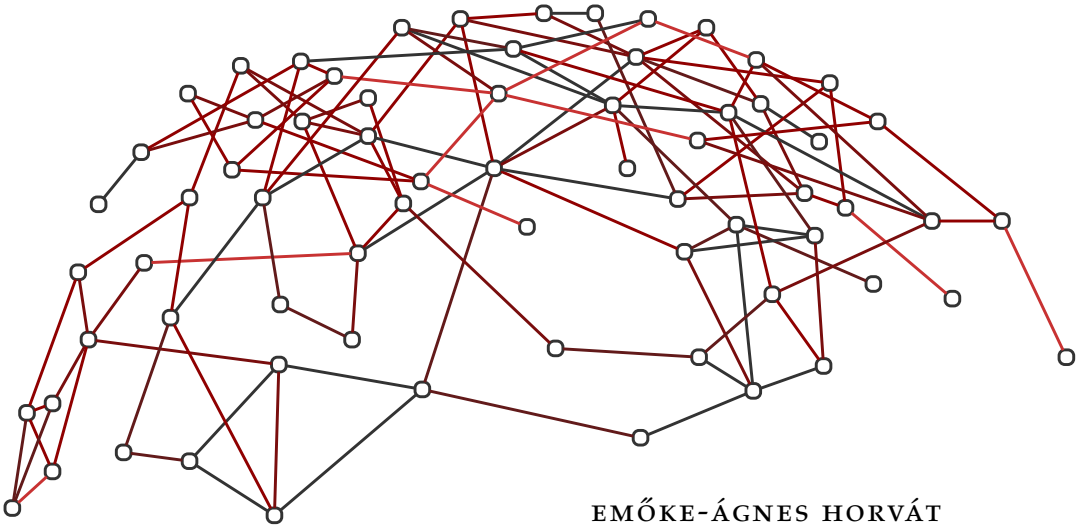
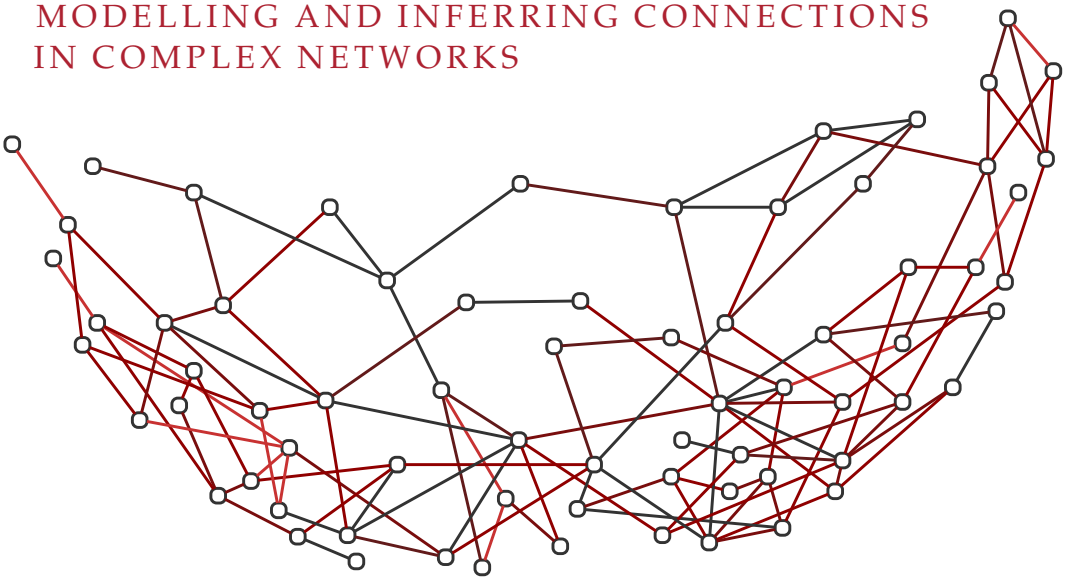


EMŐKE-ÁGNES HORVÁT

MODELLING AND INFERRING CONNECTIONS IN COMPLEX NETWORKS



DISSERTATION

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

put forward by

M.Sc. EMŐKE-ÁGNES HORVÁT

born in: Târgu Mureş, Romania

date of oral examination: December 12, 2013

MODELLING AND INFERRING CONNECTIONS IN COMPLEX NETWORKS

REFEREES: PROF. DR. DIETER W. HEERMANN

PROF. DR. KATHARINA A. ZWEIG

MODELLING AND INFERRING CONNECTIONS IN COMPLEX NETWORKS

EMŐKE-ÁGNES HORVÁT

ABSTRACT

Network phenomena are of key importance in the majority of scientific disciplines. They motivate the desire to better understand the implications of interactions between connected entities. In the focus of this thesis are two of the most prominent tasks in the research of such phenomena: the modelling and the inference of connections within networks. In particular, I provide a systematic framework for using the topology and unifying characteristics of networks from fields as diverse as biology, sociology, and economics to predict and validate connections. I build on existing random graph models and node similarity measures, which I then employ in both unsupervised and supervised machine learning approaches. Furthermore, I present novel methods for identifying the statistically significant connections in network settings that involve multiple types of entities and connections—a crucial element of modelling, which most available methods fail to address.

To demonstrate the potential of these new tools, I use them to filter networks that were constructed from large-scale noisy data generated by biological experiments as well as records of online social activity. Subsequently, I predict previously unobserved connections within these networks and evaluate the performance of the developed tools based on ground truth data. In further data sets without direct evidence for the connections in the network, a second, bipartite network serves as proxy for the analysis. Specifically, in an e-commerce setting I use connections between products and customers to deduce similarities between the products based on customer behaviour. In an analysis of high-throughput screening data on the other hand, I utilize relations between proteins and experimental conditions to identify potential functional affinities among the proteins.

The findings presented here show that the computational prediction of connections can both help researchers gain a better understanding of costly large-scale data and guide further experimental design. The thesis demonstrates the potential of a network analytic approach to modelling and inference on multiple applications, such as the uncovering of possible privacy issues in the context of online social networking platforms and the optimization of drug development in cancer treatment.

ZUSAMMENFASSUNG

Netzwerkphänomene sind in einer Vielzahl von wissenschaftlichen Disziplinen von zentraler Wichtigkeit und motivieren die Bestrebung, die Interaktionen zwischen vernetzten Entitäten besser zu verstehen. Im Fokus dieser Dissertation stehen zwei der prominentesten Probleme bei der Erforschung solcher Phänomene: die Modellierung und die Inferenz von Beziehungen innerhalb von Netzwerken. Hier präsentiere ich ein systematisches Rahmenwerk, das auf Basis der Topologie und der einheitlichen Merkmale von Netzwerken aus so unterschiedlichen Bereichen wie Biologie, Soziologie und Ökonomie neue Beziehungen vorhersagt. Aufbauend auf Zufallsgraphenmodellen und Ähnlichkeitsmaßen für Knoten, verwende ich zu diesem Zweck sowohl überwachtes als auch unüberwachtes maschinelles Lernen. Weiterhin enthält diese Dissertation neuartige Methoden zur Identifikation von statistisch signifikanten Beziehungen in solchen Netzwerken, die aus mehreren unterschiedlichen Arten von Entitäten und Beziehungen bestehen—sie stellen somit ein zentrales Element der Modellierung dar, das den meisten verfügbaren Ansätzen bisher fehlt.

Um das Potential dieser neu entwickelten Methoden zu demonstrieren, verwende ich sie zum Filtern von Netzwerken aus verrauschten Daten, die durch großangelegte biologische Experimente beziehungsweise aus Aufzeichnungen von Aktivitäten in sozialen Online-Netzwerken erzeugt wurden. Weiterhin sage ich in diesen Netzwerken unbeobachtete Kanten vorher und bewerte die Leistung der dafür verwendeten Methoden anhand eines Goldstandards. In weiteren Datensets, in denen die Beziehungen nicht direkt nachweisbar sind, dient ein zweites, bipartites Netzwerk als Proxy für die Analyse. Im Besonderen benutze ich Beziehungen zwischen Kunden und Produkten im elektronischen Handel um Ähnlichkeiten zwischen Produkten herzuleiten, sowie den Zusammenhang von Proteinen und experimentellen Bedingungen aus Hochdurchsatz-Verfahren um potentielle funktionale Abhängigkeiten zwischen Proteinen zu bestimmen.

Die in dieser Dissertation präsentierten Ergebnisse zeigen, dass die rechnergestützte Vorhersage von Beziehungen Wissenschaftlern sowohl zu einem besseren Verständnis von großen Datensets verhelfen, als auch beim Design weiterer Experimente Anwendung finden kann. Die Resultate unterstreichen zum einen das Potential eines netzwerkanalytischen Ansatzes im Data-Mining in einer Vielzahl von Anwendungsmöglichkeiten, sowie zum anderen die Implikationen solcher Analysemöglichkeiten für die Privatsphäre von Internetnutzern und die Medikamentenentwicklung in der Pharmaforschung.

ACKNOWLEDGMENTS

For their contributions to my PhD studies, I owe many thanks to:

Prof. Katharina Zweig for her trust, dedication, and constant support; for the things she taught me and for being a real *Doktormutter* to me.
Prof. Fred Hamprecht for his unconditional helpfulness and for the great idea and setup of the Facebook project.
Prof. Dieter Heermann for kindly agreeing to review this work on such a short notice.
Prof. Özgür Sahin, Dr. Jitao David Zhang, Dr. Stefan Uhlmann, and Dr. Michael Hanselmann for being inspiring collaborators and co-authors, as well as for the knowledge and experiences they shared with me.
Andreas Spitz for his devotion to our Karl Steinbuch project and for developing the ability to make sense of even my least organized ideas.
Andreas Spitz, Wolfgang Schlauch, Dr. Bo Morgan, Dr. Michael Hanselmann, and Gabriell Máté for their time and energy spent proof-reading this thesis. Their comments improved this document considerably.
The "Netzwerker Gruppe" for making me feel welcome at not just one but two universities.
The members of my two adoptive groups, the Image Processing and Modeling and the Computer Vision groups, for the entertaining discussions during lunches and after hours.
Karsten Staack, Henrik Schäfer, Lyubov Nakryyko, Angela Eigenstetter, and Sophie Abendschein for enriching the Heidelberg everyday life with their cute and hilarious comments.
Gabriell Máté for countless helpful discussions and constant encouragement.
Dr. Michael Winckler, Oktavia Klassen, Sarah Steinbach, Tanja Kohl, Ria Lynott, Sabine Kluge, Jan Keese, Jürgen Moldenhauer, and Markus Ridinger for their kind and effective assistance in technical, administrative, and press-related issues.
Prof. Zoltán Toroczkai and Melinda Varga for their great hospitality and the insightful discussions during my visit at the University of Notre Dame.
My loving family and dear friends for their understanding and for helping me with whatever I needed.
My parents for never teaching me the meaning of *impossible*.

This research was supported by the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences.

CONTENTS

When we mean to build,
We first survey the plot, then draw the model;

— William Shakespeare
King Henry IV (1597), Part II, Act 1, Scene 3

1 INTRODUCTION

Interdisciplinary work

The majority of scientific disciplines have in recent years been faced with a flood of data, such as the results of high-throughput screenings in biology and chemistry, detection experiments in particle physics, traces of human interactions on social networking platforms in sociology, or consumer data in e-commerce [41, 197]. This abundance of diverse and high-dimensional data promises new insights to scientists in the involved fields, yet also requires increasingly sophisticated analytical methods, which are not readily available within the repertoires of the individual disciplines. Recent endeavours to handle the emerging need for such frameworks consist of tackling research questions from one field with methods developed within another [269]. This approach often provides solutions to key problems and thereby transforms interdisciplinary research into one of the crucial ingredients of scientific progress.

Specific questions posed in this thesis

During my PhD studies, the challenges posed by the plethora of data motivated me to conduct innovative research at the frontier between multiple traditional disciplines. Accordingly, in this thesis I adopt an analytical and algorithmic approach to answer questions from fields as diverse as systems biology, sociology, and economics: *How could protein–protein interactions obtained from noisy experimental data be filtered computationally? What role do microRNAs play in the regulation of a set of proteins responsible for cell proliferation in breast cancer? Can negative product ratings serve to improve the recommendation systems of online stores? To what extent can social network platforms be used to deduce offline acquaintances between non-members?* At first glance, these questions seem to have very little in common. As I show in this thesis however, their solutions share similarities that lead to a systematic framework. Within their respective disciplines, they can be answered only partly through a detailed examination of the individual components of the system under study like proteins, people, or films. Usually, the data does not have the desired level of detail, such as in the case of user information on social networking platforms and in online stores. Alternatively, comprehensive information may be too expensive to acquire, as gene regulation and protein interaction data for instance. Thus, instead of only focusing on the individual components, the interactions between them have to be taken into account

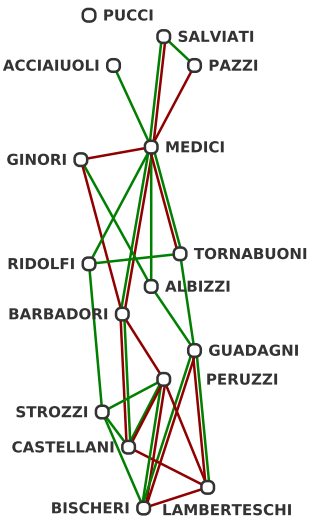
Focusing on complex interactions

to obtain a better understanding of the overall system. Subsequently, traditional discipline-specific approaches offer possibilities for verification.

The statistical physics approach

Components of various systems can be seen as particles, whose interactions reflect the "microscopic" rules that govern their behaviour. As such, they are analogous to a thermodynamic system and methods from statistical physics can be used to deduce "macroscopic" properties [12, 202]. The idea of approaching a social, biological, or economic system within this framework presumes that the components can be meaningfully reduced to simple entities with well-known behaviour. Then, lattice models like the Ising model can be employed to analyse the dynamics of the system [203, ch. 13]. Lattice models are very successful in statistical physics as they enable the study of real-world phenomena, such as phase transitions and critical behaviour [203, ch. 12]. For many real-world systems however, the interactions considered by such models are too restrictive. While the components of the systems under consideration in this thesis are fairly complicated, novel insights can be deduced from their nontrivial interaction patterns.

Network analysis



Example of a network: marriage (green) and business (red) connections among fifteenth-century Florentine families [46]

A promising modelling approach that takes into account interdependencies between the components of a system is *network analysis* [266, 42, 191]. A network consists of a finite set of entities that represent the components of the system and a set of pairwise interactions between them, which are the connections. Formally, a network is represented by a graph. A social system can for instance be modelled as the network of relationships between a set of individuals, the physical interactions between proteins form a protein–protein network, and consumers who buy or rate products can be represented as a product–customer network.

Network analysis itself is an interdisciplinary field that combines methods from statistical physics [12, 60, 191], graph theory [37, 42], and statistics [108, 129] into a generalized mathematical framework for the representation, measurement, and modelling of interconnected systems. Its fundamentally discipline-independent tools are able to tackle problems from various fields and have already led to a range of advances in biology [26, 31, 51], medicine [52, 28], technology [25, 261], economics [228, 74], and sociology [266, 40]. For instance network analysis provided the explanation for the small-world phenomenon [267], models for epidemic spreading in population networks [29, ch. 9], insights into the architecture and stability of ecological networks [32], analytical tools for studying the controllability of natural and technological networks [155], and a better understanding of the structure of neural networks [52].

However, the enthusiasm and the rapid growth of developments in this field caused several of the proposed methods to be misapplied and misinterpreted. Usually, these methods *1)* require a careful,

problem-specific adaptation, and *2)* need to be extended to both the research question at hand and the properties of the studied network, such as multiple types of connections, different entity attributes, or temporal information. These two steps are inherent to network analytic endeavours from beginning to completion. Thus, they are crucial from the stage when the real-world system is modelled as a network to the interpretation of the results [54, 281]. This makes a systematic framework, combining a sound theoretical approach to modelling with appropriate techniques for inference, indispensable for obtaining reliable insights.

1.1 A FIRST GLANCE AT MODELLING AND INFERRING CONNECTIONS

Universality in terms of topology

The analysis of diverse networks reveals that despite the many differences in the underlying systems, the networks themselves often obey similar mathematical rules and have several common properties [15, 191, 24]. The universality of different structural characteristics, such as degree distributions [25, 72], degree correlations [168], motifs [232, 176, 14], and communities [94, 206, 83], are used as a basis for studying diverse phenomena. Assuming a correlation between the topology of the network and the mechanisms that govern the formation of individual interactions between the entities [129], it is possible to infer unobserved or future connections (prediction), and detect spurious connections in noisy data (filtering or validation). These tasks are highly relevant in market basket analysis for example, where the inference of the future behaviour of individuals is the key element of recommender systems. Similarly, advances in biomedicine require the identification of probable interactions whose direct measurement is either technically infeasible or very expensive. In this case, the inference of connections has the potential for filtering noisy data as well as handling incomplete data by guiding experiments toward the most probable candidate interactions.

Types of considered networks and approaches used for inference

In this thesis, I tackle the problem of inferring connections in diverse settings through several different methods. The considered biological, social, and economic networks contain either one or multiple types of connections. One of my key contributions is a newly developed framework for modelling different types of connections as well as their influences on each other, as presented in this thesis. Regarding the deployed approaches, I either use the topology of the observed network between the entities to validate and predict connections, or rely on additional relational data about the entities under consideration. In the latter case, I use a bipartite network as proxy to deduce connections between one of the entity types, whose connectivity is not directly observable. Whereas direct inference is well suited for systems that can be modelled by non-bipartite networks, the infer-

ence based on a proxy network is the preferred option for data with an inherent bipartite structure. Due to the lack of information about the entities other than their connectivity, such as age, location, and occupation of people or 3D structure of proteins, the inference tasks in this thesis are based on topological similarity measures. In addition to surveying classic measures, I use random graph models in a null hypothesis approach to assess the topological similarity of two given entities. Thereby, I quantify the reliability or the likelihood of their potential connection. In an unsupervised learning setting, candidate connections are scored according to these measures. Based on data sets for which a ground truth is available, I then identify those measures with the highest predictive power as well as use them in a supervised learning setting.

Outlook on the main finding of the thesis and its implications

The obtained results show that given the appropriate method, connections can be inferred in a broad range of different networks based solely on the network's topology. This finding has implications on the one hand for our society, since in addition to manifold amenities in optimizing human interactions and organization, it raises inconspicuous privacy implications. On the other hand, the validation and prediction of connections in biological data enable confirming experimental evidences and guide future endeavours.

1.2 OUTLINE AND CONTRIBUTIONS OF THIS THESIS

The thesis consists of two parts. The first details both classic and newly developed methods that form the theoretical framework utilized to tackle the specific research questions posed in the second part, which contains a set of real-world applications. The dependencies between the different chapters are sketched in Figure 1.

For ease of reading, in the remainder of this thesis, I use the word "we" as a substitute for the reader and myself, the interested scientific community, the authors of the articles I co-authored, or simply myself. To highlight my own contributions, I specify them as so-called *thesis points* (TP) in the subsequent list that outlines the content of the individual chapters. The thesis points are marked accordingly also throughout the text.

Part I Theoretical framework

Chapter 2 presents the language, the concepts, and the principles on which the methodological contributions of this thesis are based. Besides the network analytic measures and models, it introduces basic aspects from graph theory, physics, and machine learning.

Chapter 3 discusses classic and generalized random graph models. It details the fixed degree sequence model and provides algorithms for generating graphs according to this model by Markov

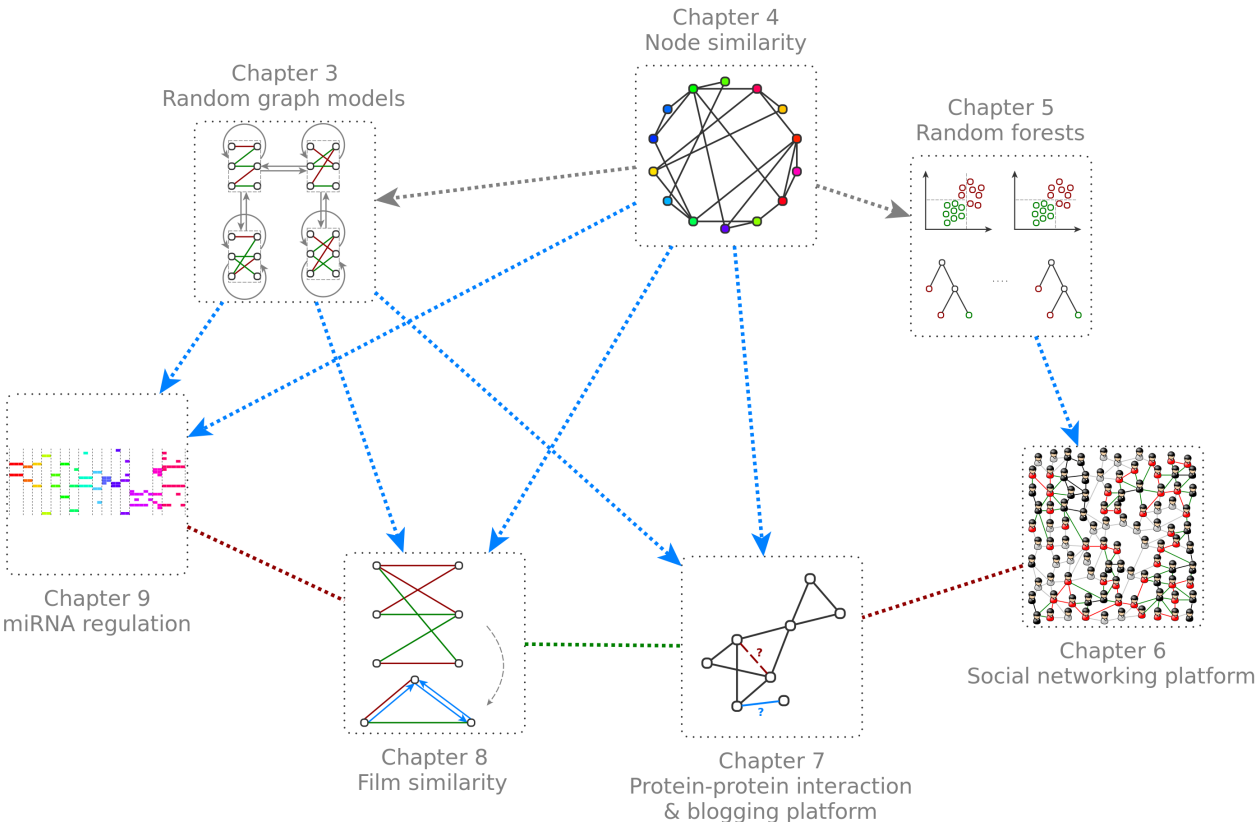


Figure 1: Connections between the individual chapters. The methods introduced in Chapters 3–5 form the theoretical basis underlying the applications presented in Chapters 6–9 (blue). Node similarity measures are used as features in random forests and as test statistic for random graph models (grey). Chapters 6 and 7 focus on non-bipartite graphs that contain one type of connection, whereas Chapters 8 and 9 analyse bipartite graphs with multiple types of connections (red). For the networks considered in Chapters 7 and 8, ground truth connections are available (green).

chain Monte Carlo sampling. Finally, it demonstrates how random graphs are used in hypothesis testing and for the construction of benchmark graphs. This chapter contains the following contributions:

TP1 the fixed degree sequence model for bipartite graphs with multiple types of connections (Section 3.2.1)

TP2 the algorithm for generating networks from this model (Section 3.3.4)

Chapter 4 surveys a set of classic node similarity measures. In it, I discuss their shortcomings and propose to assess the statistical significance of the similarity of two given entities based on their expected similarity in the fixed degree sequence model. This chapter contains the following contributions:

TP3 a comparative survey of the most relevant node similarity measures (Section 4.2 and Section 4.3)

TP4 extension of node similarity measures to multiple types of connections (Section 4.4)

Chapter 5 introduces a machine learning tool, the random forest classifier, as a supervised learning approach to inferring connections.

Part II Applications

Chapter 6 addresses a privacy concern of online social networks. I show how to infer relationships between non-members of a social networking platform such as Facebook by using topological node similarity measures and the random forest classifier. This chapter contains the following contributions:

TP5 development of a set of member acquisition models that partition the users of real-world Facebook friendship networks into members and non-members (Section 6.2)

TP6 extraction of relevant features for training the random forest (Section 6.3)

Chapter 7 provides a systematic assessment of node similarity in networks generated from noisy experimental data or user-declared information. In this chapter, I evaluate a set of node similarity measures based on ground truth data sets for networks as different as protein–protein interaction and user friendship data on an online blogging platform. The chapter illustrates how to validate and predict connections in non-bipartite graphs and contains the following contribution:

TP7 experimental comparison of node similarity measures and evidence that the newly introduced measure based on the fixed degree sequence model consistently outperforms all considered measures (Section 7.3)

Chapter 8 explores new possibilities for recommender systems based on negative ratings. The proposed network analytic point of view introduces novel perspectives that hold promise in exploiting the long tail, i.e. niche market products. In this chapter, I compute film similarities by transforming the film–user network into a film–film network, in which a pair of films is connected if their co-rating pattern is statistically significant. Then, I illustrate the extension of a method for networks with one type of connections to the case of multiple types of connections. The chapter contains thus the following contributions:

TP8 evidence that the proposed algorithm is robust against random noise when tested on artificial data sets (Section 8.2)

TP9 confirmation on different ground truth data sets that the network of positive co-ratings can be used to detect similar films (Section 8.3.2)

TP10 study of the potential of additional, mixed co-rating patterns for improving the detection of similar films, as well as the necessary criteria for the success of this approach (Section 8.3.4)

TP11 a framework for the study of the "similarity landscape" of films through the incorporation of details about the films beyond user ratings (Section 8.4)

Chapter 9 reports the analysis of data from high-throughput screening experiments that monitor the effect of all known human miRNAs on a set of proteins. In this chapter, I show how to uncover potential regulatory patterns by accounting for the generally mild regulation effect of miRNAs and the noise that is inherent to high-throughput experiments. In analogy to the market basket analysis setting, I assess the statistical significance of the co-regulation patterns based on a proper null model. Due to the lack of ground truth information, the computational predictions have been validated experimentally by our collaborators. The chapter thus contains the following contributions:

TP12 detection of connections between pairs of proteins belonging to the same functional module and pairs of miRNAs from the same seed sequence-defined family (Section 9.3)

TP13 identification of miRNAs with tumour suppressing potential (Section 9.3.2)

TP14 release of the freely available software implementation of the key algorithm (Section 9.5)

Chapter 10 concludes this thesis by summarizing the most important findings and discussing future challenges and prospects of the presented approaches.

1.3 PUBLICATIONS RELATED TO THIS THESIS

As listed below and marked at the corresponding places throughout the text, some of the ideas and figures presented in this thesis have already been published or the corresponding article is in preparation. Parts included and adapted from the articles were written by myself.

In the Facebook project, I was relevantly involved in designing the experiments and elaborating the experimental setting. I performed the data preprocessing, the ground truth imputation, as well as the feature selection and extraction. I prepared the main part of the illustrations for the two papers, wrote selected parts of Reference [117]

and Reference [118] in its entirety. Finally, I conducted some of the associated press work.

Concerning the analysis of networks with multiple types of connections, I wrote the encyclopedia essay that surveys existing representations, measures, and models under the supervision of K.A. Zweig [116].

News and Views: M. Malumbres, miRNAs versus oncogenes: the power of social networking, *Molecular Systems Biology, 8:569 (2012)*

In the miRNA regulation project, I analysed the data and co-developed the network biology part of the study in Reference [258]. This contribution of the paper was especially noted in the *News and Views* article that accompanied it in the journal [162]. In addition to providing software and visualizations, I wrote the corresponding methodological paper [119], for which I developed and implemented the used method and performed the robustness analysis under the supervision of K.A. Zweig. The additional software was written by A. Spitz within a practical student training that I advised. I co-wrote the application note [238] and created the website that hosts the tool.

I joined the Netflix project with Reference [282] that shows the superiority of the fixed degree sequence model over a simpler approximation. While for this publication I had only a supportive role, References [114, 115] contain experiments that were conceived, designed, and performed by myself. The two papers extend the fixed degree sequence model to the case of multiple types of connections and were written by me.

Finally, for the article in preparation [120] that compares existing node similarity measures and suggests a new measure based on the fixed degree sequence model, I contributed the formal comparison of the node similarity measures and the parts relevant for this thesis, namely the analyses of the non-bipartite data sets.

LIST OF THESIS-RELEVANT PUBLICATIONS

[117] **E.Á. Horvát**, M. Hanselmann, F.A. Hamprecht, and K.A. Zweig. One plus one makes three (for social networks). *PLOS ONE*, 7(4):e34740, 2012.

[118] **E.Á. Horvát**, M. Hanselmann, F.A. Hamprecht, and K.A. Zweig. You are who knows you: Predicting links between non-members of Facebook. In T. Gilbert, M. Kirkilionis, and G. Nicolis, editors, *Proceedings of the European Conference on Complex Systems 2012*, Springer Proceedings in Complexity, pages 309–316. Springer, 2013.

[116] **E.Á. Horvát** and K.A. Zweig. Multiplex networks. In *Encyclopedia of Social Network Analysis and Mining*. Springer, to appear.

[258] S. Uhlmann, H. Mannsperger, J.D. Zhang, **E.Á. Horvát**, C. Schmidt, M. Küblbeck, F. Henjes, A. Ward, U. Tschulena, K.A. Zweig, U. Korf, S. Wiemann, and Ö. Sahin. Global microRNA

level regulation of EGFR-driven cell cycle protein network in breast cancer. *Molecular Systems Biology*, 8:570, 2012.

[119] **E.Á. Horvát**, J.D. Zhang, S. Uhlmann, Ö. Sahin, and K.A. Zweig. A network-based method to assess the statistical significance of mild co-regulation effects. *PLOS ONE*, 8(9):e73413, 2013.

[238] A. Spitz, K.A. Zweig, and **E.Á. Horvát**. SICOP: identifying significant co-interaction patterns. *Bioinformatics*, 29(19):2503–2504, 2013.

[282] K.A. Zweig and **E.Á. Horvát**. How to evaluate co-occurrences of products in market-baskets from real-world applications. In *Proceedings of the Mini-conference on Applied Theoretical Computer Science*, 2010.

[114] **E.Á. Horvát** and K.A. Zweig. One-mode projections of multiplex bipartite graphs. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 598–605, 2012.

[115] **E.Á. Horvát** and K.A. Zweig. A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs. *Social Network Analysis and Mining*, online first.

[120] **E.Á. Horvát**, A. Spitz, A. Gimmler, T. Stoeck, and K.A. Zweig. Validating and assessing low intensity interactions in complex networks, in preparation.

Part I

THEORETICAL FRAMEWORK

Before tackling discipline-specific research questions, in this part we establish a methodological framework that comprises:

1. the concepts needed for the mathematical formulation of the tackled problems,
2. the related modelling tools and evidence of their correctness, and
3. the key algorithms we later employ.

This framework contains both the methodological contributions of this thesis and established tools from various disciplines we build upon. For ease of reading, we do not rigorously separate these two, but combine them into a coherent framework.

2

PRELIMINARIES

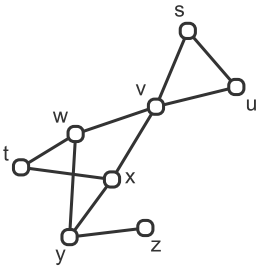
This chapter contains the technical background needed for the theoretical topics covered in this thesis. First, fundamental notions of graph theory (Section 2.1) and network analysis (Section 2.2) are discussed, followed by an overview of existing measures and models for networks that contain multiple types of connections (Section 2.3). After a short detour to Monte Carlo sampling, Markov chains (Section 2.4), and simple graph exploration methods (Section 2.5), basic aspects of statistical learning in networks (Section 2.6 and Section 2.7) are introduced.

2.1 GRAPH BASICS

We start by establishing definitions from graph theory which are necessary for the formulation of the problems and approaches throughout the thesis[1].

GRAPHS AND SUBGRAPHS Formally, a network is represented as a *graph* that consists of a set of *nodes* $\mathcal{V}$ modelling the entities and a set of *edges* $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ representing the connections[2]. Accordingly, a graph G is denoted by the tuple $G := (\mathcal{V}, \mathcal{E})$. Wherever it is not immediately clear from context, we specify that $\mathcal{V}(G)$ is the node set of G, while $\mathcal{E}(G)$ is its edge set. We say that an edge (v, w) is *incident* to its endpoints v and w, while the two nodes v and w connected by an edge are called *adjacent*.

A graph $H := (\mathcal{V}', \mathcal{E}')$ is called a *subgraph* of G if H contains a subset of nodes $\mathcal{V}'$ along with the subset of edges $\mathcal{E}'$ that connect them in G ($\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$). A graph is *complete* if there is an edge between *all* possible pairs of nodes. A complete subgraph is called a *clique*.

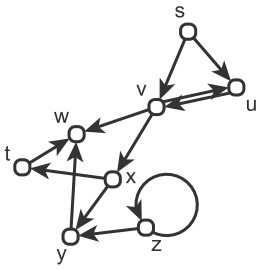


Undirected graph $\mathcal{V}' = \{s, u, v\}$ $\mathcal{E}' = \{(s, u), (s, v), (u, v)\}$ *The subgraph* $H = (\mathcal{V}', \mathcal{E}')$ *is a clique*

TYPES OF GRAPHS If the edge set $\mathcal{E}$ contains unordered pairs of nodes, the graph is said to be *undirected* or *symmetric*, otherwise it is *directed* from a *source* node to a *target* node. For an undirected edge between nodes v and w we write (v, w), while we refer to directed edges by $(v \to w)$, where v is the source and w is the target. The following definitions apply to undirected graphs, unless explicitly specified otherwise.

1 The majority of the following definitions are based on those provided by Kolaczyk [129, ch. 2].

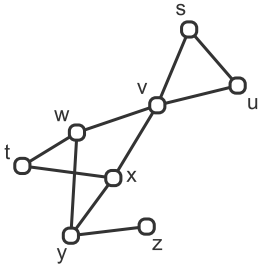
2 Note that different disciplines use different terms for what we call nodes and edges. Physicists for instance use the terms *sites* and *bonds*, computer scientists call them *vertices* and *links*, while in the social sciences *actors* and *ties* or *connections* are used.



Directed graph $(z \to z)$ *is a self-loop* $(u \to v)$ *and* $(v \to u)$ *are mutual edges*

If the set $\mathcal{E}$ is a multi-set, and thus it contains more than one edge between the same pair of nodes, we call the duplicate edges *multi-edges*. *Self-loops* are edges of the form $(v, v) \in \mathcal{E}$ or $(v \to v) \in \mathcal{E}$, as they connect a node to itself. Graphs without self-loops and without multi-edges are said to be *simple*. Directed graphs may have two edges $(v \to w)$ and $(w \to v)$ with opposite direction between the same pair of nodes. In this case, the two edges are said to be *mutual*. A graph that contains mutual edges is nevertheless considered to be simple as long as it does not contain multiple edges in the same direction and self-loops.

Graphs can be *weighted* if the edges have values assigned to them. The *weights* of the edges are given by the function $\omega : \mathcal{E} \to \mathbb{R}$. By convention, in an *unweighted* graph edges are considered to have a weight of 1.



$\{t, x, y, w, t\}$ *is a cycle* $\{s, u, v\}$ *is a triangle* $\{x, y, z\}$ *is a connected triple Shortest paths between* s *and* t *are* $\{s, v, w, t\}$ *or* $\{s, v, x, t\}$ *The length of the shortest path between* s *and* t *is* 3 *The diameter of the graph is* 4

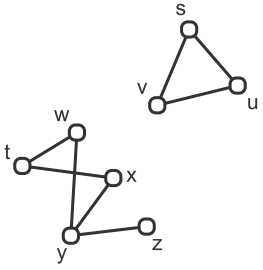
PATHS AND DISTANCES A sequence of nodes $p\alpha := \{v_1, v_2, \ldots, v_k\}$ such that $v_i \in \mathcal{V}$ and $(v_i, v_{i+1}) \in \mathcal{E}$ $\forall 1 \leqslant i < k$ is called a *path*. A *cycle* (also called *circuit*) is defined as a path that contains at least three nodes and for which $v_1 = v_k$, i.e. the path ends at its starting node. A cycle containing three edges is a *triangle*, while a subgraph of three nodes connected by two edges is called a *connected triple*.

The *length* of a path is obtained by summing over the weights of the edges in the path:

$$l(p\alpha) := \sum_{i=1}^{k-1} \omega((v_i, v_{i+1})) \tag{2.1}$$

In an unweighted graph, this length is equivalent to the number of edges in the path. For instance, the length of the longest path in a connected triple is 2. The *shortest path* between two nodes v and w is the path with the minimal possible length among all paths between v and w. The length of the shortest path is called the *(geodesic) distance* between v and w. The *diameter* of a graph is the largest distance between any two of its nodes. The set of nodes at distance 1 from a given node v forms its neighbour set (or neighbourhood) denoted by $\mathcal{N}(v)$.

In a directed graph, a path is a sequence of nodes $pd := \{v_1, v_2, \ldots, v_k\}$ such that $v_i \in \mathcal{V}$ and $(v_i \to v_{i+1}) \in \mathcal{E}$ $\forall 1 \leqslant i < k$.



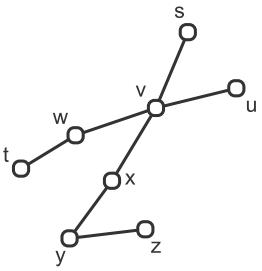
Graph with two components

TREES AND BIPARTITE GRAPHS A subgraph in which there exists a path between all pairs of nodes is called a *weak connected component*. Hereafter also referred to as *component*. A *strongly connected component* is defined for directed graphs and presumes the existence of a directed path between any two nodes in the component.

A connected graph without cycles is called a *tree* and contains exactly $|\mathcal{V}| - 1$ edges. One node may be designated as the *root* of the tree.

Non-root nodes with just one incident edge are called *leaves*, all other nodes are said to be *internal nodes*. A set of trees is said to be a *forest*.

An undirected graph without cycles of odd length is called *bipartite*, otherwise it is said to be *non-bipartite*. The nodes of a bipartite graph can be partitioned into two disjoint sets $\mathcal{L}$ and $\mathcal{R}$. A bipartite graph can therefore be described by the tuple $B := (\mathcal{L} \cup \mathcal{R}, \mathcal{E} \subseteq \mathcal{L} \times \mathcal{R})$, where edges $(v, w) \in \mathcal{E}$ exist only between nodes $v \in \mathcal{L}$ and $w \in \mathcal{R}$.

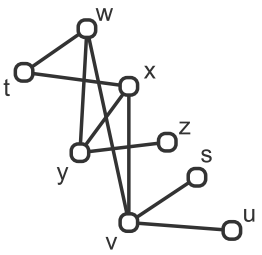


Tree: if s *is the root,* v, w, x, y *are internal nodes and* t, u, z *are leaves*

DEGREES AND DEGREE SEQUENCES The cardinality of the neighbour set of v is called its *degree*, $d(v) := |\mathcal{N}(v)|$. In directed graphs, we differentiate between the *in-degree* of a node $d^i(v)$ (the number of its incoming edges $(w \to v) \in \mathcal{E}, w \in \mathcal{V}$) and the *out-degree* $d^o(v)$ (the number of its outgoing edges $(v \to w) \in \mathcal{E}, w \in \mathcal{V}$).

The *degree sequence* of an undirected non-bipartite graph is defined as the ordered sequence of the degrees $\mathcal{D}(\mathcal{V}) := \{d(v_1), d(v_2), \ldots, d(v_{|\mathcal{V}|})\}$. The concept can be naturally extended to directed graphs by differentiating between the in- and the out-degree sequences $\mathcal{D}^i(\mathcal{V})$ and $\mathcal{D}^o(\mathcal{V})$, respectively. Bipartite graphs also have two sequences, $\mathcal{D}(\mathcal{L})$ and $\mathcal{D}(\mathcal{R})$, one for each of the two disjoint node sets. A graph is said to be *regular* if all of its nodes have the same degree.

A sequence is *graphical* if there exists at least one simple graph with this particular sequence. Any such graph is said to be a *realization* of the given sequence. Graphical sequences are defined analogously for bipartite graphs with the remark that each bipartite graph realizes a *pair* of degree sequences.

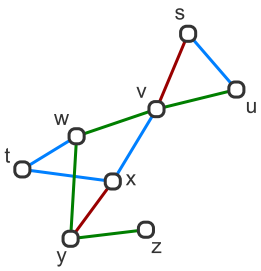


Bipartite graph
$\mathcal{L} = \{t, y, v\}$
$\mathcal{R} = \{w, x, z, s, u\}$
$d(v) = 4$
$\mathcal{D}(\mathcal{L}) = \{2, 3, 4\}$
$\mathcal{D}(\mathcal{R}) = \{3, 3, 1, 1, 1\}$

ADJACENCY MATRICES Often, it is useful to represent a graph $G = (\mathcal{V}, \mathcal{E})$ by its *adjacency matrix* A of dimension $|\mathcal{V}| \times |\mathcal{V}|$. The adjacency matrix of an undirected, unweighted graph has entries $A_{vw} := 1$ if $(v, w) \in \mathcal{E}$ and $A_{vw} := 0$ otherwise. In this case, the entries of the matrix denote the presence or absence of an edge, while the matrix is symmetric and binary.

Certain operations on the adjacency matrix provide additional information about G. Relevantly for this thesis, the row sum $A_{v+} = \sum_w A_{vw}$ is equal to the degree $d(v)$ of node v. Due to the symmetry of the matrix of undirected graphs, the row and column sums (the latter denoted by $A_{+v} = \sum_w A_{wv}$) are equal: $A_{v+} = A_{+v}$.

The adjacency matrix of a bipartite graph $B = (\mathcal{L} \cup \mathcal{R}, \mathcal{E})$ has the dimension $|\mathcal{L}| \times |\mathcal{R}|$. Thus, the row sums yield the degree sequence of the nodes in $\mathcal{L}$, while the column sums provide the degree sequence for $\mathcal{R}$. The adjacency matrix of a directed, unweighted graph is asymmetric and has the entries $A_{vw} = 1$ if there is an edge $(v \to w) \in \mathcal{E}$. In case of weighted graphs, the adjacency matrix also captures the weights: $A_{vw} := \omega((v, w))$ if $(v, w) \in \mathcal{E}$ and $A_{vw} := 0$ otherwise.



Multiplex graph
$|\Omega| = 3$
$\tilde{G}_{\gamma=red} = (\{s, v, x, y\}, \{(s, v), (x, y)\})$
$\mu(v, w) = 1$
$d_{\gamma=green}(v) = 2$
$d_{\gamma=red}(v) = 1$
$\mathcal{V}' = \{s, t, u, v, w, x\}$
$\mathcal{D}_{\gamma=blue}(\mathcal{V}') = \{1, 2, 1, 1, 1, 2\}$

MULTIPLEX GRAPHS *Multiplex graphs* include multiple types of edges between the same set of nodes[3]. Let Ω denote the set of edge types. A multiplex non-bipartite graph is then defined as $\tilde{G} := (\mathcal{V}, \tilde{\mathcal{E}})$, where $\tilde{\mathcal{E}} = \cup_{\gamma \in \Omega} \tilde{\mathcal{E}}_\gamma$ denotes the set of edges of different types. The so-called *supersociomatrix* [266, p. 81–83] representation of such a graph is a tensor $A_{vw\gamma}$ of dimension $|\mathcal{V}| \times |\mathcal{V}| \times |\Omega|$ that stores the adjacency matrix for each edge type $\gamma \in \Omega$. The subgraph $\tilde{G}_\gamma$ of $\tilde{G}$ is induced by the edge type γ and contains the edge set $\tilde{\mathcal{E}}_\gamma$ alongside the nodes that are incident to the edges from $\tilde{\mathcal{E}}_\gamma$.

The *multiplicity* $\mu(v, w) := \sum_\gamma A_{vw\gamma}$ of an edge between v and w counts the number of different edges between the two. In the special case where there is at most one type of edge admitted between any pair of nodes (the Figure to the left shows such a graph), the tensor $A_{vw\gamma}$ can be aggregated to a weighted adjacency matrix whose entries encode the edge type. The degree of a node $v \in \mathcal{V}$ with respect to edge type $\gamma \in \Omega$ is denoted by $d_\gamma(v)$ and is equal to the number of its adjacent edges of type γ. Given γ, the degree sequence $\mathcal{D}_\gamma(\mathcal{V})$ represents the ordered sequence $\{d_\gamma(v_1), d_\gamma(v_2), \ldots, d_\gamma(v_{|\mathcal{V}|})\}$.

We denote multiplex bipartite graphs by $\tilde{B} = (\mathcal{L} \cup \mathcal{R}, \tilde{\mathcal{E}} = \cup_{\gamma \in \Omega} \mathcal{E}_\gamma)$. The adaptation of the above concepts to the bipartite case is straightforward.

Various real-world networks contain multiple types of connections. Modelling them as non-multiplex graphs is incomplete and thus misleading. With the emerging need for a multiplex framework, this research area is very prolific at the time of writing and it can be expected to play an important role in the future. Thus, after presenting some of the most well-known concepts in network analysis in Section 2.2, we briefly overview existing efforts to extend these to multiplex graphs in Section 2.3.

2.2 NETWORK ANALYTIC MEASURES AND CONCEPTS

In the following, we review a core set of useful notions that are widely used for characterizing the topology of networks[4]. As presented here, they hold for undirected non-bipartite graphs. Whenever relevant for this thesis, we provide the formulations for directed and/or bipartite graphs.

AVERAGE DEGREE One of the most basic aggregated measures is the *average degree*. According to the *handshaking lemma*, in an undi-

3 Alternatively, they are called *multirelational* (for instance in Reference [68]) or *multilayered* (for instance in Reference [160]).

4 Several definitions provided in this section are based on those formulated by Kolaczyk [129, ch. 4].

rected graph $G = (\mathcal{V}, \mathcal{E})$ holds that $\sum_{v \in \mathcal{V}} d(v) = 2|\mathcal{E}|$. From this, the average degree $\langle d \rangle$ computed as the degree sum per node equals:

$$\langle d \rangle := \frac{2|\mathcal{E}|}{|\mathcal{V}|} \tag{2.2}$$

where $\langle \cdot \rangle$ denotes the average.

DENSITY Another simple metric that summarizes the topological characteristics of the network is the *density* δ. It is a scaled version of the average degree and for a simple undirected graph G it is defined as the ratio of the number of edges to the total number of possible edges:

$$\delta(G) := \frac{|\mathcal{E}|}{\binom{|\mathcal{V}|}{2}} \tag{2.3}$$

If $|\mathcal{E}| \in \mathcal{O}(|\mathcal{V}|)$, i.e. the number of edges is asymptotically bounded by a function that is linear in the number of nodes, the graph is said to be *sparse*. Otherwise, it is *dense*. In addition to defining the density of the entire graph, it is often instructive to look at the density of a given subgraph, thereby measuring how close the subgraph is to being a clique.

CLUSTERING COEFFICIENT Watts and Strogatz defined the density of a the neighbour set $\mathcal{N}(v)$ of node v, denoted by $cc(v)$, as the *local clustering coefficient* of v [267]. This measure quantifies the probability that the neighbours of a node v are connected themselves. Note that based on Equation 2.3, the local clustering coefficient is undefined for nodes with degree 0 or 1 and is set to 0 by definition. The *average local clustering coefficient* of a graph is computed as the average over all of its nodes:

$$cc(G) := \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} cc(v) \tag{2.4}$$

The local clustering coefficient can be also expressed in terms of the number of triangles $N_{\triangle}(v)$ in which v is participating and the number of connected triples $N_3(v)$ in which two edges are incident to v:

$$cc(v) := \frac{N_{\triangle}(v)}{N_3(v)} \tag{2.5}$$

for nodes v with $N_3(v) > 0$.

