the ones which were rated were about 60% as well. Thus, the construction of the data set used may have an influence as well.

Overall, even though for some movie series or TV series the SIM approach works well enough to show other elements of the series, it does not work for all of them. Smaller series, older series, series spanning a long time may benefit more from the more sophisticated $z*$ approach. Still, it is interesting to see that a small change in the equation yields this much of an improvement when using the SIM to assess similarities.

13.1 THE MODEL TO USE AND OPEN PROBLEMS

The results show for directed and undirected graphs quite similar results. When using the CFG, there is the possibility that multiple edges or self-loops are included. Self-loops do not contribute anything to most network analytic measures. Handling multiple edges between nodes does need a careful consideration since multiple edges between two nodes are not necessarily intended. Researchers who use this model sometimes drop these "bad edges". Depending on the degree distribution of the graph under investigation, this may or may not be a uniform loss in the degree sequence, but for many real-world graphs, it will be a non-uniform loss since there are usually much less high degree nodes than low-degree nodes. The loss may influence measures but usually is not considered when analyzing any graph. For both variants of the CFG it is important to check several things before applying them to analyze a graph.

1. How do measures and algorithms that are applied handle multiple edges/self-loops?
2. Are the measures degree-dependent in any way?
3. Is the loss of the property to be exact (i.e. using the FDSM) worth the speed-up by using the CFG? If not, i.e., if multiple edges or self-loops occur and the current graph is dismissed to start generating a new one, is it a speed-up?

The SIS is not the fastest model and it is not considered very often. The results indicate that it is not important whether a uniform or a degree dependent choice of neighbors is used in subgraph counting or network motif analysis. Since the run-time estimate $\mathcal{O}\left(n^3\right)$ to generate one graph is, under the assumption that $n \sim m$, worse than $\mathcal{O}\left(nm\log\left(m\right)\right)$, the runtime of the analysis using the FDSM, the SIS is not the best choice for analysis.

The SIM is the mathematical approximation of the CFG. The assumption that each edge can exist with some probability regardless of whether a node has already some edges is the principle idea. From this idea, all equations which were used in this chapter can be developed. It is possible to predict with the equations how many multiple edges and

2 http://www.imdb.com

3 http://www.rottentomatoes.com/

how many self-loops a graph that is generated with the CFG has. For directed graphs, the results are twofold; first, global estimates of the number of motifs are possible and most of the time within a standard deviation of the results of the baseline model, the FDSM. For the graphs that showed problems, the node-based analysis indicated that the problems stem from very few nodes. These nodes have high degrees and participate quite often in the subgraphs of interest. The analysis in the last part confirms this finding. For nodes with high degree, the equations overestimate their worth, as soon as the degree of a node contributes non-linearly. Furthermore, equations that take into account the repetition of a node (e.g. (u, u, v, w)), have been developed. These are more exact, but can be quite cumbersome to calculate.

Therefore, when using the SIM, several things have to be taken into consideration

1. The degree sequence should be not (too) skewed.
2. There should be only few to no nodes violating

$$\left(k_v^2 < \sum_{u \in V} k_u\right) \wedge \left(j_v^2 < \sum_{u \in v} j_v\right) \wedge \left((k_v + j_v)^2 < \sum_{u \in V} k_v + j_v\right)$$

3. When the result of the equation and the value of the real-world graph are very different (depending on the measure used), use a model that generates only simple graphs to check the validity of the result.

Of course, the equations can misestimate the number of subgraphs and yield results which are similar to the result of the real-world graph, which would be unfortunate. But, since the approach of using equations to approximate real-world systems is not uncommon, my confidence that the results of the SIM and the FDSM are similar or even coincide is high. Still, the Complete subgraph together with the Ythan Estuary graph and the Little Rock graph show that it is never bad to be careful.

Part V

SUMMARY

14

SUMMARY AND CONCLUSIONS

Network analysis calculates measures on graphs. While usually in social network analysis the measures are calculated and interpreted with experts regarding the network, another approach is to calculate the measure and compare it first with other graphs, before interpreting it. This procedure ensures that the measure can take other values, very different values in fact. If the other graphs have all the same value regarding a measure, it is not important enough to discuss this measure. The problem is, with which graphs should the comparison take place? Whenever a larger graph ($n > 10000$) is compared with a small graph ($n < 1000$), the probability that the results are different is high — especially when the assumption of $m = \mathcal{O}(n)$ holds [71]. Finding enough graphs that are similar in size is hard. Thus, random graph models are used to achieve this goal. The interesting question is, which random graph model to use?

14.1 SUMMARY

Historically, the $\mathcal{G}(n, m)$ model was enough to compare graphs to, but nowadays graphs have very different structures and degree sequences, such that analyses based on a comparison to this model yields the graph as exceptional, no matter what.

This was shown in Sections 4.1, 4.2. To do so, we used once a graph with a Poissonian degree distribution that was randomly generated as well. To compare this to other graphs, which have one important feature in common, the degree sequence, two different null models and their corresponding algorithms were used. The configuration model and the fixed degree sequence model. For a graph with an unskewed Poissonian degree distribution, it did not seem to be important which null model is used. However, when a graph with a skewed degree distribution is analyzed, the CFG showed very different results from the FDSM. As was observed later, this was due to a non-uniform distribution of multiple edges and self-loops. The CFG is the only null model used that allows for such edges. Still, the results of the analysis on real-world graphs based on purely global measures showed the CFG is a fast and plausible alternative to the FDSM, at least for some measures. As soon as we changed the analysis to a local measure and took the average/sum of this result, the CFG was not accurate. Now, one might complain that the average distance is a local measure as well since for any node all distances are computed and then the averaging starts. However, distances involve the whole component a node is embedded in, not only the direct neighbors. Thus, distances are considered as a global measure, while the average neighbor degree is considered as a local measure. Moreover, for this measure, as well as the co-occurrence, the CFG did fail to produce results similar to our baseline model, the FDSM.

The erased configuration model that generates graphs based on the configuration model and erases multiple edges and self-loops afterward fared even worse in the local measures.

For the network motif analysis, the CFG and the ECFG are both not very useful. A good example is the Feed-Forward Loop; in the real-world E. coli graph the number of Feed-Forward Loops is 42; the z-score that is produced by the CFG would be only 4.01, which is too low to consider the Feed-Forward Loop as extraordinary. The FDSM has a z-score of 10.05, the ECFG 13.41. Thus, one may consider the ECFG as a null-model for motif analysis as well. However, considering the edge-loss in graphs with skewed degree distributions, it cannot be used as a reasonable model.

To the best of my knowledge, the sequential importance sampling has been investigated the first time on such a scale. The choice of the next neighbor to a node is important. When the next neighbor is chosen based on the remaining degree of the possible neighbors, the generated graphs are more dis-assortative than when generated with the FDSM. The DSIS connects low degree more often to high degree nodes than other nodes, due to the simple fact that the algorithm starts out with low degree nodes and attempts to connect them to other nodes. Thus, often edges are build between low and high degree nodes, which skews measures such as the average neighbor degree to high values for many low degree nodes. Thus, the DSIS should be avoided whenever local measures are concerned. For global measures, one has to be aware of this as well. The diameter is a global measure and for graphs generated from the DSIS, it was always low. This observation together with the preferences described above leads to the hypothesis that the DSIS may not be the null model to use for statistical comparison. The second variant of the SIS, choosing neighbors uniformly at random, generates diameters and distances just as the FDSM. Additionally, it also generates the other measures tested on undirected graphs as the FDSM does. It can be a replacement of the FDSM, but the algorithm is slow. Thus, this model can be used, but in general, a speed-up cannot be guaranteed.

Regarding network motif analysis, it is harder to tell which of the two is more useful. Sometimes, the USIS scores better regarding the D-value of the Kolmogorov-Smirnoff two-sample test, sometimes the DSIS. Overall, both models have results close to the FDSM. However, the fact that the worst-case run-time of the FDSM is better than of the SIS shows that the FDSM is the algorithm to use.

The last model, the simple independence model, is a very basic probability model. It works very well with the $\mathcal{G}(n, m)$-model and other graphs that have an unskewed degree sequence. As soon as the degree sequence is skewed, it does not work very well. Measures such as the diameter and distance cannot be calculated with the simple methods suggested, which is due to the use of averages to calculate the measures. Using these averages may work on some graphs, but as shown in Chapter III, it does not work very well on real-world graphs. The average neighbor degree is estimated as bad as in the CFG, which is understandable since it is an approximation of the same. The model does not match well with the FDSM. For the co-occurrence the results are twofold—calculating the co-occurrence with the basic equation by Zweig and Kaufmann [108] is still off, but the approximation of Newman [66] is a good match for the global co-occurrence. Looking at the co-occurrence locally, the results are again off, and thus, only usable as approximations.

Considering network motif analysis, the equations of the SIM work well enough for many graphs and many subgraphs to skip analysis with the FDSM. Luckily, several graphs showed severe problems on some of the subgraphs, such that an investigation why it does not work

was started. On the data sets used, only one indicator aligned well enough with the bogus results, $\forall v \in V : \left(k_v^2 < \sum_{u \in V} k_u\right) \wedge \left(j_v^2 < \sum_{u \in V} j_u\right) \wedge \left((k_v + j_v)^2 < \sum_{u \in V} k_u + j_u\right)$. Since this test is the only one that had satisfying results and the test is not based on a comparison to results of a sampling algorithm, this test may be promising for future analysis. The standard deviation was calculated explicitly, but an approximation seems to suffice.

14.2 CONCLUSIONS TO DRAW

To analyze a graph by comparing it to a appropriate null model is not the fastest way, but it offers a statistical way to decide whether something is exceptional or if it could happen at random. Since the generation of the random graphs is not cheap, it has to be considered which approach should be used. For this, a simple guideline can be the following:

- Analyze the degree distribution of the graph.
  - not skewed—consider using the CFG or the ECFG
  - skewed—do not use the CFG or the CFG
- Test the percentage of nodes u for that $k_u^2 > \sum_{v \in V} k_v$
  - not skewed—consider using the CFG or the ECFG
  - skewed—do not use the CFG or the CFG
- Global or local measure?
  - global and the test show you may use the CFG? Consider using the SIM, CFG, or the ECFG
  - global and the test show you should not use the CFG? Use the FDSM or the USIS
  - local—use the FDSM

Surprisingly, the SIM is as good as not mentioned in the guideline. That is because not many measures are defined using the SIM, and even worse, variances based on this model are hard to calculate. It may be possible to calculate the variance more simple by using generating functions [69], but for combinations of nodes, as practiced in the network motif analysis (Section 11.3)), generating functions will most likely not be easier to handle. Furthermore, whenever the CFG can be used, the SIM can be used to approximate the expected value of a measure. When an equation to calculate the standard deviation exists, it can easily be applied. When no equation for this is given, a more thorough analysis based on simulations with an appropriate null model is necessary.

In the network motif analysis part I showed that it is possible to calculate standard deviation, but only when assuming the events as uncorrelated—which clearly, they are not. Still, as an approximation, it is good enough.

14.3 FUTURE WORK

The SIM is incredibly powerful—but also quite dangerous to use. Dangerous in the sense of "Whenever one trusts in results without verifying with a proper null model, it may end costly". The ease with which equations can be developed was shown several times

throughout this thesis. That not every equation is correct for every graph, was shown as well. Thus, using the SIM as a first approximation and verifying results with an appropriate null model seems more reasonable. Of course, some measures should never be calculated with the SIM on real-world graphs, since some assumptions are just not applicable. The other models do all have their kinks—either a worst-case runtime that is high, multiple edges, or similar problems. It would be very useful to investigate more deeply when the SIM can be applied without causing problems. As observed in the network motif analysis part, even though the degree sequences of some graphs do not follow the rule $\forall v \in V : \left(k_v^2 < \sum_{u \in V} k_u\right) \wedge \left(j_v^2 < \sum_{u \in V} j_u\right) \wedge \left((k_v + j_v)^2 < \sum_{u \in V} k_u + j_u\right)$, not all subgraph occurrences are estimated badly. Even some of the graphs that do not follow the rule show quite good expectation values. A deeper investigation is necessary to get a more clear decision factor on when the SIM or the CFG may be applied. Additionally, mathematically more experienced people should take a closer look at the problem with the standard deviation.

Another necessary proof that is still missing is for the algorithm of the FDSM, i.e., the mixing time. There have been many investigative attempts to lower the number of necessary swaps, to lower the number of samples, and so on, but a proof is still missing. Without such a proof, the run-time estimate is usually taken on the large side. The best example for this is in Milo et. al [63], where the experiment they used was stable after $\sim 1\,|E|$ edge swaps, but they still chose $100\,|E|$ to be on the safe side.

Besides this, it would be very useful to discuss the equations to estimate the number of subgraphs with some researchers from the field of Biology. This field uses network motif analysis the most often. Thus, it would be good to know whether they think it is applicable. If they deemed it applicable, it would be very useful to have an easy to control software, which allows researchers from the field to estimate the number of subgraphs, without bothering them with math.

BIBLIOGRAPHY

[1] John C. Almack. "The influence of intelligence on the selection of associates". In: *School and Society* 16 (1922), pp. 529–530.

[2] Lars Backstrom et al. "Four degrees of separation". In: *Proceedings of the 4th Annual ACM Web Science Conference*. ACM. 2012, pp. 33–42.

[3] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[4] Richard Bellman. *On a routing problem*. Tech. rep. DTIC Document, 1956.

[5] Edward A Bender and E Rodney Canfield. "The asymptotic number of labeled graphs with given degree sequences". In: *Journal of Combinatorial Theory, Series A* 24.3 (1978), pp. 296–307.

[6] Annabell Berger and Matthias Müller-Hannemann. "Uniform sampling of digraphs with a fixed degree sequence". In: *Graph theoretic concepts in computer science*. Springer. 2010, pp. 220–231.

[7] Etienne Birmele et al. "Detecting local network motifs". In: *Electronic Journal of Statistics* 6 (2012), pp. 908–933.

[8] Joseph Blitzstein and Persi Diaconis. "A sequential importance sampling algorithm for generating random graphs with prescribed degrees". In: *Internet Mathematics* 6.4 (2011), pp. 489–522.

[9] Béla Bollobás and Oliver Riordan. "The diameter of a scale-free random graph". In: *Combinatorica* 24.1 (2004), pp. 5–34.

[10] Béla Bollobás and Andrew Thomason. "Random graphs of small order". In: *North-Holland Mathematics Studies* 118 (1985), pp. 47–97.

[11] Michele Borassi et al. "Fast diameter and radius BFS-based computation in (weakly connected) real-world graphs: With an application to the six degrees of separation games". In: *Theoretical Computer Science* 586 (2015), pp. 59–80.

[12] I.N. Bronstein and K.A. Semendjajew. *Taschenbuch der Mathematik, 25*. H. Deutsch, 1991, p. 690.

[13] Richard A Brualdi. "Matrices of zeros and ones with fixed row and column sum vectors". In: *Linear algebra and its applications* 33 (1980), pp. 159–231.

[14] Christian Brugger et al. "Exploiting Phase Transitions for the Efficient Sampling of the Fixed Degree Sequence Model". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM. 2015, pp. 308–313.

[15] Sonia Cafieri, Pierre Hansen, and Leo Liberti. "Loops and multiple edges in modularity maximization of networks". In: *Physical Review E* 81.4 (2010), p. 046102.

[16] Pankaj Choudhary and Upasna Singh. "Article: A Survey on Social Network Analysis for Counter-Terrorism". In: *International Journal of Computer Applications* 112.9 (Feb. 2015), pp. 24–29.

