Generalised Networks for Protein Interaction Analysis

Florian Klimm
Kellogg College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2018

Abstract

Generalised Networks for Protein Interaction Analysis
Florian Klimm — Trinity 2018

Protein interaction networks (PINs) are mathematical representations of interactions between proteins within organisms. Studying their properties can give insights into biological functions and the importance of proteins, and it can therefore aid in drug-discovery. Graphs are the most common mathematical object used to represent PINs. In this thesis, we investigate generalised mathematical representations of PINs. In particular, we examine *multilayer networks* (MLNs) and *node-weighted networks*. These mathematical objects allow the construction of temporal PINs and tissue-specific PINs by combining gene-expression data with PINs. We introduce *promiscuity* as an information-theoretical measure of a node's distribution of neighbours across different layers in MLNs. We examine promiscuity in synthetic networks and tissue-specific PINs and find that the vast majority of proteins are not cell-type specific. Integrating temporal gene-expression data with PINs allows us to create temporal PINs in the form of MLNs. We investigate an eigenvector-based temporal centrality in a temporal PIN of yeast during the cell cycle. We thereby examine the change of proteins' importance over time, which reflects their activity during the cell cycle. We then discuss the detection of community structure in node-weighted networks. For synthetic networks, we show that considering node weights can alter detected community structure. We combine a human PIN with gene-expression data to construct tissue-specific PINs and investigate their community structure. Comparing the detected communities with gene-ontology information, we find some tissue-specific functions of these PINs. Overall, the case studies in this thesis suggest that MLN and node-weighted networks are suitable for the integration of protein-interaction data with other biological data sets.

Acknowledgements

This thesis was made possible through the help — in large and in small — of numerous people, to all of which I am grateful.

First and foremost, I thank my supervisors Mason A. Porter, Charlotte M. Deane, and Jonny Wray for their endless support. Your comments in countless meetings, Skype calls, and drafts shaped this thesis and my mind and so made me a better researcher. For this education provided by you, I will be forever grateful.

Whilst I am grateful to many fellow students at Oxford, I would like to particularly thank A. Roxana Pamfil, for many fruitful discussions and for sharing her thoughts and code with me; Se-Wook Oh, for commenting on drafts; OPIG, for much cake and fun; and Bernadette J. Stolz and Tammo Rukat, for being great flatmates. My conversations with Marc Wiedermann and Benjamin F. Maier were often insightful and always great fun. The discussions with my *Transfer of Status* examiners, Gesine Reinert and Mariano Beguerisse Díaz, and *Confirmation of Status* examiners, Jotun Hein and Felix Reed-Tsochas, and *Viva* examiners Heather Harrington and James Wakefield were extremely helpful and greatly improved this thesis. I thank Philip K. Maini for taking care of much Oxonian paperwork on my behalf. My thank also goes to Danielle S. Bassett and Peter J. Mucha, who introduced me to network science. Without meeting both of you, I would not have ended up doing a DPhil. Furthermore, I am grateful to my former supervisors, Jürgen Kurths and Gorka Zamora-López, for their ongoing support.

For the last four years in Oxford, Kellogg College and its Middle Common Room were central to my social life and helped to keep me sane while thinking a lot about networks. Thanks to everybody who supported me in this way.

My studies were made financially possible by the EPSRC and MRC under grant number EP/L016044/1 with further contributions from e-Therapeutics. Throughout my studies I received travel support from Kellogg College, the American Mathematical Society, e-Therapeutics, the University of California, the Mathematical Biosciences Institute, the Max Planck Institute for the Physics of Complex Systems, and the Mathematical Institute, for which I am very thankful.

Finally, a heartfelt thank you to my parents, for being always there for me; to my brother and his marvellous family, for their uplifting emails; and to Sarah, for everything.