Conversely, the *global clustering coefficient* of a graph $tr(G)$, also known as *transitivity* in the social network literature, is defined as the ratio of the number of triangles to the number of connected triples in the graph:

$$tr(G) := \frac{3N_{\triangle}(G)}{N_3(G)} \tag{2.6}$$

where $N_{\triangle}(G) := 1/3 \sum_{v \in \mathcal{V}} N_{\triangle}(v)$ is the number of triangles in the graph , and $N_3(G) := \sum_{v \in \mathcal{V}} N_3(v)$ is the total number of connected triples.

CENTRALITY INDICES *Centrality indices* aim to quantify the "importance" of individual nodes in a network [131, 124, 132]. Depending on the specific network and the research question at hand, several notions of importance are possible. Accordingly, a plethora of diverse indices has been proposed in different areas over the years.

The simplest of them is the degree itself and is called in this context *degree centrality*. Alternatively, the centrality of a node can be measured based on its location with respect to the other nodes of the graph. For instance, *closeness centrality* is defined as the inverse of the total distance between a certain node and all others, while its *betweenness centrality* is given by the fraction of shortest paths between *all* pairs of nodes that pass through it.

GRAPH PARTITIONING A *partition* $\mathcal{C} := \{C_1, \ldots, C_k\}$ of a graph $G = (\mathcal{V}, \mathcal{E})$ is a decomposition of $\mathcal{V}$ into k non-empty and disjoint subsets $C_i \subseteq \mathcal{V}$ ($C_i \neq \emptyset$ and $C_i \cap C_j = \emptyset \; \forall i \neq j$) such that $\cup_{i=1}^{k} C_i = \mathcal{V}$.

Graph partitioning algorithms (also known as *clustering* or *community detection algorithms*) attempt to find a partition of cohesive subsets such that the edge sets $\mathcal{E}(C_i, C_j)$ connecting nodes in C_i to nodes in C_j are small in comparison to the edge sets $\mathcal{E}(C_i, C_i)$ connecting nodes within C_i, $\forall 1 \leqslant i, j \leqslant k$. In other words, in a good partition there are many edges within each subset and relatively few between the subsets. Based on this general idea of optimal partitioning, multiple definitions are possible and their formalization often leads to NP-hard problems [226]. For a review on existing heuristic approaches see for example Reference [83].

Several methods are based on the idea of *hierarchical clustering*. This implies a greedy approach, in which candidate partitions are modified such as to minimize a specified cost function either by successive merging of the partitions (*agglomerative* methods) or by their consecutive splitting (*divisive* methods). The cost function is deduced from the chosen definition of cohesiveness and is often related to the dissimilarity or distance of the nodes.

Of note is also one of the most popular approaches to the clustering of spatially embedded nodes, the *k-means* algorithm [108, p. 454–455]. It starts with guessing k cluster centers. Then it alternates the following two steps until convergence: *1)* for each node the closest cluster center is identified, and *2)* each center is updated to the average position of all nodes that are closest to it. This is known as the *nearest neighbour* rule.

Due to the fundamental and crucial assumption that cohesive subsets of nodes share relevant characteristics that are not revealed per se, but are only reflected in the topology of a network, various disciplines are interested in the results of such partitioning algorithms. Examples range from community detection in social sciences [94, 205, 206] to the identification of complexes in protein interaction networks [230, 198] or functional units in metabolic networks [212] in biology.

MODULARITY The *modularity* of a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ was introduced by Newman and Girvan [192, 190] and is defined based on the matrix U of dimension $k \times k$ with entries U_{ij} that represent the fraction of edges in the original graph that connect nodes from C_i with nodes from C_j as follows:

$$mod(\mathcal{C}) := \sum_{i=1}^{k} (U_{ii} - U_{i+}U_{+i}) \tag{2.7}$$

The row sum $U_{i+} = \sum_j U_{ij}$ and the column sum $U_{+i} = \sum_j U_{ji}$ are equal to the number of edges that connect to nodes in C_i. Note that for the computation of the matrix U it is important to ensure that each edge is counted only once without appearing both above and below the diagonal in U. We give a detailed interpretation of this formulation based on a null model argument in Section 3.4.1.

ASSORTATIVE MIXING Considering a grouping of the nodes based on some categorical characteristic such as the gender or ethnicity of individuals in a social network, *assortative mixing* is a measure of network diversity that quantifies the tendency of nodes to connect to other nodes with a similar characteristic. Based on the quantities we introduced for modularity (see Equation 2.7), the assortative mixing coefficient can be defined as:

$$am(\mathcal{C}) := \frac{\sum_i U_{ii} - \sum_i U_{i+}U_{+i}}{1 - \sum_i U_{i+}U_{+i}} \tag{2.8}$$

A perfectly assortative network has a coefficient of 1, while a coefficient of 0 indicates that no mixing is observed under the chosen null model[5] (see Chapter 3 for detailed explanations regarding null models in network analysis).

It is also common practice to measure assortativity based on ordinal characteristics. In a social network for instance, it might be interesting to quantify assortative mixing according to the age of the individuals [189]. Similarly instructive is the *assortativity by degree* λ [187],

5 Note that the coefficient does not reach -1 in the case of perfect disassortativity. For a discussion see Reference [189].

which is defined as the correlation between the degrees x and y of nodes that share an edge[6] ($\forall (v, w) \in \mathcal{E} : d(v) = x, d(w) = y$):

$$\lambda(G) := \frac{\sum_{x,y} xy(u_{xy} - u_{x+}u_{+y})}{\sigma[x]\sigma[y]} \tag{2.9}$$

where $\sigma[x]$ and $\sigma[y]$ denote the standard deviations corresponding to the distributions of u_{x+} and u_{+y}, respectively. In directed networks, nodes have both an incoming (in) and an outgoing (out) degree, which results in the four degree correlations in–in, in–out, out–in, and out–out [85].

TRIADIC CLOSURE AND HOMOPHILY *Triadic closure* is one of the basic organizing principles in social networks. It states that when individuals B and C have a common friend A, then it is likely that they become friends as well. Whenever this closure is missing, sociologists talk about *structural holes* [38; 74, 60–61]. The assumed mechanism underlying this phenomenon is simply that there are numerous occasions for A to introduce B and C [74, p. 44–46]. One of the measures that capture the prevalence of this effect is the clustering coefficient presented above.

Homophily on the other hand, refers to the phenomenon that, as the phrase "Birds of feather flock together" suggests, people tend to associate with others who are similar to them [172]. The phenomenon has been confirmed repeatedly based on characteristics such as age, gender, social class and role. It suggests that given the friendship and the implied similarity of A and B as well as that of A and C, the individuals B and C are likely to be similar to each other as well. It is thus probable that B and C become friends, even if neither of them is aware that the other knows A. Homophily tests are usually performed by comparing the observed network with a network that does not show this effect [74, p. 77–83].

The principles of triadic closure and homophily drive the formation of connections in social networks and thus represent an important basis for predicting future edges. Due to triadic closure, a new edge appears in the network for purely topological reasons. According to the homophily effect, contextual factors beyond the topology of the network lead to the emergence of edges. This distinction becomes relevant when designing indicators that quantify the similarity of two nodes and thus the probability that an edge will form between them (see Chapter 4).

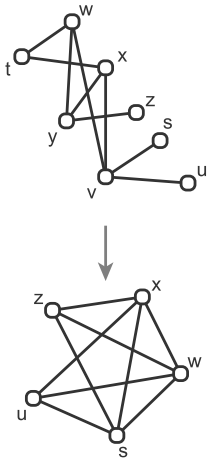
ONE-MODE PROJECTION OF BIPARTITE GRAPHS A bipartite structure is a common property of many real-world networks, such as agents who are affiliated with societies, customers who buy, rent, or rate products, and authors who write scientific papers. These are

6 Read more about the Pearson correlation coefficient in Section 4.2.

modelled by bipartite graphs, which are also called *two-mode* or *affiliation* graphs in social network analysis.

The *one-mode projection* of bipartite graphs onto either set of entities (such as societies, products, and articles) is a well-established approach for their analysis. When projecting $B = (\mathcal{L} \cup \mathcal{R}, \mathcal{E})$ onto its node set $\mathcal{L}$, B is transformed into a *new* non-bipartite graph $G = (\mathcal{L}, \mathcal{E}')$, whose edges $\mathcal{E}'$ are deduced from the connection patterns to the nodes from $\mathcal{R}$.

In the most simple approach, such a projection is obtained by connecting each pair of nodes from $\mathcal{L}$ that share at least one common neighbour in $\mathcal{R}$ [45, 267, 185, 27, 113]. There are two main drawbacks to this approach: structurally very different bipartite networks can result in the same projection (for an example illustrating this problem see Reference [138]) and in most real-world networks the projection can be expected to be uninformatively dense. In a product–customer bipartite graph for instance, there are typically some customers who have bought, rented, or rated almost every product, and thus induce a giant connected component in the resulting projection that contains the majority of the products.



One-mode projection of a bipartite graph

To further distinguish between the connections in such a dense projection, several suggestions have been made to weight them [186, 192, 163, 145, 210, 278]. However, these methods remain problem-specific and do not provide a statistical significance assessment. To address this issue, alternative approaches have been developed [184, 254, 280, 283]. Among the aims of this thesis is to extend the most accurate method to multiplex bipartite graphs (see Chapter 8 and Chapter 9).

2.3 MINING MULTIPLEX NETWORKS

This section is an adapted excerpt from

E.Á. Horvát and K.A. Zweig, Multiplex networks, *to appear in Encyclopedia of Social Network Analysis and Mining, Springer*

The study of multiplex network data goes back to the beginning of social network analysis—for a literature review see Reference [266, p. 719–21]. However, most of the considered networks were generated from small-scale observations or were manually curated from historical evidence. Today, the abundance of large-scale data sets in the humanities, biology, medicine, technology, or from social networking platforms has created a growing need for large-scale multiplex network models. Although multiplex network representations seem to be a natural and conceptually easy generalization, their analysis requires nontrivial modelling decisions and to date only a few network analytic methods have been properly extended.

POSSIBLE REPRESENTATIONS OF MULTIPLEX NETWORKS Besides the supersociomatrix representation of multiplex networks (see Figure 2A), there are two basic possibilities for their analysis:

1. The first is to consider the individual graphs that are induced by each type of edge (see Figure 2B). These graphs can then

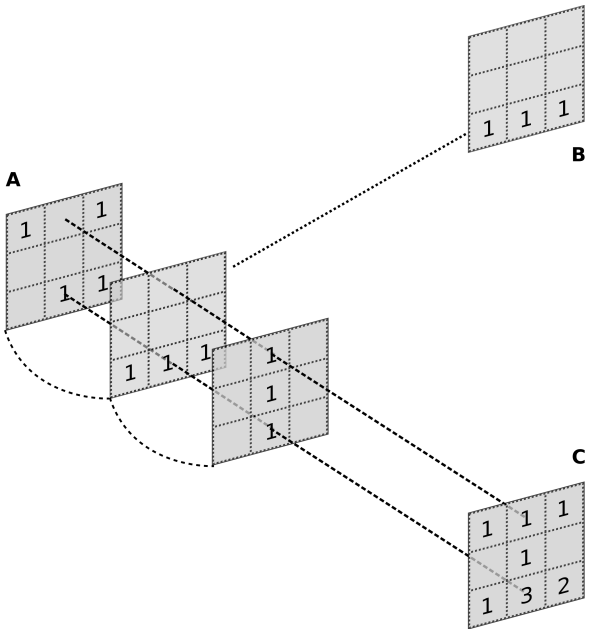


Figure 2: Possible representations of multiplex networks: (A) the supersociomatrix representation, (B) the individual representation, and (C) the aggregated representation. Figure reprinted from [116].

be used to compare network properties such as the position of the same node in each of them or the global structure of the different networks on the same set of nodes.

2. Alternatively, the multiplex network can be transformed into a simple graph (*simplex network*). For this transformation, the individual adjacency matrices are collapsed, such that any pair of nodes is connected in the simplex network if it is linked by at least one edge in any of the matrices. The edges can be weighted, for example, by counting the number of different connections that were aggregated (see Figure 2C) or by computing a weighted sum of their respective weights [146, 126]. The obtained simplex network can then be analysed using standard network analytic measures (see Section 2.2). This approach is viable whenever the overall relationship between two nodes becomes stronger as more types of connections are present between them. As an example, consider employees who are communicating via email, memos, telephone, fax, and through personal encounters. When analysing the volume of information propagated between these people, it can be assumed that the medium used for communication is irrelevant. It is thus sufficient to regard an aggregated version of the network.

Ignoring the dependencies across the different connections in individual graphs or aggregating them into a single connection, however, often leaves the social context underexploited and may result in inac-

curate interpretations [160]. This is the case, for example, when the multiplex network contains friendship and enmity relations. If an aggregate representation of this network is clustered, the result could be groups that consist of people who are partly befriended and who partly dislike each other. If the different connections are interdependent, analysing them in individual graphs will also result in misleading findings. *Structural balance theory*, for example, predicts that every triangle in a social network contains an even number of enmity relations [266, p. 220–233]. Based on the chosen representation, different types of methods can be applied to analyse multiplex network data.

MEASURES AND METHODS Most network analytic measures allow for different generalizations to multiplex networks. Even the most simple measure, the degree, can be generalized in at least three different ways: either by counting the total number of edges of any type incident to a given node, or by taking the degree of the given node in the aggregated network, or by separating the connections by type such that every node has as many degrees as different types of connections. Thus, in the case of degree centrality, one kind of degree must be preferred over the others, according to the specific problem [160].

Most other centrality measures rely on a suitable definition of the distance of two nodes, i.e. the minimal length of any path between them. It can be meaningful to allow a path to use either all or just a subset of the connections represented in the given multiplex network. Alternatively, the path can be defined separately for each of the connections. Based on a suitable distance metric, most centrality measures, such as the betweenness or closeness centrality, can be easily defined.

The framework for analysing large-scale multiplex networks is not yet standardized. Although the types of results that can be obtained are not yet clear, there exist some illustrative studies of large-scale multiplex networks such as the one performed on a network of players in an online game [245, 246]. The results reveal how cross-network analysis provides additional insight into the organizational principles and dynamics of the network. This information has been obtained by computing the overlap and the correlations between the networks corresponding to different edge types.

2.4 MARKOV CHAIN MONTE CARLO METHODS

Individual graphs can be considered random objects drawn from a collection of graphs, a so-called *ensemble*. As we will see later, this is a useful way of thinking about graphs whenever we need to statistically assess their properties. As the ensemble of graphs is potentially very large, we can not feasibly enumerate it, but only explore it by

constructing a representative sample. To understand how this works (see Chapter 3), the following basic ideas of *Markov chains* and *Monte Carlo sampling* are necessary[7].

MARKOV CHAINS Let $\mathcal{S} = \{s_0, s_1, s_2, \dots\}$ denote a finite collection of *states* that form a so-called *state space*. The states are associated with probabilities T_{ij} (known as *transition probabilities*), which quantify how likely it is to move from the current state i to another state j. Additionally, it is required that $T_{ij} \geqslant 0$ and $\sum_j T_{ij} = 1$. The *Markov chain* is defined for a system of states and associated transition probabilities as a discrete-time stochastic process without "memory". This means that the probability of being in state s at time t can be modelled as a function of the immediate predecessor state only:

$$P(s_t = j | s_{t-1} = i, s_{t-2} = h, \dots) = P(s_t = j | s_{t-1} = i) = T_{ij} \quad (2.10)$$

The state graph corresponding to a Markov chain

for all states of the state space. Note that a Markov chain can be associated with a weighted directed graph, the so-called *state graph*, whose nodes correspond to the states and whose edges are weighted by the individual transition probabilities.

PROPERTIES OF MARKOV CHAINS The chain is called *irreducible* if any state can be reached from any other in a finite number of transitions:

$$T_{ij}^q = P(s_{t+q} = j | s_t = i) > 0 \ \forall i, j \in \mathcal{S} \text{ and finite } q \quad (2.11)$$

where T_{ij}^q denotes the q step transition probability, which is the probability that starting in state i the current state is j after q time steps. Accordingly, the state graph must be strongly connected.

Irreducibility, aperiodicity, and ergodicity

An irreducible chain is said to be *aperiodic* if there is a state i for which the greatest common divisor of the length of the cycles that contain it is 1. A more stricter condition, which is sufficient for the purposes of this thesis is that $\exists i : T_{ii} > 0$. This means that the chain is not constrained to cycle through the states in a periodic manner. Equivalently, the state graph of an aperiodic chain is non-bipartite. A finite state Markov chain that is both irreducible and has an aperiodic state is called *ergodic*[8].

Consider a random process in which, starting from an initial state i, a fixed number q of steps is performed according to the transition probabilities. Then, $\{T_{ij}^q\}_{j \in \mathcal{S}}$ is a probability distribution on $\mathcal{S}$, quantifying the probability that the process stops in j after q steps. Let π be a probability distribution on $\mathcal{S}$ such that the probability of state i is $\pi_i \geqslant 0$ and $\sum_i \pi_i = 1$. If this distribution satisfies the equation:

$$\pi_j = \lim_{q \to \infty} T_{ij}^q \ \forall i, j \in \mathcal{S} \quad (2.12)$$

7 This summary is based on References [157; 129, p. 27; 137, p. 32–33; 167, ch. 14; 34].

8 A state is said to be ergodic, if it is both aperiodic and *positive recurrent* in the sense that it is possible to return to it in a finite number of steps.

than it is called a *limiting stationary distribution*. In other words, the stationary distribution no longer changes over time as more and more transitions are being performed, i.e. $\pi = \pi T$. Irreducible and aperiodic chains have such a well-defined stationary distribution.

The existence of a stationary distribution

An alternative way to test for the existence of a stationary distribution is achieved by requiring that the Markov chain fulfils the *detailed balance* condition[9]. First, we introduce the notion of a *reversible* Markov chain. A chain is reversible if $\exists \pi : \pi_i T_{ij} = \pi_j T_{ji}$, i.e. moves along the chain are performed forwards and backwards with equal probability. Based on this, we observe that reversibility implies detailed balance, meaning that any sequence of steps is equally likely to be chosen as its reversed sequence.

Reversible chains and the detailed balance condition

The theory of Markov chains states that the stationary distribution of a reversible chain is proportional to the degree of each state in the underlying state graph. For instance, a uniform stationary distribution thus requires that all states have the same degree, or equivalently, that the state graph is undirected (symmetric) and regular [157].

Sampling from an underlying distribution

RANDOM WALKS AND MONTE CARLO SAMPLING Given a Markov chain, a *random walk* can be performed on it by starting from an initial state and randomly picking each successive state based on the transition probabilities. To sample from a distribution, a Markov chain can be defined on the state space S with the target distribution as its stationary distribution π. Then, according to the principle of *Monte Carlo sampling*, performing a large enough number of random walk steps, i.e. visiting independent and identically distributed states, will produce samples whose distribution approaches π, regardless of the chosen initial state. As the number of samples grows, the sample distribution converges to the actual distribution.

Mixing time of a Markov chain

The *mixing time* τ of a random walk on a chain is defined as the minimum number of steps for which the *variation distance* is small. The variation distance measures how close the distribution is to the stationary distribution after q steps. Thus, the mixing time of the chain is the time it takes for it to reach its stationary distribution.

In summary, Monte Carlo sampling can be used to generate samples from Markov chains that were constructed in such a way as to have the distribution we want to sample from as their stationary distribution. This allows for a broad applicability in physics, molecular biology, ecology, and statistics. Moreover, it is of key importance for sampling random graphs from a given ensemble as discussed in Section 3.3.1.

9 Conversely, a chain that satisfies the detailed balance condition is necessarily ergodic.

2.5 GRAPH EXPLORATION

Basic graph exploration methods are relevant for this thesis when modelling the acquisition of new members by social networking platforms (see Chapter 6). A few wide-spread possibilities are reviewed in the following[10].

BREADTH FIRST SEARCH (BFS) A *search* on a graph can be imagined as a gradual process of moving outward from a given node v. In *breadth first search*, the order of "visiting" the other nodes is as follows. First, nodes adjacent to v are reached (i.e. those that are one hop away from it), then nodes that are two hops away from it, and so on, until all reachable nodes have been visited [66, p. 594–601]. Through this, a tree with root v is constructed and the path between v and some node w in this tree corresponds to the shortest path between the two nodes in the original graph.

DEPTH FIRST SEARCH (DFS) The *depth first search* from node v is carried out by recursively visiting one of the nodes adjacent to v (for instance w), followed by one of w's adjacent nodes, and so on. Whenever there are no more reachable nodes from a given node, a backtracking step is performed to the most recently visited node that still has unvisited adjacent nodes [66, p. 603–610]. A tree is constructed in a similar manner to BFS with the difference that this tree is likely to degenerate into a chain (especially for dense graphs). Thus, distances in the tree generated by DFS do not necessarily relate to shortest paths in the original graph.

RANDOM WALK (RW) Similarly to the random walk on a state graph used to sample graphs belonging to a given ensemble, random walks can be used to explore one particular graph. Given this graph and a starting node v, one of its adjacent nodes w is selected at random. We then move on to w; then one of w's adjacent nodes, and so on. The (random) sequence of nodes selected this way is considered a random walk on the graph [157].

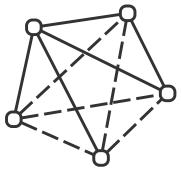
EGO-NETWORKS (EN) Popular mainly in social network analysis, an *ego-network* is a subgraph centered around a given node v and contains the node itself, its adjacent nodes $\mathcal{N}(v)$, and all edges between them [166].

10 Graph exploration as defined here should not be confused with *graph sampling*. In addition to sampling a complete graph from an ensemble of graphs, one frequently needs to sample subgraphs from large real-world networks that can not be dealt with in their entirety. This is the case for instance when the underlying graph is prohibitively large for visualization or for the desired analysis [129, ch. 5].

RANDOM SELECTION (RS) Each node of the graph is chosen randomly with the same probability until a sample of the required size is obtained.

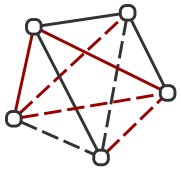
2.6 INFERRING GRAPH TOPOLOGY

Inference problems on real-world networks represent exciting challenges and aim at predicting unobserved parts of a graph. Although they are researched intensively nowadays, there is so far no coherent body of formal results on these problems. Instead, most efforts are centered around specific prediction tasks [129, ch. 7].



Original graph with edges (straight lines) and non-edges (dashed lines)

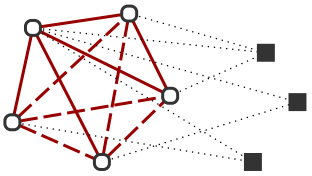
INFERRING EDGES In line with current research trends, in this thesis we concentrate on inferring the presence or absence of connections between a given, fixed set of nodes. More precisely, we focus on predicting individual edges between two given nodes. A typical example of the prediction tasks we deal with is the following. Given a social network of mutually confirmed connections, *1)* infer whether two individuals are likely to meet and form a friendship in the near future or *2)* predict if their connection exists in reality, but is not present in the network (for instance due to our inability to observe it). In other words, we predict a future or a hidden edge connecting the two individuals. To tackle this problem, we have the following two possibilities.



Edge prediction: unobserved edges and non-edges that need to be inferred are coloured red

A. We can use the topology of the observed network between the individuals and reason that, based on the effects of homophily and triadic closure, the structure of their neighbourhoods makes it likely that they in fact already know each other or will meet soon. This approach is commonly referred to as *edge* or *link prediction* (see for instance References [90, 147, 62]).

B. We can rely on additional relational data involving the individuals. For instance, knowing about their affiliation to groups, societies, or organizations indicates their potential connectedness. In other words, given data that can be modelled by a bipartite graph between individuals and their affiliations, a projection of this graph onto the individuals results in information about the likelihood that they are connected. This procedure can be seen as the inference of an *association network* and formally, it is equivalent to the one-mode projection of a bipartite graph (cf. Section 2.2).



Association network inference: none of the edges and non-edges is observed (red lines); connections to a different type of nodes are available instead (black dotted lines)

Edge inference works well in practice if the overall topology of the graph correlates in a nontrivial aspect of the driving forces of social dynamics. Thus, the structure of the observed network is assumed to be predictive and allows reasoning about connections that are not

(yet) visible, but—as we will show in this thesis—can be deduced by means of data mining and machine learning.

SUPERVISED VERSUS UNSUPERVISED LEARNING Existing statistical learning approaches are categorized into *supervised* and *unsupervised learning*. For an overview see Reference [108, ch. 1,2,14].

1. In the case of supervised learning, the aim is to predict (for instance in a medical setting) an output variable (sick or healthy) based on a set of *features* (such as blood pressure, temperature, and other clinical indicators). The output variable often can be interpreted as a *label*. The feature values corresponding to the same entity (an individual) form an *example*. Based on a *ground truth* set of examples, for which both the features and the output variable are known, a *prediction algorithm* models the data and generalizes such as to provide a sensible output for new examples for which only the features are provided.

2. In an unsupervised learning setting, only the features are observed without knowledge of the output variable. Thus, in this case an algorithm is needed that organizes and categorizes the examples based on their similarity, such that further examples can be assigned to one of the categories based on how similar they are to previously seen examples. The prototypical unsupervised learning task is partitioning (or clustering).

Depending on the availability of ground truth data, edge inference can be tackled by both learning approaches as sketched in the following and detailed later (see Chapter 4 and Chapter 5).

EDGE INFERENCE AS A CLASSIFICATION PROBLEM *Classification* is a typical supervised learning task [218, ch. 8; 167, ch. 6]. When inferring connections between individuals as in the case of social networks, the examples are pairs of people. The learning is based on features that characterize each pair (like the number of common acquaintances or shared interests) and labels given by the discrete *class* assignments (friend or non-friend). Note that each pair of individuals is either friend or non-friend (i.e. each example belongs to exactly one class) and there are both friends and non-friends among the pairs of individuals (i.e. the classes occurring in the examples cover the complete space of possible outputs).

We assume a systematic relationship between features and labels and that the examples are representative for their classes. Based on this assumption, we use the ground truth data to model the friend and non-friend classes and to design a *classifier* that is able to assign new pairs of individuals to one of the two classes, based on their features. We proceed by partitioning the ground truth data into *training* and *validation* (or *test*) sets. The classifier is built using the training

data and tested on the validation data. Naturally, the goal is to obtain good classification performance on the validation data. For more details about classification as well as an increasingly popular approach to it see Chapter 5.

EDGE INFERENCE USING SCORING METHODS In an unsupervised approach, the aim is to find effective *scoring methods* in order to establish whether it is likely to exist an edge between two nodes based on this scoring function [129, p. 201–202]. In the example of social networks, a score is computed for each pair of individuals in such a way that, just as the features in the case of supervised learning, it quantifies how similar the two individuals are based on their friend circles and connection patterns.

Connections between individuals are then predicted in two simple ways: *1)* either by considering all pairs of individuals to be connected for which the score is above a certain threshold t or *2)* by ordering the pairs according to their scores and inferring that the k highest-ranked pairs are connected for some fixed k. Many types of scores have been proposed in the literature. We review and compare the most well-known of them and assess their predictive power in the edge inference setting (see Chapter 4 and Chapter 7). Note that scoring functions can be incorporated as features into classification algorithms. Moreover, if ground truth data is available, we can use this information to assess the performance of the individual scores.

2.7 MEASURES OF PREDICTION PERFORMANCE

Assuming the existence of a ground truth, *performance measures* enable us to evaluate the quality of a prediction algorithm. Several different indicators have been proposed in the literature [270, p. 22–24; 218, p. 86–88]. Here we review the measures utilized later to test the results obtained for binary predictions.

Ground truth data suited for an edge prediction problem partitions the set of all possible pairs of nodes into edges (denoted by the set E) and non-edges (the set $\overline{E}$). Compared with the ground truth, a predicted edge can either belong to E and thus be a *true positive* (TP), or belong to $\overline{E}$ and be a *false positive* (FP). Analogously, a predicted non-edge might belong to $\overline{E}$ and thus be a *true negative* (TN) or belong to E and be a *false negative* (FN).

Usually, prediction in bioinformatical problems as well as in social networks is difficult because the set $\overline{E}$ is often substantially larger than E. This implicit disproportion (termed *class imbalance* in the classification jargon) has to be taken into account when choosing the performance measures for evaluation. A trivial algorithm which always predicts that a pair is a non-edge would deceivingly result in a perfect *specificity* $spec$ (the probability of predicting the absence of an edge

that truly doesn't exist). However, for most applications the *sensitivity* $sens$ (the probability of predicting an edge that truly exists) is more relevant. Thus, measures are needed which combine specificity and sensitivity in a meaningful way[11].

F-SCORE An informative measure for assessing the performance of an algorithm is the F-*score* (also known as F_1-*measure* or *balanced* F-*score*). It combines sensitivity and *positive predictive value* PPV (the fraction of predicted edges that are true, sometimes called *precision*) in such a way that they are evenly weighted [21, p. 144]:

$$F := 2 \cdot \frac{sens \cdot PPV}{sens + PPV} \tag{2.13}$$

The F-score is always in the range $[0, 1]$, and the higher the score, the better the prediction. Having no false positive and no false negative predictions would result in an F-score of 1.

THE AREA UNDER THE CURVE, AUC The receiver operating characteristic (ROC curve) is a scatter plot of $sens$ versus $1 - spec$[12]. The performance of an algorithm on a certain data set is displayed in the ROC diagram with a changing *discrimination threshold*, i.e. the prediction value that discriminates between edges and non-edges. Varying this threshold allows a trade-off between sensitivity and specificity.

The Area Under the ROC Curve (AUC) is a scalar performance measure that aggregates the prediction accuracy over all possible settings of this threshold. In other words, it quantifies the probability that true positives are assigned lower scores than true negatives by a given algorithm. A perfect predictor achieves an AUC of 1 while randomly guessing the result of a binary problem yields a value of 0.5 [81].

POSITIVE PREDICTIVE VALUE AT RANK k, PPV_k While the AUC measures the accuracy over the full range of possible discrimination thresholds, the PPV_k is based on a specific threshold: let k denote the number of edges in the validation set $k := |\mathcal{E}|$ and let all pairs of nodes of the validation set be ordered non-increasingly by their prediction value. The PPV_k is then defined as the fraction of correctly classified edges (true positives) among the k top-ranked pairs[13]. It is thus also equal to the sensitivity achieved by predicting these k examples to

11 There are numerous alternatives to the terminology introduced here: false positives are also called *type I error* or *error of the first kind*; false negatives are also called *type II error* or *error of the second kind*; sensitivity is sometimes called *recall*, *true positive rate*, or *hit rate*.

12 According to an alternative terminology, the ROC curve shows the true positive rate (i.e. the sensitivity) against the *false positive rate*.

13 The PPV_k is also known under the name *average precision at rank* k [21, p. 140].

be edges. It can be shown that if FP=FN, the specificity is linearly dependent on PPV_k:

$$1 + f \cdot (PPV_k - 1) = spec \tag{2.14}$$

where f denotes the ratio between $|\mathcal{E}|$ and $|\overline{\mathcal{E}}|$. Both the sensitivity and specificity are thus represented by the PPV_k. The higher its value, the more true edges are placed among the k highest-ranked pairs by the algorithm. Note that the baseline for PPV_k is the overall fraction of edges among all possible pairs of nodes. This would be the result of a naive algorithm in which k pairs of nodes are chosen uniformly at random and predicted to be edges.

The advantage of PPV_k is that it allows us to consider the ordering of the pairs in the prediction result. This is useful for the assessment of the output of both a scoring and a classification algorithm. This feature is used for instance in information retrieval for the definition of the *mean average precision* MAP:

$$MAP := \frac{\sum_{k=1}^{K} PPV_k \, y_k}{TP} \tag{2.15}$$

where k is the rank, K is the number of node pairs, and y_k is a label vector indicating whether the k-th pair is an edge ($y_k = 1$) or not ($y_k = 0$) [270, p. 23].

Throughout this thesis, the PPV_k is used as an indicator of performance in two settings:

A. The *local* PPV_k is computed for each node v from the ground truth individually by ranking the predicted pairs that contain v according to the prediction value and then counting the true positives among the k top-ranked predictions, where k denotes the degree of the node v.

B. The *global* PPV_k is given by the fraction of true positives among the k globally highest-ranked pairs, where k is the total number of edges in the ground truth. Note that the global PPV_k is equivalent to the overall PPV and we denote it as such to differentiate between the local and global positive predictive values.

NORMALIZED DISCOUNTED CUMULATIVE GAIN, nDCG While the PPV_k weighs the top predictions equally, DCG measures the usefulness, the so-called *gain*, based on the position in the ranking, in which pairs are sorted non-increasingly by their prediction value. The individual gains are discounted with decreasing ranks and are then added up. Thus, true positives with a higher ranking result in a larger accumulated gain [21, p. 145–150].

To achieve the effect of discounted gain, DCG introduces a logarithmic position dependency. Accordingly, the benefit of seeing a true positive at position k is $1/\log_2(k+1)$:

$$\mathrm{DCG} := \sum_{k=1}^{K} \frac{1}{\log_2(k+1)} y_k \tag{2.16}$$

where the notation is consistent with Equation 2.15. nDCG then normalizes by the ideal DCG, i.e. the value we would obtain if the k true edges were ranked at positions 1 to k, where k is the number of TPs in the data set. Similarly to PPV_k, it can be computed both locally, by averaging over all nodes from the ground truth, and globally.

2.8 SUMMARY

This chapter discussed basic methods for analysing the observed topology of networks and provided first ideas related to modelling and prediction tasks in graphs. Relying on these preliminaries, the following chapters give a more detailed and formal description of the key methods used in the applications presented in Part II.

3

RANDOM GRAPHS AS NULL MODELS

Graph models are useful tools in various settings. For instance, when studying the emergence of topological properties of real-world networks and the processes leading to their formation [267, 25], when investigating their evolution over time [141], when analysing the implications of diverse patterns with a given functional role [232], when studying the effect of diverse phenomena on networks such as spreading of epidemics [201] and tolerance to attacks [13], or when estimating their topological characteristics [129, p. 162–163]. Random graphs can be formally defined as:

Definition of a graph model

Definition 1 (Graph model) *A graph model is an ensemble of graphs $\mathcal{G}$ with a probability distribution P over the graphs in $\mathcal{G}$:*

$$\{P(G), G \in \mathcal{G}\} \tag{3.1}$$

Thus, the different graph models arise due to the various possibilities for specifying which graphs should be contained in $\mathcal{G}$ and with which probability. Hereafter, we distinguish between the ensembles associated to (simplex) non-bipartite, (simplex) bipartite, multiplex non-bipartite, and multiplex bipartite graphs. We denote them by $\mathcal{G}$, $\mathcal{B}$, $\tilde{\mathcal{G}}$, and $\tilde{\mathcal{B}}$, respectively.

Random graph models and their use in this thesis for hypothesis testing

The modelling tasks that are of interest in this thesis revolve around building null models for hypothesis testing. As we will see later, our main task can often be formulated as follows: Given a graph G_0 and a network observable η, assess the statistical significance of the value η_0 observed in G_0. In other words, quantify how unexpected η_0 is in comparison to an appropriate frame of reference [129, p. 155–156]. As a baseline for such a comparison we use *random graph models*. These are specific graph models for which the ensemble $\mathcal{G}$ contains graphs that maintain the values for a fixed set of parameters and are random in all other respects. After defining $\mathcal{G}$, we compare the value η_0 to the set of values $\{\eta_G \mid G \in \mathcal{G}\}$, thereby using the random graph model as a reference distribution for the possible values of the characteristic η. For simplicity, we choose P to be uniform on $\mathcal{G}$ such that no graph compatible with the model is preferred[1].

The best choice for the selection of parameters to be kept constant in the graphs from $\mathcal{G}$ is an important practical issue. It considerably influences the results and can lead to contradictory interpretations if not constructed carefully with respect to the considered

1 In social network analysis, the approach is also known under the name of *conditional uniform graph test* [53].

problem [17, 283]. This decision involves the well-known trade-off in modelling between generality and realism. Clearly, a null model constructed from an ensemble of random graphs with "no structure" will be easily rejected. With more structure, the model becomes more realistic. However, if too much structure is incorporated, the null model will closely resemble the observation and will lack generality [102, p. xii].

Ecologists locking horns over the null model approach

Among other factors, this trade-off has led to controversies concerning the justification of the null model approach, especially in ecology. An essay of Diamond from 1975 [70] launched an intense debate that resulted in a flood of papers which contained both conceptual objections from the application side and statistical criticism concerning theoretical and algorithmic issues [65, 215, 239, 101, 103, 63, 71]. Nowadays, hypothesis testing through null models is widely accepted and extensively used. According to Gotelli and Graves [102, p. 7]:

> "Null models reflect this natural variability in community structure and require that the 'signal' of mechanism be stronger than the 'noise' of natural variation."

Prevailing discussions within and beyond ecology revolve around the question of which null model to use and how to obtain it even for networks that are considerably larger than typical for ecological studies. In the following, we review the two main possibilities: *classic random graphs*[2] (i.e. graphs in which only the number of nodes and edges is maintained) and *generalized random graphs* (i.e. graphs which also preserve additional properties).

3.1 CLASSIC RANDOM GRAPH MODELS

The two equivalent formulations: $\mathcal{G}(n, m)$ and $\mathcal{G}(n, p)$

The simplest class of random graphs constrains only the number of nodes n and the number of edges m. This class is widely known as the *Erdős-Rényi model* [78], but it was actually previously proposed by Solomonoff and Rapoport [236], and introduced independently by Gilbert [91]. Denoted by $\mathcal{G}(n, m)$, the graphs of this ensemble are constructed from n nodes by randomly adding m edges such that no self-loops or multi-edges are introduced (i.e. the graph is kept simple). Furthermore, all possible graphs in $\mathcal{G}(n, m)$ appear with equal probability:

$$P(G) = \binom{\binom{n}{2}}{m}^{-1} \quad \forall\, G \in \mathcal{G}(n, m) \tag{3.2}$$

In a different formulation, the classic random graph model is defined to consist of all graphs with n nodes, in which all possible edges are instantiated independently with probability $p \in (0, 1)$. This

2 Termed as such in the literature for instance by Dorogovtsev and Mendes [72] and by Kolaczyk [129, p. 156].

ensemble is denoted by $\mathcal{G}(n, p)$ and all of its graphs are associated with the same probability:

$$P(G) = p^m (1-p)^{\binom{n}{2}-m} \quad \forall\, G \in \mathcal{G}(n, p) \tag{3.3}$$

It can be shown that in most investigations the $\mathcal{G}(n, m)$ and $\mathcal{G}(n, p)$ models are practically interchangeable, provided that m is close to pn [37, p. 37]. They are jointly referred to as classic random graph models and represent the class of models that are best understood mathematically, because several of their properties can be determined analytically using probabilistic arguments. For instance, it has been shown that in the limit of large n, their binomial degree distribution becomes a Poisson distribution[3]. Accordingly, the probability of a node having degree k is:

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\langle d \rangle^k e^{-\langle d \rangle}}{k!} \tag{3.4}$$

where the average degree $\langle d \rangle$ is $2m/n$ in $\mathcal{G}(n, m)$ and $p(n-1)$ in $\mathcal{G}(n, p)$[4]. For $n \to \infty$, unless $n \gg m$, the assortativity by degree λ is approximately 0, because the edges are placed randomly without regard to the degree of nodes [187]. The expected transitivity $\mathrm{tr}(G)$ is equal to p [129, p. 157]. Finally, the average shortest path length is small and grows only logarithmically with n [267].

Analytical results for the classic random graph models

These properties are largely different from what is observed in real-world networks. For instance, most social, biological, and technological networks' degree distributions have a long tail [72]; social networks are typically assortative, while many biological and technological networks show a tendency towards disassortativity [187]; most real-world networks are modular [94]; exhibit small-world characteristics, i.e. they have a small average shortest path length, but also a higher clustering coefficient than expected by pure chance [267]. Classic random graph models are therefore too general and unrealistic. They can thus only be deployed as a baseline against which structure should be defined. For the analysis of real-world networks, more involved models are required.

3.2 GENERALIZED RANDOM GRAPH MODELS

Generalized random graph models as subensembles of the classic models

There are various possible extensions to classic random graph models that make them more realistic. These involve additional constraints beyond the fixed number of nodes and edges. The imposed restrictions influence the ensemble $\mathcal{G}$, while, for simplicity, P remains the uniform distribution over $\mathcal{G}$. Accordingly, the constrained ensemble $\mathcal{G}$

3 This explains their alternative name of *Poisson random graphs* [188].

4 The equivalence of the average degree in the two formulations can be deduced in the limit of large n by observing that $m = p\binom{n}{2}$.

is a subensemble of $\mathcal{G}(n, m)$ and is thus strictly contained in it. Therefore, the generalized models can be specified through a conditional distribution on the $\mathcal{G}(n, m)$ model [129, p. 158; 53].

The most commonly chosen additional constraint is that of a *fixed degree sequence* (Section 3.2.1). Observe that by simply fixing a degree sequence $\mathcal{D} = \{d(v_1), d(v_2), \ldots, d(v_n)\}$ for an arbitrary ordering $v_1, v_2, \ldots, v_n$ of the nodes, the number of nodes n and the number of edges $m = \frac{1}{2}\sum_{i=1}^{n} d(v_i)$ are automatically prescribed. Thus, the fixed degree ensemble implicitly fulfils the requirements of the $\mathcal{G}(n, m)$ ensemble and as noted above, it is its subset. Efforts to include even more additional information in the null model, such as to preserve the number of mutual directed edges, have been made in particular in the sociological literature and are briefly discussed in Section 3.2.2.

3.2.1 *The fixed degree sequence model*

Topologically, the most straightforward way of incorporating information about the diversity of the nodes is through their degrees. The heterogeneous degree sequence observed in many real-world networks (cf. References [25, 72]) seems to be an important individual characteristic and as such it should be taken into account in a proper null model. Although there are several other possible node characteristics that also indicate diversity (such as the different centrality indices), various fields have agreed independently on the requirement that a meaningful null model should constrain the degrees of *all* individual nodes. Thus, in the fixed degree sequence model all graph characteristics are free to vary to the extent allowed by the maintained number of nodes, edges, and the preserved degree sequence. In the following, we illustrate the need for the constrained degree sequence on two representative examples, one from social networks and one from ecology.

Illustration of the importance of the degree sequences on two examples

A. In the analysis of social networks, a graph is constructed to represent existing relationships between individuals. Usually, the relationships are mutual and unique (as represented by an undirected graph without multi-edges) and the relationship of individuals to themselves is usually disregarded (there are no self-loops). The degree of a node indicates the popularity of the individual represented by it, and thus it should be contained in a representative null model.

B. In ecology, the presence or absence of species at different locations at a given point in time can be modelled by a bipartite graph without multi-edges. The chance of the occurrence of a species at a location depends on the abundance of the species and on the capacity of the location, which are coded in the degree sequence of the species and locations, respectively. There-

fore, a null model constructed for testing hypotheses about the competition between species should preserve both degree sequences of the bipartite graph.

The intuition is thus that the probability of an edge as well as the occurrence of higher order structures, such as co-occurring pairs of nodes or clusters, depends on the overall context of the nodes. Consider for instance a graph with a power-law degree sequence. In such a graph, low degree nodes ought to be connected to high degree nodes more often than expected in the $\mathcal{G}(n, m)$ ensemble. However, this tendency is predetermined by the graph topology and only an ensemble that takes into account the bias introduced by this degree sequence is able to assess structure that is not immediately implied by the power-law distribution. For illustrative examples about the effect of not maintaining the degree sequences when studying real-world market basket analysis data see References [92, 283].

Motivated by various applications, random graphs with fixed degree sequences have been widely investigated within application areas like sociology [216, 171, 53], ecology [65, 215, 239, 101, 103, 71], and biology [168, 232, 176], but also from the methodological point of view in combinatorics and graph theory [219, 130, 33, 249, 49, 50, 175], in statistics [211, 63, 61, 92, 10], and in network analysis [193, 188, 177, 16, 84, 127, 88, 254].

Several of these works approach the problem of generating graphs with prescribed degree sequences through the equivalent task of generating $(0,1)$-matrices with *fixed margin totals* (i.e. with fixed row and column sums). To see the exact correspondence between the two problems, let us consider again the two examples given above. The simple undirected[5] graphs required for modelling the two networks can be naturally represented by their binary adjacency matrices.

Generating $(0,1)$-matrices with fixed margin totals

A. The social network can be modelled by the non-bipartite graph $G = (\mathcal{V}, \mathcal{E})$ with the corresponding adjacency matrix $A(G)$ of dimension $|\mathcal{V}| \times |\mathcal{V}|$ with zeros on the diagonal. Both the row and column sums of the matrix are equal to the degree sequence $\mathcal{D}(\mathcal{V})$.

B. The ecological network modelled by a bipartite graph $B = (\mathcal{L} \cup \mathcal{R}, \mathcal{E})$ has an adjacency matrix $A(B)$ of dimension $|\mathcal{L}| \times |\mathcal{R}|$. Note that because self-loops are inadmissible in bipartite graphs, this matrix does not have structural zeros on the diagonal. Its row sums match the degree sequence $\mathcal{D}(\mathcal{L})$, while its column sums are equal to the degree sequence $\mathcal{D}(\mathcal{R})$.

Generating graphs from the ensembles $\mathcal{G}$ and $\mathcal{B}$ is thus equivalent to constructing the matrix collections $\mathcal{A}(\mathcal{G})$ and $\mathcal{A}(\mathcal{B})$ that preserve the

5 Directed graphs are also very interesting in this respect. However, they are not the focus of this thesis. Results for random directed graph models with fixed degree sequences can be found for instance in References [211, 188, 34, 79, 128].

corresponding margin totals. Several important results that have been provided in the context of these matrices can be directly transferred to the graph theoretic problem.

The cardinality of the ensembles $\mathcal{G}$ and $\mathcal{B}$

Most of the existing explorations revolve around toy graphs or anecdotal networks containing a few tens of nodes such as Darwin's finches on the Galápagos Islands[6]. Even in these cases, the number of possible graphs in the ensemble is astronomically large. For instance, for the mentioned network of 13 species of finches on 17 islands there are about $6.71 \cdot 10^{16}$ graphs with the given degree sequence [153, p. 93]. For a small artificial bipartite graph containing $20 + 20$ nodes and 20 edges such that each node has degree 1, the fixed degree ensemble contains $20! \approx 2.43 \cdot 10^{18}$ graphs. To our best knowledge, there is no closed formula to estimate the number of graphs in a given ensemble.

THE FIXED DEGREE SEQUENCE MODEL FOR MULTIPLEX GRAPHS

Thesis point 1

An increasing amount of data sets of interest contain details that can not be translated to binary edges between the nodes without significant loss of information. For instance, product–customer data from market basket analysis contains ratings with positive and negative connotation (customers like or dislike products), while gene expression data often encodes both activation and inhibition, i.e. two qualitatively very different relationships.

Although a null model approach to the analysis of such data could be very rewarding, methods for constructing null models for multiplex graphs are still missing (for an exception see Reference [272]). As one of the key contributions of this thesis we provide such a method for multiplex bipartite graphs. After explaining the intuition behind it, describing its theoretical aspects, and providing the relevant algorithm (Section 3.3), in Chapter 8 we use it to analyse the Netflix data set, while in Chapter 9 we show how it performs on an application from systems biology.

3.2.2 *Other generalized random graph models*

The idea of constructing further generalized random graph models that constrain additional characteristics beyond the degree sequence is very appealing. Research in this direction has been conducted based on directed social networks, in which relationships between individuals are not necessarily mutual. For instance, McDonald et al. have defined ensembles that maintain both the in- and out-degree sequences and the number of mutual edges [171]. Roberts has constructed ensembles that also preserve the number of edges between predefined communities [216]. In addition to the fixed degree sequence, the en-

6 Data compiled by Sanderson [223]. The original source is presumably Reference [104].

semble of Maslov et al. contains graphs with a given *correlation profile*, defined as the ratio $P_0(k, k')/P_{\mathcal{G}}(k, k')$, where $P_0(k, k')$ denotes the likelihood that two nodes with degrees k and k' are connected in the observed network, while $P_{\mathcal{G}}(k, k')$ denotes the same likelihood computed from the ensemble $\mathcal{G}$ [170].