[17] Fan Chung and Linyuan Lu. "The average distances in random graphs with given expected degrees". In: *Proceedings of the National Academy of Sciences* 99.25 (Dec. 10, 2002), pp. 15879–15882. ISSN: 1091-6490.

[18] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Rev.* 51.4 (Nov. 2009), pp. 661–703. ISSN: 0036-1445.

[19] Pilu Crescenzi et al. "On computing the diameter of real-world undirected graphs". In: *Theoretical Computer Science* 514 (2013), pp. 84–95.

[20] R.L. Cross and A. Parker. *The Hidden Power of Social Networks: Understanding how Work Really Gets Done in Organizations*. Harvard Business School Press, 2004. ISBN: 9781591392705.

[21] Charo I. Del Genio et al. "Efficient and Exact Sampling of Simple Graphs with Given Arbitrary Degree Sequence". In: *PLoS ONE* 5.4 (Apr. 2010), e10012.

[22] Edsger W Dijkstra. "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1.1 (1959), pp. 269–271.

[23] Sergey Edunov et al. *Three and a half degrees of separation*. English. Facebook. Feb. 2016.

[24] P. Erdös and A. Rényi. "On random graphs, I". In: *Publicationes Mathematicae (Debrecen)* 6 (1959), pp. 290–297.

[25] Paul Erdős. "On the evolution of random graphs". In: *Publ. Math. Inst. Hungar. Acad. Sci* 5 (1960), pp. 17–61.

[26] Scott L Feld. "Why your friends have more friends than you do". In: *American Journal of Sociology* (1991), pp. 1464–1477.

[27] Robert W Floyd. "Algorithm 97: shortest path". In: *Communications of the ACM* 5.6 (1962), p. 345.

[28] Agata Fronczak. "Exponential random graph models". In: *CoRR* abs/1210.7828 (2012).

[29] E. N. Gilbert. "Random Graphs". In: *Ann. Math. Statist.* 30.4 (Dec. 1959), pp. 1141–1144.

[30] Kwang-Il I. Goh et al. "The human disease network." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.21 (May 22, 2007), pp. 8685–8690. ISSN: 0027-8424.

[31] Leo A. Goodman. "On the Exact Variance of Products". In: *Journal of the American Statistical Association* 55.292 (1960), pp. 708–713. ISSN: 01621459.

[32] Leo A. Goodman. "The Variance of the Product of K Random Variables". In: *Journal of the American Statistical Association* 57.297 (1962), pp. 54–60. ISSN: 01621459.

[33] Mark S. Granovetter. "The Strength of Weak Ties". In: *The American Journal of Sociology* 78.6 (1973), pp. 1360–1380.

[34] Aric Hagberg and Nathan Lemons. "Fast Generation of Sparse Random Kernel Graphs". In: *PLoS ONE* 10.9 (Sept. 2015), pp. 1–12.

[35] Seifollah Louis Hakimi. "On realizability of a set of integers as degrees of the vertices of a linear graph. I". In: *Journal of the Society for Industrial & Applied Mathematics* 10.3 (1962), pp. 496–506.

[36] Frank Harary. *Graph Theory*. Addison-Wesley series in mathematics. Addison-Wesley Publishing Company, 1969. ISBN: 9780201410334.

[37] Václav Havel. "A Remark on the Existence of Finite Graphs". cze. In: *Časopis pro pěstování matematiky* 080.4 (1955), pp. 477–480.

[38] Lenwood S. Heath and Nidhi Parikh. "Generating random graphs with tunable clustering coefficients". In: *Physica A: Statistical Mechanics and its Applications* 390.23–24 (2011), pp. 4577–4587. ISSN: 0378-4371.

[39] Pim van der Hoorn and Nelly Litvak. "Phase transitions for scaling of structural correlations in directed networks". In: *Phys. Rev. E* 92 (2 Aug. 2015), p. 022803.

[40] Pim van der Hoorn and Nelly Litvak. "Upper bounds for number of removed edges in the Erased Configuration Model". In: *Algorithms and Models for the Web Graph*. Springer, 2015, pp. 54–65.

[41] Emőke-Ágnes Horvát and Katharina Anna Zweig. "A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs". In: *Social Network Analysis and Mining* 3.4 (2013), pp. 1209–1224. ISSN: 1869-5469.

[42] Hong Huang et al. "Mining Triadic Closure Patterns in Social Networks". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. Seoul, Korea: International World Wide Web Conferences Steering Committee, 2014, pp. 499–504. ISBN: 978-1-4503-2745-9.

[43] Shalev Itzkovitz et al. "Subgraphs in random networks". In: *Physical review E* 68.2 (2003), p. 026127.

[44] Svante Janson. "The probability that a random multigraph is simple". In: *Combinatorics, Probability and Computing* 18.1-2 (2009), pp. 205–225.

[45] Svante Janson et al. "The probability that a random multigraph is simple. II". In: *Journal of Applied Probability* 51 (2014), pp. 123–137.

[46] Carter Jernigan and Behram F. T. Mistree. "Gaydar: Facebook Friendships Expose Sexual Orientation." In: *First Monday* 14.10 (2009).

[47] Donald B Johnson. "Efficient algorithms for shortest paths in sparse networks". In: *Journal of the ACM (JACM)* 24.1 (1977), pp. 1–13.

[48] Nadav Kashtan et al. "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs". In: *Bioinformatics* 20.11 (2004), pp. 1746–1758.

[49] Hyunju Kim et al. "Constructing and sampling directed graphs with given degree sequences". In: *New Journal of Physics* 14.2 (2012), p. 023012.

[50] Bryan Klimt and Yiming Yang. "Introducing the Enron Corpus." In: *CEAS*. 2004.

[51] D. König. *Theorie der endlichen und unendlichen Graphen: kombinatorische Topologie der Streckenkomplexe*. Chelsea Pub. Co., 1950.

[52] Dirk Koschützki, Henning Schwöbbermeyer, and Falk Schreiber. "Ranking of network elements based on functional substructures". In: *Journal of Theoretical Biology* 248.3 (2007), pp. 471–479. ISSN: 0022-5193.

[53] Valdis Krebs. "Mapping Networks of Terrorist Cells". In: *CONNECTIONS* 24.3 (2002), pp. 43–52.

[54] D.C. LeBlanc. *Statistics: Concepts and Applications for Science*. Statistics: Concepts and Applications for Science v. 2. Jones and Bartlett, 2004. ISBN: 9780763746995.

[55] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations". In: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. KDD '05. Chicago, Illinois, USA: ACM, 2005, pp. 177–187. ISBN: 1-59593-135-X.

[56] Jure Leskovec et al. "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters". In: *Internet Mathematics* 6.1 (2009), pp. 29–123.

[57] Roy HA Lindelauf, Peter Borm, and Herbert Hamers. "Understanding terrorist network topologies and their resilience against disruption". In: (2009).

[58] Tiancheng Lou et al. "Learning to Predict Reciprocity and Triadic Closure in Social Networks". In: *ACM Trans. Knowl. Discov. Data* 7.2 (Aug. 2013), 5:1–5:25. ISSN: 1556-4681.

[59] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719, 9780521865715.

[60] Julian Mcauley and Jure Leskovec. "Discovering Social Circles in Ego Networks". In: *ACM Trans. Knowl. Discov. Data* 8.1 (Feb. 2014), 4:1–4:28. ISSN: 1556-4681.

[61] Stanley Milgram. "The small world problem". In: *Psychology today* 2.1 (1967), pp. 60–67.

[62] Ron Milo et al. "Network Motifs: Simple Building Blocks of Complex Networks". In: *Science* 298.5594 (2002), pp. 824–827.

[63] Ron Milo et al. *On the uniform generation of random graphs with prescribed degree sequences*. Dec. 2003. URL: `https://arxiv.org/abs/cond-mat/0312028` (visited on 11/19/2012).

[64] Ron Milo et al. "Response to Comment on "Network Motifs: Simple Building Blocks of Complex Networks" and "Superfamilies of Evolved and Designed Networks"". In: *Science* 305.5687 (Aug. 2004), pp. 1107–1107.

[65] J.L. Moreno and H.H. Jennings. *Who shall survive?: A new approach to the problem of human interrelations*. Nervous and mental disease monograph series. Nervous and mental disease publishing co., 1934.

[66] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010. ISBN: 978-0-19-920665-0.

[67] Mark E.J. Newman. "Mixing patterns in networks". In: *Physical Review E* 67.2 (Feb. 27, 2003), pp. 026126+.

[68] Mark EJ Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical review E* 69.2 (2004), p. 026113.

[69] Mark E.J. Newman, Duncan J. Watts, and Steven H. Strogatz. "Random graph models of social networks". In: *P Natl Acad Sci USA* 99 (2002), pp. 2566–2572.

[70] Nicholas C. Wormald and University of Newcastle (N.S.W.). Dept. of Electrical Engineering. *Some Problems in the Enumeration of Labelled Graphs*. University of Newcastle, 1978.

[71] Darko Obradovic. "Computational Social Network Analysis of Authority in the Blogosphere". PhD thesis. Technische Universität Kaiserslautern, 2012.

[72] Darko Obradovic and Maximilien Danisch. "Direct generation of random graphs exactly realising a prescribed degree sequence." In: *CASoN*. 2014, pp. 1–6.

[73] Tore Opsahl and Pietro Panzarasa. "Clustering in weighted networks". In: *Social Networks* 31.2 (2009), pp. 155–163. ISSN: 0378-8733.

[74] Arie Perliger and Ami Pedahzur. "Social network analysis in the study of terrorism and political violence". In: *PS: Political Science & Politics* 44.01 (2011), pp. 45–50.

[75] Gregory Piatetsky-Shapiro. "Discovery, analysis, and presentation of strong rules". In: *Knowledge discovery in databases* (1991), pp. 229–238.

[76] Jaideep Ray, Ali Pinar, and C. Seshadhri. "Are We There Yet? When to Stop a Markov Chain while Generating Random Graphs". English. In: *Algorithms and Models for the Web Graph*. Ed. by Anthony Bonato and Jeannette Janssen. Vol. 7323. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 153–164. ISBN: 978-3-642-30540-5.

[77] Alan Roberts and Lewis Stone. "Island-sharing by archipelago species". In: *Oecologia* 83.4 (1990), pp. 560–567.

[78] Daniel Mauricio Romero and Jon M. Kleinberg. "The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter". In: *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. 2010.

[79] Wolfgang E. Schlauch. "Untersuchung von Algorithmen zur Generierung uniform verteilter Zufallsgraphen anhand einer Grad-Sequenz". Bachelor Thesis. Technische Universität Kaiserslautern.

[80] Wolfgang E Schlauch, Emőke Ágnes Horvát, and Katharina A Zweig. "Different flavors of randomness: comparing random graph models with fixed degree sequences". In: *Social Network Analysis and Mining* 5.1 (2015), pp. 1–14.

[81] Wolfgang E. Schlauch et al. "Influence of the Null-Model on Motif Detection". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ASONAM '15. Paris, France: ACM, 2015, pp. 514–519. ISBN: 978-1-4503-3854-7.

[82] Daniel M. Schwartz and Tony (D.A.) Rouselle. "Using social network analysis to target criminal networks". English. In: *Trends in Organized Crime* 12.2 (2009), pp. 188–207. ISSN: 1084-4791.

[83] Termeh Shafie. “Random Multigraphs: Complexity Measures, Probability Models and Statistical Inference”. PhD thesis. 2012.

[84] Shai S Shen-Orr et al. “Network motifs in the transcriptional regulation network of Escherichia coli”. In: *Nature genetics* 31.1 (2002), pp. 64–68.

[85] Tong-Wook Shinn and Tadao Takaoka. “Efficient Graph Algorithms for Network Analysis”. In: *CoRR* abs/1309.3849 (2013).

[86] Georg Simmel. *Philosophie des Geldes*. Duncker & Humbolt, 1907.

[87] Andreas Spitz et al. “Assessing Low-Intensity Relationships in Complex Networks”. In: *PloS one* 11.4 (2016), e0152536.

[88] Isabelle Stanton and Ali Pinar. “Sampling graphs with a prescribed joint degree distribution using Markov chains”. In: *Proceedings of the Meeting on Algorithm Engineering & Expermiments*. Society for Industrial and Applied Mathematics. 2011, pp. 127–138.

[89] Michael Szell and Stefan Thurner. “Measuring social dynamics in a massive multiplayer online game”. In: *Social Networks* 32.4 (2010), pp. 313–329. ISSN: 0378-8733.

[90] Michael Szell and Stefan Thurner. “Social Dynamics in a Large-Scale Online Game”. In: *Advances in Complex Systems* 15.6 (2012).

[91] Tadao Takaoka. “Single Source Shortest Paths for All Flows with Integer Costs”. In: *15th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems, ATMOS 2015, September 17, 2015, Patras, Greece*. 2015, pp. 56–67.

[92] Amanda L. Traud et al. “Comparing Community Structure to Characteristics in Online Collegiate Social Networks”. In: *SIAM Rev.* 53.3 (Aug. 2011), pp. 526–543. ISSN: 0036-1445.

[93] Johan Ugander et al. “The anatomy of the facebook social graph”. In: *arXiv preprint arXiv:1111.4503* (2011).

[94] Stefan Uhlmann et al. “Global miRNA Regulation of A Local Protein Network: Case Study with the EGFR-Driven Cell Cycle Network in Breast Cancer”. In: *Molecular Systems Biology* 8 (2012), p. 570.

[95] Remco Van Der Hofstad. *Random graphs and complex networks*. 2009.

[96] Fabien Viger and Matthieu Latapy. “Efficient and simple generation of random simple connected graphs with prescribed degree sequence”. In: *Computing and Combinatorics*. Springer, 2005, pp. 440–449.

[97] Olivier J Walther and Dimitris Christopoulos. “A social network analysis of Islamic terrorism and the Malian rebellion”. In: *CEPS/INSTEAD Working Papers* 38 (2012).

[98] Pei Wang, Jinhu Lü, and Xinghuo Yu. “Identification of Important Nodes in Directed Biological Networks: A Network Motif Approach”. In: *PLoS ONE* 9.8 (Aug. 2014), e106132.

[99] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.

[100] Duncan J. Watts and Steven H. Strogatz. “Collective Dynamics of "small-world" networks”. In: *Nature* (1998).

[101] Beth Wellman. "The School Child's Choice of Companions". In: *The Journal of Educational Research* 14.2 (1926), pp. 126–132.

[102] Sebastian Wernicke. "A Faster Algorithm for Detecting Network Motifs". English. In: *Algorithms in Bioinformatics*. Ed. by Rita Casadio and Gene Myers. Vol. 3692. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2005, pp. 165–177. ISBN: 978-3-540-29008-7.

[103] Sebastian Wernicke. "Efficient Detection of Network Motifs". In: *IEEE/ACM Trans. Comput. Biology Bioinform.* 3.4 (2006), pp. 347–359.

[104] Marco Winkler. "NoSPaM Manual - A Tool for Node-Specific Triad Pattern Mining". In: *CoRR* abs/1509.03503 (2015).

[105] Marco Winkler and Joerg Reichardt. "Node-Specific Triad Pattern Mining: A Novel Tool for Complex-Network Analysis". In: *Data Mining in Networks*. 2014, pp. 605–612.

[106] Elisabeth Wong et al. "Biological network motif detection: principles and practice". In: *Briefings in Bioinformatics* 13.2 (2011), pp. 202–215.

[107] Katharina A. Zweig. "How to Forget the Second Side of the Story: A New Method for the One-Mode Projection of Bipartite Graphs". In: *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining ASONAM 2010*. 2010, pp. 200–207.

[108] Katharina Anna Zweig and Michael Kaufmann. "A systematic approach to the one-mode projection of bipartite graphs". In: *Social Netw. Analys. Mining* 1.3 (2011), pp. 187–218.