Contents

List of Figures **xi**

List of Abbreviations **xiii**

List of Symbols **xv**

1 Introduction **1**
1.1 Cellular Biology as a Complex System 1
1.1.1 Systems Biology 1
1.1.2 Complex Systems and Networks 2
1.2 Protein Interaction Networks 5
1.2.1 Challenges in Protein Interaction Analysis 9
1.3 Thesis Overview 11

2 Foundations **15**
2.1 Proteins 17
2.1.1 Gene Expression, Protein Biosynthesis, and Protein Abundance 21
2.1.2 Measurement of Gene Expression 25
2.2 Protein Interaction Networks 27
2.2.1 Measurement of Protein–Protein Interactions 29
2.2.2 Genetic Interactions 30
2.2.3 Protein-Interaction Databases 31
2.2.4 Reliability and Coverage of Protein Interaction Networks 32
2.3 Networks 34
2.4 Synthetic Network Models 40
2.4.1 Erdős–Rényi Model 40
2.4.2 Configuration Model 40
2.4.3 Stochastic Block Model 41
2.4.4 Regular Ring Lattice 42
2.5 Centrality Measures in Networks 42

2.5.1 Eigenvector-based Centralities 44
2.6 Community Detection in Networks 45
2.6.1 Modularity 46
2.7 Multilayer Networks 51
2.7.1 Monolayer Networks 53
2.7.2 Edge-coloured Multigraphs 54
2.7.3 Multiplex Networks 54
2.7.4 Temporal Networks 56
2.7.5 Community Detection in Multilayer Networks 58
2.8 Community Detection in Protein Interaction Networks 59
2.8.1 Community Detection in Monolayer Protein Interaction Networks 59
2.8.2 Community Detection in Multilayer Protein Interaction Networks 61
2.9 Gene Ontology Enrichment 62
2.9.1 Gene Ontology Data 62
2.9.2 Hypergeometric Test 64

3 'Promiscuity' of Nodes in Multilayer Networks 69
3.1 Promiscuity of Nodes 71
3.1.1 Definition of Promiscuity 71
3.1.2 Proof that Promiscuity $p_i \in [0, 1]$ 78
3.1.3 Proof that Promiscuity $p_i = 1$ in Colour-regular Edge-coloured Multigraphs 79
3.2 Promiscuity p_i in Synthetic Networks 80
3.2.1 Network Composed of an Erdős–Rényi Layer and a Ring Lattice Layer 81
3.2.2 Network Composed of Two Erdős–Rényi Layers 85
3.2.3 Conclusion for Synthetic Two-Layer Networks 87
3.3 Promiscuity p_i in Synthetic Networks under Rewiring 88
3.3.1 Network Construction 88
3.3.2 Randomization 89
3.3.3 Promiscuity p_i under Randomization 90
3.4 Promiscuity p_i in Empirical Networks 96
3.5 Promiscuity in Tissue-specific Protein Interaction Networks 100
3.5.1 Tissue-specific Transcription Factor Regulation 101
3.5.2 Tissue-specific Gene Regulation 105
3.6 Conclusions 109

4 Eigenvector Centrality in Temporal Protein Interaction Networks 113
4.1 Introduction . . . 114
4.2 The Yeast Cell Cycle . . . 116
4.3 Construction of Temporal Protein Interaction Networks . . . 118
4.4 Data . . . 122
4.5 Eigenvector-based Centrality in Temporal Networks . . . 123
4.5.1 Inter-Layer Coupling of Centrality Matrices . . . 124
4.5.2 Conditional Centrality for Temporal Multilayer Networks . . . 124
4.5.3 Singular Perturbation in the Strong-Coupling Limit . . . 125
4.6 Results . . . 127
4.6.1 Centrality in the Strong-Coupling Regime . . . 127
4.6.2 Centrality Trajectories during the Yeast Cell Cycle . . . 132
4.7 Discussion . . . 136
4.7.1 Future Directions . . . 140

5 Node-weighted Protein Interaction Networks 143
5.1 Motivation . . . 144
5.2 Model Specification . . . 147
5.2.1 Asymptotic Discussion . . . 152
5.3 Synthetic Examples . . . 154
5.4 Application to Tissue RNA-Abundance Data . . . 159
5.4.1 Choice of the Resolution Parameter γ . . . 160
5.4.2 Comparison with Null Models . . . 162
5.4.3 Comparison across Tissues . . . 166
5.4.4 Comparison of Detected Partitions with Gene Ontology Annotations . . . 168
5.5 Conclusions . . . 173