Adapting random graph models to constrain certain graph characteristics (or equivalently, fix particular sums in the adjacency matrix) allows to assess other observables in a controlled experiment. Unfortunately, understanding and verifying the mathematics behind these models is extremely challenging. In the absence of analytical results, formal verifications are way behind the pace of idea emergence and algorithm development.

3.3 GENERATING RANDOM GRAPHS WITH FIXED DEGREE SEQUENCE

For all practically relevant graph sizes, the fixed degree sequence ensemble can not be exhaustively enumerated and therefore the reference distribution of any statistic needs to be approximated. While there are a few analytical approaches based on the generating function formalism [193], typically Monte Carlo simulation methods are used to sample uniformly from the ensemble [129, p. 161]. In statistical terms, these are methods for examining distributions using permutation tests under a null model of equally likely graphs with a given degree sequence.

Realizability of a degree sequence

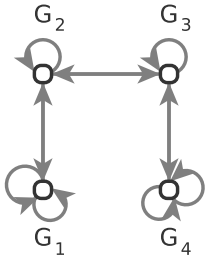
Such an approach presumes that at least one feasible realization for a given degree sequence is given. In the absence of an observed real-world network however, a realization has to be constructed. Social networks for instance can be anonymized by hiding the actual connections and revealing only the degree sequence to avoid privacy issues. Known in general as the *graph realization problem*, characterizing the existence and finding at least one graph instance with a given degree sequence is a thoroughly investigated problem in graph theory. After the more general Tutte's *f-factor theorem* [255], Erdős and Gallai gave a necessary and sufficient condition for the graphicality of a degree sequence [77], while Havel developed a greedy algorithm to construct a realization of a degree sequence as a simple undirected graph [109][7].

The following sections deal with sampling as a critical issue for the application of the fixed degree sequence model in real-world settings. They detail the used Markov chain approach and discuss some of its algorithmic aspects, showing that it can be performed efficiently and can be applied to reasonably large data sets. Section 3.3.5 then presents another wide-spread method that enables graph generation based on the degree sequence for the cases where there is no observed

7 For extensions to the directed case see References [127, 79, 88, 128].

graph. Finally, the chapter ends with a review of some basic ideas for the application of the presented algorithms that draw random graphs from an underlying fixed degree sequence model (Section 3.4).

3.3.1 *Markov chain Monte Carlo sampling*



State graph of an ensemble $\mathcal{G} = \{G_i\}_{i=1}^{4}$

Due to its simplicity and relatively low time complexity, one of the most common strategies to sample instances from a well-defined ensemble is *Markov chain Monte Carlo sampling*. For a brief summary about Markov chains and Monte Carlo sampling see Section 2.4. Within this framework, the problem of generating random graphs with fixed degree sequences lends itself to systematic analysis and facilitates algorithm development.

The basic idea is as follows. The finite number of graphs from the desired fixed degree sequence ensemble are modelled as the states of a Markov chain and we define transitions between those pairs of graphs for which a local transformation exists. Thus, the Markov chain itself can be seen as a graph of states in which two nodes are adjacent if the graphs represented by them can be transformed into each other. A sufficiently long random walk on this chain will arrive at a state that is independent from the starting state and can be considered a random sample from the ensemble. To explore the ensemble, several such random walks are performed and yield a representative sample.

This rough outline of the approach leaves several ensemble-specific details open:

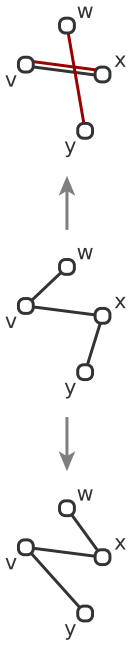
1. How to define the local transformation in such a way that ensures all graphs in the ensemble to be reachable during the random walk (the chain is irreducible, i.e. the state graph is connected)?

2. How to ensure that the chain has a stationary distribution (it is also aperiodic, i.e. the state graph is non-bipartite)?

3. How to construct the chain in such a way that ensures all graphs of the ensemble to be chosen with equal probability (the chain has a uniform stationary distribution, i.e. the state graph is undirected and regular)?

The following sections contain the definition of the corresponding chains for three different ensembles: the fixed degree sequence ensemble for undirected non-bipartite graphs, bipartite graphs, and multiplex bipartite graphs. We discuss whether the chains fulfil the requirements and present in each case a proper extension of the popular *switching algorithm* (also called *rewiring algorithm*). The idea underlying this algorithm is old [219] and represents the basis for several algorithmic and theoretical approaches to graph and matrix generation (see for instance References [211, 63, 34]).

3.3.2 *Undirected non-bipartite graphs*



Initial configuration (middle), disallowed edge swap (top), and allowed swap (bottom)

The switching algorithm uses *edge swaps* as local transformations. In undirected non-bipartite graphs, an edge swap replaces two non-adjacent edges (v, w) and (x, y) either by (v, x) and (w, y) or (v, y) and (x, w) *unless* they are already contained in the graph. Observe that such a swap maintains the degrees of all involved nodes and that it is symmetric in the sense that a swap can be undone by a reverse swap.

Consider a chain whose states are all graphs with a certain degree sequence and in which there is a transition between all those pairs of graphs that can be transformed into each other by one valid edge swap. Because the chain is reversible, its stationary distribution is proportional to the degree of each state. In general, some states have more neighbours than others. Thus, such a chain does not guarantee uniform sampling. There are two popular solutions to overcome this issue: *1)* the *Metropolis algorithm*[8] that allows converting a chain with a given stationary distribution into a chain with another stationary distribution [137, p. 70–73], and *2)* the *self-loop* or *holding method* [63, 177, 16], which introduces self-loops to ensure that all states have the same degree and the state graph is thus regular, which is the requirement for uniform sampling [34].

Several authors resent the fact that the switching algorithm as implemented with self-loops spends too much time "doing nothing" [63, 16]. However, adding self-loops does not require computing any extra information, while any improvement of the method, as well as the Metropolis algorithm, presume knowledge of the number of adjacent states for each state [92] or an estimate of the upper bound for this number [211]. Moreover, this computation is numerically expensive, as small floating point numbers must be handled appropriately to assure a sufficient accuracy. Thus, although the self-loop method needs more steps for convergence, this approach remains efficient in practice. The running time comparison performed by Gionis et al. [92] for bipartite graphs indicates that the Metropolis algorithm is less efficient than the self-loop method. Thus, in this thesis, we opt for the latter and construct the following chain, which produces the desired uniform samples.

Definition 2 (Markov chain for non-bipartite graphs) *(Berger and Müller-Hannemann) Let $\mathcal{M}(\mathcal{G}) = (\mathcal{G}, \mathcal{T}(\mathcal{G}))$ denote a state graph, where $\mathcal{G}$ is an ensemble of undirected non-bipartite graphs with a given degree sequence. The transitions $\mathcal{T}(\mathcal{G})$ are defined as follows. Two graphs $G \neq G' \in \mathcal{G}$ are connected in the state graph if there is a valid edge swap transforming one into the other. Furthermore, for each pair of non-adjacent edges (v, w) and (x, y)*

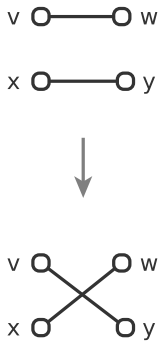
8 In the statistics and mathematics literature, the algorithm is known as the *Metropolis-Hastings algorithm*.

in G *a self-loop is introduced if* $(v,x) \in \mathcal{E}(G) \vee (w,y) \in \mathcal{E}(G) \vee (v,y) \in \mathcal{E}(G) \vee (x,w) \in \mathcal{E}(G)$.

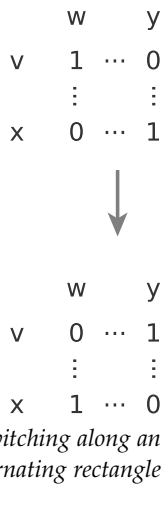
It can be shown that the needed formal conditions are satisfied, i.e. the state graph of this Markov chain is connected, non-bipartite, undirected, and regular, and thus a random walk on it results in asymptotically uniform sampling. For the proof, see for instance Reference [34, sec. 2]. Algorithm 1 is a transcription of the switching algorithm of Berger and Müller-Hannemann [34] and is adapted here to produce a sample $\mathcal{H} \subset \mathcal{G}$ of size $\kappa = |\mathcal{H}|$ as result[9].

The mixing time τ (i.e. the number of steps required to obtain an independent graph from the starting graph) has been the subject of theoretical study, but without any conclusive results (see Reference [34] and the references therein). Thus, it remains an open challenge whether these Markov chains are rapidly mixing. For the practical purposes of this thesis, we empirically check the convergence of the samples for a given real-world network and choose a "safe" τ.

3.3.3 *Bipartite graphs*



Bipartite edge swap



Switching along an alternating rectangle

Due to the prevalence of real-world networks with a bipartite structure as well as the efforts in ecology, data mining, statistics, and sociology, the generation of bipartite graphs with given degree sequences has been extensively researched. By the definition of bipartite graphs, only the edge swap interchanging (v,w) and (x,y) with (v,y) and (x,w) is allowed (where $v,x \in \mathcal{L}$ and $w,y \in \mathcal{R}$), again provided that (v,y) and (x,w) are *not* already contained in the graph. In the context of $(0,1)$-matrices, the edge swap can be viewed as switching along an *alternating rectangle*[10] in the adjacency matrix (i.e. interchanging 0s and 1s): a set of four distinct entries of the type $\{vw, vy, xy, xw\}$, such that the entries are alternately 0s and 1s. Clearly, this switching keeps the row and column sums unaltered (i.e. both degree sequences). Having defined the edge swaps, similarly to the case of non-bipartite graphs presented above, the Markov chain is constructed based on the concept of self-loops [211].

Definition 3 (Markov chain for simplex bipartite graphs) *(Rao et al.) Let* $\mathcal{M}(\mathcal{B}) = (\mathcal{B}, \mathcal{T}(\mathcal{B}))$ *denote a state graph, where* $\mathcal{B}$ *is an ensemble of bipartite simplex graphs with given degree sequences. The transitions* $\mathcal{T}(\mathcal{B})$ *connect two graphs* $B \neq B' \in \mathcal{B}$ *if there is a valid edge swap transforming one into the other. Furthermore, for each pair of non-adjacent edges* (v,w) *and* (x,y) *in* B*, where* $v,x \in \mathcal{L}$ *and* $w,y \in \mathcal{R}$*, a self-loop is introduced if* $(v,y) \in \mathcal{E}(B) \vee (x,w) \in \mathcal{E}(B)$.

9 Note that a similar algorithm can be constructed for directed graphs as well, with the difference that a small number of three-edge swaps is also needed. See References [211, 34] for details.

10 Known as a *checkerboard unit* in the ecological literature (see for example [70, 215, 239, 101, 103, 63, 71]), an *interchange* [219], or a *switching* [249].

Algorithm 1 Uniform sampling of undirected non-bipartite graphs

Input: an undirected graph $G_0 = (\mathcal{V}, \mathcal{E})$, a mixing time τ, a number κ of required graphs
Output: a sample $\mathcal{H} \subset \mathcal{G}$

```
G ← G_0
k ← 0
while k < κ do
    t ← 0
    while t < τ do
        choose uniformly at random (v, w), (x, y) ∈ E(G)
                                        ▷ a pair of non-adjacent edges
        choose with probability 1/2 between case a and b
        if case a then
            if (v, x), (w, y) ∉ E(G) then
                                        ▷ walk to an adjacent graph
                delete (v, w), (x, y) from E(G)
                add (v, x), (w, y) to E(G)
            else                                        ▷ walk a loop
            end if
        else
            if (v, y), (x, w) ∉ E(G) then
                                        ▷ walk to an adjacent graph
                delete (v, w), (x, y) from E(G)
                add (v, y), (x, w) to E(G)
            else                                        ▷ walk a loop
            end if
        end if
        t ← t + 1
    end while
    add G to H
    k ← k + 1
end while
```

The formal properties of this chain have been thoroughly studied in the mathematical literature. Irreducibility was rigorously proven for the first time by Ryser [219, th. 3.1] (alternative proofs have been given later by Rao et al. [211, th. 1], Cobb and Chen [63], and Miklós et al. [175, th. 2.2]). Aperiodicity and uniformity were also shown [211, 63]. Based on this chain, Algorithm 2 produces κ uniform and independent graphs from the ensemble with the degree sequences prescribed by B_0.

Algorithm 2 Uniform sampling of simplex bipartite graphs

Input: a simplex bipartite graph $B_0 = (\mathcal{L} \cup \mathcal{R}, \mathcal{E})$, a mixing time τ, a number κ of required graphs
Output: a sample $\mathcal{H} \subset \mathcal{B}$

```
B ← B_0
k ← 0
while k < κ do
    t ← 0
    while t < τ do
        choose uniformly at random (v, w), (x, y) ∈ E(B)
            ▷ a pair of non-adjacent edges with v, x ∈ L and w, y ∈ R
        if (v, y), (x, w) ∉ E(B) then
                                                ▷ walk to an adjacent graph
            delete (v, w), (x, y) from E(B)
            add (v, y), (x, w) to E(B)
        else                                              ▷ walk a loop
        end if
        t ← t + 1
    end while
    add B to H
    k ← k + 1
end while
```

The sequence of edge swaps is stopped heuristically after τ steps and the obtained graph is considered to be randomly drawn from the ensemble. There are several estimates for the mixing time in the bipartite case. Gionis et al. estimate a number of steps that is linear in the order of magnitude of the edges $\mathcal{O}(|\mathcal{E}|)$ to suffice for convergence [92]. To ensure that the number of performed steps is indeed sufficient, for experiments we always used a strictly super-linear function in the number of edges $\mathcal{O}(|\mathcal{E}|\log|\mathcal{E}|)$.

3.3.4 *Multiplex bipartite graphs*

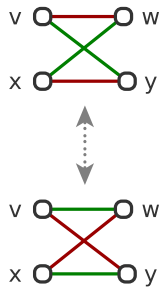
Thesis point 2

Next, we focus on bipartite graphs that contain multiple distinct types of edges. Let $\tilde{B} = (\mathcal{L} \cup \mathcal{R}, \tilde{\mathcal{E}} = \cup_{\gamma \in \Omega} \tilde{\mathcal{E}}_\gamma)$ denote a multiplex bipartite

graph that contains edges of type $\gamma \in \Omega$. If we perceive $\tilde{B}$ as the superposition of individual simplex graphs $\tilde{B}_\gamma$, then we can define for each of the graphs a separate chain as described in Section 3.3.3. Observe, however, that by doing so we allow parallel edges of different types in the superposed graph $\tilde{B}$. In other words, there can exist edges $(v, w) \in \tilde{\mathcal{E}}_\gamma$ and $(v, w) \in \tilde{\mathcal{E}}_\varphi$ with $\gamma \neq \varphi \in \Omega$, $v \in \mathcal{L}$, and $w \in \mathcal{R}$. As the examples from Chapter 8 and Chapter 9 show, for some real-world networks of interest this would be an unnatural characteristic, because there is at most one type of edge admitted between any pair of nodes, i.e. the maximal multiplicity of any edge in the graph is 1: $\max_{(v,w)\in\tilde{\mathcal{E}}} \mu(v, w) = 1$. Thus, we elaborate in the following a Markov chain and a corresponding sampling algorithm that ensure a proper sampling of the fixed degree sequence ensemble $\tilde{\mathcal{B}}$ that contains all graphs in which every node maintains its degree for each of the edge types and no parallel edges of any kind occur.

For sampling from this ensemble, we define the edge swap as follows. We choose uniformly at random two edges of the same type, e.g. $(v, w), (x, y) \in \tilde{\mathcal{E}}_\gamma$ (where $v, x \in \mathcal{L}$, $w, y \in \mathcal{R}$, and $\gamma \in \Omega$). We furthermore ensure that each edge type is sampled with a probability that corresponds to the frequency of this type in the graph. If neither (v, y) nor (x, w) is already contained in the graph, we remove $(v, w), (x, y)$ and add $(v, y), (x, w)$ instead. By always checking for the existence of *any* type of edge between (v, y) and (x, w), we do not allow for multiple edges. The multiplicity of the sampled bipartite graphs thus remains 1. We define the corresponding Markov chain as follows.

Definition 4 (Markov chain for multiplex bipartite graphs) *Let* $\mathcal{M}(\tilde{\mathcal{B}}) = (\tilde{\mathcal{B}}, \mathcal{T}(\tilde{\mathcal{B}}))$ *be a state graph, where* $\tilde{\mathcal{B}}$ *denotes the ensemble of multiplex bipartite graphs. The transitions* $\mathcal{T}(\tilde{\mathcal{B}})$ *connect two graphs* $\tilde{B} \neq \tilde{B}' \in \tilde{\mathcal{B}}$ *if there is a valid edge swap transforming one into the other. Furthermore, for each pair of non-adjacent edges of the same type* $(v, w), (x, y) \in \tilde{\mathcal{E}}_\gamma \, \forall \gamma \in \Omega$ *in* $\tilde{B}$*, where* $v, x \in \mathcal{L}$ *and* $w, y \in \mathcal{R}$*, a self-loop is introduced if* $(v, y) \in \tilde{\mathcal{E}}(\tilde{B}) \vee (x, w) \in \tilde{\mathcal{E}}(\tilde{B})$.



Example of edges that are unswappable due to edges of a different type

To ensure that this Markov chain samples the graphs from the ensemble $\tilde{\mathcal{B}}$ with equal probability, *1)* it needs to be ergodic and *2)* has to have a uniform stationary distribution.

ERGODICITY As we have seen in Section 2.4, a Markov chain is ergodic, if it is aperiodic and irreducible. Due to the introduced self-loops, aperiodicity is assured. As discussed in Section 3.3.3, the bipartite edge swap results in an irreducible chain for simplex graphs. In the multiplex case, it is possible that the irreducibility of the chain becomes violated due to the interference of the subgraphs of different types during the edge swap procedure. A rigorous proof of the necessary and sufficient conditions a graph has to fulfil such that this

situation does not occur is subject to future research. The real-world networks we deal with in this thesis are both extremely sparse and have a long-tail degree distribution. in their case, it is unlikely that we encounter such configurations.

UNIFORMITY We first observe that since the edge swaps are symmetric (i.e. they can be "undone" and thus each transition in the state graph is associated with an inverse transition) the chain is reversible. Hence, the stationary distribution π is proportional to the degrees of the nodes in the state graph. To obtain a uniform distribution, all graphs of the ensemble (the nodes of the state graph) must have the same degree. By construction, the nodes of the state graph contain self-loops for each pair of non-adjacent edges that are unswappable in the graph they represent. The degree of each node in the state graph is then equal to the total number of non-adjacent edges of the same type, which is equal for all graphs, as required. For an illustration see Figure 3.

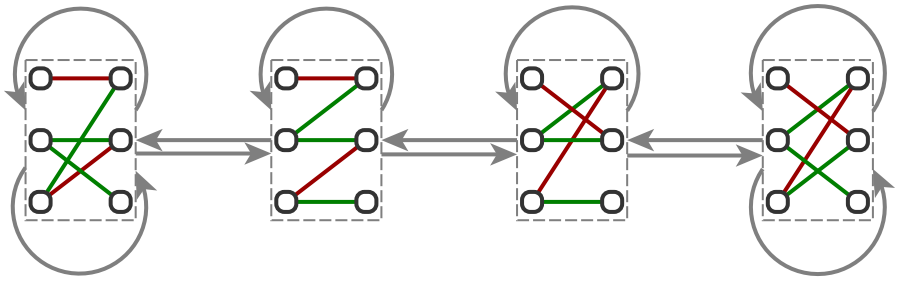


Figure 3: The state graph of the ensemble of multiplex bipartite graphs $\tilde{\mathcal{B}} = (\mathcal{L} \cup \mathcal{R}, \tilde{\mathcal{E}} = \cup_{\gamma \in \Omega} \tilde{\mathcal{E}}_\gamma)$ with $\Omega = \{\text{red,green}\}$ that realize the degree sequences $\mathcal{D}(\mathcal{L})_{\text{red}} = \{1,0,1\}$, $\mathcal{D}(\mathcal{L})_{\text{green}} = \{0,2,1\}$, $\mathcal{D}(\mathcal{R})_{\text{red}} = \{1,1,0\}$, $\mathcal{D}(\mathcal{R})_{\text{green}} = \{1,1,1\}$. The state graph is connected and aperiodic (i.e. the chain is ergodic), as well as undirected and regular (i.e. the stationary distribution of the chain is the uniform distribution).

Based on this Markov chain, we propose Algorithm 3 for sampling.

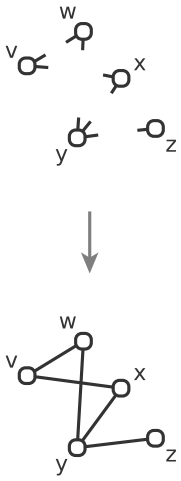
3.3.5 *The configuration model*

The major critique of the Markov chain Monte Carlo sampling concerns the unknown mixing time [128]. If the performed walk is too short, the independence of the graphs is not guaranteed and thus the resulting sample can not be used as a representative null model, because it will probably contain a bias towards the structure of the observed graph. Despite the fact that the Markov chain Monte Carlo sampling is potentially computationally intensive if high accuracy is desired, it is straightforward to implement as well as adaptable to a wide range of applications, and therefore it remains the most popular choice.

Algorithm 3 Uniform sampling of multiplex bipartite graphs

Input: a multiplex bipartite graph $\tilde{B}_0 = (\mathcal{L} \cup \mathcal{R}, \tilde{\mathcal{E}} = \cup_{\gamma \in \Omega} \tilde{\mathcal{E}}_\gamma)$, a mixing time τ, a number κ of required graphs
Output: a sample $\mathcal{H} \subset \tilde{\mathcal{B}}$

$\tilde{B} \leftarrow \tilde{B}_0$
$k \leftarrow 0$
while $k < \kappa$ **do**
  $t \leftarrow 0$
  while $t < \tau$ **do**
    choose at random an edge type $\gamma \in \Omega$ with probability $\frac{|\tilde{\mathcal{E}}_\gamma|}{|\tilde{\mathcal{E}}|}$
    choose uniformly at random $(v, w), (x, y) \in \tilde{\mathcal{E}}_\gamma$
      ▷ a pair of non-adjacent edges of the selected type
    if $(v, y), (x, w) \notin \tilde{\mathcal{E}}(\tilde{B})$ **then**
      ▷ walk to an adjacent graph
      delete $(v, w), (x, y)$ from $\tilde{\mathcal{E}}(\tilde{B})$
      add $(v, y), (x, w)$ to $\tilde{\mathcal{E}}(\tilde{B})$
    else ▷ walk a loop
    end if
    $t \leftarrow t + 1$
  end while
  add $\tilde{B}$ to $\mathcal{H}$
  $k \leftarrow k + 1$
end while



Realization of the degree sequence $\mathcal{D} = \{2, 2, 2, 3, 1\}$ by stub matching

Alternative algorithmic approaches are the stub method [180, 193, 176], sequential importance sampling [61, 10], and exact sampling [88]. From these, the most wide-spread is the stub method, which has the advantages of being conceptually uncomplicated, easy to implement, and analytically tractable through the associated configuration model. We discuss this model in the following, as we will use it to elaborate two of the node similarity measures presented in Section 4.2.

THE STUB METHOD The *stub method* is a matching algorithm that enables generating graphs with a prescribed degree sequence *without* requiring an observed graph with the given sequence. It instead builds the graph from n disconnected nodes that are assigned a number of stubs according to the desired degree sequence, by randomly matching the stubs.

Based on this procedure, in the case of highly skewed degree sequences we can expect nodes with high degree to be matched to each other more than once [129, p. 160]. This in turn changes the topological properties of the graph, especially around the nodes with high degree [169, 282, 283]. Multi-graphs are not allowed in most applications and thus the simple algorithm needs to be refined such as to exclude the formation of multi-edges and self-loops. This can be enforced either *1)* by rejecting those configurations that turn out to be multi-graphs, or *2)* by performing an additional test whenever two stubs are matched and discarding inadmissible matchings.

Both options make the originally computationally inexpensive approach almost infeasible for skewed degree sequences, because they lead to unacceptably many rejected configurations and backtracking steps, respectively. Moreover, the uniformity of sampling is not guaranteed any more. *1)* In the case of rejected configurations, Molloy and Reed argued that, under appropriate conditions on the degree sequence and in the limit of large n, the generated graphs will have equal probability [180]. *2)* When discarding inadmissible matchings, the bias in sampling becomes inevitable [191, p. 436]. Although Milo et al. report examples where this bias is sufficiently small for practical purposes [176], it remains an open question whether the approach is viable in general.

ANALYTICAL APPROXIMATIONS IN THE CONFIGURATION MODEL In the matching algorithm presented above, individual edges are considered to be independent events. This considerable simplification allows the deduction of analytical approximations that lead to the formulation of the *configuration model* [191, p. 434–445].

Consider a non-bipartite graph $G = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = n$ and $|\mathcal{E}| = m$. Based on the construction prescribed by the stub method, the probability that one stub of node v is connected to a stub of node w is $d(w)/(2m-1)$, because there are $2m-1$ possible stubs (we excluded

the one connected to v) from which exactly $d(w)$ belong to node w. Taking into account that v has $d(v)$ stubs, the probability that any of these is connected to a stub of w is $d(v)d(w)/(2m-1)$ [191, p. 439–440]. As $m \to \infty$, the probability becomes:

$$P_{vw} = \frac{d(v)d(w)}{2m} \tag{3.5}$$

Approximating the expected number of common neighbours

The number of common neighbours of v and w is defined as $|\{u \in \mathcal{V} | (v,u) \in \mathcal{E} \wedge (w,u) \in \mathcal{E}\}|$, where $|\cdot|$ denotes the cardinality of the set. To compute the number of common neighbours, we sum over all nodes u that are connected to node v with probability $P_{vu} = d(v)d(u)/2m$, while also being connected to node w with probability $P_{wu} = d(w)(d(u)-1)/2m$ [191, p. 441]:

$$\sum_u P_{vu}P_{wu} = \frac{d(v)d(w)}{(2m)^2}\sum_u d(u)(d(u)-1) = \frac{d(v)d(w)}{n}\frac{\langle d^2\rangle - \langle d\rangle^2}{\langle d\rangle^2} \tag{3.6}$$

where $\langle d\rangle$ is the average degree, $\langle d^2\rangle$ is the average square degree, and according to Equation 2.2, $2m$ can be substituted by $n\langle d\rangle$.

Based on a slightly different probabilistic argumentation, under the same purely random selection mechanism, the expected number of common neighbours of nodes v and w can also be approximated by:

$$\frac{d(v)d(w)}{n} \tag{3.7}$$

Note that this expectation value should actually be $d(v)d(w)/(n-2)$. The above simplification holds for large networks ($n \to \infty$) and is widely used [191, p. 214].

Observe that the difference between the two approximations for the number of common neighbours, Equation 3.6 and Equation 3.7, lies in the term:

$$\frac{\langle d^2\rangle - \langle d\rangle^2}{\langle d\rangle^2} = \left(\frac{\sigma[d]}{\langle d\rangle}\right)^2 \tag{3.8}$$

where $\sigma[\cdot]$ denotes the standard deviation.

The same reasoning can be used to compute the probability of an edge and the expected number of common neighbours in a bipartite graph $B = (\mathcal{L} \cup \mathcal{R}, \mathcal{E})$ with $|\mathcal{L}| = n_{\mathcal{L}}$, $|\mathcal{R}| = n_{\mathcal{R}}$, and $|\mathcal{E}| = m$. There, the probability of an edge between two nodes $v, w \in \mathcal{L}$ is:

$$P_{vw} = \frac{d(v)d(w)}{m} \tag{3.9}$$

while their expected number of common neighbours is:

$$\frac{d(v)d(w)}{n_{\mathcal{R}}}\frac{\langle d_{\mathcal{R}}^2\rangle - \langle d_{\mathcal{R}}\rangle^2}{\langle d_{\mathcal{R}}\rangle^2} \tag{3.10}$$

where $\langle d_{\mathcal{R}} \rangle$ is the average degree of the right side nodes and $\langle d^2_{\mathcal{R}} \rangle$ is their average square degree. In the simpler approximation:

$$\frac{d(v)d(w)}{n_{\mathcal{R}}} \tag{3.11}$$

The configuration model also allows approximating further quantities. As the benefits of such easily computable approximations are immense, it is crucial that we understand just how accurate they are and where they can safely be used. The expressions for the number of common neighbours presented above are relevant for this thesis, because they serve as basis for two of the used node similarity measures (i.e. the covariance and the configuration model-based similarity, see Section 4.2).

3.4 APPLICATIONS OF THE FIXED DEGREE SEQUENCE MODEL

There are several possible practical uses of the fixed degree sequence model in real-world applications. In this thesis, we use random graphs generated by Markov chain Monte Carlo sampling for the significance assessment of network observables and the detection of so-called network motifs, while we employ the stub method to create benchmark graphs. Note that although we presented the fixed degree sequence model for three types of graphs, the following analyses are directly applicable to other types of graphs as well, provided that a suitable random graph model has been defined.

3.4.1 *Significance assessment of network observables*

One of the most important tasks in various applications is to study properties of an observed network in comparison to a null model. The aim is to detect deviations from randomness and thereby to distinguish significant characteristics from expected characteristics. Topological properties of interest are for example the number and size of components, the diameter of the network, or the total number of edges connecting nodes with a given degree.

To perform such an analysis, a uniform sample from the chosen fixed degree sequence ensemble is obtained for example by one of the described algorithms. The observed value is then compared to the distribution defined by the samples. If the difference is significant, the observed value can not be explained by the structure incorporated in the null model (the given number of nodes, edges, and degree sequence), but might rather indicate a functional role, design principle, and/or evolutionary history [169]. For example, appropriate random graphs are used as frame of reference whenever we need to interpret an observed clustering coefficient for which it is impossible to determine *a priori* whether it is "high" or "low" [129, p. 164–165].

While statistics has long been dealing with significance testing [99], in the context of graphs, one often needs to assess the significance of more complex results rather than individual network observables. When testing the significance of a clustering for instance, we can compare the clustering error in the observed graph to the error of the same algorithm applied to the graphs from the ensemble. A significantly higher average error in the ensemble indicates that the observed graph is in fact strongly clustered [92].

Furthermore, the null model approach is used for quantifying the modularity of a node partition in a network (for the definition see Section 2.2). In fact, the modularity measures the difference between the total fraction of intra-cluster edges in the observed network and the *expected* fraction if edges were placed at random. Although different null models of random edge assignment are possible [206], the most popular choice is based on the configuration model, in which edges are placed at random regardless of the underlying partition [94].

3.4.2 *Detection of network motifs*

A firm assumption underlying network analysis is that a network's to follows its function [28]. It is therefore informative to look for identical replicas of small subgraphs, so-called *network motifs* [232, 176, 14], which occur more often than expected under a given null model. In this case it can be assumed that they developed due to evolution or fundamental design principles and limitations [169]. Thereby we correct for those subgraphs which occur in a network with the same basic components but an otherwise random structure. We can differentiate between several types of subgraphs for a given number of involved nodes. One of them, for the case of three nodes, is the directed feed-forward loop, in which A is influencing B and C, while C is influencing B. The feed-forward loop is a characteristic building block of protein networks and performs signal-processing tasks such as pulse generation [206]. The method for the computation of the statistical significance of a subgraph in a network was introduced by Shen-Orr et al. [232, 176]. Feed-forward loops for instance are much more common than expected in transcription networks that control gene expression. Their method can be described as follows:

1. Given a graph G and a network pattern ζ, count the number of occurrences $N_G(\zeta)$ of this pattern in the whole graph.

2. Build a set of graphs with the same degree sequence as G but otherwise randomly distributed edges. In other words, generate a sample $\mathcal{H}$ of the fixed degree sequence ensemble $\mathcal{G}$.

3. Count the number of occurrences of this pattern for all graphs $G' \in \mathcal{H}$ and compute the fraction p of graphs in which the

number of occurrences of this pattern is at least as large as in the original graph G.

This procedure is thus based on the basic principles discussed in Section 3.4.1. The resulting fraction is the empirical approximation of the real p-value. A low p-value implies that the observed occurrence of ζ is less likely to be simply caused by the structure of the data. Instead, it may hint at a functional correlation (for more details see Section 4.3).

The approach of Shen-Orr et al. for detecting network motifs is global in the sense that it takes into account subgraphs placed *anywhere* in the network[11] [232]. In various problems from data mining, molecular biology, and ecology for instance, it is far more interesting to assess the significance of a local pattern in which some of the nodes are fixed. This thesis deals in depth with one typical such pattern formed by two given nodes and their common neighbours (see Section 4.2). The idea is to count the number of common neighbours of the two nodes and obtain a context for this value by repeatedly computing it for several instances of the fixed degree sequence ensemble. While our empirical results show that there are data sets with a lot of structure (Chapter 7, Chapter 8, and Chapter 9), others have very few interesting patterns. The bipartite graph which models disorders and disease genes that are connected by known disorder–gene associations (known as the *Diseasome* [95]) for instance is so sparse that the null model approach fails to detect significant co-occurrences.

3.4.3 *Generation of benchmark graphs*

Because ground truth information that could be used for evaluation is only rarely available, results of diverse analyses are commonly tested against computer-generated benchmark graphs, for which the optimal outcome is defined by construction. Validating an algorithm on artificial graphs is a standard procedure in network analysis and has been used extensively for assessing the quality of clustering methods [94, 136]. In general, benchmark graphs should also be of moderate size and have a structure which resembles the original graph for which ground truth is lacking.

Such artificial benchmark graphs are usually built using a version of the stub method for instance, which guarantees that no multi-edges are created (see Section 3.3.5). Having one realization of the degree sequence, the switching algorithm can then be used to create further instances that can be assumed to represent equally likely random samples of the ensemble. The set of artificial graphs constructed this way enables testing algorithms for their robustness against different

11 Often, only subgraphs occurring at some minimum number of disjoint locations in the graph are declared to be motifs [129, p. 167].

types of noise, such as the random elimination or addition of observations (cf. Section 8.2 and Section 9.2).

3.5 SUMMARY

In the evaluation of network characteristics it is useful to compare the observed values against those which are expected in a null model that fits the data set at hand. By noting the extent and direction of the deviation from the reference distribution, structural biases within a given real-world network can be detected. These biases then provide useful indications of the mechanisms underlying network formation and function. In this chapter we have discussed random graph models that serve as such null models and are suitable for hypothesis testing. The key model we presented here is the fixed degree sequence ensemble, in which the number of nodes, edges, and the degree sequence are set to the values that are specified by an observed real-world network. All graphs that share these properties form the fixed degree sequence ensemble. To generate a uniform sample from this ensemble, we construct a corresponding Markov chain and perform Monte Carlo sampling on this basis.

The most relevant network characteristic for edge inference is node similarity. In the following, we review classic similarity measures and show how to assess the statistical significance of these measures by using the random graph model approach presented in this chapter.

4

NODE SIMILARITY

Similarity measures in general: the main issues and challenges

The notion of the level of similarity between individual entities is central to a variety of analytical problems arising in diverse fields, such as classification and clustering in data mining and machine learning [108, p. 79–114,453–479], searching in text retrieval [222], structure comparison in chemical information retrieval [30, 194], classification of biological species in numerical taxonomy [235, p. 116–147], or sequence alignment and comparison in bioinformatics [143, 181]. Clustering techniques for example attempt to group entities based on the supplied definition of similarity. This similarity, typically interpretable as a *distance*, is fundamental to a successful analysis—insofar as it may be more important than the choice of the clustering algorithm itself [108, p. 459].

Despite the ubiquitous need for the assessment of the alikeness of entities as a basic first step in several fundamental investigations, information about objectively verifiable similarity is only rarely available. However, there are many possibilities for quantifying implicit similarity. Usually, there is no agreement on which is the most correct or insightful measure for a given problem (not even within a certain field). Accordingly, since the late nineteenth century, a great variety of *similarity measures* have been proposed within a multitude of scientific areas. Unfortunately, the lack of communication between these fields resulted in repeated duplication of effort and inconsistency of terminology. From today's perspective, these endeavours raise two main issues:

A. Problem-specific measures have emphasized the need for domain-specific knowledge and have been usually developed to tackle individual problems. Therefore, most of the resulting measures are based on experience with distinct data sets rather than on theoretical arguments.

B. Assessing the quality of individual measures by relying on single data sets is difficult, mainly because: "the choice of a similarity coefficient is largely subjective and often based on tradition or on a posteriori criteria such as the 'interpretability' of the results", as Jackson et al. put it [123]. Moreover, Gordon aptly states that "human ingenuity is quite capable of providing a post hoc justification of dubious classifications" [100]—and not just classifications, we might add.

Definition of a similarity measure

Although similarity is a very convenient concept for humans, its formal definition is not straightforward. Here we provide its most general formulation [218, p. 13].

Definition 5 (Similarity measure) *A function* $s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *is called a similarity measure if* $\forall\, \vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$*:*

$$s(\vec{x}, \vec{y}) = s(\vec{y}, \vec{x}) \tag{4.1}$$

$$s(\vec{x}, \vec{y}) \leqslant s(\vec{x}, \vec{x}) \tag{4.2}$$

$$s(\vec{x}, \vec{z}) \leqslant s(\vec{x}, \vec{y}) + s(\vec{y}, \vec{z}) \tag{4.3}$$

$$s(\vec{x}, \vec{y}) \geqslant 0 \tag{4.4}$$

The function s *is a normalized similarity measure if additionally* $s(\vec{x}, \vec{x}) = 1$.

In this thesis, the focus is on the more limited yet less extensively researched problem of determining topological similarity between nodes in a network based on their position (Section 4.1). Unfortunately, the general difficulties outlined above also hold in the context of node similarity. This chapter is dedicated to taking a closer look at the most common existing measures (Section 4.2). Besides the measures proposed and established in different fields of research, we introduce a measure of our own (Section 4.3).

4.1 WHY AND HOW TO STUDY NODE SIMILARITY?

Main approaches and assumptions

Formulated on a global level, in terms of the question *How similar are two given nodes?*, node similarity is central to clustering. Termed on a local level, *Which other nodes are most similar to a particular node?*, it represents a key issue in recommendation. Moreover, node similarity measures that are used as scoring functions are also well-suited as predictors for the existence or formation of edges (see details in Section 4.5). Despite its broad applicability, node similarity is a concept that received little attention (for notable exceptions see References [139; 191, p. 211–220]) when compared to several other network measures that have been subject to close examination (for example the clustering coefficient [267], degree distribution [25], or centrality indices [131, 124, 132, 39]).

Two nodes can be alike in many respects such as shared external factors or similar position in the network. Here, we concentrate on node similarity based solely on network topology, meaning that no additional node attributes are taken into account (such as the age, gender, location, or occupation of individuals in a social network). In the network analysis literature, the most common approach to constructing mathematical measures for the quantification of ideas of similarity is termed *structural equivalence* and means that two nodes are considered to be similar if they share the same neighbours and

thus occupy the same place in the network [156; 38; 191, p. 211–216]. The main advantage of this idea is its generality. However, it has to be emphasized that it relies on two basic assumptions: *1)* the structure of the network reflects real information about the nodes and this type of similarity is thus well-suited for any network where the function or role of a node is related to its structural surroundings, and *2)* edges in the network indicate fundamental similarity between the nodes they connect. Given these two assumptions, the structural equivalence-based notion of similarity is appropriate for instance in the identification of functional categories or in functional prediction.

There have also been efforts to define similarity without requiring a shared neighbourhood of nodes. In a social network for instance, two students have a similar social position, even if they study at different universities and thus do not share acquaintances. To account for this within the limits of purely network topological data, the similarity of a pair of nodes can be recursively defined in terms of the similarity of their neighbours. This is known as *regular equivalence* and states that two nodes are similar if they are connected to other nodes that are similar themselves [139; 38; 191, p. 217–220].

4.2 CLASSIC NODE SIMILARITY MEASURES

Thesis point 3a

In this section we review a selection of the most well-known and long-standing similarity measures (i.e. the "fittest"), which are based on the idea of structural equivalence. We formulate them both for non-bipartite and bipartite graphs that are undirected and unweighted and thus have a binary adjacency matrix that codes the presence or absence of an edge between two nodes (cf. Section 2.1). As it turns out, in this binary case several similarity measures have alternative set theoretic and contingency table-based forms.

COMMON NEIGHBOURS, COOC Based on the intuition about various types of networks, the most basic indicator of the similarity of two nodes is the number of their common neighbours. For instance, two individuals in a social network who share several acquaintances are likely to have similar domicile, age, or activities; two films that are liked by the same people might be similar in terms of story, cast, and style; two genes regulated by the same transcription factors can be assumed to share sequence similarity or functional role; and two scientific papers that are often cited together may deal with related topics.

The number of common neighbours in a graph theoretical sense is closely related to the so-called *co-occurrence* (hence the notation `cooc`)

defined for diverse relational data[1]. For instance, in text mining terms that appear in the same document or sentence [64; 82, p. 9], in ecology species that habit the same location [102, ch. 7], or in market basket analysis items that are purchased together [152, p. 299–302] are said to co-occur.

The number of common neighbours of nodes v and w is the conjunction of their respective rows from the $(0,1)$-valued adjacency matrix A and can be perceived as the intersection of their neighbourhoods:

$$\mathrm{cooc}(v,w) := |\mathcal{N}(v) \cap \mathcal{N}(w)| \tag{4.5}$$

where $\mathcal{N}(v)$ is the neighbour set of node v and $|\cdot|$ denotes the cardinality of the set.

There is a practical alternative to this set theoretic formulation based on the observation that the number of common neighbours of two nodes is equivalent to the number of distinct paths of length two between them in the graph. Accordingly, the number of common neighbours of nodes v and w in an undirected graph with n nodes is the vw-th element of the second power of the adjacency matrix A (cf. Section 2.1). Note that the diagonal elements of A^2 contain the degrees of the individual nodes[2].

Based on these set theoretic and algebraic formulations and due to the nature of the similarity measures we discuss in the following, we define the number of common neighbours of nodes v and w in an undirected graph as the scalar product of their respective adjacency rows:

$$\mathrm{cooc}(v,w) := A_v \cdot A_w = \sum_{u=1}^{n} A_{vu} A_{wu} \tag{4.6}$$

A key property of the number of common neighbours is that it is bounded by the smaller degree of the involved nodes:

$$0 \leqslant \mathrm{cooc}(v,w) \leqslant \min\{d(v), d(w)\} \tag{4.7}$$

Thus, cooc has the shortcoming that the expectation is larger for high degree nodes sharing only a small percentage of their neighbours than for small degree nodes with a relatively large neighbourhood

1 While the number of common neighbours, as defined here, refers to two individual nodes, the co-occurrence is often used more broadly to incorporate multiple entities, i.e. a set of co-occurring nodes.

2 Citation analysis, the area that is preoccupied with the analysis of the directed network of papers citing each other, draws a distinction between two measures for the similarity of papers, both of which are actually equivalent to the number of common neighbours. The *co-citation* is equal to the number of other papers that cite them both (i.e. it is based on the incoming edges) and is computed as $A^T A$, where A^T is the transpose of the matrix A. The *bibliographic coupling* is equal to the number of other nodes to which both point (i.e. it is based on outgoing edges) and is computed as AA^T [191, p. 115–118; 270, p. 90–91].

overlap. The majority of the wide-spread similarity measures thus normalizes the number of common neighbours to compensate for the degrees such that the similarity is 1 for perfectly overlapping neighbourhoods. A prominent example of this is the Jaccard coefficient we discuss next.

JACCARD COEFFICIENT, jac One of the oldest similarity measures was proposed by Jaccard to measure the alikeness of two sets [122] and it has been heavily used ever since[3]. In a network setting, the Jaccard coefficient normalizes the intersection of the neighbour sets of nodes v and w by the cardinality of their union:

$$\mathrm{jac}(v,w) := \frac{|\mathcal{N}(v) \cap \mathcal{N}(w)|}{|\mathcal{N}(v) \cup \mathcal{N}(w)|} \tag{4.8}$$

where $\mathcal{N}(v)$ is the neighbour set of node v. Expressed in terms of the number of common neighbours and the degrees $\mathrm{d}(v)$ and $\mathrm{d}(w)$, the above formula can be rewritten as:

$$\mathrm{jac}(v,w) = \frac{\mathrm{cooc}(v,w)}{\mathrm{d}(v) + \mathrm{d}(w) - \mathrm{cooc}(v,w)} \tag{4.9}$$

The Jaccard coefficient assigns values from the interval $[0,1]$ and reaches its optimal value of 1 when the two nodes have exactly the same neighbours. Note that it is undefined for nodes with degree 0. In this case, it could be explicitly defined to be 0.

COSINE SIMILARITY, cos The cosine similarity[4] or Salton's cosine [222] gives the angular cosine distance between the n-dimensional vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$:

$$\cos(\vec{x},\vec{y}) := \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|_2 \, \|\vec{y}\|_2} \tag{4.10}$$

where $\cdot$ denotes the scalar product of two vectors and $\|\cdot\|_2$ is the *Euclidean norm* defined as $\|\vec{x}\|_2 = \sqrt{\sum_{k=1}^{n} x_k^2}$. Note that the cosine similarity is invariant against scaling of the vectors:

$$\cos(\alpha\vec{x},\vec{y}) = \cos(\vec{x},\vec{y}) \quad \forall\, \vec{x},\vec{y} \in \mathbb{R}^n \text{ and } \alpha \in \mathbb{R}_+ \tag{4.11}$$

Therefore, it is extensively used in data mining, where the relative distribution of the feature values needs to be taken into account. Its applications range from document comparison to collaborative filtering (see for instance References [163, 151]).

The cosine similarity of nodes v and w is computed from their corresponding rows in the adjacency matrix:

3 In a binary setting such as this, it is equivalent to the *Tanimoto coefficient* [248].

4 Also known as the *Ochiai coefficient* [195].

$$\cos(v,w) = \frac{A_v \cdot A_w}{\|A_v\|_2 \, \|A_w\|_2} = \frac{\sum_u A_{vu} A_{wu}}{\sqrt{\sum_u (A_{vu})^2} \sqrt{\sum_u (A_{wu})^2}} = \frac{\mathrm{cooc}(v,w)}{\sqrt{d(v)d(w)}} \tag{4.12}$$

Accordingly, the cosine similarity equals the number of shared neighbours $\mathrm{cooc}(v,w)$ when normalized by the geometric average of the degrees $d(v)$ and $d(w)$. Its value lies in the range from 0 to 1 with 1 indicating a perfect overlap between the neighbourhoods. As in the case of the Jaccard coefficient, cosine is also undefined for zero vectors and requires an explicit definition for this case.

COVARIANCE, cov The covariance is a measure of the degree of independence of two n-dimensional vectors $\vec{x}$ and $\vec{y}$ and is defined as[5]:

$$\mathrm{cov}(\vec{x},\vec{y}) := \frac{1}{n} \sum_{u=1}^{n} (x_k - \langle \vec{x} \rangle)(y_k - \langle \vec{y} \rangle) \tag{4.13}$$

where $\langle \vec{x} \rangle$ denotes the average of the elements of $\vec{x}$. If the covariance yields a large positive value, there is a strong positive linear dependency between $\vec{x}$ and $\vec{y}$. In other words, vector components with high values coincide with high component values, and low component values coincide with low component values. If it yields a large negative value, there is a strong negative linear dependency. Accordingly, high component values coincide with low component values and vice versa.