Part VI

APPENDIX

COMPARISON WITH THE RESULTS OF ITZKOVITZ ET AL.

The comparison of our own results with the results of other groups is one of the most important parts of research, such that it is not skipped in this work.

The equations provided by Itzkovitz et al. [43] are in some aspect very similar to the approach developed in the thesis (see Section 12.1). Still, there are some major differences between their and the work presented here.

For the equations that are used throughout the thesis, only the in- and out-degree sequences are needed. Itzkovitz et al. use more information about a graph. They require

> $\{K_i\}_{i=1}^n$ (the number of edges outgoing from each node), $\{R_i\}_{i=1}^n$ (the number of incoming edges at each node), and mutual degree $\{M_i\}_{i=1}^n$ (the number of mutual edges at each node).

Getting information about mutual edges requires additional effort, but it is not too complex (max. $\mathcal{O}\left(n^2\right)$) compared to other measures.

In their paper, they used partially the same subgraphs as we did in our research, but they used different graphs. Searching for them online did yield only non-existing web-pages; comparing their results to some graphs that have the same name, but are not necessarily the same, is not useful. Instead, we apply their equations as well as our equations to several data sets that were analyzed before. For those subgraphs where the sampling process was used the results are displayed as well.

In Table 1, the equations given by Itzkovitz et al. and the equations developed by me are shown. Interestingly, some equations coincide. Still, since Itzkovitz et al. have to keep track of in-, out-, and mutual degree, even the results of these base equations will be different.

In the Tables 2 to 4 the results are displayed. In the first table, the equations are applied to the E. coli graph. From these results, the results of Itzkovitz et al. seem to coincide in almost all cases. Only some entries are not defined, due to the fact that M is 0 for all nodes. The next table analyzes the S. cerevisiae graph; here, the results get curious. There are several subgraphs for which the equations of Itzokvitz et al. predict negative values; the other set of equations yields small values, which can be considered as a sign that it is very implausible that this subgraph would appear anyway. Thus, negative values can be ignored up to now. Stranger still, there are two subgraphs for which our equations predict almost no occurrence, while the equations of Itzkovitz et al. predict that there are several of these subgraphs. The analysis of the Little Rock graph is the last pointer that the equations of Itzkovitz et al. are not to be used. While the analysis with the FDSM yielded an expected occurrence of ~ 5 complete subgraphs, the SIM yields an estimate of 1.64; the equation of Itzkovitz et al. yields 100 as many. Other expected values are misestimated as well. Consider the result of the equation for the Double-Join, for which the equation of Itzkovitz et al. estimates a much lower value than our equation. Our equation was very close to the result of the sampling process. Thus, our equations seem to yield better results than the ones of Itzkovitz et al. Additionally, we have a way to calculate standard deviations, which is not discussed in their paper.

Subgraph	Equation [43]	Equation$_{\text{SIM}}$
	$n\frac{\langle K(K-1)\rangle}{2}$	$n\frac{\langle k^2\rangle-\langle k\rangle}{2}$
	$n\langle KR\rangle$	$n\langle kj\rangle$
	$n\langle KM\rangle$	$\langle kj\rangle\frac{\langle k^2j\rangle-\langle kj\rangle}{\langle k\rangle^2}$
	$n\frac{\langle R(R-1)\rangle}{2}$	$n\frac{\langle j^2\rangle-\langle j\rangle}{2}$
	$\frac{\langle K(K-1)\rangle\langle KR\rangle\langle R(R-1)\rangle}{\langle K\rangle^3}$	$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\langle kj\rangle\left(\langle j^2\rangle-\langle j\rangle\right)}{\langle k\rangle^3}$
	$\frac{\langle KM\rangle^2\langle R(R-1)\rangle}{2\langle K\rangle^2\langle M\rangle}$	$\frac{\left(\langle j^2\rangle-\langle j\rangle\right)\left(\langle k^2j\rangle-\langle kj\rangle\right)^2}{2m\langle k\rangle^3}$
	$n\langle RM\rangle$	$\frac{\langle kj\rangle\left(\langle kj^2\rangle-\langle kj\rangle\right)}{\langle k\rangle^2}$
	$n\frac{\langle M(M-1)\rangle}{2}$	$\frac{\langle kj\rangle^2\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)}{2m\langle k\rangle^3}$
	$\frac{\langle KR\rangle^3}{3\langle K\rangle^3}$	$\frac{\langle kj\rangle^3}{3\langle k\rangle}$
	$\frac{\langle KM\rangle\langle RM\rangle\langle RK\rangle}{\langle K\rangle^2\langle M\rangle}$	$\frac{\langle kj\rangle\left(\langle k^2j\rangle-\langle kj\rangle\right)\left(\langle kj^2\rangle-\langle kj\rangle\right)}{m*\langle k\rangle^3}$
	$\frac{\langle RM\rangle^2\langle K(K-1)\rangle}{2\langle K\rangle^2\langle M\rangle}$	$\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle kj^2\rangle-\langle kj\rangle\right)^2}{2m\langle k\rangle^3}$
	$\frac{\langle KM\rangle\langle RM\rangle\langle M(M-1)\rangle}{\langle K\rangle\langle M\rangle^2}$	$\frac{\left(\langle k^2j\rangle-\langle kj\rangle\right)\left(\langle kj^2\rangle-\langle kj\rangle\right)\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)}{m^2\langle k\rangle^3}$
	$\frac{\langle M(M-1)\rangle^3}{6\langle M\rangle^3}$	$\frac{\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)^3}{6m^3\langle k\rangle^3}$

Table 1: Equations of Itzkovitz et al. [43] and own equations 11.18 together with the corresponding subgraph. Observe, that some do have the same equation.

Subgraph	Itzkovitz	Own
	4819.00	4819.00
	202.00	202.00
	0.00	1.11
	269.00	269.00
	7.49	7.49
	-	0.01
	0.00	0.17
	0.00	$4.69 \cdot 10^{-3}$
	0.02	0.02
	-	$9.60 \cdot 10^{-3}$
	-	$3.57 \cdot 10^{-2}$
	-	$1.53 \cdot 10^{-5}$
	-	$3.96 \cdot 10^{-8}$

Table 2: E. coli

Subgraph	Itzkovitz	Own
	11 843.00	11 841.00
	325.00	327.00
	−4.00	1.01
	871.00	873.00
	11.58	11.61
	$3.16 \cdot 10^{-2}$	$8.37 \cdot 10^{-2}$
	4.00	0.14
	2.00	$2.23 \cdot 10^{-3}$
	0.01	0.01
	$-1.18 \cdot 10^{-2}$	$4.39 \cdot 10^{-3}$
	0.04	$2.23 \cdot 10^{-2}$
	$-3.81 \cdot 10^{-3}$	$6.51 \cdot 10^{-6}$
	0.17	$1.63 \cdot 10^{-8}$

Table 3: S. cerevisiae

Subgraph	Itzkovitz	Own
	24 386.00	25 532.00
	25 674.00	28 597.00
	1832.00	2007.03
	52 878.00	54 553.00
	9675.77	10 270.57
	152.05	268.71
	2890.00	4781.86
	1022.00	140.83
	412.17	502.52
	116.46	335.61
	174.50	713.90
	108.71	62.70
	167.65	1.64

Table 4: Little Rock

1 ON THE CONFIGURATION MODEL

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$1.00 \cdot 10^{1} \pm 4.07$	$1.02 \cdot 10^{1} \pm 5.78$	6.03 ± 3.56
Silwood	9.13 ± 5.82	7.78 ± 6.75	4.63 ± 3.83
St. Martin Island	$1.06 \cdot 10^{1} \pm 4.22$	9.90 ± 5.70	5.11 ± 2.99
Ythan Estuary	$3.99 \cdot 10^{1} \pm 1.15 \cdot 10^{1}$	$5.34 \cdot 10^{1} \pm 2.72 \cdot 10^{1}$	$1.90 \cdot 10^{1} \pm 7.20$
Little Rock	$2.81 \cdot 10^{2} \pm 3.80 \cdot 10^{1}$	$2.60 \cdot 10^{2} \pm 4.38 \cdot 10^{1}$	$1.40 \cdot 10^{2} \pm 2.28 \cdot 10^{1}$
Grassland	$1.25 \cdot 10^{-1} \pm 3.73 \cdot 10^{-1}$	$2.05 \cdot 10^{-1} \pm 7.83 \cdot 10^{-1}$	$1.20 \cdot 10^{-1} \pm 3.94 \cdot 10^{-1}$
s208	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	0.00 ± 0.00
s420	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	$2 \cdot 10^{-2} \pm 1.40 \cdot 10^{-1}$
Gnutella 08.08.2002	$5.35 \cdot 10^{-1} \pm 7.34 \cdot 10^{-1}$	$3.91 \cdot 10^{2} \pm 1.59 \cdot 10^{2}$	$4.90 \cdot 10^{-1} \pm 7.35 \cdot 10^{-1}$
Gnutella 09.08.2002	$4.85 \cdot 10^{-1} \pm 7.55 \cdot 10^{-1}$	$1.80 \cdot 10^{2} \pm 1.41 \cdot 10^{2}$	$3.30 \cdot 10^{-1} \pm 6.33 \cdot 10^{-1}$
E .coli	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$
S. cerevisiae	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$

Table 1: Number of Double-Joins found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$1 \cdot 10^{-1}$	$4 \cdot 10^{-1}$	$3.35 \cdot 10^{-1}$
St. Martin Island	$1.65 \cdot 10^{-1}$	$5.45 \cdot 10^{-1}$	$4.05 \cdot 10^{-1}$
Silwood	$1.55 \cdot 10^{-1}$	$3.75 \cdot 10^{-1}$	$2.85 \cdot 10^{-1}$
Ythan Estuary	$2.85 \cdot 10^{-1}$	$7.35 \cdot 10^{-1}$	$7.65 \cdot 10^{-1}$
Little Rock	$2.80 \cdot 10^{-1}$	$9.75 \cdot 10^{-1}$	$9.30 \cdot 10^{-1}$
Grassland	$3.50 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$	$2.50 \cdot 10^{-2}$
s208	0.00	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
s420	$2.50 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$	$1 \cdot 10^{-2}$
Gnutella 08.08.2002	1.00	$3.16 \cdot 10^{-2}$	1.00
Gnutella 09.08.2002	$7.50 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$7.50 \cdot 10^{-1}$
E .coli	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	0.00
S. cerevisiae	0.00	0.00	0.00

Table 2: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Double-Join.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$2.20 \cdot 10^{1} \pm 4.52$	$2.25 \cdot 10^{1} \pm 5.38$	$1.48 \cdot 10^{1} \pm 4.03$
Silwood	4.48 ± 2.51	6.44 ± 3.81	2.94 ± 1.94
St. Martin Island	$1.92 \cdot 10^{1} \pm 5.30$	$2.08 \cdot 10^{1} \pm 6.33$	$1.25 \cdot 10^{1} \pm 4.12$
Ythan Estuary	$5.63 \cdot 10^{1} \pm 1.04 \cdot 10^{1}$	$7.14 \cdot 10^{1} \pm 2.17 \cdot 10^{1}$	$3.21 \cdot 10^{1} \pm 7.05$
Little Rock	$4.68 \cdot 10^{2} \pm 3.11 \cdot 10^{1}$	$5.08 \cdot 10^{2} \pm 3.87 \cdot 10^{1}$	$3.49 \cdot 10^{2} \pm 2.75 \cdot 10^{1}$
Grassland	$3.60 \cdot 10^{-1} \pm 5.92 \cdot 10^{-1}$	$7.43 \cdot 10^{-1} \pm 7.10 \cdot 10^{-1}$	$4 \cdot 10^{-1} \pm 6 \cdot 10^{-1}$
s208	1.04 ± 1.02	$1.30 \pm 9.64 \cdot 10^{-1}$	$8.20 \cdot 10^{-1} \pm 8.59 \cdot 10^{-1}$
s420	1.05 ± 1.04	1.65 ± 1.11	1.11 ± 1.01
Gnutella 08.08.2002	$3.02 \cdot 10^{1} \pm 5.52$	$9.20 \cdot 10^{1} \pm 1.51 \cdot 10^{1}$	$3.06 \cdot 10^{1} \pm 5.72$
Gnutella 09.08.2002	$2.45 \cdot 10^{1} \pm 5.69$	$7.55 \cdot 10^{1} \pm 1.59 \cdot 10^{1}$	$2.42 \cdot 10^{1} \pm 5.05$
E .coli	$1.50 \cdot 10^{-2} \pm 1.22 \cdot 10^{-1}$	$1.22 \cdot 10^{-1} \pm 2.06 \cdot 10^{-1}$	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$
S. cerevisiae	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	$1.02 \cdot 10^{-1} \pm 1.86 \cdot 10^{-1}$	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$

Table 3: Number of Threecycles found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$7.50 \cdot 10^{-2}$	$5.95 \cdot 10^{-1}$	$6 \cdot 10^{-1}$
Silwood	$2.85 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$	$4.95 \cdot 10^{-1}$
St. Martin Island	$2 \cdot 10^{-1}$	$5.15 \cdot 10^{-1}$	$6.05 \cdot 10^{-1}$
Ythan Estuary	$3.95 \cdot 10^{-1}$	$8.40 \cdot 10^{-1}$	$9.05 \cdot 10^{-1}$
Little Rock	$4.45 \cdot 10^{-1}$	$9.60 \cdot 10^{-1}$	$9.90 \cdot 10^{-1}$
Grassland	$4.50 \cdot 10^{-1}$	$3.50 \cdot 10^{-2}$	$4.15 \cdot 10^{-1}$
s208	$2.58 \cdot 10^{-1}$	$7.75 \cdot 10^{-2}$	$3.30 \cdot 10^{-1}$
s420	$3.35 \cdot 10^{-1}$	$5.75 \cdot 10^{-2}$	$3.45 \cdot 10^{-1}$
Gnutella 08.08.2002	1.00	$4.84 \cdot 10^{-2}$	1.00
Gnutella 09.08.2002	1.00	$6.50 \cdot 10^{-2}$	1.00
E .coli	$2.95 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$3.05 \cdot 10^{-1}$
S. cerevisiae	$2.55 \cdot 10^{-1}$	$5 \cdot 10^{-3}$	$2.60 \cdot 10^{-1}$

Table 4: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Threecycles.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$3.50 \cdot 10^{-2} \pm 1.84 \cdot 10^{-1}$	$2.05 \cdot 10^{1} \pm 5.44 \cdot 10^{1}$	0.00 ± 0.00
Silwood	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	$1.25 \cdot 10^{1} \pm 2.99 \cdot 10^{1}$	0.00 ± 0.00
St. Martin Island	$6.50 \cdot 10^{-2} \pm 2.47 \cdot 10^{-1}$	$1.55 \cdot 10^{1} \pm 3.12 \cdot 10^{1}$	0.00 ± 0.00
Ythan Estuary	$6.15 \cdot 10^{-1} \pm 8.17 \cdot 10^{-1}$	$6.39 \cdot 10^{2} \pm 2.58 \cdot 10^{3}$	$4.50 \cdot 10^{-1} \pm 1.15$
Little Rock	5.75 ± 4.79	$1.04 \cdot 10^{2} \pm 1.01 \cdot 10^{2}$	2.27 ± 2.72
Grassland	0.00 ± 0.00	$5.03 \cdot 10^{-1} \pm 5.09 \cdot 10^{-1}$	0.00 ± 0.00
s208	0.00 ± 0.00	$7.20 \cdot 10^{-1} \pm 5.60 \cdot 10^{-1}$	0.00 ± 0.00
s420	0.00 ± 0.00	$9.25 \cdot 10^{-1} \pm 2.32$	0.00 ± 0.00
Gnutella 08.08.2002	0.00 ± 0.00	1.64 ± 3.15	0.00 ± 0.00
Gnutella 09.08.2002	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
E .coli	0.00 ± 0.00	$1.75 \cdot 10^{-1} \pm 2.95 \cdot 10^{-1}$	0.00 ± 0.00
S. cerevisiae	0.00 ± 0.00	$1.45 \cdot 10^{-1} \pm 2.63 \cdot 10^{-1}$	0.00 ± 0.00

Table 5: Number of Complete subgraphs found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$9.55 \cdot 10^{-1}$	$3.50 \cdot 10^{-2}$	$9.90 \cdot 10^{-1}$
Silwood	$8.90 \cdot 10^{-1}$	$3 \cdot 10^{-2}$	$9.20 \cdot 10^{-1}$
St. Martin Island	$9.30 \cdot 10^{-1}$	$6.50 \cdot 10^{-2}$	$9.95 \cdot 10^{-1}$
Ythan Estuary	$9.70 \cdot 10^{-1}$	$2.95 \cdot 10^{-1}$	$9.60 \cdot 10^{-1}$
Little Rock	$9.70 \cdot 10^{-1}$	$3.25 \cdot 10^{-1}$	$9.90 \cdot 10^{-1}$
Grassland	$6.35 \cdot 10^{-1}$	0.00	$6.35 \cdot 10^{-1}$
s208	$8.10 \cdot 10^{-1}$	0.00	$8.10 \cdot 10^{-1}$
s420	$8.05 \cdot 10^{-1}$	0.00	$8.05 \cdot 10^{-1}$
Gnutella 08.08.2002	$2.14 \cdot 10^{-1}$	0.00	$2.14 \cdot 10^{-1}$
Gnutella 09.08.2002	0.00	0.00	0.00
E .coli	$3.05 \cdot 10^{-1}$	0.00	$3.05 \cdot 10^{-1}$
S. cerevisiae	$2.60 \cdot 10^{-1}$	0.00	$2.60 \cdot 10^{-1}$

Table 6: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Complete subgraph.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$6.01 \cdot 10^{1} \pm 1.26 \cdot 10^{1}$	$7.05 \cdot 10^{1} \pm 1.66 \cdot 10^{1}$	$4.32 \cdot 10^{1} \pm 1.06 \cdot 10^{1}$
Silwood	6.25 ± 4.07	$1.35 \cdot 10^{1} \pm 9.45$	5.16 ± 3.66
St. Martin Island	$4.85 \cdot 10^{1} \pm 1.48 \cdot 10^{1}$	$5.99 \cdot 10^{1} \pm 2.13 \cdot 10^{1}$	$3.17 \cdot 10^{1} \pm 1.16 \cdot 10^{1}$
Ythan Estuary	$1.95 \cdot 10^{2} \pm 4.52 \cdot 10^{1}$	$3.29 \cdot 10^{2} \pm 1.27 \cdot 10^{2}$	$9.64 \cdot 10^{1} \pm 2.33 \cdot 10^{1}$
Little Rock	$3.79 \cdot 10^{3} \pm 3.06 \cdot 10^{2}$	$4.40 \cdot 10^{3} \pm 3.83 \cdot 10^{2}$	$2.58 \cdot 10^{3} \pm 2.32 \cdot 10^{2}$
Grassland	$2.40 \cdot 10^{-1} \pm 4.72 \cdot 10^{-1}$	$7.84 \cdot 10^{-1} \pm 6.69 \cdot 10^{-1}$	$5.23 \cdot 10^{-1} \pm 6.07 \cdot 10^{-1}$
s208	$1.02 \pm 9.87 \cdot 10^{-1}$	1.86 ± 1.15	1.46 ± 1.04
s420	1.22 ± 1.15	2.13 ± 1.31	1.71 ± 1.23
Gnutella 08.08.2002	$1.06 \cdot 10^{2} \pm 1.29 \cdot 10^{1}$	$4.43 \cdot 10^{2} \pm 3.75 \cdot 10^{1}$	$1.03 \cdot 10^{2} \pm 1.35 \cdot 10^{1}$
Gnutella 09.08.2002	$7.58 \cdot 10^{1} \pm 9.68$	$3.38 \cdot 10^{2} \pm 4.44 \cdot 10^{1}$	$7.52 \cdot 10^{1} \pm 1.06 \cdot 10^{1}$
E .coli	0.00 ± 0.00	$1.35 \cdot 10^{-1} \pm 2.23 \cdot 10^{-1}$	$4.75 \cdot 10^{-2} \pm 1.55 \cdot 10^{-1}$
S. cerevisiae	0.00 ± 0.00	$1.15 \cdot 10^{-1} \pm 2.04 \cdot 10^{-1}$	$4.25 \cdot 10^{-2} \pm 1.56 \cdot 10^{-1}$

Table 7: Number of Fourcycles found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$2.95 \cdot 10^{-1}$	$5.55 \cdot 10^{-1}$	$6.85 \cdot 10^{-1}$
Silwood	$4.25 \cdot 10^{-1}$	$2.10 \cdot 10^{-1}$	$5.70 \cdot 10^{-1}$
St. Martin Island	$2.40 \cdot 10^{-1}$	$5.35 \cdot 10^{-1}$	$6.35 \cdot 10^{-1}$
Ythan Estuary	$6.45 \cdot 10^{-1}$	$8.70 \cdot 10^{-1}$	$9.60 \cdot 10^{-1}$
Little Rock	$6.45 \cdot 10^{-1}$	$9.80 \cdot 10^{-1}$	1.00
Grassland	$6.55 \cdot 10^{-1}$	$3.40 \cdot 10^{-1}$	$3.15 \cdot 10^{-1}$
s208	$4.43 \cdot 10^{-1}$	$2.73 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
s420	$4.03 \cdot 10^{-1}$	$2.68 \cdot 10^{-1}$	$1.85 \cdot 10^{-1}$
Gnutella 08.08.2002	1.00	$1.21 \cdot 10^{-1}$	1.00
Gnutella 09.08.2002	1.00	$5 \cdot 10^{-2}$	1.00
E .coli	$3.50 \cdot 10^{-1}$	$9 \cdot 10^{-2}$	$2.60 \cdot 10^{-1}$
S. cerevisiae	$3.15 \cdot 10^{-1}$	$7.50 \cdot 10^{-2}$	$2.40 \cdot 10^{-1}$

Table 8: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fourcycle.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$4.60 \cdot 10^2 \pm 3.21 \cdot 10^1$	$4.22 \cdot 10^2 \pm 4.00 \cdot 10^1$	$2.51 \cdot 10^2 \pm 3.76 \cdot 10^1$
Silwood	$1.03 \cdot 10^3 \pm 9.35 \cdot 10^1$	$1.05 \cdot 10^3 \pm 1.15 \cdot 10^2$	$5.32 \cdot 10^2 \pm 7.67 \cdot 10^1$
St. Martin Island	$8.71 \cdot 10^2 \pm 4.78 \cdot 10^1$	$7.83 \cdot 10^2 \pm 7.13 \cdot 10^1$	$3.75 \cdot 10^2 \pm 5.55 \cdot 10^1$
Ythan Estuary	$3.22 \cdot 10^3 \pm 1.77 \cdot 10^2$	$4.19 \cdot 10^3 \pm 4.27 \cdot 10^2$	$1.59 \cdot 10^3 \pm 1.48 \cdot 10^2$
Little Rock	$1.67 \cdot 10^5 \pm 2.70 \cdot 10^3$	$1.89 \cdot 10^5 \pm 4.88 \cdot 10^3$	$6.62 \cdot 10^4 \pm 2.53 \cdot 10^3$
Grassland	$1.43 \cdot 10^1 \pm 6.22$	$1.90 \cdot 10^1 \pm 1.11 \cdot 10^1$	$1.02 \cdot 10^1 \pm 5.12$
s208	$6.48 \cdot 10^{-1} \pm 8.96 \cdot 10^{-1}$	$5.55 \cdot 10^{-1} \pm 8.23 \cdot 10^{-1}$	$5.25 \cdot 10^{-1} \pm 7.48 \cdot 10^{-1}$
s420	$7.73 \cdot 10^{-1} \pm 8.78 \cdot 10^{-1}$	$6.95 \cdot 10^{-1} \pm 8.61 \cdot 10^{-1}$	$6.95 \cdot 10^{-1} \pm 8.61 \cdot 10^{-1}$
Gnutella 08.08.2002	$4.61 \cdot 10^3 \pm 1.36 \cdot 10^2$	$4.65 \cdot 10^3 \pm 7.54 \cdot 10^1$	$4.42 \cdot 10^3 \pm 1.39 \cdot 10^2$
Gnutella 09.08.2002	$4.14 \cdot 10^3 \pm 1.20 \cdot 10^2$	$4.11 \cdot 10^3 \pm 8.92 \cdot 10^1$	$3.97 \cdot 10^3 \pm 1.23 \cdot 10^2$
E .coli	$6.48 \cdot 10^1 \pm 1.32 \cdot 10^1$	$6.55 \cdot 10^1 \pm 1.77 \cdot 10^1$	$5.40 \cdot 10^1 \pm 1.35 \cdot 10^1$
S. cerevisiae	$3.06 \cdot 10^2 \pm 3.50 \cdot 10^1$	$3.15 \cdot 10^2 \pm 4.24 \cdot 10^1$	$2.58 \cdot 10^2 \pm 3.07 \cdot 10^1$

Table 9: Number of Bifans found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$4.15 \cdot 10^{-1}$	1.00	$9.70 \cdot 10^{-1}$
Silwood	$1.15 \cdot 10^{-1}$	1.00	$9.95 \cdot 10^{-1}$
St. Martin Island	$5.70 \cdot 10^{-1}$	1.00	1.00
Ythan Estuary	$9.40 \cdot 10^{-1}$	1.00	1.00
Little Rock	1.00	1.00	1.00
Grassland	$2.45 \cdot 10^{-1}$	$3.15 \cdot 10^{-1}$	$4.05 \cdot 10^{-1}$
s208	$4.25 \cdot 10^{-2}$	$5.25 \cdot 10^{-2}$	$1 \cdot 10^{-2}$
s420	$5.25 \cdot 10^{-2}$	$5.25 \cdot 10^{-2}$	0.00
Gnutella 08.08.2002	$3.02 \cdot 10^{-1}$	$5.72 \cdot 10^{-1}$	$8.04 \cdot 10^{-1}$
Gnutella 09.08.2002	$2.65 \cdot 10^{-1}$	$5.50 \cdot 10^{-1}$	$6.40 \cdot 10^{-1}$
E .coli	$9.50 \cdot 10^{-2}$	$3.45 \cdot 10^{-1}$	$2.95 \cdot 10^{-1}$
S. cerevisiae	$1.10 \cdot 10^{-1}$	$5.40 \cdot 10^{-1}$	$5.45 \cdot 10^{-1}$

Table 10: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Bifan.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$3.44 \cdot 10^2 \pm 3.04 \cdot 10^1$	$3.39 \cdot 10^2 \pm 3.83 \cdot 10^1$	$2.04 \cdot 10^2 \pm 2.96 \cdot 10^1$
Silwood	$1.89 \cdot 10^2 \pm 3.29 \cdot 10^1$	$1.93 \cdot 10^2 \pm 3.88 \cdot 10^1$	$9.69 \cdot 10^1 \pm 2.32 \cdot 10^1$
St. Martin Island	$4.56 \cdot 10^2 \pm 4.52 \cdot 10^1$	$4.36 \cdot 10^2 \pm 5.73 \cdot 10^1$	$2.20 \cdot 10^2 \pm 3.32 \cdot 10^1$
Ythan Estuary	$1.73 \cdot 10^3 \pm 1.20 \cdot 10^2$	$2.18 \cdot 10^3 \pm 3.17 \cdot 10^2$	$8.04 \cdot 10^2 \pm 9.33 \cdot 10^1$
Little Rock	$6.74 \cdot 10^4 \pm 2.09 \cdot 10^3$	$5.78 \cdot 10^4 \pm 2.51 \cdot 10^3$	$2.80 \cdot 10^4 \pm 1.32 \cdot 10^3$
Grassland	7.35 ± 3.05	6.42 ± 3.86	4.81 ± 2.80
s208	1.67 ± 1.21	1.71 ± 1.22	1.59 ± 1.16
s420	2.00 ± 1.38	2.09 ± 1.47	2.02 ± 1.44
Gnutella 08.08.2002	$1.41 \cdot 10^3 \pm 7.75 \cdot 10^1$	$1.36 \cdot 10^4 \pm 1.14 \cdot 10^3$	$1.35 \cdot 10^3 \pm 7.88 \cdot 10^1$
Gnutella 09.08.2002	$1.13 \cdot 10^3 \pm 6.35 \cdot 10^1$	$1.24 \cdot 10^4 \pm 4.08 \cdot 10^2$	$1.10 \cdot 10^3 \pm 6.33 \cdot 10^1$
E .coli	1.54 ± 1.49	1.27 ± 1.41	1.06 ± 1.08
S. cerevisiae	1.78 ± 1.70	1.59 ± 1.60	1.42 ± 1.37

Table 11: Number of Biparallel subgraphs found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$1.30 \cdot 10^{-1}$	$9.80 \cdot 10^{-1}$	$9.65 \cdot 10^{-1}$
Silwood	$9.50 \cdot 10^{-2}$	$9.25 \cdot 10^{-1}$	$8.85 \cdot 10^{-1}$
St. Martin Island	$2.20 \cdot 10^{-1}$	1.00	$9.95 \cdot 10^{-1}$
Ythan Estuary	$7.15 \cdot 10^{-1}$	1.00	1.00
Little Rock	$9.85 \cdot 10^{-1}$	1.00	1.00
Grassland	$1.85 \cdot 10^{-1}$	$3.50 \cdot 10^{-1}$	$2.25 \cdot 10^{-1}$
s208	$3 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
s420	$3.25 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$2.50 \cdot 10^{-2}$
Gnutella 08.08.2002	1.00	$3.33 \cdot 10^{-1}$	1.00
Gnutella 09.08.2002	1.00	$2.65 \cdot 10^{-1}$	1.00
E .coli	$9 \cdot 10^{-2}$	$1.25 \cdot 10^{-1}$	$5.50 \cdot 10^{-2}$
S. cerevisiae	$9.50 \cdot 10^{-2}$	$1.10 \cdot 10^{-1}$	$4 \cdot 10^{-2}$