Analogously, the covariance between the adjacency rows corresponding to nodes v and w can be defined as:

$$\begin{aligned} \mathrm{cov}(v,w) &= \frac{1}{n} \sum_{u=1}^{n} (A_{vu} - \langle A_v \rangle)(A_{wu} - \langle A_w \rangle) && (4.14) \\ &= \frac{1}{n} \left(\sum_{u=1}^{n} A_{vu} A_{wu} - n \langle A_v \rangle \langle A_w \rangle \right) && (4.15) \\ &= \frac{1}{n} \left(\mathrm{cooc}(v,w) - \frac{d(v)d(w)}{n} \right) && (4.16) \end{aligned}$$

where we used Equation 4.6 and the fact that the average of the elements in row A_v is $\langle A_v \rangle = n^{-1} \sum_{u=1}^{n} A_{vu} = d(v)/n$. The normalization contained in the resulting formula suggests an intuitive way of accounting for the number of common neighbours the two nodes *would* have if both would choose their neighbours purely at random. As Equation 3.7 shows, under the assumption that the nodes "choose" independently from each other, their expected number of common neighbours is indeed $d(v)d(w)/n$.

5 The normalization factor required for sample covariance with unknown average is $n-1$. For large graphs, however, the approximation with n is sufficient.

	$A_{wu} = 1$	$A_{wu} = 0$
$A_{vu} = 1$	$\mathbf{a} = cooc(v, w)$	$\mathbf{b} = d(v) - cooc(v, w)$
$A_{vu} = 0$	$\mathbf{c} = d(w) - cooc(v, w)$	$\mathbf{d} = n + cooc(v, w) - d(v) - d(w)$

Table 1: Contingency table allowing the comparison between the neighbourhoods of nodes v and w: **a** is the number of common neighbours; **b** is the size of the exclusive neighbourhood of node v; **c** is the size of the exclusive neighbourhood of node w; and **d** is the number of nodes that are not adjacent to neither v nor w.

Note that in the case of bipartite graphs, the expected number of common neighbours of nodes $v, w \in \mathcal{L}$ is $d(v)d(w)/n_{\mathcal{R}}$, where $n_{\mathcal{R}}$ denotes the number of nodes in the opposite set (see Equation 3.11). Thus, their covariance is:

$$cov(v, w) = \frac{1}{n_{\mathcal{R}}} \left(cooc(v, w) - \frac{d(v)d(w)}{n_{\mathcal{R}}} \right) \tag{4.17}$$

This concept of normalization is analogous to a key concept in market basket analysis and is crucial for the investigation of patterns of the form *If customers bought a set of products* $\mathcal{Q}_1$ *then they also purchased the set of products* $\mathcal{Q}_2$. For two disjoint sets of products $\mathcal{Q}_1$ and $\mathcal{Q}_2$, this pattern is formulated as a so-called *association rule* $\mathcal{Q}_1 \to \mathcal{Q}_2$, meaning that if a data set contains the products in $\mathcal{Q}_1$ then it also contains the products in $\mathcal{Q}_2$. To assess the meaningfulness of such implications, Shapiro introduced a measure called *leverage* [247], which is defined as the difference between the joint probability of $\mathcal{Q}_1$ and $\mathcal{Q}_2$'s occurrence $P(\mathcal{Q}_1, \mathcal{Q}_2)$ (i.e. their co-occurrence) and their expected probability if they were appearing independently from each other:

$$P(\mathcal{Q}_1 \to \mathcal{Q}_2) = P(\mathcal{Q}_1, \mathcal{Q}_2) - P(\mathcal{Q}_1)P(\mathcal{Q}_2) \tag{4.18}$$

An equivalent formula that quantifies the excess of observed co-occurrence over the expected co-occurrence has been derived for binary data in text retrieval from a 2x2 *contingency table* representation (see Table 1). The determinant of this contingency table is proportional to the covariance: $\mathbf{ad} - \mathbf{bc} = n\, cooc(v, w) - d(v)d(w)$. For further measures deduced from this type of representation see References [75, 38].

The covariance (alongside the equivalent leverage) is intuitive and can be used successfully in many settings. However, it lacks a proper standardization and therefore it is unclear what constitutes a "good" covariance value. The Pearson correlation presented in the following suggests a possible normalization. The motivation behind this normalization is the drawback of the covariance that if a vector is multiplied by a constant factor α, then the covariance between the vector and any other vector will increase by α. The normalization used in

the Pearson correlation (i.e. the division of the covariance by the standard deviation of both vectors) compensates for this effect of constant scaling.

PEARSON CORRELATION, r The Pearson correlation or Pearson product-moment correlation is widely used to measure linear dependency between vectors and allows for several interpretations [217]. It is commonly used in traditional collaborative filtering approaches [270, p. 17] and recently also in systems biological analyses (see for example Reference [165]).

The Pearson correlation is often used to measure node similarity based on the corresponding rows of the adjacency matrix and is simply computed by rescaling the covariance [191, p. 214–215]:

$$r(v,w) := \frac{cov(v,w)}{\sigma[v]\sigma[w]} = \frac{1}{n\sigma[v]\sigma[w]}\left(cooc(v,w) - \frac{d(v)d(w)}{n}\right) \tag{4.19}$$

where $\sigma[v]$ denotes the standard deviation of the adjacency row of node v. The normalization makes the interpretation of the Pearson correlation more tractable than that of the covariance. Its values lie in the interval $[-1,1]$ with $r \approx 1$ being a strong positive correlation, $r \approx -1$ meaning a strong negative correlation, and $r \approx 0$ indicating linear independence.

Two final remarks are in order regarding the Pearson correlation. First, because the covariance takes a slightly different form for bipartite graphs, the Pearson correlation changes accordingly and becomes:

$$r(v,w) = \frac{1}{n_{\mathcal{R}}\sigma[v]\sigma[w]}\left(cooc(v,w) - \frac{d(v)d(w)}{n_{\mathcal{R}}}\right) \tag{4.20}$$

Second, observe the relationship between the Pearson correlation and the cosine similarity: for centered adjacency rows, meaning that $\langle A_v \rangle = \langle A_w \rangle = 0$, the two are equivalent to each other [217].

HYPERGEOMETRIC COEFFICIENT, hyp The cumulative hypergeometric distribution has been used extensively to measure the significance of the number of common neighbours in biochemistry and systems biology (see for instance [253, 242, 96, 243]). The hypergeometric distribution assumes a null model, according to which the neighbourhoods of the two nodes are independent, but the common neighbours are chosen without replacement. It is computed from the degrees of the two nodes and the total number of nodes as follows:

$$hyp(v,w) := -\log \sum_{c=cooc(v,w)}^{\min\{d(v),d(w)\}} \frac{\binom{d(v)}{c}\binom{n-d(v)}{d(w)-c}}{\binom{n}{d(w)}} \tag{4.21}$$

The sum gives the probability of obtaining a number of common neighbours which is at least as large as the actual number of common neighbours and thus, it can be interpreted as a p-value. Introducing the negative logarithm maps this value to $\mathbb{R}_+$.

Note that in the case of bipartite graphs with $|\mathcal{R}| = n_{\mathcal{R}}$, the hypergeometric coefficient for two nodes $v, w \in \mathcal{L}$ is:

$$\mathrm{hyp}(v,w) = -\log \sum_{c=\mathrm{cooc}(v,w)}^{\min\{d(v),d(w)\}} \frac{\binom{d(v)}{c}\binom{n_{\mathcal{R}}-d(v)}{d(w)-c}}{\binom{n_{\mathcal{R}}}{d(w)}} \tag{4.22}$$

CONFIGURATION MODEL-BASED SIMILARITY, cfm Leicht et al. suggested to normalize the observed number of common neighbours by its expected value in the configuration model [139, eq. 19]. As discussed in Section 3.3.5, the configuration model is a random graph model in which individual edges are considered to be independent events. Using Equation 3.6, we obtain the configuration model-based similarity for the nodes v and w in a non-bipartite graph as:

$$\mathrm{cfm}(v,w) := \frac{\mathrm{cooc}(v,w)}{\frac{d(v)d(w)}{n}\frac{\langle d^2\rangle - \langle d\rangle^2}{\langle d\rangle^2}} \tag{4.23}$$

where $\langle d\rangle$ is the average degree, $\langle d^2\rangle$ is the average square degree, and n denotes the number of nodes. As we compare the similarities within the same network, we can neglect the multiplicative constants [139] and obtain:

$$\mathrm{cfm}(v,w) := \frac{\mathrm{cooc}(v,w)}{d(v)d(w)} \tag{4.24}$$

Note that the configuration model-based similarity is defined analogously for bipartite graphs by taking into account Equation 3.10. After discarding the multiplicative constants, we obtain an expression that is equivalent to Equation 4.24.

The node similarity measures listed above (for an overview see Table 2) are only a few examples of the many existing possibilities. For a series of additional measures and discussions see for instance References [235, p. 116–147; 75; 218, p. 13–14]. As we see severe limitations in the null model adopted by these measures, in the following we discuss other, more accurate measures based on the fixed degree sequence model (see Section 3.2.1).

4.3 SIMILARITY BASED ON THE FIXED DEGREE SEQUENCE MODEL

Thesis point 3b

The more simplistic raw measures (such as common neighbours, Jaccard coefficient, or cosine similarity) suffer from the fact that it is unclear what constitutes "good" values, as they all depend intimately on

SIMILARITY MEASURE	FORMULA	RANGE	OPTIMAL VALUE
common neighbours, cooc	$A_v \cdot A_w$	$\mathbb{N}_0$	∞
Jaccard coefficient, jac	$\frac{cooc(v,w)}{d(v)+d(w)-cooc(v,w)}$	$[0,1]$	1
cosine similarity, cos	$\frac{cooc(v,w)}{\sqrt{d(v)d(w)}}$	$[0,1]$	1
covariance, cov	$\frac{1}{n}\left(cooc(v,w) - \frac{d(v)d(w)}{n}\right)$	$[-1,1]$	1
Pearson correlation, r	$\frac{cov(v,w)}{\sigma[v]\sigma[w]}$	$[-1,1]$	1
hypergeometric coefficient, hyp	$-\log \sum_{c=cooc(v,w)}^{\min\{d(v),d(w)\}} \frac{\binom{d(v)}{c}\binom{n-d(v)}{d(w)-c}}{\binom{n}{d(w)}}$	$\mathbb{R}_+$	∞
configuration model-based similarity, cfm	$\frac{cooc(v,w)}{d(v)d(w)}$	$[0,1]$	1
p-value, p	$\frac{\lvert\{G\in\mathcal{H} \mid cooc_G(v,w)\geqslant cooc(v,w)\}\rvert}{\lvert\mathcal{H}\rvert}$	$[0,1]$	0
z-score, z	$\frac{cooc(v,w)-\langle cooc_G(v,w)\rangle}{\sigma[cooc_G(v,w)]}$	$\mathbb{R}$	∞
presorted z-score, z^*	*see text*	$\mathbb{N}_+$	∞

Table 2: Summary of the presented similarity measures. Bold measures have a slightly different formula when adapted to bipartite graphs (see text for details).

the number of common neighbours and the degree of the nodes. Adjusted measures (like the hypergeometric coefficient, the covariance, or the Pearson correlation) try to alleviate this problem by accounting for the similarity that would be obtained under a given null model. The null models that are implied by these measures contain independence assumptions that are potentially overly simplistic in the case of many real-world networks. Based on this classic statistical approach, in the following we look at node similarity measures that are generalized to arbitrarily complicated null models. Without making any assumptions about the distribution of the similarity values, we propose a non-parametric significance assessment using permutation tests[6].

EMPIRICAL P-VALUE, p Based on the fixed degree sequence model, a common way of assessing the statistical significance of the observed number of common neighbours of v and w is to count the fraction of sampled graphs in which their number of common neighbours is at least as large as the observed value $cooc(v, w)$:

$$p(v, w) := \frac{|\{G \in \mathcal{H} \mid cooc_G(v, w) \geqslant cooc(v, w)\}|}{|\mathcal{H}|} \tag{4.25}$$

where $cooc_G(v, w)$ is the number of common neighbours of v and w in a graph $G \in \mathcal{G}$ from the ensemble $\mathcal{G}$ that contains all graphs with the same degree sequence as the observed graph (see for instance Reference [272]). Since the graph ensemble can not be exhaustively enumerated, an empirical p-value is computed based on the random sample $\mathcal{H} \subset \mathcal{G}$. This empirical measure approximates the true p-value. The observed number of common neighbours is thus corrected with the expected number of common neighbours in $\mathcal{H}$. We have presented methods for generating this sample in Section 3.3.

The smaller the p-value, the more unlikely it is that the two nodes have the observed number of common neighbours in a random graph and thus the more significant the observation.

The numerical estimation of the p-values in the tail of the distribution (i.e. where many of our points of interest lie) requires a large sample [252]. In contrast, calculating the z-scores, as described next, requires sampling the first two moments of the distribution, namely the average and the standard deviation.

z-SCORE, z Assuming Poisson distributed numbers of common neighbours, a Gaussian distribution is a good fit for nodes with large degree and thus the z-score can be used as test statistic instead of the empirical p-value. The z-score quantifies the deviation from the sample average in units of standard deviation. Applied to the number of common neighbours, in the fixed degree sequence model we obtain:

$$z(v, w) := \frac{cooc(v, w) - \langle cooc_G(v, w) \rangle}{\sigma[cooc_G(v, w)]} \tag{4.26}$$

6 For a comprehensive study of the permutation approach see the book by Good [99].

where the notation is consistent with that in Equation 4.25, $\langle\cdot\rangle$ denotes the sample average, and $\sigma[\cdot]$ denotes the sample standard deviation.

Our tests on real-world data show that the distribution of the number of common neighbours is asymptotically Gaussian for large degree nodes. For smaller degree nodes however, the distribution of the number of common neighbours may have a heavy tail. In this case, the probability of obtaining extreme z-scores can be orders of magnitude higher than in the Gaussian distribution. This could be the explanation for the astronomically large z-scores reported in the literature (see for example References [176, 272, 85]). Nevertheless, it is believed that the Gaussian approximation is frequently sufficient to gauge statistical significance [252].

PRESORTED z-SCORE, z^* For obtaining accurate p-values, a large sample is needed. Thus, a p-value that is near 0 can indicate true significance or the fact that the number of taken samples is not sufficient. In the range of low p-values, it can be thus rewarding to use the z-score, because it differentiates better than the p-value when computed based on a sample of the same size. Therefore, we propose to assess the similarity of nodes based on the combination of these two measures. Let $<_p$ and $<_z$ denote the order of the node pairs according to p-value and z-score, respectively. Then, we define the order $<_{z^*}$ as:

$$\begin{aligned}(v,w) <_{z^*} (v',w') \iff &\{(v,w) <_p (v',w')\} \vee \\ &\{(v,w) =_p (v',w') \wedge (v,w) <_z (v',w')\}\end{aligned} \tag{4.27}$$

Accordingly, z^* does not provide a similarity score for two nodes, but results in a ranking of the pairs of nodes. As we will see in Chapter 7, this combined measure outperforms the others on all tested networks.

4.4 NODE SIMILARITY IN MULTIPLEX GRAPHS

Thesis point 4

Note that the presented measures can be extended to quantify node similarity in multiplex graphs. Let $\tilde{G} = (\mathcal{V}, \tilde{\mathcal{E}} = \cup_{\gamma\in\Omega}\tilde{\mathcal{E}}_\gamma)$ denote a multiplex graph. The co-occurrence of two nodes $v, w \in \mathcal{V}$, denoted by $\mathrm{coocc}_{\gamma\varphi}(v,w)$, then equals the number of common neighbours $u \in \mathcal{V}$ they share with respect to the edge types $\gamma, \varphi \in \Omega$:

$$\mathrm{cooc}_{\gamma\varphi}(v,w) = |\{u \in \mathcal{V} | (v,u) \in \tilde{\mathcal{E}}_\gamma \wedge (w,u) \in \tilde{\mathcal{E}}_\varphi\}| \tag{4.28}$$

The co-occurrence thus signifies how often we observe an edge of type γ between v and u together with an edge of type φ between w and u. Clearly, $\mathrm{coocc}_{\gamma\gamma}(v,w)$ is equivalent to the co-occurrence of the nodes v and w computed in the simplex subgraph $\tilde{G}_\gamma$.

Based on this multiplex co-occurrence, we can then define the fixed degree sequence model-based p-value and z-score, provided that the

proper multiplex graph ensemble is used as null model. For the case of multiplex bipartite graphs see Section 3.3.4.

4.5 NODE SIMILARITY MEASURES IN EDGE INFERENCE

In the context of this thesis, the node similarity measures which we presented above are used for edge validation and prediction as follows. We compute the measures for each pair of nodes we are interested in and identify either by unsupervised or supervised learning, which pairs of nodes should be connected.

A. In the unsupervised case, individual measures serve as scoring functions [147]. The node pairs are ranked according to the chosen measure and those pairs with the highest score are predicted to be edges.

B. In the supervised setting, several measures collectively form the features, based on which a learning algorithm is trained [148].

In both cases, the underlying assumption is that there exists a meaningful correlation between the structure of the network and the mechanisms responsible for the lack of a given edge. Similarity measures based on structural equivalence thus possess predictive power in this context [129, p. 201–202]. To which extent this assumption holds and for which prediction tasks the diverse measures can be used has been scarcely addressed in the literature. In Chapter 7 we provide a contribution to this problem by empirically testing the diverse measures on networks from two very different areas and thereby address this important aspect of node similarity.

The measures discussed here assess local topology in the neighbourhood of the nodes. However, measures considering a broader structure can also be constructed. For instance, the similarity of two nodes can be quantified as the negative shortest-path distance between them [129, p. 201] or the number of paths of at most a given length [147]. Different variants based on the number of common neighbours can also be designed by weighing the common neighbours by their degree for example [7] or by their clustering coefficient as we suggest in Chapter 6.

4.6 SUMMARY

In this chapter we have presented a set of the best-known node similarity measures: common neighbours, Jaccard coefficient, cosine similarity, covariance, Pearson correlation, hypergeometric coefficient, and three futher measures based on the fixed degree sequence model. These measures rely solely on network topology and are derived from the number of common neighbours shared by the two nodes

of interest. They range from simple to those that correct for the expected value of the number of common neighbours under a given null model. While in some cases this null model is rather simplistic, we have also presented two measures that are test statistics based on the fixed degree sequence model. In addition to the expository part, we introduced a new measure that combines the latter two. As the presented similarity measures differ conceptually and in terms of the required computational effort, we argue that the choice of measure for a given problem should be subject of careful consideration and present a comparative survey in Chapter 7. In the case of edge prediction, these measures can be used individually as scoring functions or they can serve as features in a classification setting, as we show in the next chapter.

5 CLASSIFICATION WITH RANDOM FORESTS

This chapter introduces classification problems (Section 5.1), elaborates, how these can be approached using decision trees (Section 5.2), and describes the classifier algorithm called random forest (Section 5.3). Additionally, two important related issues are reviewed, namely feature selection (Section 5.4) and training schemes (Section 5.5). Although the corresponding machine learning literature is vast (see for instance Reference [108]), we focus solely on the aspects that are relevant to this thesis.

5.1 CLASSIFICATION

Classification[1] is a supervised learning task that assigns entities to classes based on a labelled feature data $D = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_e, y_e)\}$, where e denotes the number of examples, $\vec{x}_i \in \mathbb{R}^q$ is the q dimensional feature vector corresponding to entity i, and $y_i \in \{1, \ldots, c\}$, $c \geqslant 2$ is the class label of i. Using the data D, a classifier is a function $g : \mathbb{R}^q \to \{1, \ldots, c\}$ that yields a class y for a given feature vector $\vec{x}$ [218, p. 85–86]. For this thesis only *binary classification* is relevant, meaning that $c = 2$.

Wide-spread approaches

Several different classifiers exist in the literature, each with specific features and drawbacks. Among the most popular are the naive Bayes classifier, the nearest neighbours classifiers, and the support vector machine (SVM) [108, p. 21,415–423,371–376]. By default, these methods consider all features for analysis, which has a twofold drawback: *1)* not all features are actually informative and *2)* the assessment of all features requires great computational effort. To counteract this, one could rank the features by their importance based on the already observed feature data and use only a subset of the possible features [218, p. 97]. This represents the basic idea of decision trees, which is elaborated in the following.

5.2 DECISION TREES

Classification by *decision trees* has become very popular, not least due to the fact that trees are intuitive and computationally inexpensive. The cost of building a balanced binary tree of n elements is $\mathcal{O}(n \log n)$ and traversing it from its root to a leaf node is in $\mathcal{O}(\log n)$ [66, p. 286–292]. This is important in the context of classification, since the trained

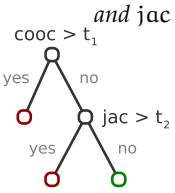
1 For a short introduction to classification and the related basic concepts see Section 2.6.

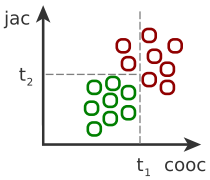
Using decision trees

decision tree classifier is queried, i.e. traversed, for each new entity that we intend to classify.

The basic idea of decision trees is to break classification down to a set of choices about each feature in turn starting at the root of the tree and progressing down to the leaves, where the final classification decision is made. In case of numeric features, each node of the tree contains a simple test of the form *Is the value of the* i*-th feature lower than threshold* t_i*?*, and all leaves contain a deterministic class label.

Decision tree constructed based on the features cooc *and* jac





Separation of the feature space in correspondence to the tree

While there are several different options for constructing the tree, all of them are based on the same principle: the tree is built in a greedy manner, starting at the root and choosing at each node the most informative feature that best enables splitting the examples into two further nodes. The amount of information can be quantified based on the *information entropy* H for instance, which describes the impurity in a set of features [167, p. 135] and is defined as:

$$H(p) = -\sum_c p_c \log_2 p_c \tag{5.1}$$

where p_c are probabilities associated with the classes. In binary classification, if p is the proportion of examples in class 1, then:

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p) \tag{5.2}$$

It can be shown that the optimal feature to pick at a node has maximal entropy, as this feature separates the examples optimally. The information gain of a feature describes how much the entropy of the training set would decrease if we selected this particular feature. Based on it, the feature is chosen that results in the highest gain.

Therefore, we employ a greedy procedure that searches the space of possible trees by choosing the feature with the highest information gain at each node, given what is already known. The tree is constructed recursively: at each node the best feature is selected and removed from the data set and the algorithm is called on the rest until there are no features left or there is only one class remaining in the data. In the first case, the most common label is added to the node, in the second, a leaf is added with that class as its label.

OVERFITTING AND PRUNING *Overfitting* is one of the major problems for statistical learning models in general and it can have a significant impact on decision trees as well. Overfitting means that instead of the underlying feature–output relation, the noise in the training data is modelled [218, p. 74]. The algorithm therefore memorizes the training data instead of finding a generalizing model. To avoid overfitting a decision tree, the size of the tree should be limited. A common advanced technique which achieves this is called *pruning*.

The idea of pruning is to compute the full tree and then reduce it while monitoring the induced error. While there are many different versions of pruning, the most basic idea works as follows: First,

the tree is built using all features. Smaller trees are then produced by iterating over the individual nodes and replacing the subtree below them with a leaf labelled with the most common label. The error of the pruned tree is calculated on the validation set and the pruning is accepted if the error is no larger than when using the original tree [167, p. 142].

5.3 RANDOM FORESTS

The *random forest* classifier was introduced by Breiman and consists of an ensemble of randomized decision trees. The basic idea behind it is to aggregate the results of a set of "weak classifiers" (simple decision trees) to form a "strong classifier" (the forest). The main advantage of the random forest is that, as opposed to single trees, it is robust to overfitting without using additional techniques like pruning [47].

The past few years have seen an increased interest in random forests. Most importantly, they have already been successfully used for edge prediction [58, 148, 265] and are thus very attractive for the purposes of this thesis.

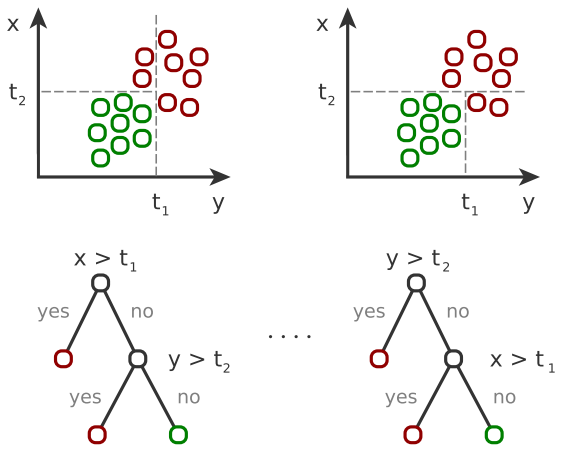


Figure 4: Two exemplary decision trees that form the random forest classifier. The two-dimensional feature space (top) is partitioned according to the splits made at the nodes of the individual trees (bottom). The final classification is based on majority votes.

Constructing the forest

As sketched in Figure 4, the random forest is constructed by drawing n_{tree} bootstrap samples from the training data. For each of the samples, a decision tree is built with the modification that at each node, rather than choosing the best binary split among *all* features, a split is selected based solely on a random sample of features. Thus, for each tree the following procedure is repeated: We start with the

root node that contains all training examples. At each node, the feature that best splits the examples in the node is selected from a randomly chosen subset of size m_{try} of the q available features (as a rule of thumb, $m_{try} = \sqrt{q}$ [47]) and an adequate threshold is found. The tests at each node of the tree are selected by first creating a set of random tests and then picking the best among them according to some quality measure (for instance the above presented information gain or the Gini index). Subsequently, the node is split and the process is recursively repeated for each of the resulting subsets until a pure subset with examples from only one class is produced [106].

Using the forest

Given a set of validation data, all trees of the forest are traversed from their roots to one of their leaves. The trees that vote for a certain class are counted. This can be interpreted as the posterior probability of a certain example belonging to a particular class, given its features. A crisp classification can be obtained by taking the majority votes.

5.4 FEATURE SELECTION

As we have seen in Chapter 4, there are several feature candidates that quantify node similarity and could therefore be used in an edge prediction setting. Usually, the question of which features to choose for a specific classification task is not easy [167, p. 221]. The *feature selection* problem is thus concerned with finding the most influential subset of features from a much larger set of potential candidates. Taking into account too many features increases the computational cost, may lower the performance due to overfitting, and entails the *curse of dimensionality*. The latter refers to the phenomenon where, as the number of features (i.e. input dimensions) grows, more data is required to assure that the algorithm generalises sufficiently well. Moreover, the number of required examples is increasing super-linearly with the number of used features [167, p. 106–108].

Feature selection typically involves looking through the available set of features and testing their usefulness, i.e. whether they are correlated to the desired output. Alternatively, the following trial-and-error procedure can also be rewarding: First, we choose subsets of the available features and use them for training. Then, we adapt the subsets successively according to the obtained performance until a satisfactory feature subset is identified. Unfortunately, finding an optimal subset is difficult, since all possible combinations need to be tested, and this in turn implies an exponential effort, causing feature selection to belong to the group of NP-hard problems [173]. All in all, it is still unclear how to best handle the trade-off between computational volume and accuracy involved in feature selection such that a

reliable separation between the classes is obtained based on the available set[2].

5.5 CROSS-VALIDATION

Based on ground truth data, the performance of a classifier can be evaluated by *cross-validation*, a training scheme that is designed to avoid overfitting. The idea here is to partition the data D into a disjoint training set D_t and validation set D_v, where $D_t, D_v \in D$, $D_t \cap D_v = \emptyset$, $D_t \cup D_v = D$. D_t is used to train the algorithm and D_v serves for validation. Since the validation data is not used for training, a high performance on it indicates that the algorithm achieves a good classification [218, p. 74–75]. The implicit assumption is of course that the two sets are independent.

A popular cross-validation scheme is the *k-fold cross-validation*. For this, D is randomly partitioned into k pairwise disjoint and approximately equal sized subsets $D_1, \ldots, D_k$ ($D_i \cap D_j = \emptyset$ and $|D_i| \approx |D_j| \forall i \neq j : i, j \in \{1, \ldots, k\}$, $\cup_{i=1}^{k} D_i = D$). Each of the subsets is used to validate the algorithm trained on the remaining $k-1$ subsets. The performance of the algorithm is then averaged over the k experiments.

5.6 SUMMARY

The edge inference problem central to this thesis can be perceived as a classification task that uses node similarity measures as features and labels from ground truth data to assign pairs of nodes to the edge or the non-edge class. Viewed as such, supervised machine learning provides a set of tools that allow us to tackle this problem. Here we opt for the conceptually simple yet powerful random forest, an approach that is based on a randomized ensemble of decision trees. Our choice is motivated by the advantages random forests have over other alternatives, such as the ability to handle large data sets and the robustness to overfitting. With this last tool, we have covered the core methods that are required to proceed to the challenging real-world problems that follow.

2 Stochastic methods such as *simulated annealing* [173] or *evolutionary algorithms* [234] approximate an optimal feature selection.

Part II

APPLICATIONS

Equipped with the above concepts and methods, we turn to selected demonstrations of their uses. The applications are chosen from diverse fields in which extensive network data awaits analysis. The concerned areas immensely profit from network modelling and edge inference, especially if the results obtained by network analytic methods are integrated into the context of knowledge acquired by discipline-specific approaches—as is shown in this part.

6

PREDICTING RELATIONSHIPS BETWEEN NON-MEMBERS OF FACEBOOK

A supervised learning approach

The work presented in this chapter has been published as

E.Á. Horvát, M. Hanselmann, F.A. Hamprecht, and K.A. Zweig, One plus one makes three (for social networks), *PLOS ONE, 7(4):e34740, 2012*

E.Á. Horvát, M. Hanselmann, F.A. Hamprecht, and K.A. Zweig, You are who knows you: predicting links between non-members of Facebook, *Proceedings of the European Conference on Complex Systems 2012, pages 309–316, Springer, 2013*

Inference of user attributes and edge prediction in online social networks are challenging tasks that have attracted the attention of many researchers over the past few years. They showed that characteristics of a given user, such as one's political preference or one's sexual orientation, can be accurately inferred based on the attributes of their friends [125, 150, 179, 277]. Previously unobserved or future relationships have also been predicted with high precision using both supervised [148, 264] and unsupervised [147] learning methods. Inference was performed either based solely on structural measures deduced from the network topology [62, 213] or by additionally taking into account the nodes' attributes [196, 107, 67, 18].

As an extension of these insights into the transparency of members of online social networking platforms, we ask here a novel question: *How many of the relationships between* non-*members could online social networks infer?* In other words: *Can we predict relationships* outside *social networking platforms from the relationships within?* We approach this question on the basis of supervised learning (cf. Section 2.6): we extract topological features based on the idea of shared neighbourhoods (cf. Section 4.2) and use random forests to build a proper prediction algorithm (cf. Chapter 5).

The problem at hand poses challenges on a technical level because there are at least fifty times more possible than realized edges (cf. References [268, 265]). Furthermore, the success of edge prediction is typically measured by cross-validation *within* the same network [279, 154]. In a graph however, the samples are usually dependent and, hence, the estimate of the performance of the prediction algorithm is overly optimistic. To evaluate the obtained results, we train and validate our algorithm on distinct Facebook networks. Our quality assessment is thus more accurate.

6.1 PROBLEM STATEMENT AND APPROACH

All members of society can be represented as nodes in an unobservable social graph. This latent graph is dynamic and extremely complex, with edges of widely differing quality (two people may be kindred, engaged, or work together, they may like or dislike each other, and so on). Given an online social network, the latent social graph is partitioned into a fraction ρ of members and $1-\rho$ of non-members

(see Figure 5). Members of the platform are connected through mutually confirmed friendship relationships. Furthermore, aiming at expanding their circle of friends, a fraction α of the platform members import their whole e-mail address-book, thereby sharing their contacts to non-members. Based on the seemingly innocuous combination of these two information types, we infer relationships between the non-members.

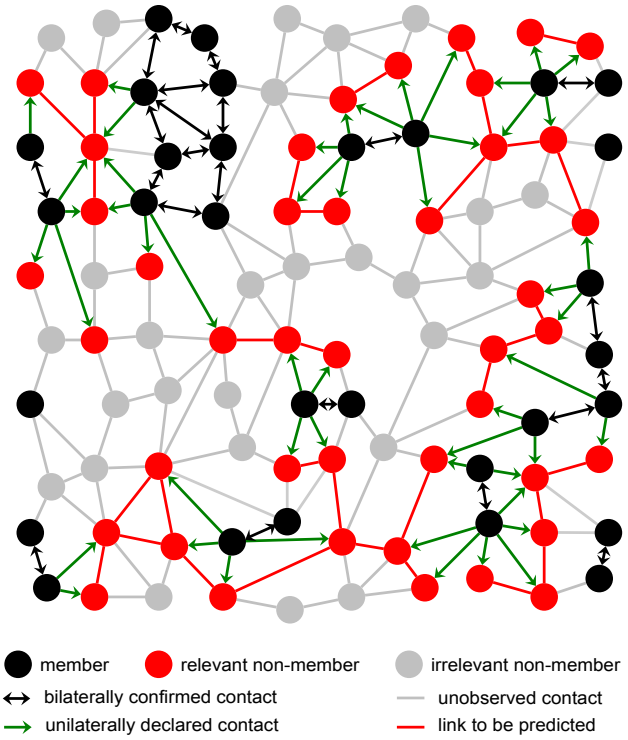


Figure 5: Division of the social network into members of a social networking platform (black nodes) and non-members. In the depicted example, a fraction of $\rho = 0.3$ (30 out of 100) individuals are members. The relevant subset of non-members consists of those who are in contact with at least one member (red nodes). A fraction of $\alpha = 0.5$ (15 out of 30) members have disclosed their e-mail contacts to non-members. The edges between members (black, bi-directed arrows) and their connections to non-members (green arrows) are used to predict edges between non-members (red lines). For the purpose of illustration, the values of ρ and α are exaggerated and the weak ties between individuals are omitted. Figure reprinted from [117].

For the very reason that the latent social graph is fundamentally unobservable, to apply a supervised machine learning procedure, the missing information needs to be imputed. The approach we choose was to use the *observed* part of a social network—for instance the Facebook network of all students at a given university—and to presume that it represents the *complete* (and unobservable) social graph of a hypothetical community. In other words, the edges in this social graph are considered the ground truth. We then proceed to partition

this community into a set of members and a set of non-members through a number of member recruitment models that represent a broad range of potential strategies by which people choose to become members. Finally, we predict the existence of edges between pairs of non-members and evaluate the quality of these predictions with respect to the ground truth. Box 1 gives an overview of the main steps.

Box 1 | *Problem statement and approach*

Data Graphs deduced from real-world Facebook data that represents the "friendships" between students of five American universities

Task Predict edges between non-members of social networking platforms such as Facebook

Framework

Step 1 | Ground truth imputation

A. Partition the nodes into members and non-members
   - Choose the desired fraction of members ρ
   - Select the member recruitment model

B. Model the probability with which a member reveals all their email contacts
   - Choose the disclosure parameter α

Step 2 | Edge prediction by supervised learning

A. Compute a set of features for all pairs of non-members who share a member acquaintance

B. Build training and test sets for learning
   - Select training scheme

C. Apply prediction algorithm (random forest classifier)

D. Assess the quality of the results

6.2 GROUND TRUTH IMPUTATION

Thesis point 5

The used data sets represent real Facebook friendship networks of students from five different US universities: `UNC`, `Princeton`, `Georgetown`, `Oklahoma`, and `Caltech` [252]. Figure 6 shows a comparison of the networks in terms of their number of nodes, average degree, density, and average local clustering coefficient (cf. Section 2.2). We partition the five networks into members and non-members. Since we do not have a clear understanding of how people decide upon joining online social networks, we consider multiple phenomena described in the literature. On the one hand, a recent analysis of the growth of Facebook

showed that the probability of a non-member joining the platform increases with the structural diversity of one's acquaintances who are members, i.e. with the number of connected components in one's Facebook neighbourhood [257]. On the other hand, there is indication that platforms recruit their members through a mixture of online mediated invitations by friends who are already members and through independent decisions by individuals who are not yet friends of a member [135].

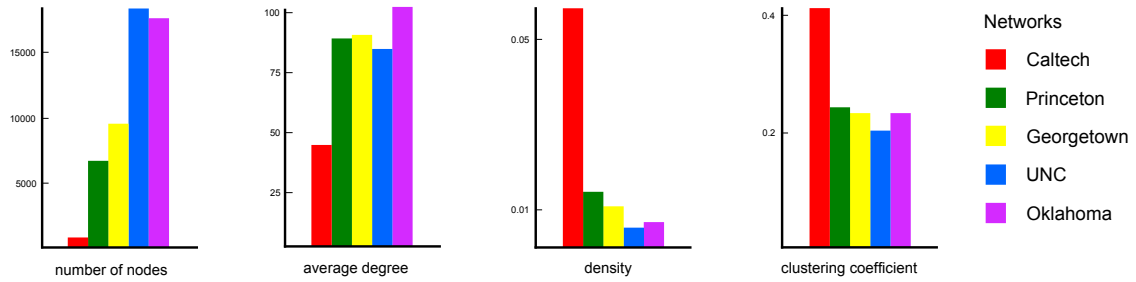


Figure 6: Comparison of the networks representing students from five different universities based on basic network analytic measures. Figure reprinted from [117].

In line with the latter investigation, we cover a wide range of possible mechanisms and use a series of different *member recruitment models* to impute the ground truth from the input data. The considered models are the following (for more details see Section 2.5): *1)* the breadth first search model (BFS): once a starting member is identified, all its friends become members, followed by all their friends and so on, *2)* the depth first search model (DFS): a randomly chosen friend of a member joins, followed by a randomly chosen friend of the new member, and so on, *3)* a random walk (RW) is started from a member and restarted as soon as someone would be chosen who already is a member, *4)* the ego-networks selection model (EN): a number of members are selected randomly and together with them, all their friends join the platform, and *5)* the random selection model (RS): people decide independently from their friends whether to become a member or not, i.e. each member is chosen randomly from the remaining non-members. Figure 7 shows the resulting partitions of a toy graph under all five models. As we will see, the specific choice of the member recruitment model does not alter our main findings.

Accordingly, the ground truth imputation for our inference problem consists in fixing the fraction of members ρ, partitioning the community into members and non-members by using one of the member recruitment models, and finally choosing the disclosure parameter α, thereby controlling for the percentage of contacts that are made public. Having devised the ground truth, we use the following experimental setting for our learning approach.

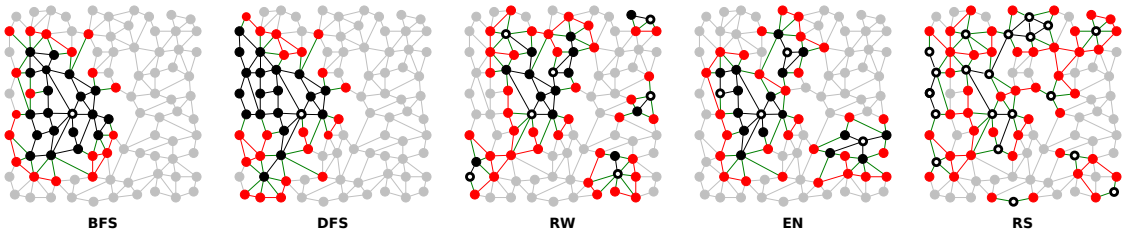


Figure 7: Membership acquisition in a toy example according to different member recruitment models. Note that real social networks exhibit more weak ties. Examples for the platform penetration value $\rho = 0.2$ show the nodes from which the propagation started (black nodes with white core). Other members are marked black and relevant non-members red; for ease of reading, arrows are not displayed, but black edges are bidirectional while green edges point from black to red nodes. With BFS and DFS the network is explored starting from one node (denoted by a white circle); with RW and EN there are several nodes from which the propagation is launched; and finally, for RS all selected nodes can be seen as starting nodes. Figure reprinted from [117].

6.3 THE EXPERIMENTAL SETTING USED FOR PREDICTION

Thesis point 6

FEATURE EXTRACTION The available Facebook networks are anonymized. Therefore, in the absence of user attributes, we base our predictions solely on topological graph features (see Chapter 4). For each pair of non-members v and w we compute 15 features deduced from the known structural properties of (online) social networks (see for instance Reference [178]). The phenomena of homophily and triadic closure (cf. Section 2.2) motivate the inclusion of a feature that counts the absolute number of common neighbours of v and w. As discussed in Section 4.2, the absolute number of common neighbours might be misleading if v has just a few neighbours, while w has many. Thus, we add normalized versions of this value such as the Jaccard coefficient. The typically high degree assortativity [187] and the significant local clustering [9] of nodes in online social networks justify considering the average degree and the clustering coefficient of the common neighbours of v and w. The community structure of social networks [94] leads us to construct several features that reflect the interconnectedness of the member side neighbours of the two nodes. As a final feature, we count the absolute number of distinct paths of length 3 between v and w. For each pair of non-members these scalars are stored in a 15 dimensional feature vector.

THE PREDICTION ALGORITHM We use the feature vectors to train a random forest classifier [47]. As described in Chapter 5, we adjust the parameters of the classifier on a training set before applying it to a validation set. We predict those pairs of non-members to be connected, for which the edge probability as determined by the algo-

rithm is higher than some threshold value. In a final step, we validate our predictions by comparing them with the ground truth. We use the area under the curve AUC and the positive predicted value of the k top-ranked predictions PPV_k, where k denotes the number of edges in the ground truth that need to be predicted (see Section 2.7) to assess the performance of the algorithm.

TRAINING SCHEMES All prior work on edge prediction with high imbalance between possible and realized future edges that we are aware of uses cross-validation (cf. Section 5.5). This represents a training scheme in which training and validation are achieved within a single network. In reality however, an algorithm that can be trained once and then allows to predict edges in independent networks is preferable (*cross-prediction*). In general, cross-validation within a single network tends to be overly optimistic in its results, since the training and the validation examples may be dependent. In addition to validating our algorithm by cross-validation, we deploy two cross-prediction schemes as well, thereby assuring the independence of the training and validation sets by learning and validating on different networks.

A. In the $4 \rightarrow 1$ cross-prediction scenario the classifier is trained on examples from *four* data sets and validated on examples from a fifth set. This scheme is less prone to overfitting, because the learning is performed on four different networks.

B. In the $1 \rightarrow 1$ cross-prediction setting the classifier is trained on *one* data set and evaluated on another. The goal here is to evaluate whether a single network contains enough characteristic patterns to obtain high-quality predictions for an entirely different network.

6.4 PREDICTION RESULTS

According to our experimental setup, imputing the ground truth requires the introduction of two parameters (the membership parameter ρ and the disclosure parameter α), as well as a member recruitment model (BFS, DFS, RW, EN, or RS). In the following, we investigate the prediction accuracies for a wide range of their combinations, using two measures (AUC and PPV_k) and three training schemes.

As argued above, the prediction performance of cross-validation should be an upper-bound to the performance of a cross-prediction approach. Thus, as a base line, Figure 8 visualizes the prediction performance as measured by the AUC, using different member recruitment models in conjunction with cross-validation on the `UNC` data set. The general pattern is that the prediction performance increases with ρ and α. In other words, the greater the percentage of members and

the higher their propensity to share their email contacts, the easier it is to predict the network between non-members. One exception to this pattern is the BFS model, whose prediction performance shows a maximum for $\rho \sim 0.5$. The behaviour of the AUC and the PPV_k is very consistent over all member recruitment models for all university data sets, implying that the exact model of the member recruitment model is not crucial.

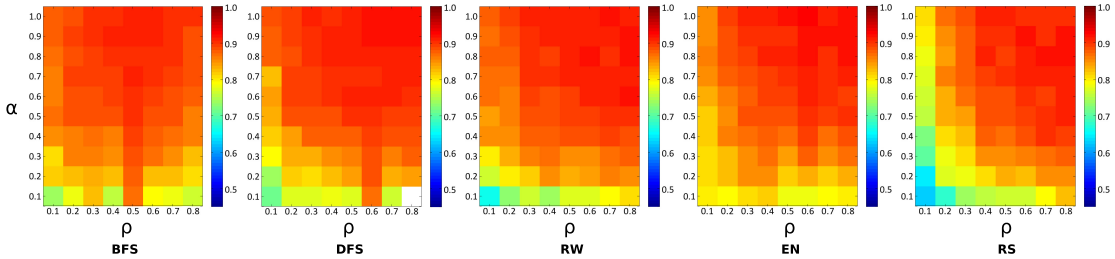


Figure 8: Prediction performance (AUC) of examples based on all member recruitment models in the cross-validation training scheme applied to UNC data. The white square denotes one data point that lacked enough data to perform the prediction. Figure reprinted from [117].

Next, we examine the performance of our algorithm with $4 \to 1$ cross-prediction for each combination of ρ and α values, all member recruitment models and all five university data sets (see Figure 9). Based on the minimal (lower triangle) and maximal (upper triangle) AUC and PPV_k values, we see that the differences between the member recruitment models are small in most cases. The AUC values are above 0.85 for all combinations with $\rho \geqslant 0.5$ and $\alpha \geqslant 0.4$ in the case of UNC, Princeton, Georgetown and Oklahoma, for all member recruitment models except the BFS. This implies that in most cases the prediction is considerably better than random guessing. The PPV_k is at least 0.4 for the same range of ρ and α and in the case of UNC, Georgetown, and Oklahoma, and for all member recruitment models except the BFS and the DFS. A value of 0.4 means that when selecting the k examples with the highest prediction values, at least 40% of them actually represent two non-members that know each other. To interpret this value correctly, we have to emphasize that our data set shows a striking class imbalance. While there is a huge number of node pairs that could be connected by an edge, there are only a few pairs which are truly connected. More precisely, depending on the chosen member recruitment model and on the ρ membership and α disclosure parameters, the ratio f between the number of edges and non-edges lies between 0.0002 and 0.03 for four out of five university networks. These values represent the baseline for PPV_k.

Finally, in the $1 \to 1$ cross-prediction setting, we evaluate how reliable the predictions are if the random forest is trained on only one network at $\rho = \alpha = 0.5$. Given the coverage of Facebook especially

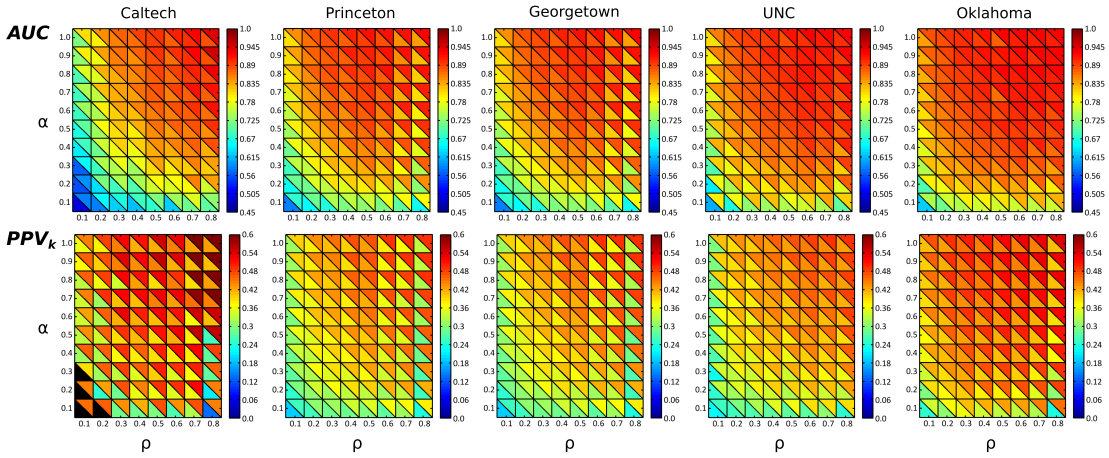


Figure 9: Minimal (lower triangle) and maximal (upper triangle) prediction performance in the $4 \to 1$ training scheme for all five member recruitment models as a function of the membership parameter ρ and the disclosure parameter α. Upper row: AUC; lower row: PPV_k; black triangles denote data points where PPV_k was smaller than the according fraction f of edges among all examples, i.e. it was worse than expected by chance. Figure reprinted from [117].

among the youngest and the heavy usage of the "friend finder" application by both novice members and experienced users of the platform, these estimates of ρ and α are rather conservative. Figure 10 shows the corresponding prediction performance. On the diagonal, we plot as reference the prediction performance when we train and validate on the same network, while the off-diagonal elements correspond to the cross-prediction case.

It can be seen that some data sets such as `Oklahoma` and `UNC` are easy to predict, while `Caltech` is difficult to predict based on any of the four other data sets. Furthermore, if the classifier is trained on `Caltech` data, the predictions are consistently the worst among all cross-predictions. The intuition behind this observation is that `Caltech` is a clear outlier among the used data sets as it is by far the smallest and the densest (cf. Figure 6).

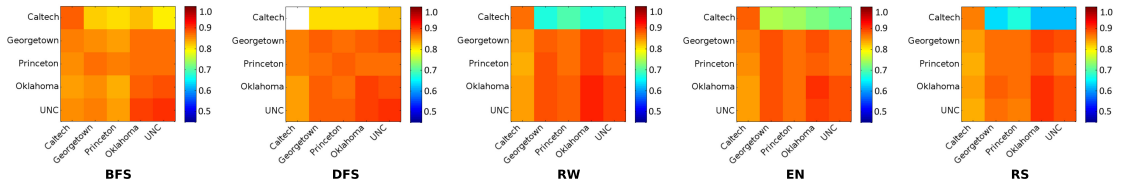


Figure 10: $1 \to 1$ cross-prediction performance: AUC values for each of the five member recruitment models when $\rho = \alpha = 0.5$. The y and x-axis show on which network the random forest was trained and validated, respectively. The white square indicates that there were too few edges to reasonably train the classifier. Figure reprinted from [117].

6.5 DISCUSSION AND CONCLUSIONS

The ground truth imputation on which the presented results are based relies on three important modelling decisions which require further discussion. First, we imply that non-members are similar to members in terms of the revealed network characteristics. Two studies from 2006 and 2009 indicate that there are in fact statistically significant differences between members and non-members among university students, yet these differences concern age, ethnicity, and gender, but not important social factors such as life satisfaction, social trust, or privacy concerns [6, 259]. Although the sociability of members and non-members was not directly assessed, these studies give no indication that members and non-member differ significantly in the structure of their contact networks.

Second, since the contact network between the members and non-members is not available for any social networking platform, we take the known Facebook friendship network as a proxy for the structure of the email contact network between members and non-members. This is justified by the fact that both belong to the large set of social networks with scale-free degree distribution, high clustering coefficient, small-world behaviour, and a positive assortativity [178, 8, 251].

Third, we only take into account pure member recruitment models which might not be realistic on their own. The surprising result, that the choice of member recruitment models does not alter the main conclusions, shows that the analysis of the pure models does not constrain the approach. Even for BFS, for which the prediction quality was worst among all models, good results are achieved. This indicates that in whichever way individuals decide to join an online social network, the unilateral disclosure by members of their contacts to non-members allows social networking platforms to gain substantial insight into the relationships of non-members. This increase in coverage due to edge prediction will be most successful if the individual members' decisions to join the network are independent. This could be exploited by the platform when developing new recruitment strategies.

Altogether, our work reveals the potential that social networking platforms have in predicting edges between non-members, based only on the connection patterns of the befriended members and their e-mail contacts to non-members. Accordingly, individuals without a profile in an online social network—such as Facebook—are not immune to data mining based on data available to the given platform. This finding is based solely on topological features, i.e. we only used contact data and no user attributes. If we had access to more comprehensive data including details about the members such as their age, location, or occupation, then our inference could be improved considerably.

6.6 SUMMARY

Could online social networks like Facebook be used to infer relationships between non-members? We showed in this chapter that the combination of relationships between members and their e-mail contacts to *non*-members provides enough information to deduce a substantial proportion of the relationships between non-members. Using topological features, we were able to predict relationship patterns that are stable over independent social networks of the same type. To obtain this result, we used a random forest classifier and applied it to data sets that are characterized by high class imbalance. Our findings are not specific to Facebook and can be applied to any other social networking platform that involves online invitations.

7

INFERRING EDGES IN BOTH BIOLOGICAL AND SOCIAL NETWORKS

An unsupervised learning approach

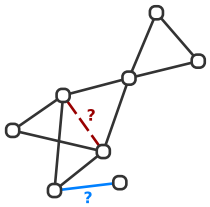
The work presented in this chapter constitutes the basis for the publication

E.Á. Horvát, A. Spitz, A. Gimmler, T. Stoeck, and K.A. Zweig, Validating and assessing low intensity interactions in complex networks, *in preparation.*

Given a complex system of interest, the first goal of a network analytic study is to find a corresponding graph representation. This mapping is not unique, but involves several modelling decisions, based on which we determine which elements of the system and which interactions should be observed or measured and included in the model [129, ch. 3, 281]. Clearly, these decisions have a strong influence on the constructed network model and hence the conclusions that are subsequently drawn from it. Assuming that we have identified the key elements of the mapping (i.e. those that constitute the node and edge set), there are further issues that must be addressed during this process and are tied to the field of research. In genomics, for instance, highly popular high-throughput experiments allow monitoring protein interaction by protein affinity experiments on the one hand and gene regulation by microarray experiments on the other. These measurements enable the construction of large graphs that provide *one* view from a single perspective on the underlying biological system. However, this perspective is likely to contain spurious interactions and to be incomplete, for example due to the fact that high-throughput experiments are error-prone and there are several interactions that have not been tested yet. A comparison of several high-throughput methods to a reference high-quality data set showed in 2002 that these methods had at that time accuracies below 20% [263]. Although less extreme, but a similar problem arises during the observation of social networks. Existing online social networking platforms record connections that are casually defined by their members, without validation of any sort. Additionally, these networks are evolving quickly, because many of the connections are ephemeral.

Due to these problems in measurement and data collection, the validation of observed connections as well as the prediction of unobserved or future edges is often very important in the context of both biological and social networks. In this chapter, we therefore analyse an experimentally derived protein–protein interaction network[1] and a large sample of an online social network.

The main difficulty with the prediction of edges in these networks is evaluation, because the true web of connections is usually unknown. Tedious efforts of experts to manually inspect the results of proposed



Twofold scope:
1.) validation of observed edges (blue)
2.) prediction of unobserved or future edges (red)

1 Chapter 9 deals in depth with a regulation network, i.e. a second common type of biological network obtained from high-throughput screening.

algorithms are often too expensive or simply not feasible at the investigated scales. While there are examples for a few evaluations on small social networks [274, 192], ground truth information that is preferably generated automatically and that enables large-scale quantitative evaluation is necessary [271]. Obtaining such large-scale ground truth data usually requires additional data sources. In the case of the two considered networks further data exists and can be used to generate the required ground truth. In the biological setting, interactions that have been individually confirmed by different small-scale experiments constitute the ground truth, while in the social network scenario, user-declared communities can be used to verify the existence of friendships between members of the same group.

The availability of ground truth data in the presented prediction context allows evaluating an unsupervised approach in which we compute the node similarity measures presented in Chapter 4 and use them as scoring functions to rank the pairs of nodes according to the likelihood that they are connected by an edge[2]. Based on the ground truth, we then compare the individual measures by testing their efficiency as single predictors[3]. We thereby identify those measures that improve the quality of inference for the investigated networks in particular and propose to use them for further biological and social data sets *without* ground truth information (Section 7.3). As the two networks are derived from very different settings (Section 7.1 and Section 7.2) and pose different challenges that are representative for each of these areas, we can assume that our results—if consistent over both chosen networks—will indicate a tendency that is valid for complex networks in general.

7.1 VALIDATING AND PREDICTING PROTEIN–PROTEIN INTERACTION

The function and molecular properties of individual proteins have been the focus of intense investigation for decades. In addition, one of the recent scopes of biological research is the mapping of protein-to-protein physical interactions, i.e. the construction of the *interactome* [93]. A *protein–protein interaction* is defined as the molecular docking between two proteins that co-occur in a cell *in vivo* [214]. It implies direct physical interaction and should not be confused with functional contact. It has to be noted that these physical interactions

2 Note that in the case of the two particular networks with ground truth a supervised learning approach, as the one presented in Chapter 6, could be used. However, as we have argued there, simply splitting a network into training and validation sets is problematic due to the inherent dependency between the two resulting sets. Thus, whenever we do not have several different networks of the same type as in the case of the five entirely independent, but structurally similar Facebook networks, this approach is not flawless.

3 See Section 2.6 and Section 2.7 for the preliminaries for this chapter.

are not static or permanent and not all possible interactions will occur in a cell at any time. Protein–protein interactions are essential for diverse biological processes, including the formation of macromolecular structures, cell signalling, regulation, metabolic pathways, and several physiological processes [244]. From these interactions, a protein–protein network (PPI) can be constructed, which contains the proteins as nodes and the interactions as edges [256, 198, 214].

The number of technologies that measure proteome-wide physical connections on a large scale has increased substantially recently [233]. However, even the most common and widely accepted techniques such as yeast two-hybrid (Y2H) or co-complex methods face several challenges. For example, there may exist a bias in the selection of interactions that are tested and different databases cover only subsets of all experimentally tested interactions. Therefore, strategies to improve the reliability of measured interactions are highly requested. Our contribution consists of predicting possible interactions between pairs of proteins based on the topology of the entire protein–protein interaction network. This approach does not require additional information, such as 3D structural information about the interacting proteins, and can be expected to scale better than more complicated methods. Furthermore, its financial cost is substantially lower than those of biological experiments making it a worthwhile test in any case.

The specific network we use to verify our method is constructed for the unicellular model organism *Saccharomyces cerevisiae* yeast [256]. An empirical estimate of the interactome of this yeast contains $18,000 \pm 4,500$ interactions [273]. Other estimates assume over $30,000$ interactions between roughly $6,000$ proteins [35]. These large discrepancies in the estimations of the size of the interactome occur due to the large proportion of false positives in the recorded interactions [207]. Several factors may favour false positives, i.e. the detection of biologically non-relevant interactions between proteins that never simultaneously co-occur *in vivo* [93]. A network analytic approach holds promise for evaluating the reliability of interactions based on the assumption that the overall topology of the protein–protein interaction network is characteristic. Thus, the network topology alone provides evidence based on which spurious local interactions may be detected. For a similar endeavour that is restricted to just a few node similarity measures, see Reference [96], while for an approach based on stochastic block models, see Reference [105].

To evaluate filtering and/or prediction algorithms, results of well-documented small-scale experiments can be used as reference [273]. We use the full set of protein–protein interactions available for *Saccharomyces cerevisiae* on the public repository Database of Interacting Proteins (DIP) [1], release of August 18, 2012. The data integrates information from large- and small scale experiments reported in the

literature. After removing 318 self-interactions, the built network contains 22,148 interactions between 5,078 proteins. The 3,543 interactions that are classified as *dip-quality control core* constitute the ground truth, i.e. the interactions that were verified manually in experiments.

7.2 DEDUCING HIGH-PROBABILITY ACQUAINTANCES

The notion that individuals are embedded in webs of social connections motivates networks as a straightforward representation of social systems [266]. Ever since their appearance, there has been an explosion of interest in and research about online social networking platforms even beyond the social sciences (cf. Chapter 6). Accordingly, current efforts aim at finding explanations of social phenomena at previously unobservable scales within classic social network analysis. Furthermore, as Butz and Boyle Torrey note, "the fundamental challenge in the social sciences is moving from complicated correlations to useful prediction" [55]. Here we provide solutions for edge inference in large graphs. We show the potential of our method by analysing the online social network and blogging platform called LiveJournal (LJ) [19, 271].

Within this network, individuals explicitly state their group memberships. Groups are formed based on a common external property that the members share and around which the group is organized. As the network is prohibitively large (34,681,189 connections between 3,997,962 individuals), we perform our analysis on a sample. To generate the sample, we first select all groups that have a size between 3 and 50 (inclusive). Our assumption is that larger groups do not allow a close relationship between the individuals. We then proceed by picking a starting group uniformly at random and continually increase the network by adding adjacent groups. Two groups are considered adjacent if they share at least one individual. The process is stopped once 1,000 groups are selected and results in a network of 11,755 individuals and a total of 80,023 connections. Out of these, 30,230 edges connect individuals that are members of the same group and are considered to constitute the ground truth.

7.3 INFERENCE BASED ON NODE SIMILARITY: WHICH MEASURE TO CHOOSE?

Thesis point 7

The main tasks involving the protein–protein interaction and the LiveJournal network as well as the proposed solutions are outlined in Box 2.

Box 2 | *Problem statement and approach*

Data
1. A graph deduced from protein–protein interaction data
2. A graph that models the "friendships" on the LiveJournal online blogging platform
3. A ground truth data set for each of the above graphs

Task Compare a set of topological node similarity measures (Jaccard coefficient, cosine similarity, covariance, Pearson correlation, hypergeometric coefficient, configuration model-based similarity, p-value, z-score, and the presorted z-score) based on their ability to validate available edges and predict future or unobserved edges

Framework

A. Compute node similarities for all pairs of nodes that share a common neighbour, whether they are connected by an edge or not
B. Rank the pairs of nodes by each individual similarity measure
C. Evaluate the similarity measures by comparing the top-ranked pairs to the ground truth edges

SIMILARITY CALCULATION Both in the PPI and LJ networks, we compute the node similarity measures defined in Chapter 4 for all pairs of nodes, regardless of their actual connection status. Due to the nature of the used measures, we restrict the pairs to those with at least one common neighbour. The computation of the classic similarity measures is straightforward. For the hypergeometric coefficient, the factorials are calculated by numerical approximation. The fixed degree sequence model-based measures require sampling from the ensembles of non-bipartite graphs that have the same degree sequence as the PPI and LJ network, respectively. Algorithm 1 is used for obtaining the samples. As discussed in Section 3.3.2, the theoretical estimates for the mixing time of the Markov chain are inconclusive. However, the mixing time can be compensated by the number of taken samples and we thus perform a test to check the sample size needed for the performance to converge. Figure 11 shows the resulting plot for LJ[4]. As expected, the z-score reaches convergence much faster than the p-value, since it only requires the first two moments of the distribution (cf. Section 4.3). The presorted z-score z^* deduced from the combination of the two reaches convergence before the p-value. Based on this test, we deduce that $\kappa = 100,000$ graphs are enough to form a representative sample.

RANKING PAIRS OF NODES BY THEIR SIMILARITY As described in Section 4.5, the pairs of nodes are ranked non-increasingly according to the considered similarity measures. During this analysis, the

4 The convergence test for the PPI network behaves similarly.

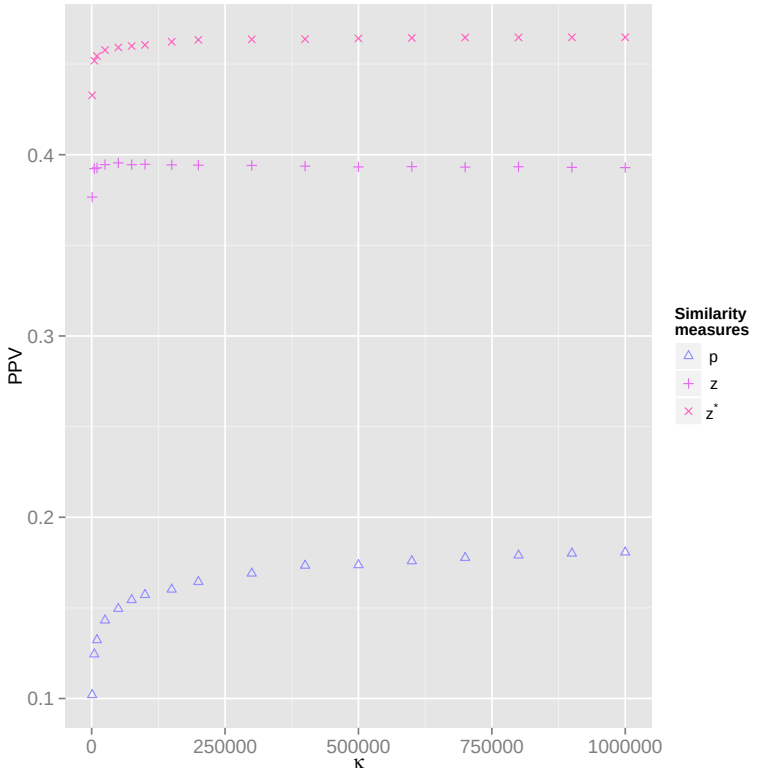


Figure 11: The performance of the similarity measures based on the fixed degree sequence ensemble, as quantified by the global PPV, plotted against growing sample size κ. Experiment conducted on the LJ network.

unnormalized number of common neighbours is omitted. Our assumption is then that there exists a meaningful correlation between the obtained rankings and the verity of the individual protein–protein interactions (in the PPI network) and acquaintances of individuals (in the LJ network). In other words, we test the hypothesis that the similarity measures indicate whether there is a real interaction or acquaintance between given pairs of nodes. To test this assumption, we compare the rankings with the respective ground truth. First, we perform a global evaluation by using the PPV and the nDCG, i.e. the standard measure for evaluating rankings (cf. Section 2.7).

CORRELATION BETWEEN THE RANKINGS AND THE VALIDITY OF EDGES Figure 12 shows large differences in the performances of the different node similarity measures. As a baseline, we use the result of a random predictor, which unsurprisingly has the worst performance of all measures. The ranking of the individual measures based on their performances is the same according to both performance measures (PPV and nDCG). The top performers for PPI are z^*, p, hyp, and cov, while for LJ z^*, jac, z, and r come out on top. The presorted z-score z^* achieves an improvement of 6% over the second-best mea-

sures for the PPI network (i.e. the p-value and hyp) and 5% for LJ (i.e. jac and z-score). There is no indication that the hypergeometric or the Jaccard coefficient would perform consistently well on different networks. Note that for both the biological and the social data set one of the similarity measures based on the fixed degree sequence model perform best after z^*. As the presorted z-score is a ranking function, which is a combination of z-score and p-value, in Chapter 8 and Chapter 9 we will rely on the z-score and the p-value.

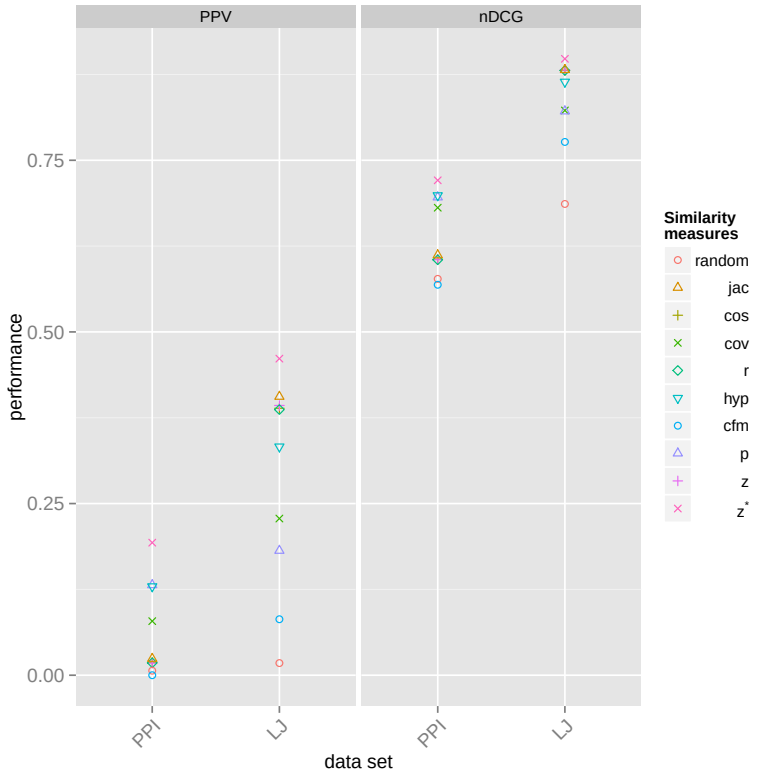


Figure 12: Performance of the node similarity measures as evaluated by PPV and nDCG. Ties in the ranking are broken at random and to account for this, we average over 100 iterations. The top performing measure for both data sets is the presorted z-score, z^*. The notation of the measures is shown in Table 2.

After analysing the overall performance of the measures in Figure 12, i.e. the accuracy of the entire ranking when compared to the ground truth, we now consider the distribution of true positives in this ranking. Figure 13 depicts the number of TPs at increasing ranks for the PPI network. We plot the fraction of interactions validated by high-confidence small-scale experiments for pairs of nodes with at least one common neighbour. The pairs are ranked in non-increasing order by the individual similarity measures. In a random baseline, the expected fraction of TPs is constant across all bins at approximately 0.007. We expect a useful measure to place a high fraction of validated interactions on top positions in the ranking. Conversely, edges

without evidence should be placed on lower positions, thus resulting in a monotonically decreasing curve. z^*, the p-value, and hyp show this desired behaviour. z^* clearly dominates the others, i.e. it ranks true edges higher than any other measure. Interestingly, the z-score alone performs rather poorly. However, when presorted with p-value, a better performance is obtained than when using the p-value alone.

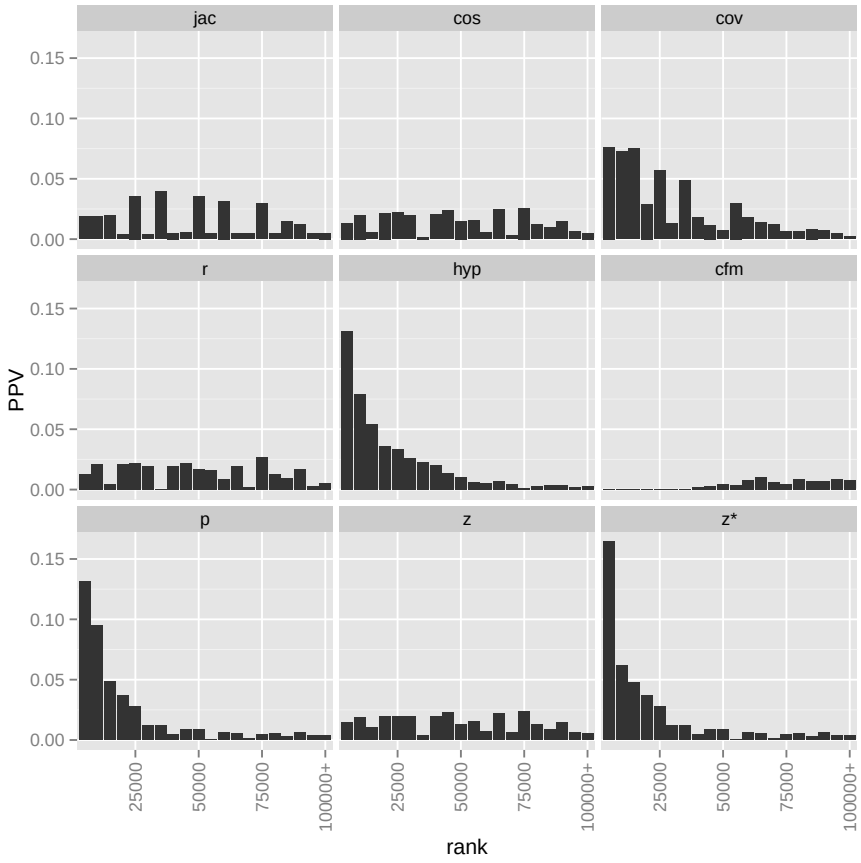


Figure 13: Global PPV at different rank intervals obtained by using various similarity measures. On the x axis, the ranks are binned linearly. The size of one bin is 5,000, while the pairs above rank 100,000 are all contained in the last bin. Ties in the p-value ranking are broken at random and to account for this, the performance is averaged over 1,000 iterations.

7.4 DISCUSSION AND CONCLUSIONS

With the above procedure, we have identified the measures that separate true edges from unreliable edges in the ranking and showed that this can be exploited to identify true positives. The benefits of such an analysis are twofold. First, we can assess the quality of the edges and thereby filter out false positives. Recall that we compute the similarity also for those pairs of nodes that are connected by an edge in the

data set. Second, we can predict interactions for which we currently have no direct supporting evidence. These are the top-ranked pairs without a connection in the original network. More importantly, exhaustive literature research or targeted experiments concerning these top predictions could further validate our findings.

Considering our selection of node similarity measures we find that:

1. Measures based on the immediate neighbourhood of the nodes (such as the Jaccard coefficient $\mathtt{jac}$, the cosine similarity $\mathtt{cos}$, and the configuration model-based similarity $\mathtt{cfm}$) indicate the likelihood of edge formation whenever it can be presumed that two nodes having more neighbours in common at a given time point will get connected later. For instance, due to the repeatedly confirmed principles of triadic closure (cf. Section 2.2), this is indeed to be expected in social network settings.

2. Measures quantifying the level of association between two nodes in terms of their connection patterns (like the covariance $\mathtt{cov}$ and the Pearson correlation coefficient $\mathtt{r}$), are expected to be meaningful predictors for instance when predicting gene regulatory networks, i.e. relationships between genes based on the effect of a series of experimental conditions on them. In this case it is usually assumed that each experiment exerts a roughly similar effect on all genes. In graph theoretic terms, the degree sequence of the experiments is Poissonian (cf. Reference [282]).

3. Finally, where issues of measurement error and degree heterogeneity are of potential concern, it is usually necessary to use the systematic statistical approach prescribed by the measures based on the fixed degree sequence model: the p-value, the z-score, and the presorted z-score z^* computed from the proper graph ensemble.

In summary, we have exploited the topology of protein–protein interaction and online social networks to rank the pairs of nodes by the likelihood that they should be connected by an edge. We have suggested a new measure that best filters out low-confidence interactions in the PPI network and thereby addressed one of the main problems with this type of data. The same measure is also suited for predicting interactions that have not been tested yet. In the social network setting, it enables detecting connections between individuals with common interests. Importantly, the inference is based solely on the graph topology and does not use any additional details about the proteins or individuals.

7.5 SUMMARY

Biological networks that are generated from noisy experimental data and online social networks deduced from user-specified information

both require a systematic way of assessing the confidence in existing edges and predicting further connections. Despite the differences, such networks share similarities in their topology. We thus proposed in this chapter a network analytic framework that relies on node similarity measures, which are equally suitable for either kind of network. We compared a set of node similarity measures and evaluated them based on ground truth data sets for networks as different as protein–protein interaction and user friendship data on a blogging platform. We showed that the similarity measures based on the fixed degree sequence model are inherently meaningful for true edges and that this information is consistent enough to enable their use for inference. Finally, we provided evidence that our new similarity measure, the presorted z-score z^*, consistently outperforms existing measures and is thus more suited than any other measure we tested for adjusting our confidence in the veracity of edges in diverse types of complex networks.

The following two chapters use the idea of node similarity based on the fixed degree sequence model to infer associations between the same type of entities in different bipartite graphs. Furthermore, they extend the method to the case of multiplex graphs.

8

EVALUATING FILM SIMILARITY IN A MARKET BASKET SETTING

From simplex to multiplex analysis

Parts of the work presented in this chapter have been published as

E.Á. Horvát and K.A. Zweig, One-mode projections of multiplex bipartite graphs, *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 598–605, 2012*

E.Á. Horvát and K.A. Zweig, A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs, *Social Network Analysis and Mining, online first.*

Data about human behaviour that was once considered to be unattainable is nowadays collected in unprecedented amounts [41]. As data acquisition methods become more efficient, the need for more sophisticated exploratory methods which allow studying this often multidimensional data becomes increasingly evident. Behind this multidimensionality often lies an inherent bipartite structure: in the social sciences for example, agents are affiliated with societies and scientists write papers, while in biology birds inhabit islands and genes are in relation to diseases. Data that can be modelled as bipartite graphs is also collected in large amounts for the purpose of *market basket analysis* and usually contains records of customers buying, renting or rating products.

The most common task associated with these latter data sets of economic interest is to relate products by quantifying their similarity based on customer behaviour. Accordingly, the product–customer bipartite graph is transformed into a non-bipartite graph between the products, i.e. it is subdued to a one-mode projection (see Section 2.2). A prominent large-scale data set of this kind is the Netflix data set that consists of millions of discrete ratings from 1 to 5 given by 480,000 distinct users to 17,770 films [5]. The availability of different ratings in the data set enables a multiplex analysis in which one can differentiate for instance between "likes" (ratings of 4 and 5) and "dislikes" (ratings of 1 or 2). To model this extra information, a multiplex network representation is needed that contains two distinct types of edges (see Figure 14A). In general, there are two straightforward options for analysing such networks: *1)* The analysis is based on the individual representation of the networks that contain different types of connections. This of course means that the interdependencies across the different connections are completely ignored. *2)* A representation is used, in which connections of different types are aggregated into a single network. For the Netflix application both of these approaches are overly simplistic: *1)* Considering the like and dislike networks independently disregards one of the key structural properties of the data set, namely that users can not simultaneously like and dislike the same film. Furthermore, the combination of likes and dislikes promises new insights. *2)* The aggregated approach leads to inaccurate interpretations, because it inevitably mixes qualitatively different connections. Therefore, a proper multiplex analysis is required.

Analytic methods handing networks that contain different types of entities and different types of connection (for instance multiplex bipartite networks) are mostly missing. Until now, research mainly concentrated on understanding only one of these two aspects. On the one hand, state-of-the-art work includes studies about the role of different entity properties and their influence on connection probability [200, 11]. On the other hand, several approaches exist for the analysis of multiplex networks (see Section 2.1 and Section 2.3). These range from methods that are firmly based on the field of classic social network analysis and advanced from small-scale questionnaire-based approaches [266, 172] to methods for large-scale analysis that granted additional insight into the organization principles [144, 246, 245], the community structure [182], and the predictability [73, 126, 146] of multiplex networks. Recently, researchers suggested frameworks for the representation and handling of multiplex networks [160] and for the extension of traditional network analytic measures to deal with multiplexity [48].

In this chapter, we demonstrate how the null model approach presented in Section 3.3.4 and Section 4.3 can be used to perform a one-mode projection based on a multiplex fixed degree sequence ensemble (Section 8.1). We test the robustness of the framework on computer generated data for which an exact ground truth is known (Section 8.2). We then apply it to the Netflix data and discuss the results in terms of the like/dislike patterns contained in the projection (Section 8.3). Furthermore, we evaluate our findings in relation to the classification of films into genres and to two ground truth data sets of similar films. After analysing the potential of negative ratings for finding significant film similarities, in a final step we use the approach to explore further aspects of film similarity beyond the similarity deduced from the rating data (Section 8.4).

8.1 MULTIPLEX ONE-MODE PROJECTION

Without loss of generality, let us consider the case of projecting the multiplex bipartite graph $\tilde{B} = (\mathcal{L} \cup \mathcal{R}, \tilde{\mathcal{E}})$ onto the node set $\mathcal{L}$. Furthermore, let Ω denote the set of edge types in $\tilde{B}$, i.e. $\tilde{\mathcal{E}} = \cup_{\gamma \in \Omega} \tilde{\mathcal{E}}_\gamma$. As we focus on graphs with maximal multiplicity 1 it holds that $\tilde{\mathcal{E}}_\gamma \cap \tilde{\mathcal{E}}_\varphi = \emptyset \; \forall \gamma \neq \varphi \in \Omega$ (cf. Section 2.1 and Section 3.3.4). The one-mode projection of $\tilde{B}$ is based on the co-occurrence $\mathrm{coocc}_{\gamma\varphi}(v, w)$ of the node pairs $v, w \in \mathcal{L}$, i.e. on the number of their common neighbours $u \in \mathcal{R}$ for which the edge (v, u) is of type γ and the edge (w, u) is of type φ. It results in a non-bipartite, weighted, and multiplex graph $\tilde{G}$ that contains edges between nodes from $\mathcal{L}$, with the types of these edges in $\Omega' = \Omega \times \Omega$. Figure 14B shows a projection containing three edge types. In this projection two nodes are connected if they have a co-occurrence of at least one in the bipartite graph from

Figure 14A. We further include a function ω that assigns a weight to each edge in $\tilde{G}$. For practical reasons, we choose this weight to express how unlikely it is to obtain the given edge by chance, i.e. it quantifies its statistical significance with respect to a null model consisting of the ensemble of multiplex bipartite graphs with the same degree sequences as $\tilde{B}$. Details of the construction of the null model have been presented in Section 3.3.4, while the choice of test statistic for the function ω was described in Section 4.3. Note that the test statistic serves as the similarity measure and the process of one-mode projection can thus be viewed as the inference of an association network between the nodes from $\mathcal{L}$.

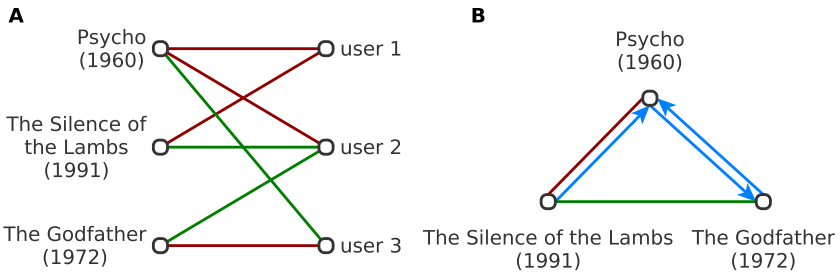


Figure 14: (A) Example of a film–user bipartite graph. Ratings that express "like" are shown as green lines, while ratings that express "dislike" appear as red lines. (B) In the multiplex projection of this graph, which contains co-rated films, there are connections between pairs of films that were both liked (green line), pairs that were both disliked (red line), and pairs where one of the films was liked (source of the blue arrow) and the other was disliked (target of the blue arrow) by the same users. The presented method assesses the statistical significance of the edges in this projection network. Figure reprinted from [115].

In summary, the method relies on the common practice in network analysis, according to which the statistical significance of topological patterns (i.e. network observables) is assessed by using the null model approach as presented in Chapter 3 and in particular in Section 3.4.1. Accordingly, we proceed as follows:

1. Given the multiplex bipartite graph $\tilde{B}$, we compute the co-occurrence $\mathrm{cooc}_{\gamma\varphi}(v, w)\ \forall\ \gamma, \varphi \in \Omega$ for all distinct node pairs v and w from the node set $\mathcal{L}$, for which the co-occurrence in $\tilde{B}$ is at least 1.

2. We then compare these observed values with the expected values in a bipartite graph from the corresponding fixed degree sequence ensemble, i.e. in which every node maintains its degree for each of the edge types and no parallel edges of any kind occur. As the ensemble cannot be fully enumerated, we compute a

large sample using a Markov chain Monte Carlo technique (see Algorithm 3).

3. We count the number of co-occurrences of v and w with respect to all possible combinations of edge types for all sampled graphs and compute as a test statistic the p-value and the z-score[1], both of which quantify the statistical significance of the edge between v and w, given their expected co-occurrence in the null model.

Note that the multiplex projection $\tilde{G}$, that is induced by a multiplex bipartite graph with $|\Omega| = n$ types of edges, will contain $|\Omega'| = \binom{n+1}{2}$ types of edges. Regardless of the number of edge types, there are two straightforward approaches for analysing the projection. First, a global-level analysis aims at identifying the overall most similar pairs of nodes within $\mathcal{L}$, based on their connection patterns to the nodes from $\mathcal{R}$. This is a transformation of the obtained complete projection into a sparser graph and involves choosing a set of thresholds $\{t_{\gamma\varphi} | \gamma\varphi \in \Omega'\}$, one for each edge type available in the projection. If using the z-score as test statistic, an edge of type $\gamma\varphi$ can then be created between all pairs of nodes with a weight of at least $t_{\gamma\varphi}$ in $\tilde{G}_{\gamma\varphi}$. Note that per definition $\tilde{G}_{\gamma\varphi}$ contains only nodes with a degree of at least 1. Choosing the proper threshold is usually done based on a rule of thumb stating that observations with $z \geqslant 2$ (in terms of the z-score) and equivalently $p \leqslant 0.05$ (in terms of the p-value) are statistically significant. Another approach is to find meaningful significance thresholds based on the topology of the subgraphs of the original graph built with different possible thresholds. As we will see in Section 9.3, this idea is very viable when analysing biological networks, but it has also been used successfully in sociology [86], chemistry [275], and physics [231]. In the following, we will use both of these methods.

Second, it is interesting to explore on a local level which nodes are the most similar neighbours for a given node, as this information can for example be used to generate recommendations. In the case of the Netflix data, this local approach enables us to validate the obtained similarities, as we describe in Section 8.3.

8.2 ROBUSTNESS ANALYSIS

Thesis point 8

First, we show the robustness of the presented method on bipartite graphs containing two types of edges. This reveals how stable the one-mode projection is with respect to a single edge type (the robustness of the projection $\tilde{G}_{\gamma\gamma}$ $\forall\gamma \in \Omega$) and with respect to the combination of edge types (the robustness of the projection $\tilde{G}_{\gamma\varphi}$ $\forall\gamma \neq \varphi$ and

1 Hereafter we report results based mainly on the z-score. Whenever this is not the case, we state it explicitly.

$\gamma, \varphi \in \Omega$). In the following we refer to the two different edge types of the bipartite graph as $+$ and $-$, i.e. $\Omega = \{+, -\}$. Accordingly, the one-mode projection contains the edge types $\Omega' = \{++, --, +-\}$. To demonstrate the robustness of the method, we use computer generated benchmark graphs for which the optimal structure of the projection is defined by construction.

8.2.1 *Construction of the artificial data*

Artificial graphs that are suitable for testing the presented multiplex one-mode projection algorithm should have an embedded ground truth. This property enables us to verify the ability of the algorithm to detect groups of nodes that are similar based on their connection patterns. Artificial graphs should also recreate one of the main difficulties that real-world data sets impose on one-mode projection algorithms, namely the heterogeneity of the degree sequences.

These two requirements can be instantiated in different ways. Existing bipartite network models aim at explaining cooperation for ecological [57] and organizational networks [220] or model affiliation [97] and scientific collaboration networks [209]. These problem-specific models are ill-suited for producing artificial graphs for large-scale testing, since they are based on processes that constrain the graph structure beyond the degree sequences. Therefore, we use a bipartite model that recreates the structure of the Netflix data while keeping the artificial graphs small and adhering to the above two requirements[2]. Our artificial graphs consist of four built-in clusters with $|\mathcal{L}| = |\mathcal{R}| = 60$ nodes and the same number of $|\mathcal{E}_+| = |\mathcal{E}_-| = 128$ edges per cluster. Each cluster has the following structure: in the node set onto which we are projecting, there are two groups of equal size that have only $+$ or $-$ edges. With this we model a film rating scenario where we have a group of films X which is liked by most users, and a group of films Y that is disliked by the majority. The degree sequences corresponding to the two edge types are the same ($\mathcal{D}_+(\mathcal{L}) = \mathcal{D}_-(\mathcal{L})$): there are 16 nodes with degree 2, 8 with degree 4, 4 with degree 8, and 2 with degree 16 (implicitly, the rest has degree 0). To keep it simple, in the other set of nodes of the bipartite graph, nodes have the same number of $+$ and $-$ edges. Accordingly, there are 32 nodes with degree $1+1$, 16 with degree $2+2$, 8 with $8+8$, and 4 with $16+16$.

8.2.2 *Results on the artificial data*

We generated an ensemble of 100 random graphs with the given degree sequences and projected each of them using the z-score as test

2 We also ran experiments on artificial data where the degree sequence of one of the node sets of the bipartite graph was more homogeneous. Results on this slightly different model are presented in Section 9.2.

statistic. We defined the ground truth (i.e. the optimal result of a meaningful computation) by a simple projection in which there is an edge between all pairs of nodes with a co-occurrence of at least 1 in the original graph. Without noise, all the edges contained in the ground truth are recognized by the algorithm. In two separate runs we then randomly add and randomly eliminate up to $\rho = 50\%$ of the edges in the artificial graphs (without preference for the edge types) and check how the applied noise affects the ability of the method to recover the ground truth.

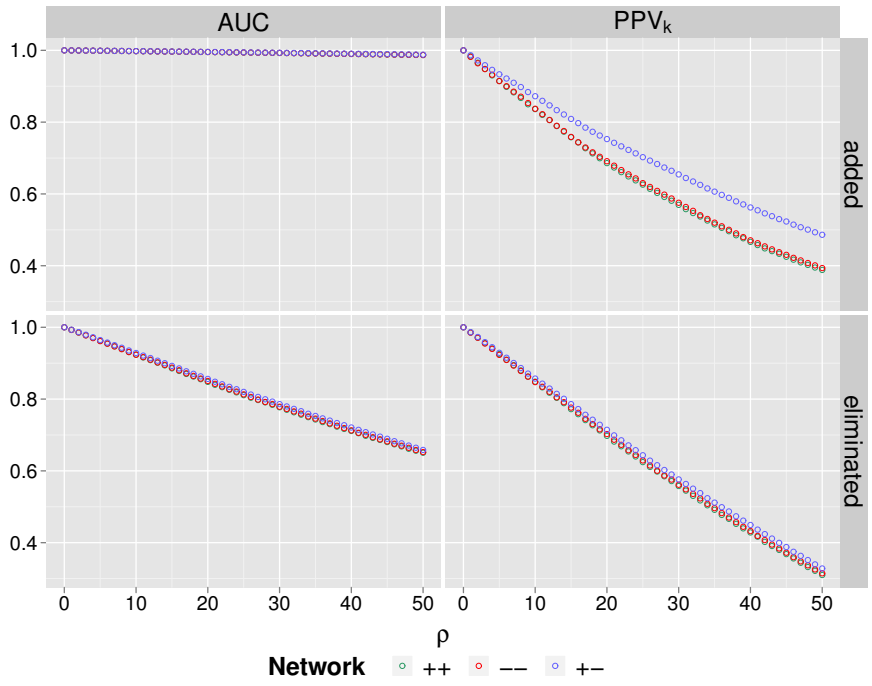


Figure 15: Robustness analysis on artificial data: AUC and PPV_k evaluate the performance of the algorithm on an ensemble of 100 artificial graphs for increasing ρ noise levels. Results are shown for added (upper row) and eliminated edges (lower row). Green data points represent the performance of recovering the $++$ network, red data points refer to the $--$ network, and blue points indicate results for the $+-$ network. Figure reprinted from [115].

Figure 15 (left) shows that the performance measured by the AUC is almost perfect for up to $\rho = 50\%$ added edges, but less accurate for eliminated edges. Since the AUC is unable to detect any changes for added edges, we include a second measure. The PPV_k is equal to the fraction of true positives among the k top-ranked node similarities, where k is the number of elements in the ground truth[3]. The PPV_k shows an inferior quality for both added and eliminated edges, as shown in Figure 15 (right). Nevertheless, the algorithm is able to detect more than 90% of the truly significant edges when $\rho \leqslant 6-8\%$ of

3 For more details about the used performance measures see Section 2.7.

edges are added or $\rho \leqslant 8\%$ are eliminated—depending on the edge type. Two trends are visible in the results:

1. adding random edges affects the quality of the algorithm less than randomly eliminating edges and
2. the different edge types show different sensibility to noise: while $++$ and $--$ edges are practically indistinguishable in terms of precision due to the symmetry of the artificial graphs, the accuracy in case of $+-$ edges is slightly better.

We conclude that the presented method is robust against random noise. The results obtained here with z-scores as test statistic are in agreement with the corresponding values obtained based on p-values [114]. The striking similarity suggests that on the data set at hand, the test statistic used for assigning weights to the edges of the projection is not a defining step as long as the proper random graph model is used for hypothesis testing. Having verified the robustness of the method, we show in the following how it performs on the Netflix data set.

Box 3 | *Problem statement and approach*

Data
1. A bipartite graph connecting films and users by edges that express like and dislike
2. Film attributes such as genre and cast
3. Two ground truth data sets consisting of similar TV shows and feature films

Task Detect films that are similar, based on the user rating patterns and their shared attributes

Framework

A. Project the film–user bipartite graph onto the set of films
   - Threshold the individual projections based on topological criteria

B. Compare the co-like, co-dislike, and like–dislike projections in terms of their ability to detect similar pairs of films based on
   - A classification of the films according to genre
   - The TV show and feature film ground truth sets

C. Explore the potential for identifying similar films based on the combination of different types of projection networks

D. Project the film–attribute bipartite graphs onto the set of films in order to include further aspects of film similarity

8.3 APPLICATION TO THE NETFLIX DATA SET

We use the presented and tested method for the projection of multiplex bipartite graphs to explore the similarity landscape of films based on pairs that are both rated similarly by the same users. Box 3 gives an overview of the adopted approach.

Due to the computational complexity of our method as well as the size of the Netflix data set, an analysis of the entire data set is not feasible. However, in a first approach, Zweig had shown that the absolute number of co-occurrences and their expected values are stable if large enough subsets are used [280][4]. We therefore restrict our investigation to a subset of the whole data set that can be projected in a reasonable time frame and contains 5,000 randomly chosen users with all their ratings. Of these ratings, 590,248 express like (+) while 152,131 ratings express dislike (−). In a first approximation, ratings of 3 are considered neutral and are omitted. During this process, we completely remove 10 users with only neutral ratings and are left with 15,206 films forming a multiplex bipartite graph with two types of edges.

We then annotate the films with genres using the comprehensive list of genres available from the Internet Movie Database (IMDb) [3]. Due to the discrepancy in the titles and release years, our string matching algorithm assigns genres to only 7,314 out of the 15,206 films. The majority of these cannot be classified into a single genre but are tagged with 2 or 3 genres. Figure 16 (left) shows the frequency of the number of genres per film.

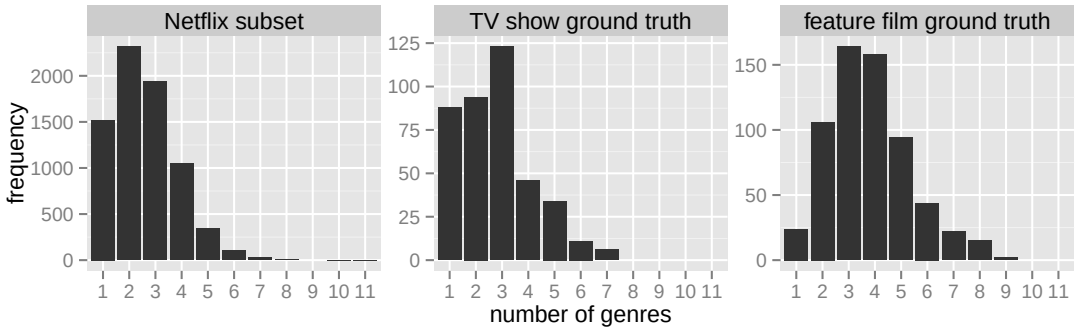


Figure 16: Histograms showing the frequency of films with a given number of genres for the considered Netflix subset (left), the genre-matched TV show ground truth (middle), and the genre-matched feature film ground truth (right). Figure adapted from [115].

4 Although this was tested on the simplex network containing only the good ratings, none of our observations suggest that the bad ratings (with values 1 and 2) are distributed differently throughout subsets of the Netflix data.

8.3.1 *Ground truth data sets*

In order to objectively assess the quality of the projection, we need a ground truth that contains films which are similar or alternatively films that are dissimilar. Constructing ground truth data sets for complex real-world data sets is a challenging task that involves several elaborate decisions and rarely yields a flawless result. While the generation of a ground truth consisting of dissimilar films does not appear to be achievable, there are several possibilities for the construction of a ground truth of similar films. A classification of the films into genres and genre combinations results in a coarse-grained clustering. Although films which belong to the same genre share thematic and stylistic similarities and users often define their preferences in terms of genres, we cannot expect all films belonging to a pure genre or a combination of genres to be similar. Recall that in this context, two films are considered similar if users who liked one of them will also like the other. A more refined grouping can be built based on the concept of film series, though. To avoid biased results due to the inherent inaccuracies in the construction of such a ground truth, we rely on two different approaches and create one ground truth for TV shows and one for feature films:

1. Zweig proposed a ground truth for TV shows based on productions that run for several years and are grouped into seasons, i.e. collections of all episodes produced during one year [280]. She extracted all films with the keyword "season" in their titles and subsequently grouped them by the remaining part of their titles. This procedure results in a usable ground truth, which contains coherent groups of TV shows like *Friends* or *Star Trek* without their spin-offs. Using this procedure and trimming the ground truth to the films from our $++$, $--$, and $+-$ projection networks yields 400–600 TV shows grouped into about 150 cliques, which are small complete graphs of films, i.e. they contain all available seasons of a series and the full set of all possible edges between them.