Table 12: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Biparallel subgraph.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$9.67 \cdot 10^{1} \pm 1.76 \cdot 10^{1}$	$9.71 \cdot 10^{1} \pm 2.60 \cdot 10^{1}$	$4.36 \cdot 10^{1} \pm 1.17 \cdot 10^{1}$
Silwood	$6.98 \cdot 10^{1} \pm 1.91 \cdot 10^{1}$	$6.87 \cdot 10^{1} \pm 2.42 \cdot 10^{1}$	$2.74 \cdot 10^{1} \pm 9.86$
St. Martin Island	$1.97 \cdot 10^{2} \pm 2.81 \cdot 10^{1}$	$2.05 \cdot 10^{2} \pm 6.58 \cdot 10^{1}$	$6.52 \cdot 10^{1} \pm 1.74 \cdot 10^{1}$
Ythan Estuary	$7.73 \cdot 10^{2} \pm 9.85 \cdot 10^{1}$	$1.65 \cdot 10^{3} \pm 5.83 \cdot 10^{2}$	$2.54 \cdot 10^{2} \pm 5.03 \cdot 10^{1}$
Little Rock	$3.92 \cdot 10^{4} \pm 1.63 \cdot 10^{3}$	$3.72 \cdot 10^{4} \pm 3.01 \cdot 10^{3}$	$1.10 \cdot 10^{4} \pm 7.71 \cdot 10^{2}$
Grassland	2.90 ± 2.23	2.92 ± 3.77	1.69 ± 1.69
s208	$1.50 \cdot 10^{-2} \pm 1.22 \cdot 10^{-1}$	$4.50 \cdot 10^{-2} \pm 2.07 \cdot 10^{-1}$	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$
s420	$2.50 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	$1.50 \cdot 10^{-2} \pm 1.22 \cdot 10^{-1}$	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$
Gnutella 08.08.2002	$3.60 \cdot 10^{1} \pm 7.50$	$3.56 \cdot 10^{1} \pm 9.35$	$3.39 \cdot 10^{1} \pm 7.71$
Gnutella 09.08.2002	$2.72 \cdot 10^{1} \pm 7.08$	$2.43 \cdot 10^{1} \pm 4.76$	$2.49 \cdot 10^{1} \pm 6.09$
E .coli	$2.25 \cdot 10^{-1} \pm 4.74 \cdot 10^{-1}$	$1.75 \cdot 10^{-1} \pm 6.51 \cdot 10^{-1}$	$1.05 \cdot 10^{-1} \pm 3.66 \cdot 10^{-1}$
S. cerevisiae	$1.05 \cdot 10^{-1} \pm 3.07 \cdot 10^{-1}$	$1.45 \cdot 10^{-1} \pm 4.73 \cdot 10^{-1}$	$1.15 \cdot 10^{-1} \pm 3.34 \cdot 10^{-1}$

Table 13: Number of In-Fans found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$1.05 \cdot 10^{-1}$	$9.50 \cdot 10^{-1}$	$8.75 \cdot 10^{-1}$
Silwood	$1 \cdot 10^{-1}$	$8.65 \cdot 10^{-1}$	$7.95 \cdot 10^{-1}$
St. Martin Island	$2.10 \cdot 10^{-1}$	1.00	$9.60 \cdot 10^{-1}$
Ythan Estuary	$8.80 \cdot 10^{-1}$	1.00	1.00
Little Rock	$4.15 \cdot 10^{-1}$	1.00	1.00
Grassland	$1.75 \cdot 10^{-1}$	$2.90 \cdot 10^{-1}$	$1.40 \cdot 10^{-1}$
s208	$3 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$
s420	$7.50 \cdot 10^{-3}$	$1.25 \cdot 10^{-2}$	$5 \cdot 10^{-3}$
Gnutella 08.08.2002	$1.87 \cdot 10^{-1}$	$1.26 \cdot 10^{-1}$	$1.89 \cdot 10^{-1}$
Gnutella 09.08.2002	$4.35 \cdot 10^{-1}$	$1.45 \cdot 10^{-1}$	$3.65 \cdot 10^{-1}$
E .coli	$1.10 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$	$3.50 \cdot 10^{-2}$
S. cerevisiae	$2.50 \cdot 10^{-2}$	$5 \cdot 10^{-3}$	$2 \cdot 10^{-2}$

Table 14: Results of the Kolmogorov-Smirnov two-sample test between the different models for the In-Fan.

Graph	FDSM	CFG	eCFG
St. Marks Seagrass	$1.36 \cdot 10^{2} \pm 2.26 \cdot 10^{1}$	$3.12 \cdot 10^{2} \pm 1.35 \cdot 10^{2}$	$6.28 \cdot 10^{1} \pm 1.57 \cdot 10^{1}$
Silwood	$3.23 \cdot 10^{2} \pm 6.55 \cdot 10^{1}$	$1.04 \cdot 10^{3} \pm 5.59 \cdot 10^{2}$	$1.36 \cdot 10^{2} \pm 3.98 \cdot 10^{1}$
St. Martin Island	$2.00 \cdot 10^{2} \pm 2.43 \cdot 10^{1}$	$3.18 \cdot 10^{2} \pm 1.02 \cdot 10^{2}$	$7.08 \cdot 10^{1} \pm 1.75 \cdot 10^{1}$
Ythan Estuary	$1.11 \cdot 10^{3} \pm 1.32 \cdot 10^{2}$	$3.86 \cdot 10^{3} \pm 1.50 \cdot 10^{3}$	$4.35 \cdot 10^{2} \pm 7.31 \cdot 10^{1}$
Little Rock	$3.29 \cdot 10^{4} \pm 1.04 \cdot 10^{3}$	$3.86 \cdot 10^{4} \pm 2.54 \cdot 10^{3}$	$1.02 \cdot 10^{4} \pm 5.76 \cdot 10^{2}$
Grassland	1.08 ± 1.14	$1.37 \cdot 10^{1} \pm 1.94 \cdot 10^{1}$	$9.25 \cdot 10^{-1} \pm 1.20$
s208	$2 \cdot 10^{-2} \pm 1.57 \cdot 10^{-1}$	3.94 ± 5.94	$3.50 \cdot 10^{-2} \pm 1.84 \cdot 10^{-1}$
s420	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	4.49 ± 6.32	$3.50 \cdot 10^{-2} \pm 1.84 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.85 \cdot 10^{1} \pm 4.49$	$2.10 \cdot 10^{2} \pm 8.31 \cdot 10^{1}$	$1.83 \cdot 10^{1} \pm 4.59$
Gnutella 09.08.2002	$1.42 \cdot 10^{1} \pm 4.18$	$1.03 \cdot 10^{2} \pm 7.27 \cdot 10^{1}$	$1.39 \cdot 10^{1} \pm 3.96$
E .coli	3.00 ± 3.56	$1.94 \cdot 10^{1} \pm 4.84 \cdot 10^{1}$	1.51 ± 2.08
S. cerevisiae	3.47 ± 3.28	$2.91 \cdot 10^{1} \pm 6.08 \cdot 10^{1}$	3.07 ± 2.84

Table 15: Number of Out-Fans found in the respective models.

Graph	FDSM - CFG	FSDM - eCFG	CFG - eCFG
St. Marks Seagrass	$8.45 \cdot 10^{-1}$	$9.50 \cdot 10^{-1}$	1.00
Silwood	$8.50 \cdot 10^{-1}$	$9.25 \cdot 10^{-1}$	$9.80 \cdot 10^{-1}$
St. Martin Island	$6.85 \cdot 10^{-1}$	1.00	1.00
Ythan Estuary	$9.85 \cdot 10^{-1}$	1.00	1.00
Little Rock	$9.35 \cdot 10^{-1}$	1.00	1.00
Grassland	$4.90 \cdot 10^{-1}$	$8 \cdot 10^{-2}$	$5.10 \cdot 10^{-1}$
s208	$4.93 \cdot 10^{-1}$	$1.75 \cdot 10^{-2}$	$4.75 \cdot 10^{-1}$
s420	$5.25 \cdot 10^{-1}$	$5 \cdot 10^{-3}$	$5.20 \cdot 10^{-1}$
Gnutella 08.08.2002	1.00	$6.15 \cdot 10^{-2}$	1.00
Gnutella 09.08.2002	$7.50 \cdot 10^{-1}$	$6 \cdot 10^{-2}$	$7.50 \cdot 10^{-1}$
E .coli	$2.45 \cdot 10^{-1}$	$2.10 \cdot 10^{-1}$	$3.05 \cdot 10^{-1}$
S. cerevisiae	$2.45 \cdot 10^{-1}$	$7.50 \cdot 10^{-2}$	$2.55 \cdot 10^{-1}$

Table 16: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Out-Fan.

2 ON THE SEQUENTIAL IMPORTANCE SAMPLING MODEL

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$1.00 \cdot 10^{1} \pm 4.07$	$1.01 \cdot 10^{1} \pm 3.17$	8.54 ± 3.93
Silwood	9.13 ± 5.82	6.74 ± 5.62	7.68 ± 5.59
St. Martin Island	$1.06 \cdot 10^{1} \pm 4.22$	$1.03 \cdot 10^{1} \pm 4.05$	7.47 ± 4.25
Ythan Estuary	$3.99 \cdot 10^{1} \pm 1.15 \cdot 10^{1}$	$3.65 \cdot 10^{1} \pm 7.29$	$3.42 \cdot 10^{1} \pm 9.41$
Little Rock	$2.81 \cdot 10^{2} \pm 3.80 \cdot 10^{1}$	$2.79 \cdot 10^{2} \pm 3.41 \cdot 10^{1}$	$2.81 \cdot 10^{2} \pm 4.56 \cdot 10^{1}$
Grassland	$1.25 \cdot 10^{-1} \pm 3.73 \cdot 10^{-1}$	$2 \cdot 10^{-2} \pm 1.40 \cdot 10^{-1}$	$9 \cdot 10^{-2} \pm 2.86 \cdot 10^{-1}$
s208	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	0.00 ± 0.00	$6 \cdot 10^{-2} \pm 2.37 \cdot 10^{-1}$
s420	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	0.00 ± 0.00	0.00 ± 0.00
Gnutella 08.08.2002	$5.35 \cdot 10^{-1} \pm 7.34 \cdot 10^{-1}$	$5.20 \cdot 10^{-1} \pm 7.55 \cdot 10^{-1}$	$6.40 \cdot 10^{-1} \pm 7.94 \cdot 10^{-1}$
Gnutella 09.08.2002	$4.85 \cdot 10^{-1} \pm 7.55 \cdot 10^{-1}$	$5.20 \cdot 10^{-1} \pm 8.42 \cdot 10^{-1}$	$4.20 \cdot 10^{-1} \pm 6.19 \cdot 10^{-1}$
E .coli	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	0.00 ± 0.00	$5 \cdot 10^{-2} \pm 2.18 \cdot 10^{-1}$
S. cerevisiae	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	0.00 ± 0.00	0.00 ± 0.00

Table 17: Number of Double-Joins found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$1.05 \cdot 10^{-1}$	$1.95 \cdot 10^{-1}$	$3 \cdot 10^{-1}$
Silwood	$2.50 \cdot 10^{-1}$	$2.65 \cdot 10^{-1}$	$2.20 \cdot 10^{-1}$
St. Martin Island	$1.35 \cdot 10^{-1}$	$3.95 \cdot 10^{-1}$	$3.90 \cdot 10^{-1}$
Ythan Estuary	$2.70 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$	$2.30 \cdot 10^{-1}$
Little Rock	$1.45 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Grassland	$9 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$7 \cdot 10^{-2}$
s208	$5 \cdot 10^{-3}$	$5.50 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
s420	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	0.00
Gnutella 08.08.2002	$2.17 \cdot 10^{-2}$	$6.34 \cdot 10^{-2}$	$8 \cdot 10^{-2}$
Gnutella 09.08.2002	$1.50 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$4 \cdot 10^{-2}$
E .coli	$5 \cdot 10^{-3}$	$4.50 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
S. cerevisiae	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	0.00

Table 18: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Double-Join.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$2.20 \cdot 10^{1} \pm 4.52$	$2.34 \cdot 10^{1} \pm 4.87$	$1.88 \cdot 10^{1} \pm 5.38$
Silwood	4.48 ± 2.51	3.72 ± 2.52	2.55 ± 1.45
St. Martin Island	$1.92 \cdot 10^{1} \pm 5.30$	$2.09 \cdot 10^{1} \pm 4.86$	$2.12 \cdot 10^{1} \pm 6.37$
Ythan Estuary	$5.63 \cdot 10^{1} \pm 1.04 \cdot 10^{1}$	$4.68 \cdot 10^{1} \pm 1.05 \cdot 10^{1}$	$4.69 \cdot 10^{1} \pm 8.91$
Little Rock	$4.68 \cdot 10^{2} \pm 3.11 \cdot 10^{1}$	$5.35 \cdot 10^{2} \pm 2.97 \cdot 10^{1}$	$5.24 \cdot 10^{2} \pm 3.17 \cdot 10^{1}$
Grassland	$3.60 \cdot 10^{-1} \pm 5.92 \cdot 10^{-1}$	$5.40 \cdot 10^{-1} \pm 7.41 \cdot 10^{-1}$	$4.40 \cdot 10^{-1} \pm 7.66 \cdot 10^{-1}$
s208	1.04 ± 1.02	$8.40 \cdot 10^{-1} \pm 8.21 \cdot 10^{-1}$	$9.20 \cdot 10^{-1} \pm 9.66 \cdot 10^{-1}$
s420	1.05 ± 1.04	$1.17 \pm 8.61 \cdot 10^{-1}$	1.30 ± 1.09
Gnutella 08.08.2002	$3.02 \cdot 10^{1} \pm 5.52$	$3.13 \cdot 10^{1} \pm 5.09$	$3.07 \cdot 10^{1} \pm 5.50$
Gnutella 09.08.2002	$2.45 \cdot 10^{1} \pm 5.69$	$2.45 \cdot 10^{1} \pm 4.60$	$2.39 \cdot 10^{1} \pm 4.83$
E .coli	$1.50 \cdot 10^{-2} \pm 1.22 \cdot 10^{-1}$	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	0.00 ± 0.00
S. cerevisiae	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	0.00 ± 0.00	0.00 ± 0.00

Table 19: Number of Threecycles found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$2.15 \cdot 10^{-1}$	$3.65 \cdot 10^{-1}$	$4.70 \cdot 10^{-1}$
Silwood	$2.65 \cdot 10^{-1}$	$5.05 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$
St. Martin Island	$1.65 \cdot 10^{-1}$	$2.40 \cdot 10^{-1}$	$1.30 \cdot 10^{-1}$
Ythan Estuary	$4.50 \cdot 10^{-1}$	$5.10 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
Little Rock	$7.45 \cdot 10^{-1}$	$6.95 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$
Grassland	$1.10 \cdot 10^{-1}$	$1.30 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
s208	$1.58 \cdot 10^{-1}$	$9.75 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
s420	$1.78 \cdot 10^{-1}$	$1.18 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.20 \cdot 10^{-1}$	$8.27 \cdot 10^{-2}$	$8 \cdot 10^{-2}$
Gnutella 09.08.2002	$1.15 \cdot 10^{-1}$	$7.50 \cdot 10^{-2}$	$1 \cdot 10^{-1}$
E .coli	$1.50 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$	$3 \cdot 10^{-2}$
S. cerevisiae	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	0.00

Table 20: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Threecycle.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$3.50 \cdot 10^{-2} \pm 1.84 \cdot 10^{-1}$	$1.50 \cdot 10^{-1} \pm 6.54 \cdot 10^{-1}$	0.00 ± 0.00
Silwood	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	0.00 ± 0.00	0.00 ± 0.00
St. Martin Island	$6.50 \cdot 10^{-2} \pm 2.47 \cdot 10^{-1}$	0.00 ± 0.00	0.00 ± 0.00
Ythan Estuary	$6.15 \cdot 10^{-1} \pm 8.17 \cdot 10^{-1}$	$7.50 \cdot 10^{-1} \pm 1.30$	$8.10 \cdot 10^{-1} \pm 1.64$
Little Rock	5.75 ± 4.79	5.94 ± 4.43	4.77 ± 4.67
Grassland	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
s208	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
s420	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Gnutella 08.08.2002	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Gnutella 09.08.2002	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
E .coli	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
S. cerevisiae	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 21: Number of Complete subgraphs found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$5 \cdot 10^{-2}$	$3.50 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
Silwood	$3 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	0.00
St. Martin Island	$6.50 \cdot 10^{-2}$	$6.50 \cdot 10^{-2}$	0.00
Ythan Estuary	$2.15 \cdot 10^{-1}$	$2.15 \cdot 10^{-1}$	$5 \cdot 10^{-2}$
Little Rock	$6.50 \cdot 10^{-2}$	$1.20 \cdot 10^{-1}$	$1.70 \cdot 10^{-1}$
Grassland	0.00	0.00	0.00
s208	0.00	0.00	0.00
s420	0.00	0.00	0.00
Gnutella 08.08.2002	0.00	0.00	0.00
Gnutella 09.08.2002	0.00	0.00	0.00
E .coli	0.00	0.00	0.00
S. cerevisiae	0.00	0.00	0.00

Table 22: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Complete subgraph.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$6.01 \cdot 10^{1} \pm 1.26 \cdot 10^{1}$	$6.07 \cdot 10^{1} \pm 1.23 \cdot 10^{1}$	$5.19 \cdot 10^{1} \pm 1.38 \cdot 10^{1}$
Silwood	6.25 ± 4.07	5.08 ± 3.64	4.37 ± 4.27
St. Martin Island	$4.85 \cdot 10^{1} \pm 1.48 \cdot 10^{1}$	$4.98 \cdot 10^{1} \pm 1.31 \cdot 10^{1}$	$5.54 \cdot 10^{1} \pm 1.55 \cdot 10^{1}$
Ythan Estuary	$1.95 \cdot 10^{2} \pm 4.52 \cdot 10^{1}$	$1.63 \cdot 10^{2} \pm 3.34 \cdot 10^{1}$	$1.61 \cdot 10^{2} \pm 3.65 \cdot 10^{1}$
Little Rock	$3.79 \cdot 10^{3} \pm 3.06 \cdot 10^{2}$	$4.51 \cdot 10^{3} \pm 2.55 \cdot 10^{2}$	$4.41 \cdot 10^{3} \pm 3.46 \cdot 10^{2}$
Grassland	$2.40 \cdot 10^{-1} \pm 4.72 \cdot 10^{-1}$	$3.20 \cdot 10^{-1} \pm 5.81 \cdot 10^{-1}$	$3.70 \cdot 10^{-1} \pm 7.70 \cdot 10^{-1}$
s208	$1.02 \pm 9.87 \cdot 10^{-1}$	1.45 ± 1.00	$1.27 \pm 8.59 \cdot 10^{-1}$
s420	1.22 ± 1.15	$9.40 \cdot 10^{-1} \pm 1.01$	$7.50 \cdot 10^{-1} \pm 8.87 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.06 \cdot 10^{2} \pm 1.29 \cdot 10^{1}$	$1.06 \cdot 10^{2} \pm 1.17 \cdot 10^{1}$	$1.07 \cdot 10^{2} \pm 1.08 \cdot 10^{1}$
Gnutella 09.08.2002	$7.58 \cdot 10^{1} \pm 9.68$	$7.73 \cdot 10^{1} \pm 1.02 \cdot 10^{1}$	$7.66 \cdot 10^{1} \pm 9.61$
E .coli	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
S. cerevisiae	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 23: Number of Fourcycles found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$9 \cdot 10^{-2}$	$2.35 \cdot 10^{-1}$	$3.10 \cdot 10^{-1}$
Silwood	$1.50 \cdot 10^{-1}$	$3.85 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$
St. Martin Island	$1.55 \cdot 10^{-1}$	$2.45 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$
Ythan Estuary	$3.70 \cdot 10^{-1}$	$4.05 \cdot 10^{-1}$	$1.80 \cdot 10^{-1}$
Little Rock	$7.90 \cdot 10^{-1}$	$6.60 \cdot 10^{-1}$	$2.50 \cdot 10^{-1}$
Grassland	$4 \cdot 10^{-2}$	$6 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
s208	$1.83 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
s420	$1.68 \cdot 10^{-1}$	$1.88 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Gnutella 08.08.2002	$9.78 \cdot 10^{-2}$	$1.84 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$
Gnutella 09.08.2002	$7.50 \cdot 10^{-2}$	$1.50 \cdot 10^{-1}$	$1 \cdot 10^{-1}$
E .coli	0.00	0.00	0.00
S. cerevisiae	0.00	0.00	0.00

Table 24: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Fourcycle.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$4.60 \cdot 10^2 \pm 3.21 \cdot 10^1$	$4.64 \cdot 10^2 \pm 3.51 \cdot 10^1$	$4.69 \cdot 10^2 \pm 2.94 \cdot 10^1$
Silwood	$1.03 \cdot 10^3 \pm 9.35 \cdot 10^1$	$1.11 \cdot 10^3 \pm 9.11 \cdot 10^1$	$1.10 \cdot 10^3 \pm 8.02 \cdot 10^1$
St. Martin Island	$8.71 \cdot 10^2 \pm 4.78 \cdot 10^1$	$9.09 \cdot 10^2 \pm 5.85 \cdot 10^1$	$9.01 \cdot 10^2 \pm 4.66 \cdot 10^1$
Ythan Estuary	$3.22 \cdot 10^3 \pm 1.77 \cdot 10^2$	$3.41 \cdot 10^3 \pm 1.79 \cdot 10^2$	$3.30 \cdot 10^3 \pm 1.17 \cdot 10^2$
Little Rock	$1.67 \cdot 10^5 \pm 2.70 \cdot 10^3$	$1.74 \cdot 10^5 \pm 1.83 \cdot 10^3$	$1.75 \cdot 10^5 \pm 1.81 \cdot 10^3$
Grassland	$1.43 \cdot 10^1 \pm 6.22$	$1.88 \cdot 10^1 \pm 7.45$	$1.47 \cdot 10^1 \pm 4.80$
s208	$6.48 \cdot 10^{-1} \pm 8.96 \cdot 10^{-1}$	$7.20 \cdot 10^{-1} \pm 6.94 \cdot 10^{-1}$	$1.70 \cdot 10^{-1} \pm 5.11 \cdot 10^{-1}$
s420	$7.73 \cdot 10^{-1} \pm 8.78 \cdot 10^{-1}$	$7.90 \cdot 10^{-1} \pm 8.75 \cdot 10^{-1}$	$4.90 \cdot 10^{-1} \pm 6.71 \cdot 10^{-1}$
Gnutella 08.08.2002	$4.61 \cdot 10^3 \pm 1.36 \cdot 10^2$	$4.64 \cdot 10^3 \pm 1.38 \cdot 10^2$	$4.66 \cdot 10^3 \pm 1.37 \cdot 10^2$
Gnutella 09.08.2002	$4.14 \cdot 10^3 \pm 1.20 \cdot 10^2$	$4.15 \cdot 10^3 \pm 1.33 \cdot 10^2$	$4.15 \cdot 10^3 \pm 1.31 \cdot 10^2$
E .coli	$6.48 \cdot 10^1 \pm 1.32 \cdot 10^1$	$7.58 \cdot 10^1 \pm 1.42 \cdot 10^1$	$7.83 \cdot 10^1 \pm 2.05 \cdot 10^1$
S. cerevisiae	$3.06 \cdot 10^2 \pm 3.50 \cdot 10^1$	$3.28 \cdot 10^2 \pm 3.79 \cdot 10^1$	$3.27 \cdot 10^2 \pm 3.56 \cdot 10^1$

Table 25: Number of Bifans found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$1.15 \cdot 10^{-1}$	$2.70 \cdot 10^{-1}$	$2.50 \cdot 10^{-1}$
Silwood	$4.85 \cdot 10^{-1}$	$3.30 \cdot 10^{-1}$	$2.40 \cdot 10^{-1}$
St. Martin Island	$2.95 \cdot 10^{-1}$	$3.05 \cdot 10^{-1}$	$1.70 \cdot 10^{-1}$
Ythan Estuary	$4.70 \cdot 10^{-1}$	$2.75 \cdot 10^{-1}$	$4 \cdot 10^{-1}$
Little Rock	$9.30 \cdot 10^{-1}$	$9.70 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Grassland	$3.80 \cdot 10^{-1}$	$1.65 \cdot 10^{-1}$	$3.50 \cdot 10^{-1}$
s208	$1.23 \cdot 10^{-1}$	$3.28 \cdot 10^{-1}$	$4.50 \cdot 10^{-1}$
s420	$2.25 \cdot 10^{-2}$	$1.43 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.10 \cdot 10^{-1}$	$1.78 \cdot 10^{-1}$	$1 \cdot 10^{-1}$
Gnutella 09.08.2002	$8 \cdot 10^{-2}$	$1.20 \cdot 10^{-1}$	$8 \cdot 10^{-2}$
E .coli	$3.55 \cdot 10^{-1}$	$3.65 \cdot 10^{-1}$	$1.40 \cdot 10^{-1}$
S. cerevisiae	$3.05 \cdot 10^{-1}$	$3.20 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$

Table 26: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Bifan.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$3.44 \cdot 10^2 \pm 3.04 \cdot 10^1$	$3.29 \cdot 10^2 \pm 3.45 \cdot 10^1$	$3.28 \cdot 10^2 \pm 3.14 \cdot 10^1$
Silwood	$1.89 \cdot 10^2 \pm 3.29 \cdot 10^1$	$1.74 \cdot 10^2 \pm 3.66 \cdot 10^1$	$1.80 \cdot 10^2 \pm 3.93 \cdot 10^1$
St. Martin Island	$4.56 \cdot 10^2 \pm 4.52 \cdot 10^1$	$4.06 \cdot 10^2 \pm 5.40 \cdot 10^1$	$3.98 \cdot 10^2 \pm 4.14 \cdot 10^1$
Ythan Estuary	$1.73 \cdot 10^3 \pm 1.20 \cdot 10^2$	$1.53 \cdot 10^3 \pm 1.25 \cdot 10^2$	$1.52 \cdot 10^3 \pm 1.37 \cdot 10^2$
Little Rock	$6.74 \cdot 10^4 \pm 2.09 \cdot 10^3$	$5.56 \cdot 10^4 \pm 1.62 \cdot 10^3$	$5.61 \cdot 10^4 \pm 1.83 \cdot 10^3$
Grassland	7.35 ± 3.05	6.93 ± 3.36	6.37 ± 3.05
s208	1.67 ± 1.21	1.73 ± 1.07	2.04 ± 1.81
s420	2.00 ± 1.38	2.35 ± 1.34	2.58 ± 1.58
Gnutella 08.08.2002	$1.41 \cdot 10^3 \pm 7.75 \cdot 10^1$	$1.38 \cdot 10^3 \pm 7.65 \cdot 10^1$	$1.40 \cdot 10^3 \pm 7.29 \cdot 10^1$
Gnutella 09.08.2002	$1.13 \cdot 10^3 \pm 6.35 \cdot 10^1$	$1.12 \cdot 10^3 \pm 6.14 \cdot 10^1$	$1.13 \cdot 10^3 \pm 6.24 \cdot 10^1$
E .coli	1.54 ± 1.49	1.84 ± 1.53	2.04 ± 1.57
S. cerevisiae	1.78 ± 1.70	1.59 ± 1.41	1.77 ± 1.73

Table 27: Number of Biparallel subgraphs found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$2.90 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$	$1 \cdot 10^{-1}$
Silwood	$3.30 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
St. Martin Island	$5.35 \cdot 10^{-1}$	$5.45 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
Ythan Estuary	$6.55 \cdot 10^{-1}$	$7.05 \cdot 10^{-1}$	$2.20 \cdot 10^{-1}$
Little Rock	1.00	1.00	$2 \cdot 10^{-1}$
Grassland	$1.30 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$	$1.80 \cdot 10^{-1}$
s208	$6 \cdot 10^{-2}$	$2.33 \cdot 10^{-1}$	$2.50 \cdot 10^{-1}$
s420	$1.23 \cdot 10^{-1}$	$1.83 \cdot 10^{-1}$	$8 \cdot 10^{-2}$
Gnutella 08.08.2002	$2.31 \cdot 10^{-1}$	$9.48 \cdot 10^{-2}$	$2 \cdot 10^{-1}$
Gnutella 09.08.2002	$1.55 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$	$1 \cdot 10^{-1}$
E .coli	$1.30 \cdot 10^{-1}$	$1.95 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$
S. cerevisiae	$5 \cdot 10^{-2}$	$9.50 \cdot 10^{-2}$	$1 \cdot 10^{-1}$

Table 28: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Biparallel subgraph.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$9.67 \cdot 10^{1} \pm 1.76 \cdot 10^{1}$	$8.98 \cdot 10^{1} \pm 1.81 \cdot 10^{1}$	$9.04 \cdot 10^{1} \pm 1.35 \cdot 10^{1}$
Silwood	$6.98 \cdot 10^{1} \pm 1.91 \cdot 10^{1}$	$6.40 \cdot 10^{1} \pm 1.70 \cdot 10^{1}$	$7.00 \cdot 10^{1} \pm 2.66 \cdot 10^{1}$
St. Martin Island	$1.97 \cdot 10^{2} \pm 2.81 \cdot 10^{1}$	$1.54 \cdot 10^{2} \pm 2.75 \cdot 10^{1}$	$1.57 \cdot 10^{2} \pm 2.96 \cdot 10^{1}$
Ythan Estuary	$7.73 \cdot 10^{2} \pm 9.85 \cdot 10^{1}$	$6.46 \cdot 10^{2} \pm 9.68 \cdot 10^{1}$	$6.05 \cdot 10^{2} \pm 1.00 \cdot 10^{2}$
Little Rock	$3.92 \cdot 10^{4} \pm 1.63 \cdot 10^{3}$	$3.02 \cdot 10^{4} \pm 1.33 \cdot 10^{3}$	$3.10 \cdot 10^{4} \pm 1.59 \cdot 10^{3}$
Grassland	2.90 ± 2.23	3.15 ± 2.48	2.54 ± 1.40
s208	$1.50 \cdot 10^{-2} \pm 1.22 \cdot 10^{-1}$	0.00 ± 0.00	0.00 ± 0.00
s420	$2.50 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	0.00 ± 0.00	$6 \cdot 10^{-2} \pm 2.37 \cdot 10^{-1}$
Gnutella 08.08.2002	$3.60 \cdot 10^{1} \pm 7.50$	$3.66 \cdot 10^{1} \pm 7.54$	$3.71 \cdot 10^{1} \pm 7.04$
Gnutella 09.08.2002	$2.72 \cdot 10^{1} \pm 7.08$	$2.73 \cdot 10^{1} \pm 6.70$	$2.78 \cdot 10^{1} \pm 7.18$
E .coli	$2.25 \cdot 10^{-1} \pm 4.74 \cdot 10^{-1}$	$1.40 \cdot 10^{-1} \pm 3.47 \cdot 10^{-1}$	$3.80 \cdot 10^{-1} \pm 5.79 \cdot 10^{-1}$
S. cerevisiae	$1.05 \cdot 10^{-1} \pm 3.07 \cdot 10^{-1}$	$2.10 \cdot 10^{-1} \pm 4.07 \cdot 10^{-1}$	$1.50 \cdot 10^{-1} \pm 3.57 \cdot 10^{-1}$

Table 29: Number of In-Fans found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$3.05 \cdot 10^{-1}$	$2.40 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$
Silwood	$2.15 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$1.40 \cdot 10^{-1}$
St. Martin Island	$5.95 \cdot 10^{-1}$	$5.95 \cdot 10^{-1}$	$1.60 \cdot 10^{-1}$
Ythan Estuary	$5.70 \cdot 10^{-1}$	$6.30 \cdot 10^{-1}$	$2.70 \cdot 10^{-1}$
Little Rock	1.00	1.00	$3.40 \cdot 10^{-1}$
Grassland	$1.05 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$	$1.80 \cdot 10^{-1}$
s208	$1.50 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$	0.00
s420	$2.25 \cdot 10^{-2}$	$3.75 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
Gnutella 08.08.2002	$6.85 \cdot 10^{-2}$	$1.56 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
Gnutella 09.08.2002	$7 \cdot 10^{-2}$	$1.25 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$
E .coli	$6 \cdot 10^{-2}$	$1.30 \cdot 10^{-1}$	$1.90 \cdot 10^{-1}$
S. cerevisiae	$1.05 \cdot 10^{-1}$	$4.50 \cdot 10^{-2}$	$6 \cdot 10^{-2}$

Table 30: Results of the Kolmogorov-Smirnov two-sample test between the different models for the In-Fan.