2. Constructing a ground truth for films other than TV shows is possible by using the comprehensive list of feature film series available on Wikipedia [2]. Due to inconsistencies and duplicate entries inherent to user-generated data, the extraction of this compilation required manual postprocessing before being matched with the Netflix films by title and release year. This results in a ground truth of around 750 films forming around 280 cliques, among which we find the *James Bond* films or the films of the *Three Colours* series.

We consider the two presented ground truth data sets to be directed graphs. This enables us in a subsequent analysis to properly deal with

film similarities detected by our method which are not symmetric. For instance, if we find that film B is most similar to film A, the reverse is not necessarily true and there can be a film C which is most similar to film B.

Since we use the ground truth data sets to evaluate the results of our algorithm, we require them to be representative for the considered Netflix subset. An indication of this property is provided in Figure 16: the frequency of films with a given number of genres in the entire Netflix subset and their frequency in the ground truth data sets show similar trends.

In the following, we perform a one-mode projection for the subset of the Netflix data. We then analyse the resulting projection to identify interesting structures in its topology and to understand the relevance of these patterns in detecting similar pairs of films.

8.3.2 *Characterization of the multiplex projection*

Thesis point 9

We perform a one-mode projection of the Netflix subgraph by generating 10,000 random graphs with the same degree sequences as the original graph. Each of the graphs is created from the previously sampled graph through a sequence of edge swaps by using Algorithm 3. The corresponding multiplex projection contains film pairs that are both liked ($++$ edges forming the co-like network), both disliked ($--$ edges in the co-dislike network) or rated antagonistically by the users ($+-$ edges of the like–dislike network).

Zweig and Kaufmann [283] provided a global analysis based on leverage[5] for the simplex network based on like ratings. However, since leverage is bounded from above by the minimum degree of the two considered films, edges which contain a film with low degree can never be as highly ranked as edges between two films with high degrees. The normalization of leverage as incorporated in the z-score may overcome this problem and improve the detection of similar films. In order to analyse our projection at a global level with the z-score, we choose a threshold for each of the three individual networks ($++$,$--$, and $+-$) based on their topology.

Co-like edges are a certain indicator of similarity, while co-dislike edges only partly indicate similarity. For instance, significantly many users disliked both volumes of *Kill Bill*, obviously a pair of similar films. However, there were several users who disliked both *Gulliver's Travels* (the adaptation of a fantastical tale as a *family* film) and *Chinatown* (Polanski's *neo-noir*, a mixture of *mystery* and *drama*)—a pair of conceptually very different films. Assuming that the co-rating behaviour of the users—if properly denoised—indicates which films are indeed similar, we expect both of these types of edges to be transitive. In order to find a significance threshold we therefore con-

5 For details about the equivalent covariance see Section 4.2.

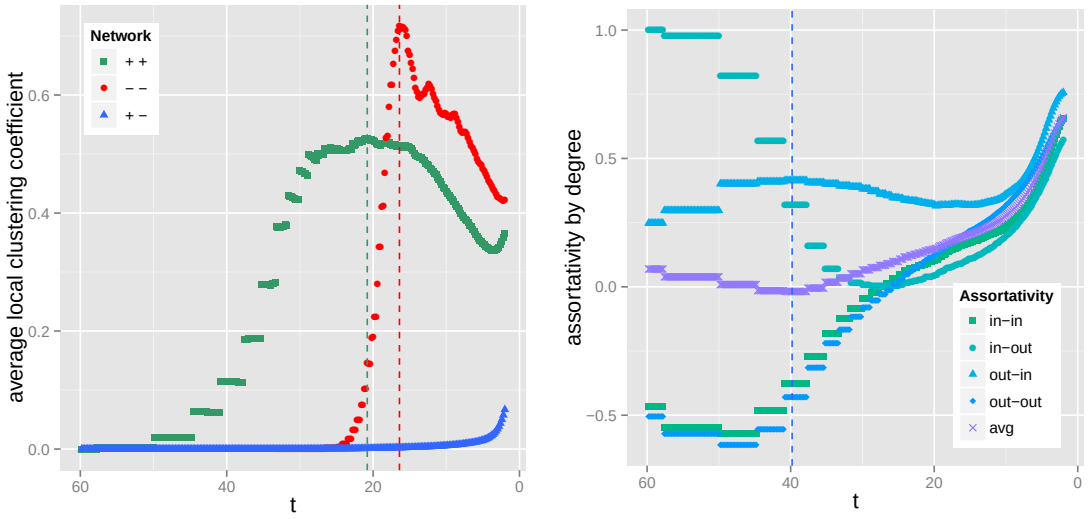


Figure 17: Deducing meaningful significance level thresholds t for the different networks in the multiplex Netflix projection. Left: average local clustering coefficient of the projections which contain film similarities equal to or higher than the candidate z-score thresholds. Results for the various types of networks (++,−−,+−) are color coded (green, red, blue). The thresholds which are indicated by the dashed vertical lines ($t_{++} = 20.8$, $t_{--} = 16.4$) are selected based on this plot. Right: assortativities by degree and their average for the directed +− network at different threshold candidates. The threshold $t_{+-} = 40$ (marked by a dashed vertical line) is chosen based on this plot. Figure reprinted from [115].

sider the clustering coefficient, which quantifies the probability that neighbours of a film are connected themselves (cf. Section 2.2). Monitoring the average local clustering coefficient for subgraphs of $\tilde{G}_{++}$ and $\tilde{G}_{--}$ at varying z-score thresholds t as shown in Figure 17 (left), we see nontrivial changes in topology that indicate which threshold candidates are meaningful[6]. Accordingly, we choose $t_{++} = 20.8$ and $t_{--} = 16.4$ to be the optimal thresholds, as they mark a clear maximum in the average clustering coefficient of the particular networks, suggesting a strong increase in interconnectedness.

The +− network shows a very low clustering with monotonic increase, indicating that there is no specific trend in the way users couple liked and disliked films (see Figure 17, left). Ideally, this directed network should contain hubs: if users liked (or disliked) the majority of films from a given group of similar films and disliked (or liked) one film from the group, then the projection should contain +− edges between this outlier and the rest of the group. The presence of hubs, i.e. nodes with high degree connected to nodes with low degree, is

6 This approach can be extended to other structural measures besides the clustering coefficient. We provide an example of this extension to the number of components and component density in Section 9.3.1.

	t	$\lvert\mathcal{L}\rvert$	$\lvert\mathcal{E}'\rvert$	δ	$\lvert\mathcal{C}\rvert$	$\mathcal{C}_{max}(\%)$	cc	λ
$\tilde{G}_{++}$	20.8	6,248	57,846	0.003	358	81.16	0.53	0.61
$\tilde{G}_{--}$	16.4	3,444	129,097	0.022	532	32.87	0.72	0.97
$\tilde{G}_{+-}$	40.0	2,490	4,267	0.001	314	41.16	0.00	*

*	in	out
in	−0.3760	0.3191
out	0.4140	−0.4297

Table 3: Basic network statistics of the multiplex Netflix projection broken down to the networks induced by the different edge types. Shown are the z-score threshold t, the number of nodes in the node set we project onto $\lvert\mathcal{L}\rvert$ and edges $\lvert\mathcal{E}'\rvert$, the density δ, the number of components $\lvert\mathcal{C}\rvert$, the percentage of nodes in the largest component $\mathcal{C}_{max}(\%)$, the average local clustering coefficient cc, and the assortativity by degree λ which in the case of the directed $+-$ network consists of four correlation values (in–in, in–out, out–in, and out–out).

not reflected by the clustering coefficient. We thus use another aggregated measure for finding the threshold, namely the assortativity by degree as defined for directed networks (see Section 2.2). We expect the number of hubs to be maximal when the network is disassortative, i.e. all assortativities are minimized. Thus, based on Figure 17 (right) we set the threshold to $t_{+-} = 40$.

As a first step in the analysis of the thresholded projection, we focus on the basic statistics of the individual networks. As summarised in Table 3, all three are extremely sparse. In terms of the number of nodes, the $--$ network is smaller than the $++$ network, partly due to the lower number of dislike edges in the bipartite graph in general. The $--$ network is also the most clustered and has an almost perfect assortativity by degree. Besides having the highest clustering coefficient, the $--$ network is also composed of more components than the $++$ network. This is surprising, because the intuition is that co-like edges are more specific than co-dislike edges and we thus expect that they would cause the formation of many tightly connected components. The fact that the $--$ network has more components than the $++$ network could be the result of two factors. On the one hand, blockbusters which are co-liked with films belonging to different groups of similar films bridge these groups, thereby destroying the clustering. On the other hand, systematic recommendation of several films by Netflix's recommendation engine which are falsely classified into a group of similar films results in a co-disliked clique of these films.

Finally, the $+-$ network is built with the strictest threshold and it is characterized by a small number of edges and no clustering. We

establish the significance of these +− edges later, based on the multiplex projection consisting of the ++, −−, and +− edges. So far, we note that it is informative to extract a pair of films A and B that are in a mutual like–dislike relation, as they are an indication that users either liked A and disliked B or vice versa. From these pairs we can deduce the films that were considered controversial by the Netflix users at the time of data collection. The +− subnetwork with the mutual edges contains mainly classics, block busters, and popular films of the considered time frame. Interestingly, it seems that this network reveals subcultures within the popular genres. As examples, Figure 18A shows a component of musicals from the 40s/50s and Figure 18B one of horror and mystery films from the 80s/90s.

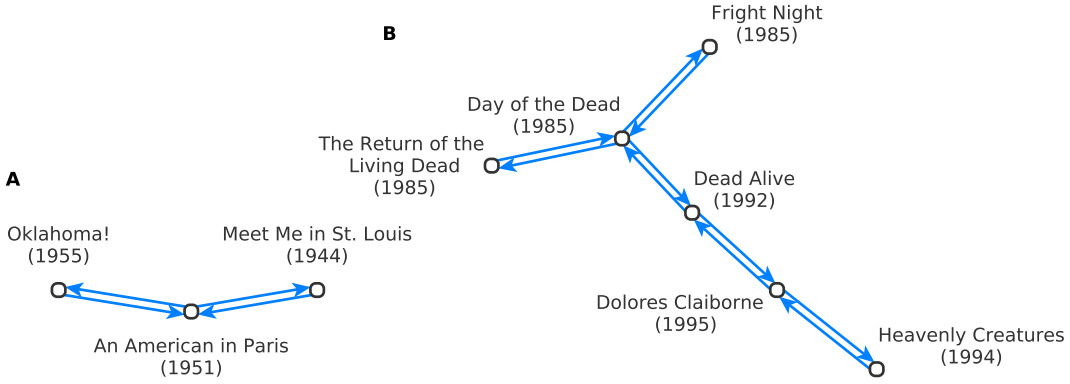


Figure 18: Two components of the like–dislike subgraph. Only the mutual edges are shown. The projection was created by using p-values and applying a threshold of $t_{+-} = 0.004$. Figure adapted from [114].

On a side note, we could technically also project the *aggregated* network, i.e. the original bipartite graph that contains both edges expressing like and dislike. However, due to the inability of the aggregated network to differentiate between the distinct connotation of the edge types, the resulting projection is uninformative with respect to the ++ and −− edges and misleading regarding the +− edge type [114]. For example, such a projection contains an edge between *Tootsie* and *Rosemary's Baby*. This edge between a romantic comedy and a horror film is deceiving: the multiplex projection reveals that this connection should be actually of type +− because significantly many users liked *Tootsie* and at the same time disliked *Rosemary's Baby*, a distinction that is otherwise lost due to aggregation.

SHORTCOMINGS OF THE GLOBAL-LEVEL ANALYSIS To evaluate the multiplex projection, we use the ground truth for TV shows and feature films as described above. Based on these, we know for each film in the ground truth the number of k other similar films our algorithm should find. For each such film we then separately rank

	individual		uniform	
	TV shows	feature films	TV shows	feature films
$\tilde{G}_{++}$	0.70	0.31	0.76	0.47
$\tilde{G}_{--}$	0.06	0.04	0.11	0.05
$\tilde{G}_{+-}$	0	0	0	0

Table 4: Average local PPV_k for the TV show and feature film ground truth for the projection obtained with strict individual thresholds and a standard uniform threshold.

its neighbours from the three projection networks decreasingly by z-scores and compare the k top-ranked films from this list to the films contained in the ground truth clique. To compensate for possible bias in the rankings due to ties, we use 10 rounds of randomization for tie braking. By this procedure we have computed the local PPV_k, i.e. the fraction of truly similar pairs of films among the k highest ranked pairs, where k is the degree of each individual film from the ground truth. The PPV_k values averaged over all considered films are shown in Table 4 under the "individual" header. According to this, the $++$ network is fairly good at detecting similar films, while edges deduced from the $--$ network provide only a weak inference. As expected, the most significant like–dislike edges *never* indicate meaningful film similarities.

There is a considerable difference in the performance of our algorithm with respect to the two different ground truth data sets: 70% of the detected film similarities are correct for TV shows, while only 31% are correct for feature films. Possible explanations are that *1)* fans of TV series are more loyal or have clearer preferences; *2)* for many TV shows it is difficult to understand a single episode without having watched the majority of previous episodes; *3)* feature films belonging to a series are separated by a larger time span (a couple of years as compared to a few months) and are also more independent in terms of story, style, and crew; *4)* inaccuracies in the ground truth are more prominent for feature films than for TV shows, i.e. the feature film ground truth we constructed by matching the films from Netflix with the user-generated Wikipedia lists is qualitatively worse than the more straightforwardly constructed TV show cliques.

Finally, we compute the PPV_k values in a similar manner based on the projection with the less stringent standardized threshold $t_{++} = t_{--} = t_{+-} = 2$. The resulting $++$ and $--$ rankings better resemble the cliques from the ground truth, meaning that with the strict thresholds we excluded some of the correct film similarities, as shown in Table 4 under the "uniform" header. This indicates that for the application at hand, a local analysis is preferable over a global one.

8.3.3 *A coarse-grained analysis based on genres*

A one-mode projection of bipartite graphs is often used as a method for finding groups of similar objects, i.e. it results in a clustering of the set of nodes we project onto. In the following, we consider the similarity landscape of the films generated by the multiplex projection and grouped by the most intuitive, coarse-grained classification of films, namely into genres. This classification, although arguable in its exactness due to evolving genre definitions and the flexible borderlines between individual genres, is a good opportunity for identifying tendencies in the placement of the different edge types and to check whether the detected film similarities are meaningful beyond the available ground truth sets.

Without the strict individual thresholds, the projection with the uniform threshold $t_{++} = t_{--} = t_{+-} = 2$ on the z-scores is substantially larger and requires a higher-level investigation. For this purpose we aggregate films which belong to the same genre into nodes of a *genre network*. Edges of different types between the films are divided into those connecting films of the same genre (*intra-genre edges* represented by self-loops in the genre network) and edges connecting films from different genres (*inter-genre edges* represented by normal edges in the genre network). As like–dislike edges emphasise the antagonistic perception of two films, their direction is of little importance at this stage and is thus disregarded. The individual edges between the films are grouped according to the genres the films belong to and then counted. The resulting value is normalized by the number of possible connections between all films of the given genres. Due to the fact that the same films belong to multiple genres, the number of possible inter-genre edges for genres A and B containing n_A and n_B edges respectively is given by:

$$n_A \cdot n_B - \frac{n \cdot (n+1)}{2}$$

where n is the number of films in the intersection. For the directed case of the $+-$ edges, the number of possible edges is doubled. Given the projection network, we obtain the fraction of realised edges in the way described above and assign this value as weight to the edges of the genre network. Figure 19 shows the 20 highest-valued edges for each edge type in this genre network. With the exception of *fantasy*, which has two self-loops, all genres with self-loops fall into one of the following two categories:

1. Genres with triple self-loop. Here the $++$ and $--$ self-loops indicate agreement within the individual genres, while the $+-$ intra-genre edge suggests divided opinions among users. Thus, films belonging to these genres seem to be highly debated.

2. Genres with a single self-loop. Single self-loops stand for undivided opinion. These are mostly classical Hollywood genres that have passed their golden times, such as *musical*, *western*, and *film-noir*.

We note that the weights of the ++ edges are one order of magnitude higher than the other two. Most of the genres have a small ++ degree. *Film-noir* is an exception as it is mostly watched by connoisseurs and is one of the most conservative Hollywood genres with no inclination to renewal. *Animation* has the highest +− degree and it is a genre with an extremely different target audience and evidently distanced from genres like *crime*, *action*, or *war*. As opposed to the other two networks, the −− network shows striking triangles linking thematically and stylistically connected genres like *science fiction–horror–thriller* or *adventure–family–fantasy*. A ++ and −− edge between the same two genres, as in the case of *science fiction–horror* and *family–fantasy–animation*, suggests more closely related genres.

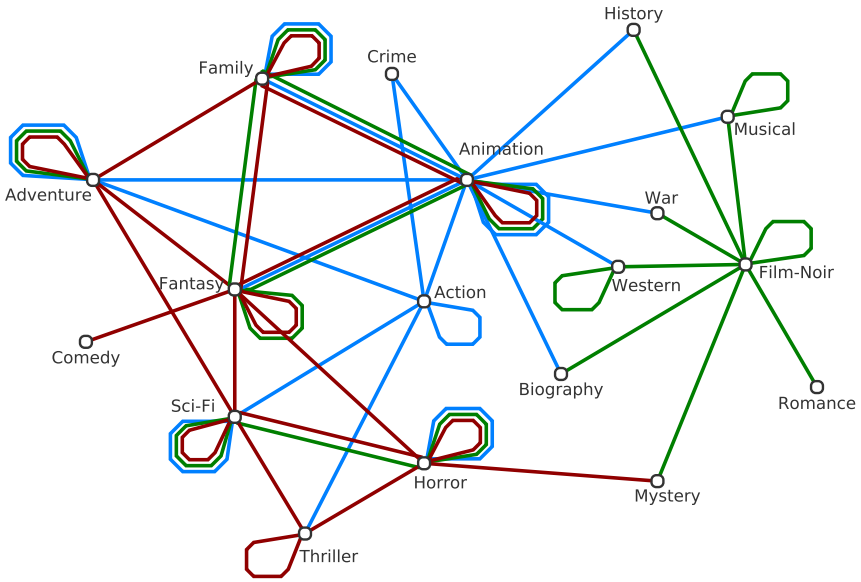


Figure 19: The genre network containing the top 20 edges for each edge type. Nodes are contractions of all films of the multiplex projection with the given genre, while edges are an aggregation of all edges of a given type between two genres. The direction of the like–dislike edges was disregarded. Colour coding as before. Figure reprinted from [115].

Assuming that films which belong to a genre represent a cohesive unit, we expect that the realized number of edges normalized by the number of possible edges is higher within a genre than it is between different genres. To test this, we consider the genre network as constructed above and, for each genre, count how often the normalized number of its intra-genre edges exceeds the number of inter-genre

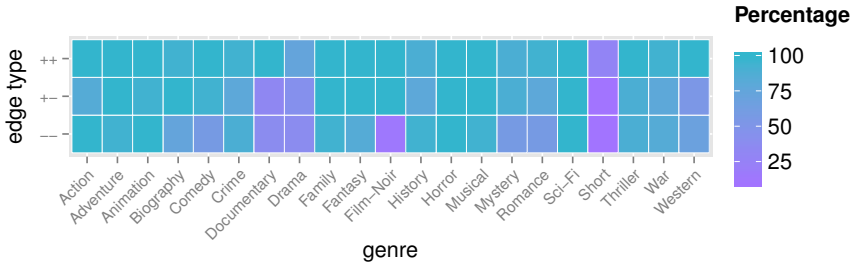


Figure 20: Comparison of the frequency of realized intra-genre edges to the frequency of realized inter-genre edges, deduced from the complete genre network. For each type of edge and each genre, we show the percentage of other genres to which the considered genre has a lower frequency of realized edges than to itself. Figure reprinted from [115].

edges with respect to all other genres. By this, we obtain the percentage of other genres to which the considered genre has a lower frequency of realized edges than to itself. Figure 20 shows the obtained percentages for the different genres and types of edges. In agreement with our expectation, the percentages are highest for the ++ network, reinforcing the observation that user preferences are correlated with the genres. However, the classification of films into genres is acceptable as a first approximation for a ground truth only at the qualitative level, because this classification does not allow us to decide which of the edges that we used during our analysis actually represent similar films. In order to perform a quantitative analysis, we turn in the following to our ground truth of film similarity.

LOCAL RANKINGS PER GENRE Using only those films from the ground truth data sets that could be annotated with a genre, we compute the local PPV_k for each of these films. When computing the PPV_k for film X, k denotes the number of other films that are similar to X according to the ground truth. Figure 21 shows the local PPV_k values averaged over the films belonging to the given genres. Accordingly, there are clear differences in the performance of detecting pairs of similar films belonging to the distinct genres. For instance, based on our ground truth data sets, pairs of similar *family* and *romance* films are rather hard to identify, while detecting pairs of similar *action* films proves to be easier. With the exception of *westerns*, the detection for TV shows is better than for feature films and the difference between the two values can rise to more than 0.4. One of the prerequisites for a good performance of our method is that users rate the films belonging to a given genre in a consistent way[7]. Intuitively, smaller and more specific genres thus have better chances of attaining a good

7 This is verified for most genres, see Figure 20.

performance. Interestingly, our results indicate that even large genres such as *action* films can perform well. For instance, there is a much higher agreement between users with regard to *action* films than there is for *romance* films.

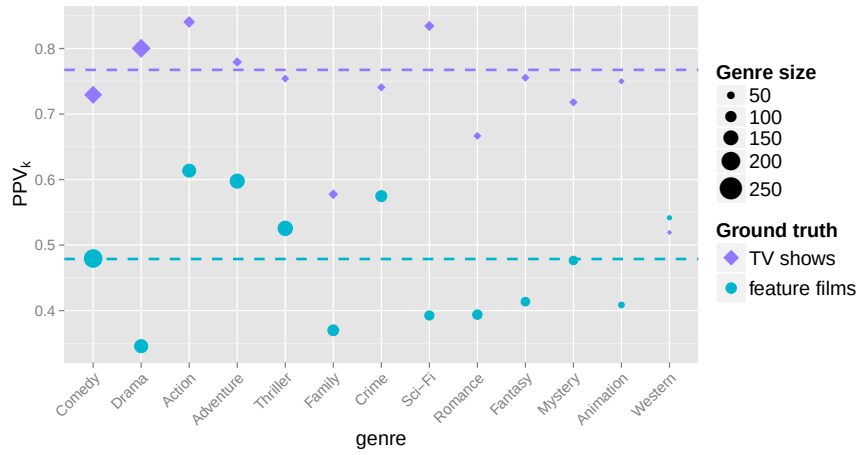


Figure 21: Quality of the rankings for the different genres as measured by the local PPV_k. Shown are results on the TV shows (purple diamonds) and feature films (blue dots) with genre annotations. The performance of the algorithm on the entire TV show and feature film ground truth is marked with dashed lines (cf. Table 4). The size of data points indicates the number of films with the given genres. Figure reprinted from [115].

8.3.4 *The role of the co-dislike and the like–dislike networks*

Thesis point 10

The trends we presented above motivate ideas on how to improve the film similarities deduced from the $++$ network by using the $--$ and $+-$ networks. The key observation for this is that the multiplex projection contains *overlapping edges*, i.e. different types of edges between the same pairs of films. Co-dislike edges indicate similarity, and thus could be used to augment the results obtained from co-like edges through the aggregation of overlapping $++$ and $--$ edges, in a first approach for instance by adding their respective z-scores. Like–dislike edges indicate dissimilarity and therefore they could be used to annihilate or weaken co-like edges which do not connect truly similar films, for example by subtracting their z-scores. In the following experiment we test these two ideas. We progressively decrease the z-score threshold for the $--$ (or $+-$) edges, thereby obtaining a larger fraction of these networks in every step. We then update the rankings of the $++$ network according to the newly introduced $--$ (or $+-$) edges, as a proof of concept for example by creating the sum of the z-scores of overlapping edges (in the case of the $--$ network) or by eliminating overlapping edges (in the case of the $+-$ network).

For each threshold value, we compute the average local PPV_k and monitor its progression.

Using the multiplex projection in such a way requires the existence of overlapping edges. When considering the Netflix subset, the fraction of co-like edges that overlap with co-dislike edges is 7%, while 11% of co-like edges overlap with like–dislike edges. Due to the fact that this overlap is relatively small and the sets of $--$ and $+-$ edges are not complete, we did not find this approach to produce a significant and consistent improvement of the results for our data set.

An ideal scenario which would enable such an analysis is illustrated by a subgraph taken from the multiplex projection. It contains two ground truth cliques with differing target audiences (see Figure 22). The provocative action-comedy series parodying well-known spy films and the popular series of family films are both eligible for clear fan bases as well as opponents. Thus, they are co-liked and co-disliked by a significant number of users. While the *Austin Powers* films form a perfect $++$ and $--$ clique, sequels of *Home Alone*, which have a greater time span between each other (up to 12 years between their release dates), fail to keep their audience. The additional like–dislike edge between the first two films of the *Home Alone* series indicates a frequently observed phenomenon, namely that fans of a production can be disappointed by the follow up of a beloved film. The remaining $+-$ edges lie exclusively between the two cliques and are evidence of viewers from both cliques stating their dislike of the other.

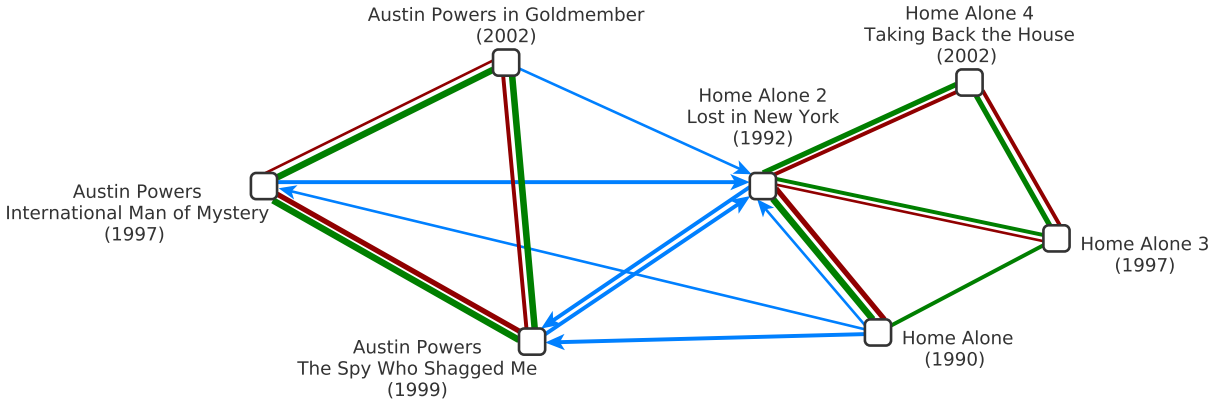


Figure 22: Exemplary subgraph of the multiplex projection showing the *Austin Powers* and the *Home Alone* films. The two almost perfect $++$ (green) and $--$ (red) cliques are connected by $+-$ edges (blue arrow from the liked film to the disliked one) indicating the differing tastes of their individual fan bases. The thickness of the edges relates to their z-score. Figure reprinted from [115].

8.4 FILM SIMILARITY BEYOND THE MARKET BASKET SETTING

Thesis point 11

Projecting the film–user bipartite graph sheds light on one aspect of the similarity of films, namely the similarity deduced from co-ratings. Provided that the data set underlying the graph is comprehensive enough, such an approach can be very rewarding, as this notion of similarity is highly relevant for recommender systems and constitutes the basis for collaborative filtering [270, p. 63–64,169–175].

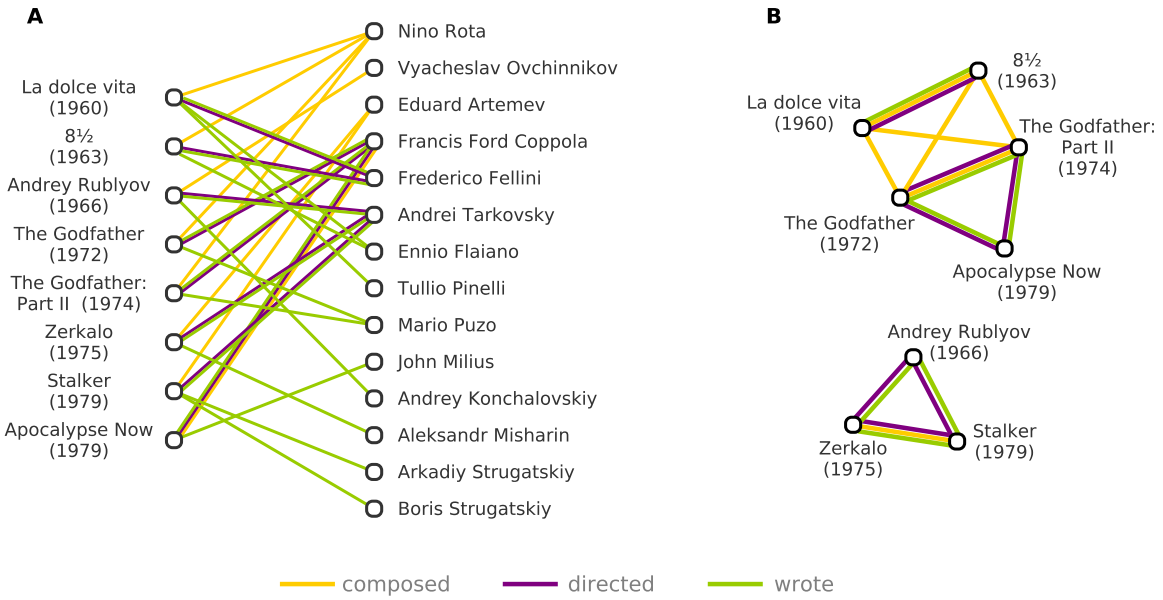


Figure 23: Using attributes of films to obtain a more nuanced notion of film similarity. (A) Excerpt from the film–composer, film–director, and film–writer bipartite graphs. (B) Exemplary subgraphs of the multiplex projection.

Nevertheless, the presented framework enables studying various other aspects of film similarity as well. For instance, we can obtain a more nuanced notion of similarity by using further discrete attributes of films that can be represented as different bipartite graphs, in which the films constitute the node set $\mathcal{L}$, while the diverse attributes form the node set $\mathcal{R}$ (see Figure 23A). One such possible attribute is film genre. As it provides an intrinsic indication of the similarity of films, we have already used it as a coarse-grained ground truth. Further attributes such as the cast (e.g. the actors), the creators (e.g. the director, writer, and composer), and the technical specifications (e.g. colour, sound, and film format) are available on the Internet Movie Database [3]. The fixed degree sequence model is a meaningful choice because it takes into account the popularity of an actor (in the film–actor graph), the productivity of the crew (in the film–main creators graph), and the adoption of a certain technology (in the film–technical

specifications graph). Projecting these bipartite graphs individually results in a multiplex film network in which films are connected whenever their similarity in terms of these aspects is statistically significant. Figure 23B shows exemplary subgraphs from the obtained multiplex projection. The possibilities of incorporating this manifold similarity into a recommender system are rewarding avenues for future research.

8.5 DISCUSSION AND CONCLUSIONS

The algorithm for the one-mode projection of multiplex bipartite graphs we presented in this chapter extends a simplex projection method [280, 283]. The adaptation of the simplex method consists of making the random graph model sensitive to the edge types. This is a necessary step which assures the existence of a proper null model for establishing the relevance of the edges in the projection. When computing a multiplex projection of a similar subset of the Netflix data set, we observe that the projection of the aggregated network (the bipartite network with no differentiation between like and dislike) inevitably mixes up edges with different connotations. This results in misleading similarities, showing that a more nuanced random graph model is needed.

WHICH TEST STATISTIC TO USE AS SIMILARITY MEASURE? Several possible statistical tests can be used for assessing the significance of edges in the resulting projection (cf. Chapter 4). The use of leverage (or the covariance) leads to meaningful film similarities at the top of the global ranking [280, 283]. However, the highest-rated pairs were all blockbusters, since leverage is bounded by the smaller degree of the individual films and thus disfavours less popular films. Another widely used statistical test, the p-value, does not directly depend on the degrees of the nodes in the bipartite graph. Still, in the case of the Netflix data set, it was not well-suited for a global-level analysis, because it assigned maximal significance to a relatively large amount of edges in the projection, unless the number of samples was infeasibly high [114]. Thus, in this chapter we mainly used the z-score, which is a normalised version of leverage and best differentiates between the edges in the projection.

HOW ROBUST IS THE METHOD? We showed that the method is very stable based on artificial data, for both p-value [114] and z-score, as long as the number of added or removed edges lie in the range of noise we expect from market basket analysis data. Future research will have to show whether the algorithm is more sensible in the case of more difficult applications (e.g. nodes with extremely low/high degree).

WHY ARE OUR RESULTS BETTER ON TV SHOWS THAN ON SERIES OF FEATURE FILMS? We presented a framework for the analysis and evaluation of the multiplex similarity landscape of films. For this purpose, we used the classification of films into genres and two ground truth data sets constructed based on the concept of film series. Compiling sound ground truth data sets is difficult (already at a conceptual level) due to the manifold aspects and the inherent subjectivity of film similarity. While the episodes of a TV show are usually released within a short time span and the variance in terms of their story, director, and cast is negligible, ground truth-construction for feature films lacks these criteria and is thus more limited. On a technical level, information about series could be attained only for a subset of the films present in the multiplex one-mode projection because we aggregated data from diverse sources (Netflix, Internet Movie Database, and Wikipedia). The ground truth data sets suggest that similarities between TV shows are easier to detect than those between feature film series. Two possible explanations for this discrepancy are *1)* the more "faithful" fanbase of TV shows, i.e. sequels of TV shows are rated more consistently throughout our data set than series of feature films and *2)* the less accurate ground truth for feature films. Our genre analysis revealed that detecting similar films that belong to some of the genres is harder than for others. This was typically the case for broader categories like *animation*, which is a technique rather than a genre and contains films as diverse as the Disney productions and Japanese animes with mature content. Additionally, the performance for a genre increases not with its size, but rather with the consistency of the ratings given to the films it contains.

GLOBAL OR LOCAL ANALYSIS? Applying the presented method to the Netflix data revealed that a global-level analysis is only partly meaningful. Since there is no ground truth that would contain the globally most similar pairs of films, we use the local ground truth for verification instead. We found that some of the pairs of similar films contained in our ground truth data sets were assigned lower z-scores by our algorithm than other pairs, which were not similar based on the ground truth. These correct pairs were thus eliminated from the network through the overly strict thresholding, leaving pairs of outliers among the highest-ranked edges. Their presence can be explained through statistical reasoning, as there always exist random ratings that can not be differentiated from truly significant ones. This problem is not specific to our method, but an inherent issue of any statistical approach with the same purpose.

On the local level we found that edges of the co-like network are well-suited for detecting similar films. The co-dislike network on the other hand is too unspecific to be used by itself, while the like–dislike network finds only pairs which are not similar according to the ground

truth data sets. Although our attempts of using the latter two networks did not improve our results reliably on the considered data set, we argue that a more comprehensive record of user ratings expressing dislike (resulting in a more complete set of co-dislike and like–dislike edges) would contribute more reliable information to our efforts.

FUTURE WORK The main limitation of our method can be found in the range of nodes with a low-degree in the bipartite network. Future work in this direction could therefore include the acquisition of data from other sources to compensate for scarce statistics.

Interesting questions arise from the analysis of the Netflix data as well. For instance, here we did not take into account the temporal aspect of the data. Netflix collected the ratings between October, 1998 and December, 2005 and the release years of the considered films vary from 1896 to 2005. It is an open question how these informations could be used to improve the results.

Furthermore, we projected a bipartite graph with two types of edges. Nevertheless, the presented method can be generalised to more types of edges. Thus, instead of concentrating on ratings that express like and dislike, we could directly project the bipartite graph that contains an edge type for each of the five distinct ratings, possibly allowing for a more nuanced analysis. Moreover, this approach could be refined by using results from the psychology of online rating systems for developing normalization schemes to be applied to the ratings of individual users.

We conclude that beyond the specific problem tackled here, the analysis of multiplex networks with different types of entities is an important area of future research in complex network analysis that has not yet been explored to its full potential. The versatile method presented here is not limited to the Netflix application. As we will see in the following Chapter 9, a similar approach can be used for a biological data set and enables us to identify biomolecules that hinder the growth of an especially lethal type of breast cancer.

8.6 SUMMARY

Inference of association networks in general and the one-mode projection of multiplex bipartite graphs in particular is a key tool in analysing data with inherent bipartite structure. In this chapter, we presented a framework that is based on the null model approach and uses a multiplex fixed degree sequence ensemble as reference for significance assessment. We showed the robustness of the method on artificial data before applying it to a real-world network of user ratings for films, namely the Netflix data set. Based on the assumption that co-ratings of films contain information about the films' similar-

ities, we analysed the multiplex projection as an approximation of the similarity landscape of the films. In addition to comparing the projection to the coarse-grained classification of films into genres, we validated the resulting similarities based on ground truth data sets containing film series. Our analysis confirmed that the network of positive co-ratings can be used to detect similar films. Furthermore, we explored the potential of additional, mixed co-rating patterns in improving the detection of similarities and highlighted necessary criteria for this approach. Based on additional data about the crew, cast, and technical specifications of films, we detected further aspects of film similarity using the same framework. We expect that the generality of the method can be further exploited for multiplex bipartite graphs and in the next chapter we demonstrate this on a very different application from systems biology.

9

ASSESSING THE STATISTICAL SIGNIFICANCE OF MILD CO-REGULATION

From computational modelling to experimental validation

The work presented in this chapter has been published as

S. Uhlmann, H. Mannsperger, J.D. Zhang, E.Á. Horvát, C. Schmidt, M. Kühlbeck, F. Henjes, A. Ward, U. Tschulena, K.A. Zweig, U. Korf, S. Wiemann, and Ö. Sahin, Global microRNA level regulation of EGFR-driven cell cycle protein network in breast cancer, *Molecular Systems Biology, 8:570, 2012*

E.Á. Horvát, J.D. Zhang, S. Uhlmann, Ö. Sahin, and K.A. Zweig, A network-based method to assess the statistical significance of mild co-regulation effects, *PLOS ONE, 9(8):e73413, 2013*

A. Spitz, K.A. Zweig, E.Á. Horvát, SICOP: identifying significant co-interaction patterns, *Bioinformatics, 29(19):2503–2504, 2013*

High-throughput screening is a well-established tool for large-scale experiments and provides an overview of how different cellular variables change under various conditions. Such experiments monitor for instance the alteration of protein levels due to the presence of different transcription factors and changed environmental conditions like starvation or enhanced radiation [36]. Biological or chemical perturbations that specifically influence single gene expression—including small interference RNAs (siRNAs) or microRNAs (miRNAs)—have been used alongside protein assays to systematically study the relationship between gene expression and function [221]. miRNAs are a large class of small non-protein-coding RNAs that usually (but not exclusively [260]) function as negative regulators. It is known that they play an essential role in the development and maintenance of many diseases: for example, they are tumour suppressors or oncogenes (oncomirs) in various types of cancer [56, 80, 110, 159, 134, 204, 158]. There are over 2,000 mature human miRNAs registered in the miRBase release 19 [4, 224] and these can target over 60% of the mammalian genes [87] whose corresponding proteins display diverse functions.

Until recently, large-scale experiments designed to investigate regulatory relationships between miRNAs and protein-coding genes have been used to either study one or few miRNAs against a large number of genes (on the transcriptomic [149] or the proteomic [20, 229] level), or test a library of miRNA mimics or inhibitors against one or few genes [140]. In either approach, univariate analysis prevalent in high-throughput analysis [161] has been frequently applied to rank targets or perturbations, for instance by z-score or p-value, in order to interpret the results. It is known that large-scale experiments often come with the trade-off that not all of the results are very reliable [183]: the preparation of cells and tissues, variances in the chip, detection mediated by antibodies, and sensors that quantify signals are all independent sources of noise. To avoid false-positive results, a strict threshold on these values assures that only those effects are reported that have a low probability to be caused by random or non-functional fluctuation around the resting level, e.g. due to handling or measuring errors. It has however been confirmed that many of the protein regulating effects of the whole human genome miRNA (miRome) are mild [20, 229, 258]. These mild effects can only be detected if observa-

tions with a low significance are also included in the analysis, which in turn increases false-positive results.

This problem of detecting mild regulation effects was the motivation behind our computational approach: based on the methodological framework presented in Chapter 3 and Chapter 4, it is computationally feasible to determine whether the number of shared co-regulation conditions of two proteins is statistically significant or not. The implication is then that if two proteins are co-regulated by a significant number of regulating conditions, these regulation effects have a higher chance to be true-positive regulating effects than their individual z-scores suggest. Furthermore, by identifying pairs of proteins that are significantly co-regulated, experimentalists can make hypotheses of functional relationships following the *guilt-by-association* principle [240, 208].

In this chapter, we present the network-based method and give the details needed for applying it in diverse biological settings. For instance, we discuss when to use it (noisy data containing mild effects) and which decisions are required in order to apply it (especially concerning the choice of meaningful significance thresholds). The idea was motivated by the specific biological question raised in a high-throughput study conducted by our collaboration partners: *How to map regulatory network structures in the EGFR-driven signalling system modulated by human miRNAs?* Subsequently to our analysis, our collaborators provided experimental validation for several of the predictions obtained with our method.

Besides determining co-regulation patterns, the framework is generally applicable to any biological data set that contains two types of interacting entities. In network terms, the data set must have a bipartite structure. The method is similar to the method used for the Netflix application (see Section 8.1) and is based on the theoretical prerequisites presented in Section 3.3.4, Section 3.4.1, and Section 4.3. Its main feature is its robustness against noise, which we demonstrate here on artificial data sets that emulate a possible biological structure (Section 9.2). The advantage of artificial data sets is that they can be constructed in such a way that the ground truth (i.e. the true positive and negative results to be found by an optimal algorithm) can easily be determined. As in the case of the market basket application, we show on artificial graphs that mimic the structure of the biological data that the method is robust against random elimination and random addition of observations. These model two typical sources of noise in biological data.

Furthermore, this chapter presents the analysis of a real data set between the genome-wide human miRNAs (miRome) and a subset of proteins in the EGFR-driven signalling system in an *in vitro* model of human breast cancer. We provide key features of co-regulated miRNAs (Section 9.3.1), report the results for protein co-regulation

(Section 9.3.2), and discuss the general applicability of the method in systems biology (Section 9.4). Finally, we provide the open-source software implementation SICOP (SIgnificant CO-interaction Patterns) available under a GPL licence [237] (Section 9.5). Box 4 gives an overview of the tasks and approaches used in this chapter.

Box 4 | *Problem statement and approach*

Data High-throughput screening data that monitors the effect of various miRNAs on a selection of proteins

Task Identify statistically significant co-regulation patterns and detect miRNAs that are potential drug targets

Framework

A. Build multiplex miRNA–protein bipartite graphs that differentiate between up- and down-regulation and correspond to different stringency levels
B. Project the bipartite graphs onto the set of miRNAs
   - Threshold the individual multiplex projections based on topological criteria
   - Compare the groups of co-down-regulated miRNAs with the miRNA families
C. Project the bipartite graphs onto the set of proteins
   - Construct a consensus graph from statistically significantly co-regulated proteins
   - Identify proteins that belong to the same functional module
   - Detect miRNAs that co-target the cell cycle proteins

9.1 FROM REGULATION GRAPHS TO CO-REGULATION GRAPHS

9.1.1 *Building a bipartite graph model from protein array data*

The high-throughput data was obtained by transfecting cells from the human breast cancer line MDA-MB-231 with a library of 810 miRNA mimics. The level of 26 different proteins from the EGFR-signalling pathway was then measured to monitor the effect of various miRNAs on them. The data was processed such as to result in a z-score for each pair of miRNA and protein, which quantifies the change in the expression level with regard to the protein's resting level.

To build the basic bipartite graph, we determine a hard threshold t_B. Given the data and the threshold t_B, the bipartite graph model contains an edge between any pair of miRNA and protein if the absolute value of the corresponding observed z-score is at least as large as t_B. Note that these edges are unweighted, i.e. all of the edges are

treated equally after this step, regardless of the value of the original z-score. However, we differentiate between those edges with a positive z-score (*up-regulation*) and those edges with a negative z-score (*down-regulation*). Figure 24A–C shows schematically how the protein array data is transformed into an unweighted bipartite graph. Alternatively, different thresholds can be used to filter up- and down-regulations.

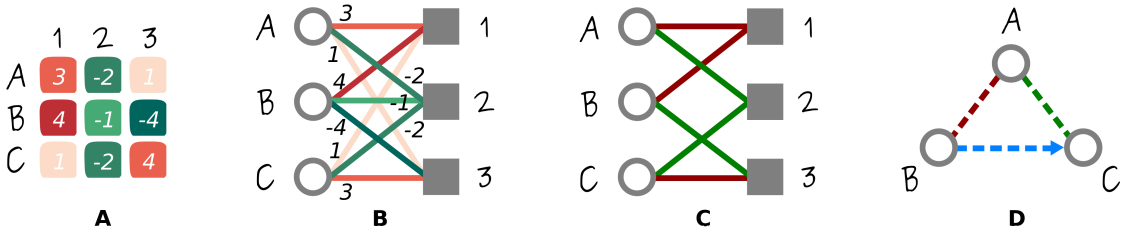


Figure 24: Converting the normalized z-score array into a bipartite graph and illustration of the co-regulation patterns of interest. (A) Exemplary array depicting the normalized z-scores of the change in expression level for proteins A, B, and C when cells are transfected with miRNAs 1, 2, and 3. The z-scores are specified by the white labels. (B) The corresponding bipartite graph where z-scores are represented by weighted edges; the weights are shown as labels on the edges. (C) After applying a threshold t_B to the weights, only some connections are retained. In this case t_B equals 1.96, corresponding to a p-value of 0.05. Edges with a positive weight (up-regulation) are shown in red, edges with a negative weight (down-regulation) in green. (D) Protein co-regulation graph based on the co-regulation patterns as described in the text. Colours denote the co-regulation pattern: the red edge denotes co-up-regulation; the green edge denotes co-down-regulation; the blue, directed edge from B to C indicates that B is down-regulated while C is up-regulated by the same miRNA. Figure reprinted from [119].

The higher the z-score threshold, the smaller the probability that the change in the protein level is merely a random fluctuation, and subsequently the fewer edges are present in the bipartite graph. As stated above, the goal is to understand mild regulation effects, which can only be analysed if the threshold is moderately low. In the following, we choose three thresholds: 2.58 (corresponding to an unadjusted, two-sided p-value of 0.01), 1.96 (p-value of 0.05), and 1.64 (p-value of 0.10). The unweighted bipartite graph that results from thresholding the weighted bipartite graph at t_B is henceforth called the *regulation graph* B at t_B (see Figure 25).