Graph	FDSM	uSIS	dSIS
St. Marks Seagrass	$1.36 \cdot 10^{2} \pm 2.26 \cdot 10^{1}$	$1.29 \cdot 10^{2} \pm 1.93 \cdot 10^{1}$	$1.30 \cdot 10^{2} \pm 2.05 \cdot 10^{1}$
Silwood	$3.23 \cdot 10^{2} \pm 6.55 \cdot 10^{1}$	$3.31 \cdot 10^{2} \pm 5.93 \cdot 10^{1}$	$3.24 \cdot 10^{2} \pm 1.12 \cdot 10^{2}$
St. Martin Island	$2.00 \cdot 10^{2} \pm 2.43 \cdot 10^{1}$	$1.83 \cdot 10^{2} \pm 2.14 \cdot 10^{1}$	$1.81 \cdot 10^{2} \pm 2.02 \cdot 10^{1}$
Ythan Estuary	$1.11 \cdot 10^{3} \pm 1.32 \cdot 10^{2}$	$1.11 \cdot 10^{3} \pm 1.69 \cdot 10^{2}$	$1.06 \cdot 10^{3} \pm 1.22 \cdot 10^{2}$
Little Rock	$3.29 \cdot 10^{4} \pm 1.04 \cdot 10^{3}$	$3.00 \cdot 10^{4} \pm 1.05 \cdot 10^{3}$	$3.03 \cdot 10^{4} \pm 9.68 \cdot 10^{2}$
Grassland	1.08 ± 1.14	1.69 ± 1.75	$8.60 \cdot 10^{-1} \pm 7.62 \cdot 10^{-1}$
s208	$2 \cdot 10^{-2} \pm 1.57 \cdot 10^{-1}$	$2.60 \cdot 10^{-1} \pm 1.41$	0.00 ± 0.00
s420	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	$5 \cdot 10^{-2} \pm 2.18 \cdot 10^{-1}$	$1 \cdot 10^{-1} \pm 3 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.85 \cdot 10^{1} \pm 4.49$	$2.01 \cdot 10^{1} \pm 5.19$	$2.03 \cdot 10^{1} \pm 5.00$
Gnutella 09.08.2002	$1.42 \cdot 10^{1} \pm 4.18$	$1.43 \cdot 10^{1} \pm 4.05$	$1.48 \cdot 10^{1} \pm 4.64$
E .coli	3.00 ± 3.56	3.05 ± 3.88	4.51 ± 4.91
S. cerevisiae	3.47 ± 3.28	3.63 ± 2.60	2.92 ± 2.44

Table 31: Number of Out-Fans found in the respective models.

Graph	FDSM - uSIS	FSDM - dSIS	uSIS - dSIS
St. Marks Seagrass	$1.65 \cdot 10^{-1}$	$1.40 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$
Silwood	$2.25 \cdot 10^{-1}$	$2.80 \cdot 10^{-1}$	$3.30 \cdot 10^{-1}$
St. Martin Island	$4.05 \cdot 10^{-1}$	$4.30 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
Ythan Estuary	$1.35 \cdot 10^{-1}$	$3.35 \cdot 10^{-1}$	$4.20 \cdot 10^{-1}$
Little Rock	$8.80 \cdot 10^{-1}$	$8 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
Grassland	$1.55 \cdot 10^{-1}$	$1.05 \cdot 10^{-1}$	$2 \cdot 10^{-1}$
s208	$6.25 \cdot 10^{-2}$	$1.75 \cdot 10^{-2}$	$8 \cdot 10^{-2}$
s420	$2 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$5 \cdot 10^{-2}$
Gnutella 08.08.2002	$1.98 \cdot 10^{-1}$	$1.51 \cdot 10^{-1}$	$6 \cdot 10^{-2}$
Gnutella 09.08.2002	$6 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$9 \cdot 10^{-2}$
E .coli	$9 \cdot 10^{-2}$	$2.25 \cdot 10^{-1}$	$2.40 \cdot 10^{-1}$
S. cerevisiae	$1.30 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$	$1.70 \cdot 10^{-1}$

Table 32: Results of the Kolmogorov-Smirnov two-sample test between the different models for the Out-Fan.

3 ON THE FASTER OPTION

Graph	FDSM	SIM
St. Marks Seagrass	$1.00 \cdot 10^{1} \pm 4.07$	9.14 ± 3.02
Silwood	9.13 ± 5.82	8.78 ± 2.96
St. Martin Island	$1.06 \cdot 10^{1} \pm 4.22$	9.59 ± 3.10
Ythan Estuary	$3.99 \cdot 10^{1} \pm 1.15 \cdot 10^{1}$	$6.56 \cdot 10^{1} \pm 8.10$
Little Rock	$2.81 \cdot 10^{2} \pm 3.80 \cdot 10^{1}$	$2.69 \cdot 10^{2} \pm 1.64 \cdot 10^{1}$
Grassland	$1.25 \cdot 10^{-1} \pm 3.73 \cdot 10^{-1}$	$1.39 \cdot 10^{-1} \pm 3.73 \cdot 10^{-1}$
s208	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	$6.96 \cdot 10^{-3} \pm 8.35 \cdot 10^{-2}$
s420	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	$4.73 \cdot 10^{-3} \pm 6.88 \cdot 10^{-2}$
Gnutella 08.08.2002	$5.35 \cdot 10^{-1} \pm 7.34 \cdot 10^{-1}$	$5.95 \cdot 10^{-1} \pm 7.71 \cdot 10^{-1}$
Gnutella 09.08.2002	$4.85 \cdot 10^{-1} \pm 7.55 \cdot 10^{-1}$	$3.98 \cdot 10^{-1} \pm 6.30 \cdot 10^{-1}$
E .coli	$5 \cdot 10^{-3} \pm 7.05 \cdot 10^{-2}$	$8.19 \cdot 10^{-3} \pm 9.05 \cdot 10^{-2}$
S. cerevisiae	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	$8.37 \cdot 10^{-3} \pm 9.15 \cdot 10^{-2}$

Table 33: Number of Double-Joins found in the respective models.

Graph	FDSM	SIM
St. Marks Seagrass	$2.20 \cdot 10^{1} \pm 4.52$	$2.22 \cdot 10^{1} \pm 4.71$
Silwood	4.48 ± 2.51	5.16 ± 2.27
St. Martin Island	$1.92 \cdot 10^{1} \pm 5.30$	$1.97 \cdot 10^{1} \pm 4.43$
Ythan Estuary	$5.63 \cdot 10^{1} \pm 1.04 \cdot 10^{1}$	$6.67 \cdot 10^{1} \pm 8.17$
Little Rock	$4.68 \cdot 10^{2} \pm 3.11 \cdot 10^{1}$	$5.03 \cdot 10^{2} \pm 2.24 \cdot 10^{1}$
Grassland	$3.60 \cdot 10^{-1} \pm 5.92 \cdot 10^{-1}$	$3.79 \cdot 10^{-1} \pm 6.16 \cdot 10^{-1}$
s208	1.04 ± 1.02	$9.72 \cdot 10^{-1} \pm 9.86 \cdot 10^{-1}$
s420	1.05 ± 1.04	1.07 ± 1.03
Gnutella 08.08.2002	$3.02 \cdot 10^{1} \pm 5.52$	$3.11 \cdot 10^{1} \pm 5.57$
Gnutella 09.08.2002	$2.45 \cdot 10^{1} \pm 5.69$	$2.44 \cdot 10^{1} \pm 4.94$
E .coli	$1.50 \cdot 10^{-2} \pm 1.22 \cdot 10^{-1}$	$1.97 \cdot 10^{-2} \pm 1.40 \cdot 10^{-1}$
S. cerevisiae	$1 \cdot 10^{-2} \pm 9.95 \cdot 10^{-2}$	$1.00 \cdot 10^{-2} \pm 1.00 \cdot 10^{-1}$

Table 34: Number of Threecycles found in the respective models.

Graph	FDSM	SIM
St. Marks Seagrass	$3.50 \cdot 10^{-2} \pm 1.84 \cdot 10^{-1}$	$3.32 \cdot 10^{-2} \pm 1.82 \cdot 10^{-1}$
Silwood	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	$2.41 \cdot 10^{-2} \pm 1.55 \cdot 10^{-1}$
St. Martin Island	$6.50 \cdot 10^{-2} \pm 2.47 \cdot 10^{-1}$	$4.79 \cdot 10^{-2} \pm 2.19 \cdot 10^{-1}$
Ythan Estuary	$6.15 \cdot 10^{-1} \pm 8.17 \cdot 10^{-1}$	5.77 ± 2.40
Little Rock	5.75 ± 4.79	1.64 ± 1.28
Grassland	0.00 ± 0.00	$8.17 \cdot 10^{-8} \pm 2.86 \cdot 10^{-4}$
s208	0.00 ± 0.00	$4.33 \cdot 10^{-8} \pm 2.08 \cdot 10^{-4}$
s420	0.00 ± 0.00	$6.50 \cdot 10^{-9} \pm 8.06 \cdot 10^{-5}$
Gnutella 08.08.2002	0.00 ± 0.00	$1.96 \cdot 10^{-5} \pm 4.42 \cdot 10^{-3}$
Gnutella 09.08.2002	0.00 ± 0.00	$4.33 \cdot 10^{-6} \pm 2.08 \cdot 10^{-3}$
E .coli	0.00 ± 0.00	$3.96 \cdot 10^{-8} \pm 1.99 \cdot 10^{-4}$
S. cerevisiae	0.00 ± 0.00	$1.63 \cdot 10^{-8} \pm 1.28 \cdot 10^{-4}$

Table 35: Number of Complete subgraphs found in the respective models.

Graph	FDSM	SIM
St. Marks Seagrass	$6.01 \cdot 10^{1} \pm 1.26 \cdot 10^{1}$	$6.75 \cdot 10^{1} \pm 8.21$
Silwood	6.25 ± 4.07	9.64 ± 3.10
St. Martin Island	$4.85 \cdot 10^{1} \pm 1.48 \cdot 10^{1}$	$5.74 \cdot 10^{1} \pm 7.58$
Ythan Estuary	$1.95 \cdot 10^{2} \pm 4.52 \cdot 10^{1}$	$2.93 \cdot 10^{2} \pm 1.71 \cdot 10^{1}$
Little Rock	$3.79 \cdot 10^{3} \pm 3.06 \cdot 10^{2}$	$4.32 \cdot 10^{3} \pm 6.57 \cdot 10^{1}$
Grassland	$2.40 \cdot 10^{-1} \pm 4.72 \cdot 10^{-1}$	$2.97 \cdot 10^{-1} \pm 5.45 \cdot 10^{-1}$
s208	$1.02 \pm 9.87 \cdot 10^{-1}$	1.04 ± 1.02
s420	1.22 ± 1.15	1.18 ± 1.09
Gnutella 08.08.2002	$1.06 \cdot 10^{2} \pm 1.29 \cdot 10^{1}$	$1.06 \cdot 10^{2} \pm 1.03 \cdot 10^{1}$
Gnutella 09.08.2002	$7.58 \cdot 10^{1} \pm 9.68$	$7.67 \cdot 10^{1} \pm 8.76$
E .coli	0.00 ± 0.00	$5.74 \cdot 10^{-3} \pm 7.57 \cdot 10^{-2}$
S. cerevisiae	0.00 ± 0.00	$2.33 \cdot 10^{-3} \pm 4.83 \cdot 10^{-2}$

Table 36: Number of Fourcycles found in the respective models.

Graph	FDSM	SIM
St. Marks Seagrass	$4.60 \cdot 10^2 \pm 3.21 \cdot 10^1$	$4.63 \cdot 10^2 \pm 2.15 \cdot 10^1$
Silwood	$1.03 \cdot 10^3 \pm 9.35 \cdot 10^1$	$1.19 \cdot 10^3 \pm 3.46 \cdot 10^1$
St. Martin Island	$8.71 \cdot 10^2 \pm 4.78 \cdot 10^1$	$9.09 \cdot 10^2 \pm 3.01 \cdot 10^1$
Ythan Estuary	$3.22 \cdot 10^3 \pm 1.77 \cdot 10^2$	$5.04 \cdot 10^3 \pm 7.10 \cdot 10^1$
Little Rock	$1.67 \cdot 10^5 \pm 2.70 \cdot 10^3$	$2.01 \cdot 10^5 \pm 4.48 \cdot 10^2$
Grassland	$1.43 \cdot 10^1 \pm 6.22$	$3.46 \cdot 10^1 \pm 5.88$
s208	$6.48 \cdot 10^{-1} \pm 8.96 \cdot 10^{-1}$	$6.68 \cdot 10^{-1} \pm 8.17 \cdot 10^{-1}$
s420	$7.73 \cdot 10^{-1} \pm 8.78 \cdot 10^{-1}$	$8.49 \cdot 10^{-1} \pm 9.21 \cdot 10^{-1}$
Gnutella 08.08.2002	$4.61 \cdot 10^3 \pm 1.36 \cdot 10^2$	$4.70 \cdot 10^3 \pm 6.86 \cdot 10^1$
Gnutella 09.08.2002	$4.14 \cdot 10^3 \pm 1.20 \cdot 10^2$	$4.22 \cdot 10^3 \pm 6.50 \cdot 10^1$
E .coli	$6.48 \cdot 10^1 \pm 1.32 \cdot 10^1$	$9.26 \cdot 10^1 \pm 9.63$
S. cerevisiae	$3.06 \cdot 10^2 \pm 3.50 \cdot 10^1$	$3.49 \cdot 10^2 \pm 1.87 \cdot 10^1$

Table 37: Number of Bifans found in the respective models.

Graph	FDSM	SIM
St. Marks Seagrass	$3.44 \cdot 10^2 \pm 3.04 \cdot 10^1$	$3.53 \cdot 10^2 \pm 1.88 \cdot 10^1$
Silwood	$1.89 \cdot 10^2 \pm 3.29 \cdot 10^1$	$2.15 \cdot 10^2 \pm 1.46 \cdot 10^1$
St. Martin Island	$4.56 \cdot 10^2 \pm 4.52 \cdot 10^1$	$4.57 \cdot 10^2 \pm 2.14 \cdot 10^1$
Ythan Estuary	$1.73 \cdot 10^3 \pm 1.20 \cdot 10^2$	$2.43 \cdot 10^3 \pm 4.93 \cdot 10^1$
Little Rock	$6.74 \cdot 10^4 \pm 2.09 \cdot 10^3$	$5.89 \cdot 10^4 \pm 2.43 \cdot 10^2$
Grassland	7.35 ± 3.05	6.41 ± 2.53
s208	1.67 ± 1.21	1.67 ± 1.29
s420	2.00 ± 1.38	2.00 ± 1.41
Gnutella 08.08.2002	$1.41 \cdot 10^3 \pm 7.75 \cdot 10^1$	$1.41 \cdot 10^3 \pm 3.75 \cdot 10^1$
Gnutella 09.08.2002	$1.13 \cdot 10^3 \pm 6.35 \cdot 10^1$	$1.14 \cdot 10^3 \pm 3.37 \cdot 10^1$
E .coli	1.54 ± 1.49	1.46 ± 1.21
S. cerevisiae	1.78 ± 1.70	1.80 ± 1.34

Table 38: Number of Biparallel subgraphs found in the respective models.

Graph	FDSM	SIM
St. Marks Seagrass	$9.67 \cdot 10^{1} \pm 1.76 \cdot 10^{1}$	$8.62 \cdot 10^{1} \pm 9.28$
Silwood	$6.98 \cdot 10^{1} \pm 1.91 \cdot 10^{1}$	$7.46 \cdot 10^{1} \pm 8.64$
St. Martin Island	$1.97 \cdot 10^{2} \pm 2.81 \cdot 10^{1}$	$1.72 \cdot 10^{2} \pm 1.31 \cdot 10^{1}$
Ythan Estuary	$7.73 \cdot 10^{2} \pm 9.85 \cdot 10^{1}$	$1.64 \cdot 10^{3} \pm 4.05 \cdot 10^{1}$
Little Rock	$3.92 \cdot 10^{4} \pm 1.63 \cdot 10^{3}$	$3.63 \cdot 10^{4} \pm 1.91 \cdot 10^{2}$
Grassland	2.90 ± 2.23	2.60 ± 1.61
s208	$1.50 \cdot 10^{-2} \pm 1.22 \cdot 10^{-1}$	$1.57 \cdot 10^{-2} \pm 1.25 \cdot 10^{-1}$
s420	$2.50 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	$1.21 \cdot 10^{-2} \pm 1.10 \cdot 10^{-1}$
Gnutella 08.08.2002	$3.60 \cdot 10^{1} \pm 7.50$	$3.72 \cdot 10^{1} \pm 6.10$
Gnutella 09.08.2002	$2.72 \cdot 10^{1} \pm 7.08$	$2.77 \cdot 10^{1} \pm 5.26$
E .coli	$2.25 \cdot 10^{-1} \pm 4.74 \cdot 10^{-1}$	$1.77 \cdot 10^{-1} \pm 4.20 \cdot 10^{-1}$
S. cerevisiae	$1.05 \cdot 10^{-1} \pm 3.07 \cdot 10^{-1}$	$1.97 \cdot 10^{-1} \pm 4.43 \cdot 10^{-1}$

Table 39: Number of In-Fans found in the respective models.

Graph	FDSM	SIM
St. Marks Seagrass	$1.36 \cdot 10^{2} \pm 2.26 \cdot 10^{1}$	$1.23 \cdot 10^{2} \pm 1.11 \cdot 10^{1}$
Silwood	$3.23 \cdot 10^{2} \pm 6.55 \cdot 10^{1}$	$3.46 \cdot 10^{2} \pm 1.86 \cdot 10^{1}$
St. Martin Island	$2.00 \cdot 10^{2} \pm 2.43 \cdot 10^{1}$	$1.81 \cdot 10^{2} \pm 1.35 \cdot 10^{1}$
Ythan Estuary	$1.11 \cdot 10^{3} \pm 1.32 \cdot 10^{2}$	$2.15 \cdot 10^{3} \pm 4.64 \cdot 10^{1}$
Little Rock	$3.29 \cdot 10^{4} \pm 1.04 \cdot 10^{3}$	$3.52 \cdot 10^{4} \pm 1.88 \cdot 10^{2}$
Grassland	1.08 ± 1.14	2.38 ± 1.54
s208	$2 \cdot 10^{-2} \pm 1.57 \cdot 10^{-1}$	$2.76 \cdot 10^{-2} \pm 1.66 \cdot 10^{-1}$
s420	$3 \cdot 10^{-2} \pm 1.71 \cdot 10^{-1}$	$2.20 \cdot 10^{-2} \pm 1.48 \cdot 10^{-1}$
Gnutella 08.08.2002	$1.85 \cdot 10^{1} \pm 4.49$	$2.01 \cdot 10^{1} \pm 4.48$
Gnutella 09.08.2002	$1.42 \cdot 10^{1} \pm 4.18$	$1.51 \cdot 10^{1} \pm 3.89$
E .coli	3.00 ± 3.56	2.44 ± 1.56
S. cerevisiae	3.47 ± 3.28	3.64 ± 1.91

Table 40: Number of Out-Fans found in the respective models.

BIFAN EQUATION REVISITED - CONTINUED

$$
\begin{aligned}
&\frac{1}{4m^4}\sum_{u,v,w,x\in V}\left(k_u^2-k_u\right)\left(k_v^2-k_v\right)\left(j_w^2-j_w\right)\left(j_x^2-j_x\right)\\
&=\frac{1}{4m^4}\sum_{u,v,w\in V}\left(k_u^2-k_u\right)\left(k_v^2-k_v\right)\left(j_w^2-j_w\right)\left(n\left(\langle j^2\rangle-\langle j\rangle\right)-\left(\frac{j_u^2-j_u}{n}\right)-\left(\frac{j_v^2-j_v}{n}\right)-\left(\frac{j_w^2-j_w}{n}\right)\right)\\
&=\frac{n\left(\langle j^2\rangle-\langle j\rangle\right)}{4m^4}\sum_{u,v}\left(k_u^2-k_u\right)\left(k_v^2-k_v\right)\left(n\left(\langle j^2\rangle-\langle j\rangle\right)-\left(\frac{j_u^2-j_u}{n}\right)-\left(\frac{j_v^2-j_v}{n}\right)\right)\\
&-\frac{1}{4nm^4}\sum_{u,v\in V}\left(k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u\right)\left(k_v^2-k_v\right)\left(n\left(\langle j^2\rangle-\langle j\rangle\right)-\left(\frac{j_u^2-j_u}{n}\right)-\left(\frac{j_v^2-j_v}{n}\right)\right)\\
&-\frac{1}{4nm^4}\sum_{u,v\in V}\left(k_u^2-k_u\right)\left(k_v^2j_v^2-k_v^2j_v-k_vj_v^2+k_vj_v\right)\left(n\left(\langle j^2\rangle-\langle j\rangle\right)-\left(\frac{j_u^2-j_u}{n}\right)-\left(\frac{j_v^2-j_v}{n}\right)\right)\\
&-\frac{1}{4nm^4}\sum_{u,v\in V}\left(k_u^2-k_u\right)\left(k_v^2-k_v\right)\left(n\left(\langle j^4\rangle-2\langle j^3\rangle+\langle j^2\rangle\right)-\left(\frac{j_u^4-2j_u^3+j_u^2}{n}\right)-\left(\frac{j_v^4-2j_v^3+j_v^2}{n}\right)\right)\\
&=\frac{n^2\left(\langle j^2\rangle-\langle j\rangle\right)^2}{4m^4}\sum_{u\in V}\left(k_u^2-k_u\right)\left(n\left(\langle k^2\rangle-\langle k\rangle\right)-\left(\frac{k_u^2-k_u}{n}\right)\right)\\
&-\frac{\left(\langle j^2\rangle-\langle j\rangle\right)}{4m^4}\sum_{u\in V}\left(k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u\right)\left(n\left(\langle k^2\rangle-\langle k\rangle\right)-\left(\frac{k_u^2-k_u}{n}\right)\right)\\
&-\frac{\left(\langle j^2\rangle-\langle j\rangle\right)}{4m^4}\sum_{u\in V}\left(k_u^2-k_u\right)\left(n\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)-\left(\frac{k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u}{n}\right)\right)\\
&-\frac{\left(\langle j^2\rangle-\langle j\rangle\right)}{4m^4}\sum_{u\in V}\left(k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u\right)\left(n\left(\langle k^2\rangle-\langle k\rangle\right)-\left(\frac{k_u^2-k_u}{n}\right)\right)\\
&+\frac{1}{4n^2m^4}\sum_{u\in V}\left(k_u^2j_u^4-2k_u^2j_u^3+k_u^2j_u^2-k_uj_u^4+2k_uj_u^3-k_uj_u^2\right)\left(n\left(\langle k^2\rangle-\langle k\rangle\right)-\left(\frac{k_u^2-k_u}{n}\right)\right)\\
&+\frac{1}{4n^2m^4}\sum_{u\in V}\left(k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u\right)\left(n\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)\right.\\
&\qquad\left.-\left(\frac{k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u}{n}\right)\right)\\
&-\frac{\left(\langle j^2\rangle-\langle j\rangle\right)}{4m^4}\sum_{u\in V}\left(k_u^2-k_u\right)\left(n\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)-\left(\frac{k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u}{n}\right)\right)\\
&+\frac{1}{4n^2m^4}\sum_{u\in V}\left(k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u\right)\left(n\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)\right.\\
&\qquad\left.-\left(\frac{k_u^2j_u^2-k_u^2j_u-k_uj_u^2+k_uj_u}{n}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{4n^2m^4}\sum_{u\in V}\left(k_u^2-k_u\right)\left(n\left(\langle k^2j^4\rangle-2\langle k^2j^3\rangle+\langle k^2j^2\rangle-\langle kj^4\rangle+2\langle kj^3\rangle-\langle kj^2\rangle\right)\right.\\
&\qquad\left.-\left(\frac{k_u^2j_u^4-2k_u^2j_u^3+k_u^2j_u^2-k_uj_u^4+2k_uj_u^3-k_uj_u^2}{n}\right)\right)\\
&-\frac{\left(\langle j^4\rangle-2\langle j^3\rangle+\langle j^2\rangle\right)}{4m^4}\sum_{u\in V}\left(k_u^2-k_u\right)\left(n\left(\langle k^2\rangle-\langle k\rangle\right)-\left(\frac{k_u^2-k_u}{n}\right)\right)\\
&+\frac{1}{4n^2m^4}\sum_{u\in V}\left(k_u^2j_u^4-2k_u^2j_u^3+k_u^2j_u^2-k_uj_u^4+2k_uj_u^3-k_uj_u^2\right)\left(n\left(\langle k^2\rangle-\langle k\rangle\right)-\left(\frac{k_u^2-k_u}{n}\right)\right)\\
&+\frac{1}{4n^2m^4}\sum_{u\in V}\left(k_u^2-k_u\right)\left(n\left(\langle k^2j^4\rangle-2\langle k^2j^3\rangle+\langle k^2j^2\rangle-\langle kj^4\rangle+2\langle kj^3\rangle-\langle kj^2\rangle\right)\right.\\
&\qquad\left.-\left(\frac{k_u^2j_u^4-2k_u^2j_u^3+k_u^2j_u^2-k_uj_u^4+2k_uj_u^3-k_uj_u^2}{n}\right)\right)\\
&=\frac{\left(\langle k^2\rangle-\langle k\rangle\right)^2\left(\langle j^2\rangle-\langle j\rangle\right)^2}{4\langle k\rangle^4}-\frac{\left(\langle k^4\rangle-2\langle k^3\rangle+\langle k^2\rangle\right)\left(\langle j^2\rangle-\langle j\rangle\right)^2}{4m^2\langle k\rangle^2}\\
&\quad-4\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle j^2\rangle-\langle j\rangle\right)\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)}{4m^2\langle k\rangle^2}\\
&\quad+4\frac{\left(\langle j^2\rangle-\langle j\rangle\right)\left(\langle k^4j^2\rangle-2\langle k^3j^2\rangle+\langle k^2j^2\rangle-\langle k^4j\rangle+2\langle k^3j\rangle-\langle k^2j\rangle\right)}{4m^4}\\
&\quad+4\frac{\left(\langle k^2\rangle-\langle k\rangle\right)\left(\langle k^2j^4\rangle-2\langle k^2j^3\rangle+\langle k^2j^2\rangle-\langle kj^4\rangle+2\langle kj^3\rangle-\langle kj^2\rangle\right)}{4m^4}\\
&\quad-6\frac{\left(\langle k^4j^4\rangle-2\langle k^4j^3\rangle+\langle k^4j^2\rangle-2\langle k^3j^4\rangle+4\langle k^3j^3\rangle-2\langle k^3j^2\rangle+\langle k^2j^4\rangle-2\langle k^2j^3\rangle+\langle k^2j^2\rangle\right)}{4m^4n^2}\\
&\quad+2\frac{\left(\langle k^2j^2\rangle-\langle k^2j\rangle-\langle kj^2\rangle+\langle kj\rangle\right)^2}{4m^4}\\
&\quad-\frac{\left(\langle k^2\rangle-\langle k\rangle\right)^2\left(\langle j^4\rangle-2\langle j^3\rangle+\langle j^2\rangle\right)}{4m^2\langle k\rangle^2}+\frac{\left(\langle k^4\rangle-2\langle k^3\rangle+\langle k^2\rangle\right)\left(\langle j^4\rangle-2\langle j^3\rangle+\langle j^2\rangle\right)}{4m^4}
\end{aligned}
$$

PUBLICATIONS

1 JOURNAL ARTICLES

1. "Different flavours of randomness—when to use which null-model to assess statistical significance?"

   In this paper, the fixed degree sequences model and the configuration model are used in a comparative analysis on undirected graphs. While the results for global measures, such as the diameter, coincide, local measures, such as the co-occurrence or the average neighbor degree, do not. This analysis was based on an observation by Zweig and Kaufmann [108], as well as Hórvat and Zweig [41]; they observed similar behavior on bipartite graphs.

2. "Motif detection speed up by using equations based on the degree sequence"

   Due to my Bachelor thesis, I was interested in network motif analysis and whether the configuration model achieves plausible results regarding this topic. Out of curiosity, the simple independence model was used for the same task, assuming that the result would be off. Since the equation yielded almost the same result as the fixed degree sequence model, a more thorough analysis was due. The idea to calculate the standard deviation as it is done was due to critique of one reviewer and several long discussions of how to approach this with Katharina Zweig.

2 CONFERENCES

1. "Social Network Analysis and Gaming: Survey of the Current State of Art"

   An overview of different approaches to analyze massive multiplayer online games. Since this overview was presented at a Serious Games convention, it was highlighted what this research may learn from network analysis on Big Data.

2. "Influence of the Null-Model on Motif Detection" A conference version of "Motif detection speed up by using equations based on the degree sequence". This was the first paper that used the SIM approach to calculate the occurrence of motifs; it did this without calculating the standard deviation, since these equations were only a curios by-product of the actual topic, the comparison of the configuration model to the standard model of network motif analysis, the fixed degree sequence model.

3 OTHER

1. "Dealing with Null-Models"—SunbeltXXXIII

   A talk on the danger of not realizing which null-model is appropriate for analysis.

2. "Quick or exact - When to use which random graph model to asses statistical significance"—NetSci2013

   Talk based on random graph models, including more theory, due to the audience being from the field of network scientists.

Wolfgang Schlauch

Curriculum vitae

Education

09.2012–06.2016 **PhD Student**, *Technische Universität Kaiserslautern*, Kaiserslautern.

03.2011–04.2012 **Master of Science**, *Technische Universität*, Kaiserslautern.

04.2007–10.2010/11 **Bachelor of Science**, *Technische Universität*, Kaiserslautern.

Dissertation

Title *Recognition and rating of subgraphs for dynamical complex communication networks*

Advisor Prof. Dr. Katharina Anna Zweig

Summary Usually, network analysis consists out of measuring something and discussing it with professionals. In this thesis, another point of view is highlighted. Measures on a real-world graph can be compared to samples of randomly generated graphs. I investigate which model is appropriate and show with statistical analysis that not all models are equally applicable.

Master Thesis

Title *Organizational Social Network Analysis*

Advisor Prof. Dr. Andreas Dengel

Summary Developement and deployment of a questionnaire used to analyze the work-relations in a (small) company. Results of this questionnaire show that almost no reciprocal connection exists and that there are hubs of knowledge and knowledge distribution.

Bachelor Thesis

Title *Analysis of algorithms which generate graphs based on a fixed degree sequence*

Advisor Prof. Dr. Andreas Dengel

Summary Implementation of several algorithms to generate random graphs; additionally, development of experiments to check whether they are generated uniformly at random by using ideas akin to PageRank.