9.1.2 *Multiplex co-regulation patterns*

In the setting described above, we are interested in the statistically significant co-regulation of either the proteins or the protein-regulating conditions (hereafter: miRNAs). As in the case of the film–user

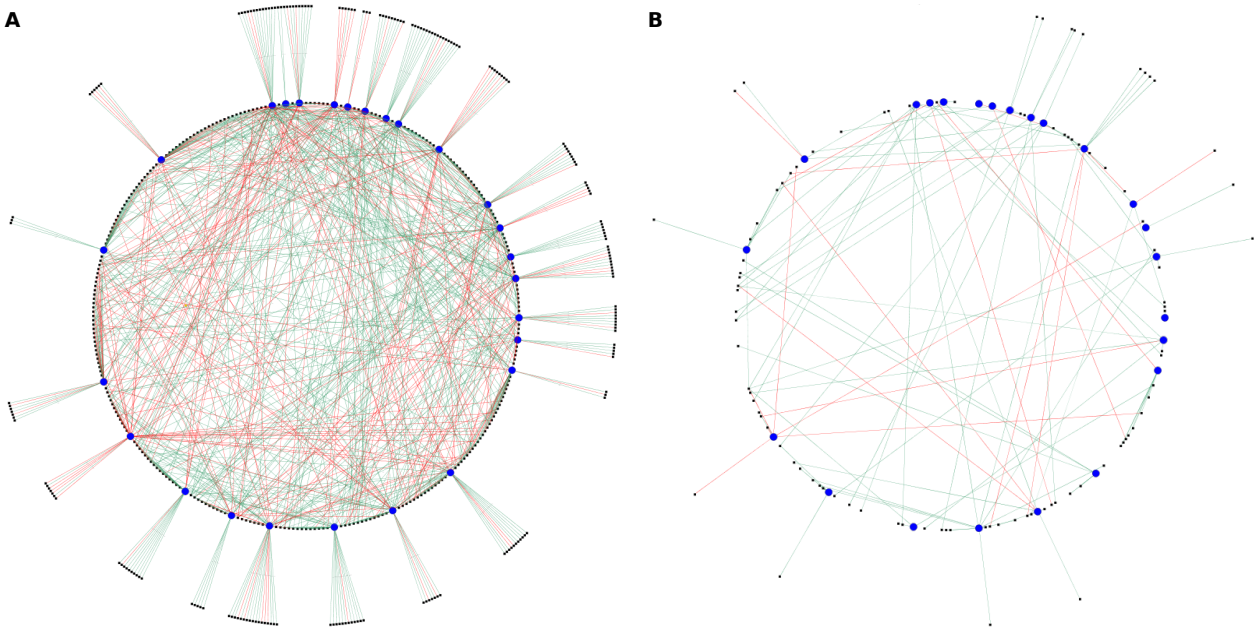


Figure 25: Bipartite graph models of the whole-genome miRNA regulation data of the EGFR/cell cycle proteins. (A) Dense miRNA–protein interaction network at the z-score threshold of 1.96 ($p = 0.05$). Blue nodes on the inner circle represent the proteins and black nodes indicate the miRNAs. While green edges between a miRNA and a protein show down-regulation of that protein by the miRNA, red edges show the up-regulation of the protein by the given miRNA. miRNAs that regulate more than one protein are located on the inner circle, while those that regulate only one protein are placed on the outer circle. (B) A much sparser miRNA–protein interaction network obtained from a more stringent threshold: $z = 3.29$ ($p = 0.001$). Figure adapted from [258].

bipartite graph model from Chapter 8, the protein–miRNA graph contains two different types of edges (up- and down-regulation effects) and thus, its one-mode projection (or in other words, the generated association network) displays the following connections that can be defined for both proteins and miRNAs, as illustrated by Figure 24D:

1. *Co-up-regulation*: A and B are both up-regulated by the same miRNA 1, represented by the two red edges connecting A and B to 1;

2. *Co-down-regulation*: A and C are both down-regulated by the same miRNA 2, represented by the two green edges connecting A and C to 2;

3. *Antagonistic regulation*: B is down-regulated by miRNA 3 while C is up-regulated by it. This antagonistic co-regulation is denoted by a *directed* edge (represented by an arrow) from B to C.

Note that in principle, each pair of proteins or miRNAs could be connected by all four types of co-regulation patterns and thus be connected by all four possible edges (red, green, and a blue edge in either direction). In reality, we expect that two proteins or miRNAs are either *1)* in only one relationship, or *2)* at the same time co-up-regulated and co-down-regulated (connected by one green and one red edge), or *3)* reversely co-regulated (blue edges in both directions). Next, we present why and how our method for the projection of multiplex bipartite graphs discussed in Section 3.3.4 and already applied to a market basket analysis data set Chapter 8 can be used to assess the statistical significance of co-regulations.

9.1.3 *Inference of the association network by finding significant co-regulation patterns*

Given z-scores from a large-scale protein regulation experiment and a threshold t_B on the observations to be included into the graph model, the number of co-regulating miRNAs can be computed for each pair of proteins. Vice versa, the number of co-regulated proteins can be computed for each pair of miRNAs. Based on this information, we want to understand whether the resulting numbers are actually significant or might *1)* be just a random effect caused by noise, *2)* occur simply due to some of the proteins showing extreme variation in their level, or *3)* result from many miRNAs targeting a central protein by both direct interference as well as indirect effects propagated through the gene regulatory network. All of these problems can be mitigated by assessing the probability that this number of co-regulating miRNAs is observed in graphs with the given degree sequence. Only those numbers which are unlikely to be the result of this null model will then be accepted as significant. The main idea behind overcoming the first problem is that filtering randomly missing edges or randomly added edges will not induce significant numbers of miRNAs. The second problem, namely proteins with an erratically jumping abundance level, will mainly induce random edges in the network. The random model can cope with both types of problems since a node with a higher degree will also have higher numbers of co-regulating miRNAs in the model. The third problem is that miRNAs with many indirect effects induce proteins with high degree. Their co-regulations are corrected however by the same noise-filtering effect.

Our method uses as null model for significance assessment the fixed degree sequence model adapted to bipartite graphs that contain two types of edges: those corresponding to up-regulation and those corresponding to down-regulation (Figure 26A). Recall that in this case we need to maintain both the degree sequences of the up-regulations and the degree sequences of the down-regulations (see Section 3.3.4). The edge type specific degree sequence of each protein and

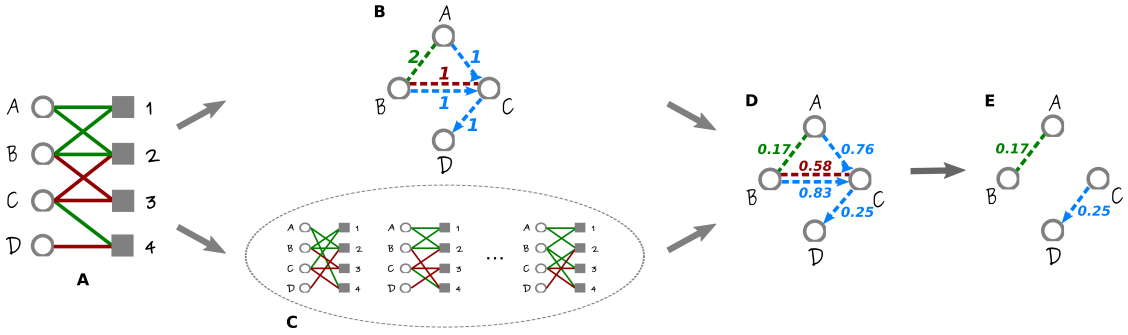


Figure 26: Pipeline of the algorithm. (A) We define the initial bipartite graph, (B) count the observed number of co-regulations, (C) simulate a sample of random bipartite graphs which define the expected number of co-regulations, (D) build the protein co-regulation graph where the weight of the edges indicates the p-value assigned to the co-regulation of a given protein pair, and (E) consider each co-regulation with a p-value smaller than or equal to a threshold t_P statistically significant. Figure adapted from [119].

each miRNA in the bipartite graph is then fixed while the edges of the same type are swapped (Figure 26C). This is achieved by the Markov chain Monte Carlo procedure discussed in Section 3.3.4. Figure 26 sketches the main steps of the method.

9.2 ROBUSTNESS ANALYSIS

For the kind of question at hand, namely the co-regulation behaviour of proteins under various experimental conditions, there is, to our knowledge, no large data set where the correct result is known. We thus build artificial data sets for which the ground truth is defined by construction and test our method against them (cf. Section 3.4.3). This approach is often used in the clustering of networks to test the performance of clustering algorithms [44, 43]. Note that in Section 8.2 we showed the robustness of the same method on artificial graphs that resembled the structure of the film–user data set. The benchmark graphs used here mimic the structure of the biological data under study.

9.2.1 *Construction of the artificial data*

In the considered data set, there is a strong imbalance between the number of proteins and the number of miRNAs (26 versus 810). Moreover, their degree sequences show a large variance. Constructing artificial graphs that best resemble this topology involves several modelling decisions. For illustration purposes, we formulate the simplifying assumptions behind the construction of the artificial graphs in terms of *artificial proteins* and *artificial miRNAs*: *1)* There are groups

of artificial proteins that are either co-up- or co-down-regulated by a subset of artificial miRNAs. *2)* Such a group of up-regulated artificial proteins and a group of down-regulated artificial proteins are antagonistically regulated by some subset of artificial miRNAs. *3)* Each group of artificial miRNAs is responsible for up-regulating exactly one group of artificial proteins and down-regulating another group of artificial proteins. *4)* Additionally, the regulation effect of the artificial miRNAs is assumed to be half up- and half down-regulations. Note however, that real-world data might be biased towards one of the edge types. For instance, in the biological data set at hand, miRNAs have a preference for down-regulation.

To model these assumptions, we build artificial graphs consisting of five modules with 16 nodes on the left side and 60 nodes on the right side, where the left side represents the artificial proteins and the right side the artificial miRNAs. In each module, there are 8 artificial proteins that are up-regulated and 8 that are down-regulated by the artificial miRNAs in the same module. Each of these modules represents one group of artificial proteins that are up-regulated, and another group of artificial proteins that are down-regulated by the same group of artificial miRNAs. Figure 27A sketches the structure of a single module. The degree distributions of the artificial proteins and artificial miRNAs are chosen to be similar to the ones in our biological data set: The degree distribution of the artificial miRNAs is strongly skewed, i.e. four of the nodes have degree 16, 8 nodes have degree 8, 16 nodes have degree 4, and 32 nodes have degree 2, while artificial proteins have a Poisson degree distribution.

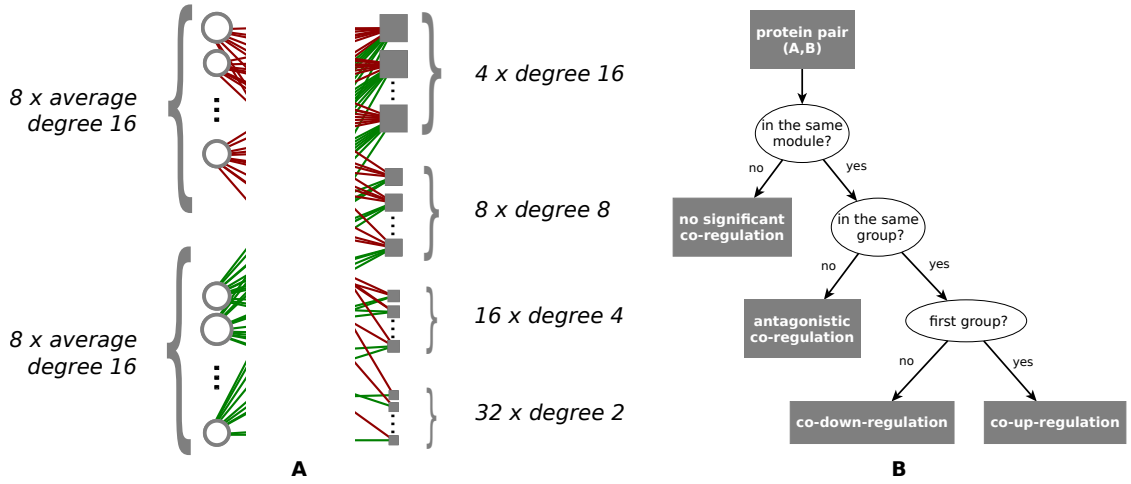


Figure 27: Structure of the artificial data. (A) Sketch of one module of an artificial graph. The degree of artificial proteins/miRNAs is proportional to size of the circles/squares. (B) Decision tree illustrating the principle behind the construction of the ground truth. Figure adapted from [119].

For these artificial graphs, when projecting to the artificial protein side, the ground truth is the following: within each of the modules all artificial proteins of the first group are significantly co-up-regulated, while all artificial proteins of the second group are significantly co-down-regulated. For any pair consisting of one artificial protein from the first and one from the second group, we require the algorithm to detect a significant antagonistic co-regulation directed from the second group to the first (see Figure 27B).

Defining a ground truth for the projection to the artificial miRNA side is not equally straightforward due to the presence of artificial miRNAs with a high degree, which will inherently be involved in non-significant co-regulations as well. However, since the robustness of the projection onto the artificial miRNA side is also highly relevant, we test the stability of the obtained artificial miRNA co-regulations with increasing noise.

To show the stability of our method, the artificial data is further perturbed to model two types of noise that are typical for biological data:

A. False-negative observations, i.e. the miRNA does regulate the protein's level but the change is too low due to random fluctuations, measuring errors, or simple handling errors. In this case, the regulation is not included in the regulation graph model and is thus a *missing edge*.

B. False-positive observations. By lowering the threshold of the original z-scores we *add edges* to the bipartite graph which are unlikely to represent significant regulations.

These two types of noise are modelled by the random elimination of a percentage ρ of edges ($0 < \rho \leqslant 100$), and the random addition of a percentage ρ of edges ($0 < \rho \leqslant 100$). The quality of the algorithm is measured by its ability to find the structure embedded in the original, artificial graph despite the presence of noise.

9.2.2 *Experiments on the artificial data*

We construct 100 artificial graphs with this predefined modular structure. In this section, whenever we refer to artificial proteins or artificial miRNAs, we use the terms *protein* and *miRNA*. Each artificial graph is projected twice: first to the protein side and then to the miRNA side. To assess the statistical significance of the co-regulations, a sample of $\kappa = 10,000$ random graphs is used. Based on the projection onto the protein side (with an easily definable ground truth), our aim is to assess how well our algorithm recovers the built-in modular structure of the ground truth projection. Then, based on both projections, we test the robustness of the algorithm against elimination and addition of randomly chosen edges. To quantify the precision of the

algorithm for different noise levels, we use the F-score and the PPV_k as performance measures (for details see Section 2.7). Figure 28 shows the performance of the algorithm when projecting onto the protein side (upper half) and when projecting onto the miRNA side (lower half). There are three patterns of interest for the protein case: when both proteins are up-regulated, both proteins are down-regulated, or one is up- while the other is down-regulated. For miRNAs, we only have two patterns: the antagonistic co-regulation pattern is omitted due to the lack of miRNA pairs in the original graphs that would antagonistically co-regulate proteins.

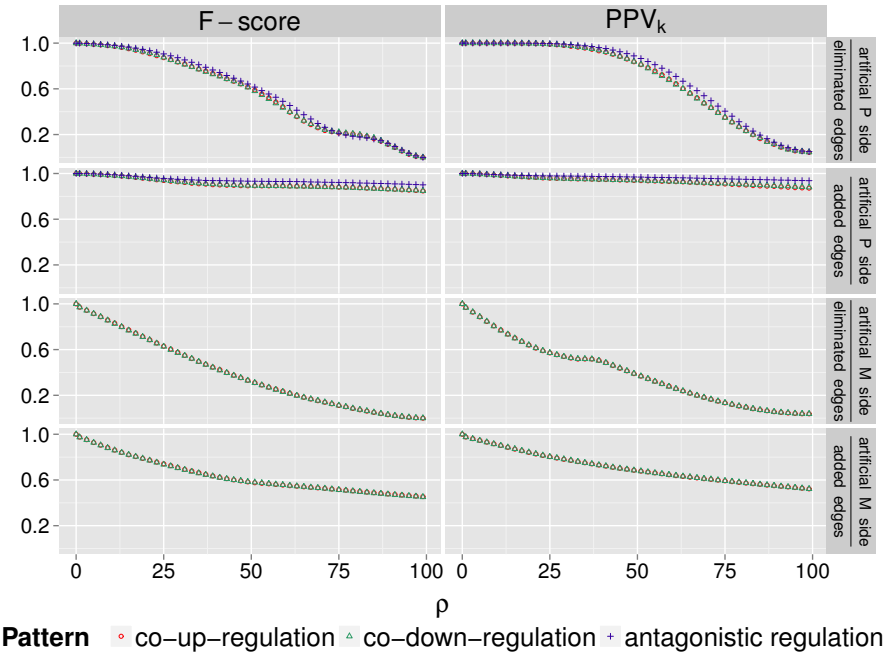


Figure 28: F-score and PPV_k evaluating the performance of our algorithm on artificial data sets for increasing noise levels ρ. Results are shown for eliminated and added edges when projecting onto the artificial protein and the artificial miRNA side. Red data points represent the performance of predicting co-up-regulation, green data points refer to co-down-regulation, and blue ones to antagonistic co-regulation. Figure reprinted from [119].

As the results of both measures suggest, the algorithm recovers the protein modules perfectly in the absence of noise. As the noise increases, the performance decays slowly. When projecting onto the protein side, gradual elimination of all edges in the bipartite graph ($\rho = 0\%$ to 100%) covers the whole range of possible prediction qualities. Accordingly, the F-score drops from 1 to 0 at $t_P = 0.05$ (the threshold used for determining the significance level of the edges that are included in the projection). The PPV_k decreases from 1 to about 0.04 (for the co-up- and co-down-regulation patterns) and to 0.05 (for the antagonistic co-regulation). These values are the baseline for this

measure, i.e. the proportion of true positives among all samples. Up until the point where 20% of all edges are eliminated, the PPV_k is almost perfect, while the F-score is above 0.9 for all considered patterns. Thus, the algorithm compensates well for noise. The prediction accuracies when projecting onto the miRNA side show similar tendencies: for 22% noise, the PPV_k is about 0.6, while the F-score is 0.67.

The addition of edges exerts a milder effect on the prediction quality. Thus, for as many as $\rho = 100\%$ added edges, there are still many correct predictions. In this range, when projecting onto the protein side, the PPV_k is above 0.88 and the F-score exceeds 0.85 for all patterns. Projecting onto the miRNA side results in lower, yet still convincing accuracies: the PPV_k remains above 0.52, while the F-score always exceeds 0.45. This is reassuring, as it means that one can still find significant co-regulation patterns even when also including mild effects into the bipartite regulation graph.

Although the two chosen quality measures are conceptually different, the resulting performance plots are relatively similar. The general trend is that, for low noise values, the PPV_k scores higher than the F-score. This is due to the different thresholds the two measures use. While PPV_k uses a threshold that is innate to the graph (the number of edge samples k), for the F-score we fix the threshold according to the rule of thumb $t_P = 0.05$. This emphasises that the proper choice of t_P for the algorithm is crucial and needs further consideration. Overall, we conclude that the algorithm is robust against both investigated types of noise. Having validated it on artificial data, we proceed to the analysis of a real biological data set.

9.3 RESULTS ON THE BIOLOGICAL DATA SET

Thesis point 12

As described above, the chosen biological data set contains the effect of a genome-wide library of miRNA mimics on the expression of 26 proteins in the EGFR-driven cell cycle pathway in a breast cancer cell line. Proteins are typically regulated by multiple miRNAs and miRNAs generally modulate, directly and/or indirectly, the expression of many proteins (see Figure 25). Given these complex interactions between proteins and miRNAs, it is challenging to differentiate mild biological effects from technical fluctuations and to identify regulatory patterns. Our algorithm can be used to detect on the one hand those pairs of proteins which are systematically co-targeted by a set of miRNAs, and on the other hand those pairs of miRNAs which systematically co-target a set of proteins. In the following, we present results for both cases.

9.3.1 *Consistently co-regulated miRNAs versus their families*

First, we search for miRNA pairs which simultaneously and significantly regulate the same proteins, i.e. we project the bipartite graph onto the miRNA side. Out of the obtained three projections, one for each co-regulation type, we focus on the biologically most relevant miRNA co-regulation pattern, namely the co-down-regulation. A similar analysis can be performed on the other two projections consisting of co-up-regulations and antagonistic regulations.

The robustness analysis discussed above suggests that the choice of t_P is one of the subtleties of the method that may influence the performance of the algorithm considerably. Thus, we first discuss this final step of the algorithm (see Figure 26E). When interpreting the result of a statistical analysis, it is common practice to choose the threshold for the significance level by some rule of thumb. For instance, it is widely accepted to define the significance level as 0.05 or 0.01. In contrast to this arbitrary choice of threshold, a trial and error approach is possible: one can set different thresholds and choose the best parameter by validating the results against prior knowledge or experiments, i.e. by using an *external reference approach*. Since external references might be difficult to obtain, we suggest the use of intrinsic properties like the network topology to automatically determine threshold candidates[1]. The idea behind this *internal reference approach* is motivated by the core assumption in the analysis of biological networks, namely that a network's function is reflected by its structure [139, 28]. To find the significance threshold, one can thus use a general criterion that relies on network analytic reasoning and results in a network-specific threshold that is chosen based on the structure of the network rather than just on the underlying problem. In an ideal setting, the two methods (the external and internal reference approaches) can be combined in order to maximize the efficiency of the predictions.

To choose a proper threshold for miRNA co-regulations, we propose the internal reference approach and base the decision on intrinsic information deduced from the underlying graph. Thus, we search for an appropriate threshold by inspecting the topology of the subgraphs built with different possible thresholds. Topological features of interest are[2]:

1. the number of edges normalized by the maximum number of edges,
2. the number of connected components,
3. the component density of the subgraphs normalized by the maximum number of components, where the density of a compo-

1 We have already presented an application of this procedure in Section 8.3.2.
2 For more details about these topological features see Section 2.1 and Section 2.2.

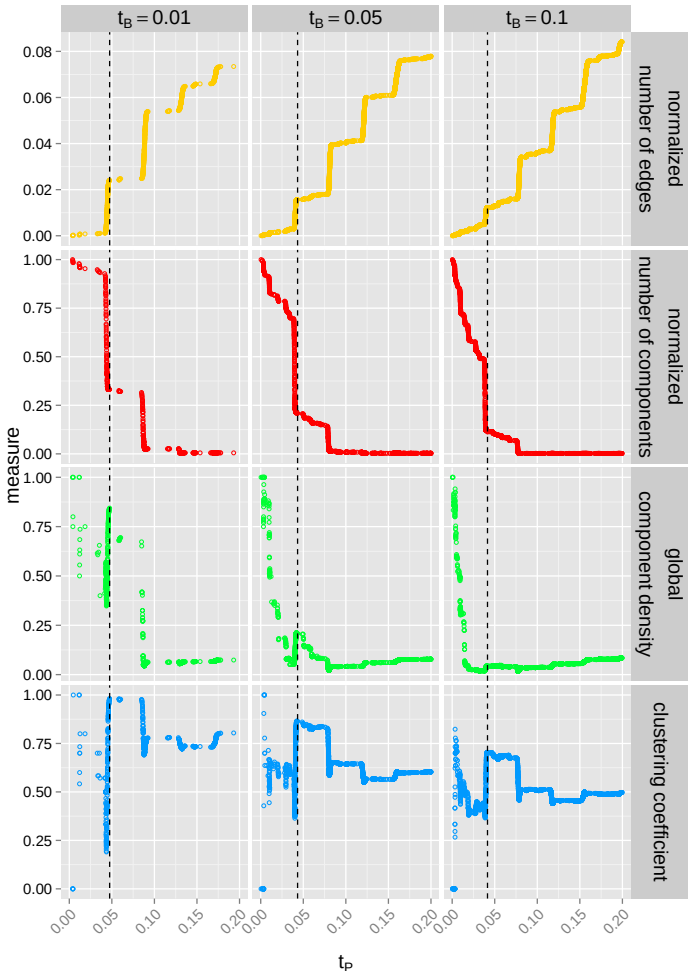


Figure 29: Deducing meaningful t_P significance thresholds from topological measures. Shown are four measures against the t_P thresholds for the p-values in the projection. The projections onto the miRNA side are constructed from the bipartite graphs with thresholds t_B corresponding to a p-value of 0.01, 0.05 and 0.10. Figure reprinted from [119].

nent is defined as the total number of its edges divided by the number of possible edges, and

4. the local clustering coefficient that quantifies the probability that any two of a node's neighbours are connected themselves. The clustering coefficient of a graph is the average local clustering coefficient of its nodes.

As shown in Figure 29, monitoring these features at varying threshold levels, we observe nontrivial changes in the structure of the subgraphs, indicating the more informative threshold candidates. The thresholds are considered optimal when there is a strong increase or local maximum in the average local clustering coefficient and in the global component density, while the number of components is

t_B	0.10	0.05	0.01
t_P	0.0440	0.0459	0.0509
number of miRNAs	437	322	151
number of groups	33	42	31

Table 5: Properties of the co-down-regulation projections obtained from bipartite graphs with different t_B thresholds. Shown are the significance thresholds t_P for the edges in the corresponding co-down-regulation graphs alongside the number of miRNAs and groups of size > 1 obtained at those thresholds. Table reprinted from [119].

still considerable. With respect to miRNA co-regulation, these criteria assure increased transitivity and best reveal the local connection patterns of the individual miRNAs. Accordingly, for our data we choose the t_P thresholds shown in Table 5. Interestingly, for this data set, the thresholds for the statistical significance of the co-regulations do not differ considerably for altered significance levels t_B of the edges in the bipartite graph.

Analysing the effect of the bipartite graph threshold on the resulting co-regulation graphs, we observe that as t_B gets stricter, these projections contain a decreasing number of miRNAs that are grouped in several components of size > 1 (see Table 5). First, to reinforce the assumption that the algorithm detects miRNA groups which have similar regulation patterns, we return to the bipartite graph model and analyse it with respect to the newly acquired grouping of the miRNAs. As shown in Figure 30, based on the number of proteins that are co-targeted by the miRNAs contained in the found groups, we can differentiate between three types of groups:

1. Groups of miRNAs that target one single protein (section I in Figure 30A). Although they do not provide new biological insights, these groups are reassuring findings since they obviously satisfy the criterion of non-random co-regulation.

2. Groups of miRNAs that have 2 to 8 protein targets (section II in Figure 30A and magnified in Figure 30B). These groups represent nontrivial co-regulations and should be central to further experimental investigations aimed at finding candidates for new tumour suppressors.

3. One larger group that contains several miRNAs with multiple targets (section III in Figure 30A). Here, the interconnectedness in the bipartite graph is highly complex and requires further research. For instance, the group could be split up by lowering the projection threshold t_P or using a subsequent clustering algorithm which detects subgroups based on the p-values assigned to each miRNA pair.

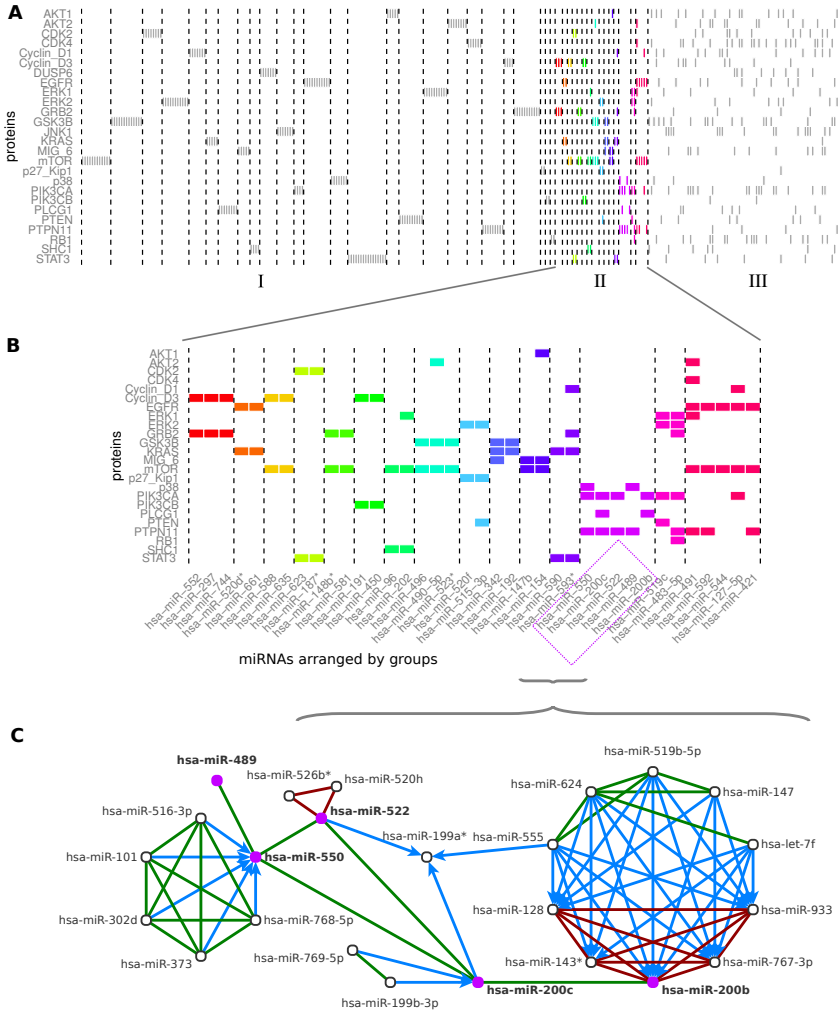


Figure 30: (A) miRNA groups obtained by our algorithm from the bipartite graph with $t_B = 0.05$ and $t_P = 0.0459$. Each square represents a down-regulation in the bipartite graph. Shown are the groups with one exclusive protein target (section I), with 2 to 8 targets (section II, coloured regulations), and with multiple targets (section III). (B) Magnification of section II containing nontrivial co-regulations. In accordance with (A), colours indicate the different groups. (C) The group containing hsa-miR-489, hsa-miR-522, hsa-miR-200c, hsa-miR-550, and hsa-miR-200b (purple nodes) together with the co-regulating miRNAs. Co-down-regulation is shown in green, co-up-regulation in red, while antagonistic regulation is coloured blue and is directed from the down-regulating miRNA to the up-regulating miRNA. Figure adapted from [119].

Figure 30C shows an exemplary excerpt of the co-regulation graph with typical patterns for the entire graph. The subgraph is constructed around the five miRNAs belonging to one of the found groups by the addition of the co-regulating miRNAs. Accordingly, co-up- and co-down-regulations define tightly connected clusters. Antagonistic co-regulations occur *between* these clusters, systematically connecting

enriched miRNA family	miRNAs of the family that are in the group as well	p_{hyp}
mir-99	hsa-miR-100, hsa-miR-99a, hsa-miR-99b	0.001
let-7	hsa-let-7f, hsa-let-7f-1*, hsa-let-7f-2*, hsa-let-7g*, hsa-let-7i*	0.029
mir-146	hsa-miR-146a, hsa-miR-146b	0.005
mir-221	hsa-miR-221, hsa-miR-222	0.011
mir-29	hsa-miR-29a, hsa-miR-29c	0.001
mir-506	hsa-miR-509-3-5p, hsa-miR-510	0.018
mir-8	hsa-miR-200b, hsa-miR-200c	0.001
mir-515	hsa-miR-515-3p, hsa-miR-520f	0.021

Table 6: miRNA groups identified by the algorithm in which the families are significantly over-represented. For our analysis, we consider the seed sequences of the groups obtained at the regulation stringency threshold $t_B = 0.05$. The statistical significance of over-representation was assessed by a hypergeometric test, resulting in the p-value p_{hyp}. Table adapted from [119].

co-down-regulated clusters with co-up-regulated clusters, i.e. consistently with their direction.

We expect that the membership of the miRNAs in the identified groups is biologically meaningful. To test this, we analyse the groups in relation to the assignment of miRNAs into families according to their *seed sequence*—a non-disrupted subsequence between the 2nd and 7th bases of the mature miRNA, which is believed to be decisive for RNA binding. Specifically, we compare the seed sequences of miRNAs belonging to the same group. To quantify the similarity of two miRNAs, we use the *edit distance* of their seed sequences, i.e. the minimum number of alterations (insert, delete, or exchange of a nucleotide) required to transform one sequence into the other [143]. The similarity of the miRNAs which the algorithm places in the same group is then defined as the average pairwise edit distance between the miRNAs. To test whether the sequence similarity within a given group is statistically significant, we conduct simulations with bootstrapping. In some of the cases, the edit distances suggest a significant similarity between the sequences in the identified groups. As shown in Table 6, a hypergeometric test reveals that for $t_B = 0.05$ there are 8 over-represented families in the groups. Four of these families are reported to be oncogene or tumour suppressors in breast cancer, while two of them, miR-99 and miR-506, play a role in prostate/head-and-neck cancer and melanoma, respectively. Thus, by using the al-

gorithm, we can extract miRNAs and families which have already established roles in the pathogenesis of breast cancer. This suggests the ability of our algorithm to identify the potentially most pathologically-relevant miRNAs.

9.3.2 *Co-regulation of proteins from the same functional module*

So far, little is known about the physiological relevance of protein pairs that are co-regulated by miRNAs. It is believed that such co-regulations within a network confer signalling robustness (e.g. dampening and buffering effects) and can mediate the crosstalk of different signalling pathways [121]. To better understand the mechanisms underlying co-regulation, we now project the miRNA–protein interaction network to the protein side, thereby identifying significant co-regulations of proteins.

As discussed in Section 9.1.1, the bipartite graph is determined by the choice of the threshold t_B. However, depending on how many and which edges are added to the bipartite model, different co-regulations might appear to be significant. To limit this effect, we combine the results obtained for different choices of t_B by building a *consensus graph* based on the co-regulations that are significant over a broad range of selected threshold combinations. To keep only those co-regulations that are statistically significant under different t_B thresholds, we choose three relatively relaxed z-score thresholds for t_B, namely 1.28, 1.64, and 1.96. We project the three resulting bipartite graphs to the protein side and, similarly to the miRNA case, monitor the change of the average clustering coefficient with increasing t_P to find threshold candidates. From these we deduce a set of threshold levels. The consensus graph corresponding to a given level then consists of the edges that are contained in all three graphs at the given level.

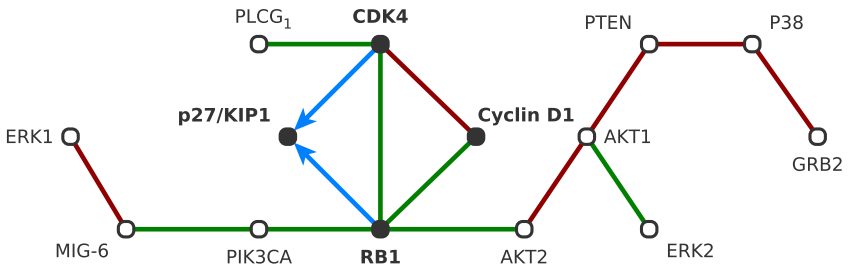


Figure 31: Consensus graph of co-regulated proteins. Cell cycle proteins are CDK4, p27/KIP1, Cyclin D1, and RB1 (black nodes). The green edges between the proteins show the co-down-regulation of two proteins while the red edges show the co-up-regulation. The directed blue edges indicates that the source node is down-regulated, whereas the target node is up-regulated. Figure adapted from [258].

rank	miRNA	rank	miRNA
1	**miR-124**	6	**miR-147**
2	miR-892b	7	miR-491
3	miR-124*	8	miR-342
4	**miR-193-3p**	9	miR-518e*
5	miR-17-3p	10	miR-769-5p

Table 7: Ranking of the miRNAs based on their frequency in the consensus graphs. miRNAs that were chosen for experimental validation are highlighted in bold.

The obtained consensus graph with the most stringent threshold is shown in Figure 31. We observe that most edges between the cell cycle proteins CDK4, p27/KIP1, Cyclin D1, and RB1 are present in the consensus graph, indicating that their co-regulation is robust. Furthermore, the miRNAs that co-regulate the cell cycle proteins show an interesting pattern, which is biologically relevant: they up-regulate the expression of the CDK inhibitor protein p27/Kip1, while down-regulating CDK4 and RB1 which would, in turn, inhibit cell cycle progression. Furthermore, PIK3CA and AKT2 are co-down-regulated with RB1, while PLCG1 is significantly co-down-regulated with CDK4 by dozens of miRNAs. Overall, these results indicate that the proteins, which belong to the same known functional module (here: EGFR-driven cell cycle module) are systematically co-regulated by certain miRNAs.

Thesis point 13

Finally, we wish to understand which miRNAs are responsible for these protein co-regulations. For this purpose, we consider again the bipartite graph model and for each pair of proteins with a statistically significant co-regulation, we record the miRNAs with which they are connected at the specific threshold t_B in the bipartite graph. By this procedure we obtain for each miRNA the frequency of its participation in the consensus graph. Ranking then the miRNAs according to this frequency, we obtain those that have the most important regulating role altogether. Table 7 shows the 10 miRNAs located at the top of the ranking. miR-124 (rank 1), miR-193a-3p (rank 4), and miR-147 (rank 6) are among those with the highest frequency, indicating that they should, when overexpressed, alter cell cycle progression patterns. Wet laboratory experiments of our collaborators have confirmed the validity of this prediction [258].

9.4 ADVANTAGES OF OUR METHOD OVER EXISTING APPROACHES

High-throughput studies that aim at exploiting regulatory networks between two types of biological entities have become feasible due to technological development and community efforts. Recently, as

the ENCODE project reached its milestone, several data sets and accompanying papers were published (for a review see Reference [76]), providing data in various settings that can be modelled as bipartite graphs. Some examples are transcription factors binding to DNA promoter regions [89], gene-coding RNAs and co-transcriptional long non-coding RNAs [250], as well as single-nucleotide polymorphisms (SNPs) and diseases [227]. Despite their distinct nature, all these data sets can be analysed by our algorithm to identify significant co-interaction patterns. Previous approaches of finding such patterns include various clustering methods, most prominently hierarchical clustering and k-means clustering (cf. Section 2.2). Our approach differs from these methods in four important aspects:

1. It applies thresholding when building the bipartite graph model. We reckon that this step can be both advantageous and risky. By using a hard threshold, on the one hand we filter out noise, but on the other we may disregard potentially useful information by eliminating edges. However, the robustness test on artificial graphs suggests that our method is highly robust against randomly added edges (noise included due to a loose threshold) or eliminated edges (relevant regulation lost because of a strict threshold). This gives us flexibility in choosing the threshold t_B, suggesting that small deviations of the threshold have no considerable impact on the results obtained by our algorithm.

2. The co-regulation graph with the threshold t_P is selected by tracing changes in the graph characteristics with respect to the threshold choice. Instead of relying on rules of thumb, this allows for a threshold-selection which retains a maximum of information obtainable from the primary data.

3. Classical hierarchical clustering returns a tree in which each biological entity (e.g. miRNA) is connected to another entity via an internal node. The k-means clustering results in groups of nodes without internal edges. In comparison, our method provides an intuitive way of understanding co-regulations within the groups.

4. For each identified co-regulation, it reports an empirical p-value which quantifies the likelihood of observing the given co-regulation pattern in random graphs. This is neither the case for hierarchical clustering nor for k-means. Therefore, our method makes it possible to compare the statistical significance of the co-regulations within one network as well as between different networks. Comparing significant co-regulation patterns instead of comparing top hits may help in revealing the mechanisms underlying observations of interest, as pathway and network analysis have demonstrated in microarray analysis [241].

Besides these classical clustering methods, weighted correlation network analysis (WGCNA) has been proposed [276] and successfully applied in gene expression microarray analysis [262]. WGCNA assumes a scale-free topology of the underlying network. In contrast, our method does not make any assumption regarding the structure of the data. Thus, we believe that it offers an unbiased analysis as compared to WGCNA. A thorough comparison between our method and other existing approaches on different bipartite data sets with ground truth represents the main direction for future research. An exemplary comparison of our method with the Pearson correlation of the expression values, i.e. one of the standard methods for evaluating gene co-expression [165], showed that our algorithm outperforms this on artificial data sets: When identifying co-regulated proteins from data sets containing 50% noise in form of added edges, the Pearson correlation achieves a PPV of 0.85, while our method has a performance of 0.96.

9.5 THE SOFTWARE SICOP

Thesis point 14

In order to enable experimentalists to use our method on their own data, we provide an open source tool that implements the algorithm. SICOP accepts several common input formats and supports different output formats to facilitate additional analysis and visualization. The key features of SICOP include a user-friendly interface, easy installation, and platform-independence.

Existing tools for the analysis of bipartite graphs like the `R` packages *bipartite* [71] or *networksis* [10] are mainly tailored to the purpose of understanding principles in community ecology. Systems biology applications pose three important challenges that these tools cannot cope with: *1)* large-scale experiments that are often prone to noise, *2)* mild interaction effects that are difficult to detect, and *3)* simultaneously observed distinct types of interactions. The presented method overcomes these issues and SICOP implements it as an easy-to-use client-side tool. The most important functionalities of SICOP are:

DATA IMPORT FROM DIVERSE INPUT FILE FORMATS The tool accepts a list of the observed interactions stored in a text file, a matrix containing the measured level of all interactions in a `csv` file (comma separated value), or a graph representation of the network in a `gml` file (graph mark-up language) as produced by graph editing programs such as yEd.

SIMPLEX AND DUPLEX NETWORK SUPPORT AND PRECOMPUTATIONAL EDGE FILTERING Given the experimental data, SICOP first constructs the bipartite graph model. If the observed interactions are assigned weights, the user may add a threshold to filter them. If there

are two types of interaction, the user can treat them as a single type (simplex network data) or use both of them (duplex network data).

STATISTICAL SIGNIFICANCE ASSESSMENT OF CO-INTERACTIONS SICOP detects patterns in the bipartite graph that are significant under a null model defined by the fixed degree sequence ensemble associated with the original graph (see Section 3.3.4). The statistical significance of the individual co-regulations is quantified by a z-score or a p-value (see Section 4.3).

MULTIPLE DATA EXPORT FORMATS The same edge selection options are available when exporting the data as when importing it. Thus, the user may create and store multiple networks with different threshold values corresponding to different p-values or z-scores. The obtained co-regulation networks may be exported in any of the input file formats or alternatively as `graphml`.

HIGH CONFIGURABILITY Besides the key functionalities described above, SICOP allows more confident users to modify the parameter values. The default values are based on the theoretical and empirical considerations presented in Chapter 3 and are automatically adjusted to the size of the input data. Increasing the pre-set values enhances accuracy but comes at the cost of additional computational time. See Reference [238] or consult the manual and the download page for further information [237].

SICOP can be applied to a wide range of data sets, such as transcription factors binding to DNAs, gene-coding RNAs interacting with co-transcriptional non-coding RNAs, genes in relation with diseases, or diseases and their symptoms. Designed as a flexible tool, it offers an effortless way to better explore such data.

9.6 CONCLUSIONS

Since the early days of genetics and molecular biology, it has been noted that proteins can be regulated by more than one regulator and one regulator may in turn affect several proteins. In many situations, a regulator or a given experimental condition exerts only a mild effect on an observed protein, which might be difficult to differentiate from a random fluctuation. To address this complication, this chapter discussed a network analytic method, which is rooted in the observation that if proteins are "collaborating" with each other to coerce a common biological function, then this should be reflected in the way they are co-regulated. Based on this assumption, we search for pairs of proteins or protein-regulating agents, which are significantly co-regulated under many different experimental conditions.

In a biological system with many layers of regulatory networks, co-regulations may contribute to the robustness of the system, since the regulation can be resistant to partial losses of functional members due to gene deletion, mutation, or stochastic expression regulations. Understanding co-regulation is vital in establishing an effective and stable modulation of the molecular target and is thus important for cellular engineering and drug research.

Given a complex interconnected system of proteins and regulators, our method finds statistically significant co-regulations. We have shown on artificial data sets that systematic co-regulations are detected even in the presence of random noise in the form of eliminated or added regulations. To test the algorithm on a real biological data set, we applied it to the EGFR-driven cell cycle system regulated by miRNAs.

Focusing on miRNA co-regulation, we showed with sequence analysis and miRNA family enrichment analysis that the theory, according to which miRNA targeting is sequence-dependent, indeed partially explains the observed co-regulations obtained by the method. However, the results of the algorithm show that even miRNAs with distinct seed regions can induce strong co-regulations, which may be caused by the co-targeting of upstream transcription factors or separate targeting of canalized pathways. This indicates the complexity of the miRNA regulatory machinery, since miRNAs from different families may target different genes while yielding the same output. To tackle this complexity, further experiments are needed, such as profiling gene expression by over-expressing miRNAs of the same groups. Our results do not only yield proteomic evidences that sequence similarity of miRNAs determine their targets, but also provide hypotheses of other types of co-targeting that can be tested experimentally. Thus, potential therapeutic applications have to consider miRNA sets with similar co-regulation patterns. Based on our observations, we therefore argue that systematic approaches examining regulations between two biological components (miRNA and EGFR pathway proteins in our case) can be essential to the detection of co-regulation patterns and in the design of multiplex targeting strategies.

Concerning protein co-regulation, we could show that there are consistent regulatory patterns in which miRNAs simultaneously and significantly co-regulate several proteins, which act in the same functional module. This approach also enabled us to identify and subsequently validate experimentally three miRNAs (miR-124, miR-147, and miR-193a-3p) as novel tumour suppressors that co-target the EGFR-driven cell cycle proteins and inhibit cell cycle progression and proliferation in breast cancer.

The results obtained on the EGFR-driven cell cycle system transfected with miRNAs illustrate merely one systems biology context in which our method can be used. The method can be applied as

long as the system of interest can be modelled as a bipartite graph *and* the research question can be meaningfully approached in terms of co-occurrences of nodes of the same type. The statistically significant co-occurrences identified by our method are expected to unravel functional groups which could be profitably analysed from this perspective.

9.7 SUMMARY

Interactions between various types of molecules that regulate crucial cellular processes are extensively investigated by high-throughput experiments and require dedicated computational methods for the analysis of the resulting data. In many cases, this data can be represented as a bipartite graph because it describes interactions between entities of two different types such as the influence of different experimental conditions on cellular variables or the direct interaction between receptors and their activators/inhibitors. One of the major challenges in the study of such noisy data sets is the statistical evaluation of the relationship between two entities of the *same* type—a task known in the literature as the one-mode projection of bipartite graphs or as the inference of association networks.

In this chapter we have presented a method for the detection of pairs of entities with a statistically significant relationship. We showed the stability of the proposed method on artificial data sets: when randomly adding and deleting observations, we obtained reliable results even with noise exceeding the level that can be expected in large-scale experiments. Subsequently, we illustrated the viability of the method based on the analysis of a proteomic screening data set to reveal regulatory patterns of human microRNAs that target proteins in the EGFR-driven cell cycle signalling system. Since statistically significant relationships may indicate functional synergy, they hold promise in drug target identification and therapeutic development. To reduce the amount of required experiments, we also provide an implementation of our algorithm that offers an effortless way of exploring such data to its full potential and is thus a flexible, novel tool in the arsenal of high-throughput screening analysis.

10

CONCLUSIONS AND OUTLOOK

Inferring connections in diverse large networks

In the work covered in this thesis, we addressed the problem of inferring connections in various networks. We identified mild co-regulation effects in biological networks, filtered physical interactions between proteins, predicted social ties based on online social networks, and deduced film similarities in a market basket setting. Despite the very diverse nature of the studied systems, we described them in terms of a common network analytic framework. We modelled the underlying systems as graphs that contained one or more types of edges between one or more types of nodes. Exploiting these models to their full potential required selecting, combining, adapting, and developing novel exploratory methods and algorithmic solutions. Accordingly, we compared a set of node similarity measures, which are based on structural equivalence, and showed that the measures which use the fixed degree sequence model as null model are better suited for assessing node similarity than their popular alternatives. Our experiments on biological and social networks confirmed that there exists a meaningful correlation between the topology of the graph and the mechanisms responsible for the formation of edges. This finding emphasizes our best-performing measure as a valuable unsupervised scoring method in exploratory settings. On the other hand, for scenarios in which ground truth information is available, we proposed the use of similarity measures to train a random forest classifier, thereby approaching the edge prediction problem in a supervised manner. To address the challenges represented by graphs with multiple types of edges, we extended node similarity measures to this case and presented a new method for the projection of multiplex bipartite graphs in which at most one type of edge is admitted between two nodes. The adapted and developed methods proved to be highly flexible and generalizable to other applications both within and beyond the fields of the domain-specific problems around which they were created.

Methodological developments

The central findings of our projects

Regarding our project-specific findings, our method for the study of the miRNA regulation of the EGFR/cell cycle proteins detected three miRNAs as novel potential tumour suppressors, which were validated experimentally by our collaborators and shown to indeed hinder the growth of an especially lethal type of breast cancer. Our approach to the prediction of relationships between non-members of online social networking platforms correctly inferred 40% of the connections between *non*-members. This result highlights the implications of a common practice where members of such platforms provide relational data about non-members on a constant basis, often without

their consent. In the market basket setting, our identification of similar films was successful for 76% of the pairs of TV shows available in the ground truth data set and 47% of the contained pairs of feature films. The additional multiplex analysis enabled us to analyse different aspects of film similarity and contributed a number of observations and measurements about the important interplay between these aspects.

A summary of the obtained results, concluding remarks, and directions for future research concerning each individual project were already given at the end of the respective chapters. In the following, we thus focus on more general challenges that relate to our framework in a broader sense.

CHALLENGES AND PERSPECTIVES

TOPOLOGICAL NODE SIMILARITY In this thesis, we considered topological node similarity measures to validate the presence or predict the emergence of an edge between two nodes. Therefore, we only used structural details to deduce potential functional information. Although this has provided valuable insights for the considered applications, the correlation between structure and function requires further explorations, for instance by the incorporation of additional data wherever possible (for some examples see References [107, 67]). Furthermore, we only addressed those topological measures that assess local structure in the neighbourhood of the nodes. Although some studies use longer-range information [147; 129, p. 201], it is still unclear to which extent this improves inference.

MULTIPLEX NETWORKS WITH DIFFERENT TYPES OF ENTITIES As shown in this thesis, the majority of the systems of interest are only rarely reducible to graphs that contain solely a single type of edge between a single type of node [53]. This renders the study of multiplex graphs with different types of nodes indispensable, as they model ubiquitous concepts inherent to real-world systems [246, 182, 69, 98, 59]. Due to the increasing availability of large-scale data sets that record various types of connections between different types of entities, a multitude of problem-specific approaches can be expected to appear. For example, systems biology already advances integrated studies that analyse various types of interactions (for instance regulatory, structural, and catalytic) among different cellular components (genes, proteins, and metabolites) [225]. Despite these endeavours, a standardized approach consisting of rigorous and universal methods for the analysis of such networks is still missing and increasingly needed. As such, it represents a key direction for future research.

TEMPORAL ASPECTS Our focus in this thesis has been on snapshots of networks that model either instantaneous or time-averaged views on the underlying system. Taking temporal aspects into account would lead to more accurate modelling and to a better understanding of processes on these networks [112]. Like network topology, the temporal structure of connection activations can affect the dynamics of disease contagion on the network of patients or of information diffusion on a communication network for instance. Therefore, there has been an increased interest in the analysis of temporal networks in recent years [22, 133, 199, 142, 111, 182]. In a further project, independent from this thesis, we took a step in this direction. There, we modelled a citation network by a multiplex graph with time-ordered nodes and identified central nodes. While this stipend-supported project is still ongoing, it already provided insights into this type of network, not least due to the methods developed for this thesis.

Karl-Steinbuch scholarship for supporting innovative IT- and media-related projects, MFG Baden-Württemberg, Germany

CO-EVOLUTION OF NETWORKS When predicting connections, both the context and certain attributes of the involved entities may affect the formation of further connections. To assure a unified framework, we considered elements of the context and the discrete attributes in terms of networks. This allowed us to construct bipartite graph models in which the nodes of one set modelled the entities of interest, while the nodes of the other set modelled the elements of the context or the attributes. We then used the bipartite graph as a proxy to infer the connections. Moving beyond this, the combination of such bipartite graphs with graphs that model observed connections between the entities for which further predictions are desired, is a rewarding avenue for future research. For instance, the co-evolution of social and affiliation networks modelling the participation of individuals in different activities can shed light on the mechanisms behind social influence and selection [74, p. 86–88]. Similarly, in a biological setting, the joint consideration of known functional relationships among genes and gene regulation data could be rewarding. Moreover, effects on the cellular level alongside environmental and social influences can be considered to represent a key step towards personalized medicine [23]. Following this trend of jointly considering different networks, we expect an increased scientific interest in the study of their co-evolution.

INTERDISCIPLINARY WORK The highly interdisciplinary work behind this thesis builds on a broad network analytic framework which proved to be just as useful in the social sciences as in molecular biology. Although such interdisciplinary endeavours are very beneficial, the differences between the disciplines persist. In the words of Borgatti [38]:

> "To a physical scientist, network research in the social sciences is descriptive because measures of network properties are often taken at face value and not compared to expected values generated by a theoretical model such as the Erdős-Rényi random graphs. For their part, social scientists have reacted to this practice with considerable bemusement. To them, baseline models like simple random graphs seem naïve in the extreme—like comparing the structure of a skyscraper to a random distribution of the same quantities of materials."

As emphasized in this statement, network analytic procedures cannot be applied "out of the box" and require problem-specific adaptation. The methods for edge inference presented in this thesis are no different. They are based on the number of common neighbours of two individual nodes in the graph that models the studied system. Therefore, only those networks can be analysed in a similar fashion, in which the concept of similarity through shared neighbourhood is meaningful within the investigated context. Consider for example the widely studied concept of word adjacency in sentences. The probabilistic study of longer sequences of n consecutive words, the so-called n-grams, has led to advances in statistical natural language parsing [164] and provided novel insights into the evolution of grammar [174]. However, a network of words, which is constructed by connecting each pair of words that occur successively in a text, is not meaningful, since even syntactic dependencies between words span greater distances. This example shows that a network representation is not always required and stresses the need for a network modelling with appropriate definitions of entity and connection, as well as a clearly specified set of methodologies and algorithms for obtaining significant and reliable results.

After this note of caution, we emphasize that correctly used network analysis provided a novel perspective on old questions and inspired valuable new solutions for a wide range of problems. As the field itself evolved from different disciplines and becomes richer through every scientific area with which it intersects, so will the general framework presented here. I expect it to contribute to the developments that yield a better understanding and a wide-spread awareness of diverse aspects of our world.

LIST OF NOTATIONS AND ABBREVIATIONS

ζ network pattern 51

$\mathcal{G}(n, m), \mathcal{G}(n, p)$ classic random graph models 34

$\mathcal{G}$ ensemble of simplex non-bipartite graphs 33

$\mathcal{B}$ ensemble of simplex bipartite graphs 33

$\tilde{\mathcal{G}}$ ensemble of multiplex non-bipartite graphs 33

$\tilde{\mathcal{B}}$ ensemble of multiplex bipartite graphs 33

$\mathcal{H}$ sample from an ensemble of graphs 42

κ sample size 42

P probability 24

$\mathcal{S}$ set of states 24

T transition probability matrix 24

π stationary distribution of a Markov chain 25

τ mixing time of a Markov chain 25

$\mathcal{M}(\mathcal{X}) = (\mathcal{X}, \mathcal{T}(\mathcal{X}))$ Markov chain of the graph ensemble $\mathcal{X}$ 42

$\vec{x} \in \mathbb{R}^q$ q dimensional feature vector 69

$y \in \{1, \ldots, c\}$ class label 69

$D = \{(\vec{x}_i, y_i)\}_{i \in \{1,\ldots,e\}}$ labelled feature data containing e examples 69

D_t training data 73

D_v validation data 73

$g : \mathbb{R}^q \to \{1, \ldots, c\}$ classifier function 69

H information entropy 70

TP number of true positives 29

TN number of true negatives 29

FP number of false positives 29

FN number of false negatives 29

$spec$ specificity 29

$sens$ sensitivity 30

ROC receiver operating characteristic 30

AUC area under the curve 30

BIBLIOGRAPHY

[1] Database of Interacting Proteins (DIP). URL http://dip.doe-mbi.ucla.edu/dip/Download.cgi. (Cited on page 89.)

[2] Film series at Wikipedia. URL http://en.wikipedia.org/wiki/Film_series/. (Cited on page 105.)

[3] The Internet Movie Database (IMDb). Alternative interfaces. URL http://imdb.com/interfaces/. (Cited on pages 104 and 116.)

[4] miRBase version 19. URL http://www.mirbase.org/. (Cited on page 121.)

[5] The Netflix Prize. URL http://www.netflixprize.com/. (Cited on page 97.)

[6] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Privacy Enhancing Technologies*, volume 4258 of *Lecture Notes in Computer Science*, pages 36–58. 2006. (Cited on page 85.)

[7] L.A. Adamic and E. Adar. Friends and neighbours on the Web. *Social Networks*, 25(3):211–230, 2003. (Cited on page 67.)

[8] L.A. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005. (Cited on page 85.)

[9] L.A. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the Web. *First Monday*, 8(6), 2003. (Cited on page 81.)

[10] R. Admiraal and M.S. Handcock. *networksis*: A package to simulate bipartite graphs with fixed marginals through sequential importance sampling. *Journal of Statistical Software*, 24(8), 2008. (Cited on pages 37, 48, and 140.)

[11] Y.Y. Ahn, J.P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010. (Cited on page 98.)

[12] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002. (Cited on page 2.)

[13] R. Albert, H. Jeong, and A.L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000. (Cited on page 33.)

[14] U. Alon. Network motifs: theory and experimental approaches. *Nature*, 8(6):450–461, 2007. (Cited on pages 3 and 51.)

[15] L.A.N. Amaral and J.M. Ottino. Complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38 (2):147–162, 2004. (Cited on page 3.)

[16] Y. Artzy-Randrup and L. Stone. Generating uniformly distributed random networks. *Physical Review E*, 72(5):056708, 2005. (Cited on pages 37 and 41.)

[17] Y. Artzy-Randrup, S.J. Fleishman, N. Ben-Tal, and L. Stone. Comment on "Network motifs: Simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, 305(5687):1107, 2004. (Cited on page 34.)

[18] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the 4th ACM international conference on Web search and data mining*, pages 635–644, 2011. (Cited on page 77.)

[19] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54, 2006. (Cited on page 90.)

[20] D. Baek, J. Villén, C. Shin, F.D. Camargo, S.P. Gygi, and D.P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008. (Cited on page 121.)

[21] R.A. Baeza-Yates and B.A. Ribeiro-Neto. *Modern Information Retrieval. The concepts and technology behind search*. Pearson Education Ltd., Harlow, England, 2nd edition, 2011. (Cited on pages 30 and 31.)

[22] A.L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005. (Cited on page 147.)

[23] A.L. Barabási. Network medicine — From obesity to the "Diseasome". *The New England Journal of Medicine*, 357(4):404–407, 2007. (Cited on page 147.)

[24] A.L. Barabási. The network takeover. *Nature Physics*, 8(1):14, 2011. (Cited on page 3.)

[25] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. (Cited on pages 2, 3, 33, 36, and 56.)

[26] A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2): 101–113, 2004. (Cited on page 2.)

[27] A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311(3–4):590–614, 2002. (Cited on page 21.)

[28] A.L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature*, 12(1):56–68, 2011. (Cited on pages 2, 51, and 132.)

[29] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, New York, 2008. (Cited on page 2.)

[30] S.C. Basak, V.R. Magnuson, G.J. Niemi, and R.R. Regal. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Applied Mathematics*, 19(1–3):17–44, 1988. (Cited on page 55.)

[31] J. Bascompte. Disentangling the web of life. *Science*, 325(5939): 416–419, 2009. (Cited on page 2.)

[32] J. Bascompte. Structure and dynamics of ecological networks. *Science*, 329(5993):765–766, 2010. (Cited on page 2.)

[33] E.A. Bender and E.R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978. (Cited on page 37.)

[34] A. Berger and M. Müller-Hannemann. Uniform sampling of digraphs with a fixed degree sequence. In *Graph Theoretic Concepts in Computer Science*, volume 6410 of *Lecture Notes in Computer Science*, pages 220–231. Springer, 2010. (Cited on pages 24, 37, 40, 41, and 42.)

[35] N. Blow. Untangling the protein web. *Nature*, 460(7253):415–418, 2009. (Cited on page 89.)

[36] S. Boldt, K. Knops, R. Kriehuber, and O. Wolkenhauer. A frequency-based gene selection method to identify robust biomarkers for radiation dose prediction. *International Journal of Radiation Biology*, 88(3):267–276, 2012. (Cited on page 121.)

[37] B. Bollobás. *Random graphs*. Cambridge University Press, Cambridge, UK, 2nd edition, 2001. (Cited on pages 2 and 35.)

[38] S.P. Borgatti. *Encyclopedia of Complexity and System Science*, chapter 2-mode concepts in social network analysis, pages 8279–8291. Springer, New York, 2009. (Cited on pages 20, 57, 61, and 147.)

[39] S.P. Borgatti and M.G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, 2006. (Cited on page 56.)

[40] S.P. Borgatti, A. Mehra, D.J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009. (Cited on page 2.)

[41] d. boyd and K. Crawford. Six provocations for big data. In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011. (Cited on pages 1 and 97.)

[42] U. Brandes and T. Erlebach, editors. *Network Analysis: Methodological Foundations*, volume 3418 of *Lecture Notes in Computer Science*. Springer, New York, 2005. (Cited on page 2.)

[43] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. In *Proceedings of the 11th European Symposium on Algorithms*, pages 68–579, 2003. (Cited on page 127.)

[44] U. Brandes, M. Gaertler, and D. Wagner. Engineering graph clustering: Models and experimental evaluation. *ACM Journal of Experimental Algorithmics*, 12, 2007. (Cited on page 127.)

[45] R.L. Breiger. The duality of persons and groups. *Social Forces*, 53(2):181–190, 1974. (Cited on page 21.)

[46] R.L. Breiger and P.E. Pattison. Cumulated social roles: The duality of persons and their algebras. *Social Networks*, 8(3):215–256, 1986. (Cited on page 2.)

[47] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. (Cited on pages 71, 72, and 81.)

[48] P. Bródka, P. Stawiak, and P. Kazienko. Shortest path discovery in the multi-layered social network. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 497–501, 2011. (Cited on page 98.)

[49] R.A. Brualdi. Matrices of zeros and ones with fixed row and column sum vectors. *Linear Algebra and its Applications*, 33:159–231, 1980. (Cited on page 37.)

[50] R.A. Brualdi. Algorithms for constructing (0,1)-matrices with prescribed row and column sum vectors. *Discrete Mathematics*, 306(23):3054–3062, 2006. (Cited on page 37.)

[51] M. Buchanan, G. Caldarelli, P. De Los Rios, F. Rao, and M. Vendruscolo, editors. *Networks in cell biology*. Cambridge University Press, Cambridge, UK, 2010. (Cited on page 2.)

[52] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. (Cited on page 2.)

[53] C.T. Butts. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11(1):13–41, 2008. (Cited on pages 33, 36, 37, and 146.)

[54] C.T. Butts. Revisiting the foundations of network analysis. *Science*, 325(5939):414–416, 2009. (Cited on page 3.)

[55] W.P. Butz and B. Boyle Torrey. Some frontiers in social science. *Science*, 312(5782):1898–1900, 2006. (Cited on page 90.)

[56] G.A. Calin, C.D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Nocha, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich, and C.M. Croce. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 99(24):15524–15529, 2002. (Cited on page 121.)

[57] C. Campbell, S. Yang, R. Albert, and K. Sheab. A network model for plant–pollinator community assembly. *Proceedings of the National Academy of Sciences*, 108(1):197–202, 2011. (Cited on page 101.)

[58] D. Caragea, V. Bahirwani, W. Aljandal, and W.H. Hsu. Ontology-based link prediction in the LiveJournal social network. In *Proceedings of the Eighth Symposium on Abstraction, Reformulation, and Approximation*, pages 34–41, 2009. (Cited on page 71.)

[59] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, and S. Boccaletti. Emergence of network features from multiplexity. *Scientific reports*, 3:1344, 2013. (Cited on page 146.)

[60] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009. (Cited on page 2.)

[61] Y. Chen, P. Diaconis, S. Holmes, and J. Liu. Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–121, 2005. (Cited on pages 37 and 48.)

[62] A. Clauset, C. Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008. (Cited on pages 27 and 77.)

[63] G.W. Cobb and Y.P. Chen. An application of Markov Chain Monte Carlo to community ecology. *The American Mathematical Monthly*, 110(4):265–288, 2003. (Cited on pages 34, 37, 40, 41, 42, and 44.)

[64] A.M. Cohen, W.R. Hersh, C. Dubay, and K. Spackman. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics*, 6 (1):103, 2005. (Cited on page 58.)

[65] E.F. Connor and D. Simberloff. The assembly of species communities: chance or competition? *Ecology*, 60(6):1132–1140, 1979. (Cited on pages 34 and 37.)

[66] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge, Massachusetts, USA, 3rd edition, 2009. (Cited on pages 26 and 69.)

[67] D.J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52): 22436–22441, 2010. (Cited on pages 77 and 146.)

[68] D. Davis, R. Lichtenwalter, and N.V. Chawla. Supervised methods for multi-relational link prediction. *Social Network Analysis and Mining*, 3(2):127–141, 2013. (Cited on page 16.)

[69] D.A. Davis and N.V. Chawla. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PlOS ONE*, 6(7):e22670, 2011. (Cited on page 146.)

[70] J. Diamond. Assembly of species communities. In M. Cody and J. Diamond, editors, *Ecology and evolution of communities*, pages 342–444. Belknap Press of Harvard University Press, 1975. (Cited on pages 34 and 42.)

[71] C.F. Dormann, J. Fründ, N. Blüthgen, and B. Gruber. Indices, graphs and null models: Analysing bipartite ecological networks. *The Open Ecology Journal*, 2:7–24, 2009. (Cited on pages 34, 37, 42, and 140.)

[72] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002. (Cited on pages 3, 34, 35, and 36.)

[73] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009. (Cited on page 98.)

[74] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, New York, 2010. (Cited on pages 2, 20, and 147.)

[75] D. Ellis, J. Furner-Hines, and P. Willett. Measuring the degree of similarity between objects in text retrieval. *Perspectives in Information Management*, 3(2):128–149, 1993. (Cited on pages 61 and 63.)

[76] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. (Cited on page 139.)

[77] P. Erdős and T. Gallai. Graphs with prescribed degree of vertices. *Matematikai Lapok*, 11:264–274, 1960. (Cited on page 39.)

[78] P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959. (Cited on page 34.)

[79] P.L. Erdős, I. Miklós, and Z. Toroczkai. A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs. *The Electronic Journal of Combinatorics*, 17(1):R66, 2010. (Cited on pages 37 and 39.)

[80] A. Esquela-Kerscher and F.J. Slack. Oncomirs – microRNAs with a role in cancer. *Nature*, 6(4):259–269, 2006. (Cited on page 121.)

[81] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. (Cited on page 30.)

[82] R. Feldman and J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, 2007. (Cited on page 58.)

[83] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010. (Cited on pages 3 and 18.)

[84] J.G. Foster, D.V. Foster, P. Grassberger, and M. Paczuski. Link and subgraph likelihoods in random undirected networks with fixed and partially fixed degree sequences. *Physical Review E*, 76(4):046112, 2007. (Cited on page 37.)

[85] J.G. Foster, D.V. Foster, P. Grassberger, and M. Paczuski. Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences*, 107(24):10815–10820, 2010. (Cited on pages 20 and 66.)

[86] L.C. Freeman. The sociological concept of "group": An empirical test of two models. *American Journal of Sociology*, 98(1):152–166, 1992. (Cited on page 100.)

[87] R.C. Friedman, K.K.H. Farh, C.B. Burge, and D. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009. (Cited on page 121.)

[88] C.I. Del Genio, H. Kim, Z. Toroczkai, and K.E. Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLOS ONE*, 5(4):e10012, 2010. (Cited on pages 37, 39, and 48.)

[89] M.B. Gerstein et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012. (Cited on page 139.)

[90] L. Getoor and C.P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7:3–12, 2005. (Cited on page 27.)

[91] E.N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959. (Cited on page 34.)

[92] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. In *ACM Transactions on Knowledge Discovery from Data*, volume 1, 2007. (Cited on pages 37, 41, 44, and 51.)

[93] L. Giot et al. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, 2003. (Cited on pages 88 and 89.)

[94] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. (Cited on pages 3, 19, 35, 51, 52, and 81.)

[95] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, and A.L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007. (Cited on page 52.)

[96] D.S. Goldberg and F.P. Roth. Assessing experimentally derived interactions in a small world. *Proceedings of the National Acedemy of Sciences*, 100(8):4372–4376, 2003. (Cited on pages 62 and 89.)

[97] J. Gómez-Gardeñes, D. Vilone, and A. Sanchez. Disentangling social and group heterogeneities: Public Goods games on complex networks. *European Journal of Physics*, 95(6):68003, 2011. (Cited on page 101.)

[98] J. Gómez-Gardeñes, I. Reinares, A. Arenas, and L.M. Floría. Evolution of cooperation in multiplex networks. *Scientific reports*, 2:620, 2012. (Cited on page 146.)

[99] P. Good. *Permutation, parametric, and bootstrap tests of hypothesis*. Springer, New York, 3rd edition, 2005. (Cited on pages 51 and 65.)

[100] A.D. Gordon. A review of hierarchical classification. *Journal of the Royal Statistical Society. Series A (General)*, 150(2):119–137, 1987. (Cited on page 55.)

[101] N.J. Gotelli. Null model analysis of species co-occurrence patterns. *Ecology*, 81(9):2606–2621, 2000. (Cited on pages 34, 37, and 42.)

[102] N.J. Gotelli and G.R. Graves. *Null Models in Ecology*. Smithsonian Institution Press, Washington D.C., 1996. (Cited on pages 34 and 58.)

[103] N.J. Gotelli and K. Rohde. Co-occurrence of ectoparasites of marine fishes: a null model analysis. *Ecology Letters*, 5(1):86–94, 2002. (Cited on pages 34, 37, and 42.)

[104] J. Gould. *Zoology of the voyage of H. M. S. Beagle Part 3. Birds*. Smith Elder and Co., London, 1841. (Cited on page 38.)

[105] R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009. (Cited on page 89.)

[106] M. Hanselmann, U. Köthe, M. Kirchner, B.Y. Renard, E.R. Amstalden, K. Glunde, R.M.A. Heeren, and F.A. Hamprecht. Toward digital staining using imaging mass spectrometry and random forests. *Journal of Proteome Research*, 8(7):3558–3567, 2009. (Cited on page 72.)

[107] M.A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *Workshop on Link Analysis, Counter-terrorism and Security*, 2006. (Cited on pages 77 and 146.)

[108] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, 2001. (Cited on pages 2, 18, 28, 55, and 69.)

[109] V. Havel. A remark on the existence of finite graphs. *Časopis pro pěstování matematiky*, 80(4):477–480, 1955. (Cited on page 39.)

[110] L. He, X. He, S.W. Lowe, and G.J. Hannon. microRNAs join the p53 network–another piece in the tumour-suppression puzzle. *Nature Reviews Cancer*, 7(11):819–822, 2007. (Cited on page 121.)

[111] C.A. Hidalgo and C. Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008. (Cited on page 147.)

[112] P. Holme and J. Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012. (Cited on page 147.)

[113] P. Holme, F. Liljeros, C.R. Edling, and B.J. Kim. Network bipartivity. *Physical Review E*, 68(5):056107, 2003. (Cited on page 21.)

[114] **E.Á. Horvát** and K.A. Zweig. One-mode projections of multiplex bipartite graphs. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 598–605, 2012. (Cited on pages 8, 9, 103, 109, and 117.)

[115] **E.Á. Horvát** and K.A. Zweig. A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs. *Social Network Analysis and Mining*, online first. (Cited on pages 8, 9, 99, 102, 104, 107, 112, 113, 114, and 115.)

[116] **E.Á. Horvát** and K.A. Zweig. Multiplex networks. In *Encyclopedia of Social Network Analysis and Mining*. Springer, to appear. (Cited on pages 8 and 22.)

[117] **E.Á. Horvát**, M. Hanselmann, F.A. Hamprecht, and K.A. Zweig. One plus one makes three (for social networks). *PLOS ONE*, 7 (4):e34740, 2012. (Cited on pages 7, 8, 78, 80, 81, 83, and 84.)

[118] **E.Á. Horvát**, M. Hanselmann, F.A. Hamprecht, and K.A. Zweig. You are who knows you: Predicting links between non-members of Facebook. In T. Gilbert, M. Kirkilionis, and G. Nicolis, editors, *Proceedings of the European Conference on Complex Systems 2012*, Springer Proceedings in Complexity, pages 309–316. Springer, 2013. (Cited on page 8.)

[119] **E.Á. Horvát**, J.D. Zhang, S. Uhlmann, Ö. Sahin, and K.A. Zweig. A network-based method to assess the statistical significance of mild co-regulation effects. *PLOS ONE*, 8(9):e73413, 2013. (Cited on pages 8, 9, 124, 127, 128, 130, 133, 134, 135, and 136.)

[120] **E.Á. Horvát**, A. Spitz, A. Gimmler, T. Stoeck, and K.A. Zweig. Validating and assessing low intensity interactions in complex networks. in preparation. (Cited on pages 8 and 9.)

[121] M. Inui, G. Martello, and S. Piccolo. MicroRNA control of signal transduction. *Nature Reviews Molecular Cell Biology*, 11(4):252–263, 2010. (Cited on page 137.)

[122] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272, 1901. (Cited on page 59.)

[123] D.A. Jackson, K.M. Somers, and H.H. Harvey. Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist*, 133(3):436–453, 1989. (Cited on page 55.)

[124] R. Jacob, D. Koschützki, K. Lehmann, L. Peeters, and D. Tenfelde-Podehl. Algorithms for centrality indices. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological foundations*, pages 62–82. Springer, 2005. (Cited on pages 18 and 56.)

[125] C. Jernigan and B. Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009. (Cited on page 77.)

[126] P. Kazienko, K. Musial, and T. Kajdanowicz. Multidimensional social network in the social recommender system. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 41(4):746–759, 2011. (Cited on pages 22 and 98.)

[127] H. Kim, Z. Toroczkai, P.L. Erdős, I. Miklós, and L.A. Székely. Degree-based graph construction. *Journal of Physics A: Mathematical and Theoretical*, 42(39):392001, 2009. (Cited on pages 37 and 39.)

[128] H. Kim, C.I. Del Genio, K.E. Bassler, and Z. Toroczkai. Constructing and sampling directed graphs with given degree sequences. *New Journal of Physics*, 14:023012, 2012. (Cited on pages 37, 39, and 46.)

[129] E.D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, New York, 2009. (Cited on pages 2, 3, 13, 16, 24, 26, 27, 29, 33, 34, 35, 36, 39, 48, 50, 52, 67, 87, and 146.)

[130] M. Koren. Pairs of sequences with a unique realization by bipartite graphs. *Journal of Combinatorial Theory (B)*, 21(3):224–234, 1976. (Cited on page 37.)

[131] D. Koschützki, K. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski. Centrality indices. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological foundations*, pages 16–61. Springer, 2005. (Cited on pages 18 and 56.)

[132] D. Koschützki, K. Lehmann, D. Tenfelde-Podehl, and O. Zlotowski. Advanced centrality concepts. In U. Brandes and

T. Erlebach, editors, *Network Analysis: Methodological foundations*, pages 83–110. Springer, 2005. (Cited on pages 18 and 56.)

[133] G. Kossinets and D.J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006. (Cited on page 147.)

[134] J. Kota, R.R. Chivukula, K.A. O'Donnell, E.A. Wentzel, C.L. Montgomery, H.W. Hwang, T.C. Chang, P. Vivekanandan, M. Torbenson, K.R. Clark, J.R. Mendell, and J.T. Mendel. Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell*, 137(6):1005–1017, 2009. (Cited on page 121.)

[135] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 611–617, 2006. (Cited on page 80.)

[136] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110, 2008. (Cited on page 52.)

[137] D.P. Landau and K. Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge University Press, New York, 3rd edition, 2009. (Cited on pages 24 and 41.)

[138] S. Lehmann, M. Schwartz, and L.K. Hansen. Biclique communities. *Physical Review E*, 78, 2008. (Cited on page 21.)

[139] E.A. Leicht, P. Holme, and M.E.J. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006. (Cited on pages 56, 57, 63, and 132.)

[140] S.K. Leivonen, R. Mäkelä, P. Östling, P. Kohonen, S. Haapa-Paananen, K. Kleivi, E. Enerly, A. Aakula, K. Hellström, N. Sahlberg, V.N. Kristensen, A.L. Børresen-Dale, P. Saviranta, M. Perälä, and O. Kallioniemi. Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene*, 28(44):3926–3936, 2009. (Cited on page 121.)

[141] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. *Proceedings 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005. (Cited on page 33.)

[142] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007. (Cited on page 147.)

[143] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. (Cited on pages 55 and 136.)

[144] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christachis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4):330–342, 2008. (Cited on page 98.)

[145] M. Li, Y. Fan, J. Chen, L. Gao, Z. Di, and J. Wu. Weighted networks of scientific communication: the measurement and topological role of weight. *Physica A*, 350(1–2):643–656, 2005. (Cited on page 21.)

[146] N. Li and G. Chen. Multi-layered friendship modeling for location-based mobile social networks. In *Proceedings of Mobiquitous 2009*, pages 1–10, 2009. (Cited on pages 22 and 98.)

[147] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. (Cited on pages 27, 67, 77, and 146.)

[148] R.N. Lichtenwalter, J.T. Lussier, and N.V. Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 243–252, 2010. (Cited on pages 67, 71, and 77.)

[149] L.P. Lim, N.C. Lau, P. Garrett-Engele, A. Grimson, J.M. Schelter, J. Castle, D.P. Bartel, P.S. Linsley, and J.M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005. (Cited on page 121.)

[150] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th International Conference on World Wide Web*, pages 1145–1146, 2009. (Cited on page 77.)

[151] G. Linden, B. Smith, and J. York. Amazon.com recommendations. Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003. (Cited on page 59.)

[152] G.S. Linoff and M.J. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing, Inc., 2nd edition, 2004. (Cited on page 58.)

[153] J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, New York, 2001. (Cited on page 38.)

[154] W. Liu and L. Lü. Link prediction based on local random walk. *Europhysics Letters*, 89(5):58007, 2010. (Cited on page 77.)

[155] Y.Y. Liu, J.J. Slotine, and A.L. Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011. (Cited on page 2.)

[156] F. Lorrain and H.C. White. Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1): 49–80, 1971. (Cited on page 57.)

[157] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdős is Eighty (Volume 2)*, pages 1–46, 1993. (Cited on pages 24, 25, and 26.)

[158] A. Lujambio and S.W. Lowe. The microcosmos of cancer. *Nature*, 482(7385):347–355, 2012. (Cited on page 121.)

[159] L. Ma, J. Teruya-Feldstein, and R.A. Weinberg. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*, 449(7163):682–688, 2007. (Cited on page 121.)

[160] M. Magnani and L. Rossi. The ML-model for multi-layer social networks. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 5–12, 2011. (Cited on pages 16, 23, and 98.)

[161] N. Malo, J.A. Hanley, S. Cerquozzi, J. Pelletier, and R. Nadon. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*, 24(2):167–175, 2006. (Cited on page 121.)

[162] M. Malumbres. miRNAs versus oncogenes: the power of social networking. *Molecular Systems Biology*, 8:569, 2012. (Cited on page 8.)

[163] K.K. Mane and K. Börner. Mapping topics and topic bursts in PNAS. *Proceedings of the National Acedemy of Sciences*, 101(suppl 1):5287–5290, 2004. (Cited on pages 21 and 59.)

[164] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, Massachusetts, USA, 1999. (Cited on page 148.)

[165] A. Marco, C. Konikoff, T.L. Karr, and S. Kumar. Relationship between gene co-expression and sharing of transcription factor binding sites in *Drosophila melanogaster*. *Bioinformatics*, 25(19): 2473–2477, 2009. (Cited on pages 62 and 140.)

[166] P.V. Marsden. Network data and measurement. *Annual Review of Sociology*, 16:435–463, 1990. (Cited on page 26.)

[167] S. Marsland. *Machine Learning. An Algorithmic Perspective*. Machine Learning and Pattern Recognition. Chapman & Hall/CRC, 2009. (Cited on pages 24, 28, 70, 71, and 72.)

[168] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002. (Cited on pages 3 and 37.)

[169] S. Maslov, K. Sneppen, and U. Alon. Correlation profiles and motifs in complex networks. In S. Bornholdt and H. Schuster, editors, *Handbook of graphs and networks. From the genome to the Internet*, pages 168–198. Wiley-VCH, 2003. (Cited on pages 48, 50, and 51.)

[170] S. Maslov, K. Sneppen, and A. Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333: 529–540, 2004. (Cited on page 39.)

[171] J.W. McDonald, P.W.F. Smith, and J.J. Forster. Markov chain monte carlo exact inference for social networks. *Social Networks*, 29(1):127–136, 2007. (Cited on pages 37 and 38.)

[172] M. McPherson, L. Smith-Lovin, and J.M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. (Cited on pages 20 and 98.)

[173] R. Meiri and J. Zahavi. Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 171(3):842–858, 2006. (Cited on pages 72 and 73.)

[174] J.B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. (Cited on page 148.)

[175] I. Miklós, P.L. Erdős, and L. Soukup. Towards random uniform sampling of bipartite graphs with given degree sequence. *The Electronic Journal of Combinatorics*, 20(1):P16, 2013. (Cited on pages 37 and 44.)

[176] R. Milo, S.S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. (Cited on pages 3, 37, 48, 51, and 66.)

[177] R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *http://arxiv.org/pdf/cond-mat/0312028v2.pdf*, 2003. (Cited on pages 37 and 41.)

[178] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattarcharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM Sigcomm Conference on Internet Measurement*, pages 29–42, 2007. (Cited on pages 81 and 85.)

[179] A. Mislove, B. Viswanath, K.P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 251–260, 2010. (Cited on page 77.)

[180] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6(2–3):161–179, 1995. (Cited on page 48.)

[181] D.M. Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2nd edition, 2004. (Cited on page 55.)

[182] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, and J.P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010. (Cited on pages 98, 146, and 147.)

[183] D. Murphy. Gene expression studies using microarray: principles, problems, and prospects. *Advances in Physiology Education*, 26(4):256–270, 2002. (Cited on page 121.)

[184] Z. Neal. Identifying statistically significant edges in one-mode projections. *Social Network Analysis and Mining*, online first. (Cited on page 21.)

[185] M.E.J. Newman. Scientific collaboration networks I. Network construction and fundamental results. *Physical Review E*, 64(1): 016131, 2001. (Cited on page 21.)

[186] M.E.J. Newman. Scientific collaboration networks II. Shortest paths, weighted networks, and centrality. *Physical Review Letters*, 64(1):016132, 2001. (Cited on page 21.)

[187] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002. (Cited on pages 19, 35, and 81.)

[188] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. (Cited on pages 35 and 37.)

[189] M.E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):26126, 2003. (Cited on page 19.)

[190] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23): 8577–8582, 2006. (Cited on page 19.)

[191] M.E.J. Newman. *Networks. An introduction*. Oxford, 2010. (Cited on pages 2, 3, 48, 49, 56, 57, 58, and 62.)

[192] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004. (Cited on pages 19, 21, and 88.)

[193] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001. (Cited on pages 37, 39, and 48.)

[194] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity–a review. *QSAR & Combinatorial Science*, 22(9–10): 1006–1026, 2004. (Cited on page 55.)

[195] A. Ochiai. Zoogeographical studies on the soleoid fishes found in japan and its neighbouring regions. *Bulletins of the Japanese Society for Scientific Fisheries*, 22(9):526–530, 1957. (Cited on page 59.)

[196] J. O'Madadhain, J. Hutchins, and P. Smyth. Prediction and ranking algorithms for event-based network data. *ACM SIGKDD Explorations Newsletter*, 7(2):23–30, 2005. (Cited on page 77.)

[197] J. Ouellette. The mathematical shape of things to come. Quanta Magazine, October 2013. (Cited on page 1.)

[198] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005. (Cited on pages 19 and 89.)

[199] G. Palla, A.L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007. (Cited on page 147.)

[200] J. Park and A.L. Barabási. Distribution of node characteristics in complex networks. *Proceedings of the National Academy of Sciences*, 104(46):17916–17920, 2007. (Cited on page 98.)

[201] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001. (Cited on page 33.)

[202] R. Pastor-Satorras, M. Rubi, and A. Diaz-Guilera, editors. *Statistical Mechanics of Complex Networks*, volume 625 of *Lecture Notes in Physics*. Springer, Berlin, Heidelberg, 2003. (Cited on page 2.)

[203] R.K. Pathria and P.D. Beale. *Statistical Mechanics*. Academic Press, Oxford, UK, 3rd edition, 2011. (Cited on page 2.)

[204] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W.J. Haveman, and P.P. Pandolfi. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301): 1033–1038, 2010. (Cited on page 121.)

[205] M.A. Porter, P.J. Mucha, M.E.J. Newman, and C.M. Warmbrand. A network analysis of committees in the U.S. House of Representatives. *Proceedings of the National Academy of Sciences*, 102 (20):7057–7062, 2005. (Cited on page 19.)

[206] M.A. Porter, J.P. Onnela, and P.J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1166, 2009. (Cited on pages 3, 19, and 51.)

[207] C. Prieto and J. De Las Rivas. Structural domain–domain interactions: Assessment and comparison with protein–protein interaction data to improve the interactome. *Proteins*, 78(1):109–117, 2009. (Cited on page 89.)

[208] J. Quackenbush. Microarrays—guilt by association. *Science*, 302 (5643):240–241, 2003. (Cited on page 122.)

[209] J. Ramasco, S. Dorogovtsev, and R. Pastor-Satorras. Self-organization of collaboration networks. *Physical Review E*, 70 (3):036106, 2004. (Cited on page 101.)

[210] J.J. Ramasco and S.A. Morris. Social inertia in collaboration networks. *Physical Review E*, 73(1):016122, 2006. (Cited on page 21.)

[211] A.R. Rao, R. Jana, and S. Bandyopadhyay. A Markov chain Monte Carlo method for generating random (0,1)-matrices with given marginals. *The Indian Journal of Statistics*, 58(Series A):225–242, 1996. (Cited on pages 37, 40, 41, 42, and 44.)

[212] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1553, 2002. (Cited on page 19.)

[213] S. Redner. Networks: Teasing out the missing links. *Nature*, 453 (7191):47–48, 2008. (Cited on page 77.)

[214] J. De Las Rivas and C. Fontanillo. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLOS Computational Biology*, 6(6):e1000807, 2010. (Cited on pages 88 and 89.)

[215] A. Roberts and L. Stone. Island-sharing by archipelago species. *Oecologia*, 83(4):560–567, 1990. (Cited on pages 34, 37, and 42.)

[216] J.M. Roberts. Simple methods for simulating sociomatrices with given marginal totals. *Social Networks*, 22(3):273–283, 2000. (Cited on pages 37 and 38.)

[217] J.L. Rodgers and W.A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988. (Cited on page 62.)

[218] T.A. Runkler. *Data analytics. Models and algorithms for intelligent data analysis*. Springer Vieweg, Wiesbaden, 2012. (Cited on pages 28, 29, 56, 63, 69, 70, and 73.)

[219] H.J. Ryser. Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics*, 9:371–377, 1957. (Cited on pages 37, 40, 42, and 44.)

[220] S. Saavedra, F. Reed-Tsochas, and B. Uzzi. A simple model of bipartite cooperation for ecological and organizational networks. *Nature*, 457(7228):463–466, 2008. (Cited on page 101.)

[221] Ö. Sahin, C. Löbke, U. Korf, H. Appelhans, H. Sültmann, A. Poustka, S. Wiemann, and D. Arlt. Combinatorial RNAi for quantitative protein network analysis. *Proceedings of the National Academy of Sciences*, 104(16):6579–6584, 2007. (Cited on page 121.)

[222] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. (Cited on pages 55 and 59.)

[223] J.G. Sanderson. Testing ecological patterns. *American Scientist*, 88(4):332–339, 2000. (Cited on page 38.)

[224] J. Satoh. Molecular network analysis of human microRNA targetome: from cancers to Alzheimer's disease. *BioData Mining*, 5(17):1–22, 2012. (Cited on page 121.)

[225] U. Sauer, M. Heinemann, and N. Zamboni. Getting closer to the whole picture. *Science*, 316(5824):550–551, 2007. (Cited on page 146.)

[226] S.E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1): 27–64, 2007. (Cited on page 18.)

[227] M.A. Schaub, A.P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder. Linking disease associations with regulatory information in the human genome. *Genome Research*, 22(9):1748–1759, 2012. (Cited on page 139.)

[228] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D.R. White. Economic networks: The new challenges. *Science*, 325(5939):422–425, 2009. (Cited on page 2.)

[229] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008. (Cited on page 121.)

[230] T.Z. Sen, A. Kloczkowski, and R.L. Jernigan. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*, 7(1):355, 2006. (Cited on page 19.)

[231] M.Á. Serrano, M. Boguñá, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009. (Cited on page 100.)

[232] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1):64–68, 2002. (Cited on pages 3, 33, 37, 51, and 52.)

[233] B.A. Shoemaker and A.R. Panchenko. Deciphering protein–protein interactions. Part I. experimental techniques and databases. *PLOS Computational Biology*, 3(3):e42, 2007. (Cited on page 89.)

[234] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10(5): 335–347, 1989. (Cited on page 73.)

[235] P.H.A. Sneath and R.R. Sokal. *Numerical taxonomy. The principles and practice of numerical classification*. W.H. Freeman and Company, San Francisco, USA, 1973. (Cited on pages 55 and 63.)

[236] R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bulletin of Mathematical Biophysics*, 13(2):107–117, 1951. (Cited on page 34.)

[237] A. Spitz, K.A. Zweig, and **E.Á. Horvát**. SICOP resources. URL `http://cna.cs.uni-kl.de/SICOP/`. (Cited on pages 123 and 141.)

[238] A. Spitz, K.A. Zweig, and **E.Á. Horvát**. SICOP: identifying significant co-interaction patterns. *Bioinformatics*, 29(19):2503–2504, 2013. (Cited on pages 8, 9, and 141.)

[239] L. Stone and A. Roberts. The checkerboard score and species distribution. *Oecologia*, 85(1):74–79, 1990. (Cited on pages 34, 37, and 42.)

[240] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003. (Cited on page 122.)

[241] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. (Cited on page 139.)

[242] P. Sudarsanam, Y. Pilpel, and G.M. Church. Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Research*, 12(11):1723–1731, 2002. (Cited on page 62.)

[243] J. Sun, X. Gong, B. Purow, and Z. Zho. Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma. *PLOS Computational Biology*, 8(7):e1002488, 2012. (Cited on page 62.)

[244] B. Suter, S. Kittanakom, and I. Stagljar. Two-hybrid technologies in proteomics research. *Current Opinion in Biotechnology*, 19(4): 316–323, 2008. (Cited on page 89.)

[245] M. Szell and S. Thurner. Measuring social dynamics in a massive multiplayer online game. *Social Networks*, 32(4):313–329, 2010. (Cited on pages 23 and 98.)

[246] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010. (Cited on pages 23, 98, and 146.)

[247] P.N. Tan, V. Kumar, and J. Sivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4): 293–313, 2004. (Cited on page 61.)

[248] T.T. Tanimoto. An elementary mathematical theory of classification and prediction. Technical report, IBM Internal Report, 1958. (Cited on page 59.)

[249] R. Taylor. Contrained switchings in graphs. In K.L. McAvaney, editor, *Combinatorial Mathematics VIII*, volume 884 of *Lecture Notes in Mathematics*, pages 314–336. Springer Berlin Heidelberg, 1981. (Cited on pages 37 and 42.)

[250] H. Tilgner, D.G. Knowles, R. Johnson, C.A. Davis, S. Chakraborty, S. Djebali, J. Curado, M. Snyder, T.R. Gingeras, and R. Guigó. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625, 2012. (Cited on page 139.)

[251] R. Toivonena, L. Kovanena, M. Kiveläa, J.P. Onnela, J. Saramäkia, and K. Kaski. A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks*, 31(4):240–254, 2009. (Cited on page 85.)

[252] A.L. Traud, E.D. Kelsic, P.J. Mucha, and M.A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review*, 53(3):526–543, 2011. (Cited on pages 65, 66, and 79.)

[253] S. Travazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–285, 1999. (Cited on page 62.)

[254] M. Tumminello, S. Micciché, F. Lillo, J. Piilo, and R.N. Mantegna. Statistically validated networks in bipartite complex systems. *PLOS ONE*, 6(3):e17994, 2011. (Cited on pages 21 and 37.)

[255] W.T. Tutte. The factors of a graph. *Canadian Journal of Mathematics*, 4(3):314–328, 1952. (Cited on page 39.)

[256] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J.M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000. (Cited on page 89.)

[257] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012. (Cited on page 80.)

[258] S. Uhlmann, H. Mannsperger, J.D. Zhang, **E.Á. Horvát**, C. Schmidt, M. Küblbeck, F. Henjes, A. Ward, U. Tschulena, K.A. Zweig, U. Korf, S. Wiemann, and Ö. Sahin. Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Molecular Systems Biology*, 8:570, 2012. (Cited on pages 8, 121, 125, 137, and 138.)

[259] S. Valenzuela, N. Park, and K.F. Kee. Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication*, 14(4):875–901, 2009. (Cited on page 85.)

[260] S. Vasudevan, Y. Tong, and J.A. Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–1934, 2007. (Cited on page 121.)

[261] A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009. (Cited on page 2.)

[262] I. Voineagu, X. Wang, P. Johnston, J.K. Lowe, Y. Tian, S. Horvath, J. Mill, R.M. Cantor, B.J. Blencowe, and D.H. Geschwind. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351):380–384, 2011. (Cited on page 140.)

[263] C. Von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002. (Cited on page 87.)

[264] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *7th IEEE International Conference on Data Mining*, pages 322–331, 2007. (Cited on page 77.)

[265] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.L. Barabási. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1100–1108, 2011. (Cited on pages 71 and 77.)

[266] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. (Cited on pages 2, 16, 21, 23, 90, and 98.)

[267] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. (Cited on pages 2, 17, 21, 33, 35, and 56.)

[268] G.M. Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004. (Cited on page 77.)

[269] C. Willyard. The agony and ecstasy of cross-disciplinary collaboration. Science Careers, August 2013. (Cited on page 1.)

[270] G. Xu, Y. Zhang, and L. Li. *Web mining and social networking. Techniques and applications*. Springer, New York, 2011. (Cited on pages 29, 31, 58, 62, and 116.)

[271] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, number 3, 2012. (Cited on pages 88 and 90.)

[272] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R.Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Acedemy of Sciences*, 101(16):5934–5939, 2004. (Cited on pages 38, 65, and 66.)

[273] H. Yu et al. High-quality binary protein interaction map of yeast interactome network. *Science*, 322(5898):104–110, 2008. (Cited on page 89.)

[274] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977. (Cited on page 88.)

[275] L. Zahoránszky, G. Katona, P. Hári, A. Málnási-Csizmadia, K.A. Zweig, and G. Zahoránszky-Kőhalmi. Breaking the hierarchy – a new cluster selection mechanism for hierarchical clustering methods. *Algorithms for Molecular Biology*, 4(12):1–22, 2009. (Cited on page 100.)

[276] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1544–6115, 2005. (Cited on page 140.)

[277] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web*, pages 531–540, 2009. (Cited on page 77.)

[278] T. Zhou, J. Ren, M. Medo, and Y.C. Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76 (4):046115, 2007. (Cited on page 21.)

[279] T. Zhou, L. Lü, and Y.C. Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009. (Cited on page 77.)

[280] K.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 second International Conference on Advances in Social Network Analysis and Mining*, pages 200–207. IEEE Computer Society, 2010. (Cited on pages 21, 104, 105, and 117.)

[281] K.A. Zweig. Network representations of complex data. In *Encyclopedia of Social Network Analysis and Mining*. Springer, to appear. (Cited on pages 3 and 87.)

[282] K.A. Zweig and **E.Á. Horvát**. How to evaluate co-occurrences of products in market-baskets from real-world applications. In *Proceedings of the Mini-conference on Applied Theoretical Computer Science*, 2010. (Cited on pages 8, 9, 48, and 95.)

[283] K.A. Zweig and M. Kaufmann. A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, 1(3):187–218, 2011. (Cited on pages 21, 34, 37, 48, 106, and 117.)