6 Conclusions 177

A The GenLouvain Community Detection Algorithm 183
A.1 The Original Louvain Algorithm . . . 183
A.2 The GenLouvain Algorithm . . . 185

B Community Structure in a Temporal Protein Interaction Network 187
B.1 Introduction 187
B.2 Data and Preprocessing 188
B.2.1 Multilayer Network Construction 189
B.3 Community Detection 190
B.3.1 Parameter Choice 192
B.4 Conclusions 196

C Relating Modularity Maximization and Stochastic Block Models in Multilayer Networks 199

D Comparison of Node-weighted Modularity Function with Weighted Likelihood Stochastic Block Model 203

E Network Construction 207

Bibliography 212

References 213

List of Figures

1.1 Schematic representations of the network generalisations employed in this thesis . . . 12

2.1 Actin–profilin complex in *Bos taurus* . . . 22
2.2 Experimental methods in the BioGRID database . . . 30
2.3 Growth of the BioGRID database . . . 33
2.4 A network consisting of $N = 5$ nodes . . . 34
2.5 Degree distribution of a human protein interaction network and an Erdős–Rényi network . . . 37
2.6 Community structure in a PIN of HIV . . . 49
2.7 An example of the most general type of multilayer network that we use in this thesis . . . 53
2.8 A few different types of multilayer networks . . . 55
2.9 A subgraph of the domain 'biological process' . . . 63
2.10 GO enrichment with a hypergeometric test and multiple-testing correction 66

3.1 (a) Illustration of a MLN as an edge-coloured multigraph and (b) the role of some of the nodes in this network . . . 73
3.2 Promiscuity p_i and multilayer degree K_i for a synthetic network consisting of an ER layer and a ring lattice layer . . . 82
3.3 Promiscuity p_i and multilayer degree K_i for a synthetic two-layer network that consists of a sparse ER layer and a dense ER layer . . . 87
3.4 Supra-adjacency matrices of a ring-lattice MLN under randomization . 90
3.5 Promiscuity p_i of a ring-lattice MLN under randomization . . . 91
3.6 Violin plots [193] of the promiscuity distribution $P(p_i)$ across all nodes in eleven different networks . . . 98
3.7 Multilayer organisation of transcription factors in human cells . . . 103
3.8 Tissue-specific organization of genes in human cells . . . 107

4.1 The mitotic cell cycle in yeast . . . 116

4.2 Gene-expression profiles during the yeast cell cycle 119
4.3 Construction of temporal PINs from temporal gene expression profiles and a static PIN 120
4.4 Time-averaged centralities α_i and first-mover scores m_i for proteins during the yeast cell cycle 130
4.5 Centrality trajectories of SLA2 during yeast cell cycles for several values of the coupling parameter 133
4.6 Centrality trajectories during yeast cell cycles for the top three and bottom three proteins according to time-averaged centrality α_i and mover score m_i 135

5.1 We can combine a network (a) with the node weights (b) to obtain a node-weighted network (c) 147
5.2 Weight function $W(w_1, w_2)$ for three different values of the steepness parameter $s \in \{0.1, 1, 10\}$ 150
5.3 Community detection in a synthetic node-weighted network 156
5.4 To fix the resolution parameter γ for the node-weighted community detection, we use an iterative algorithm for finding an 'optimal' choice of γ 161
5.5 We compare the detected community structure in the PIN with three null models 163
5.6 The NMI of community structure in two tissues and across tissues . . . 168
5.7 The number n_{enriched} of enriched GO terms as a function of steepness s 169
5.8 The mean hierarchy level $\langle h \rangle$ of all terms that are enriched for two tissue-specific PINs as a function of the steepness s 173

B.1 Alluvial diagrams of modular structure in a temporal PIN 191
B.2 Convergence of community-detection parameters γ and ω 193
B.3 Change in size of communities 195

List of Abbreviations

DNA . . . . . Deoxyribonucleic acid

IC . . . . . Information content

ML . . . . . Maximum likelihood

MLN . . . . . Multilayer network

NMI . . . . . Normalised mutual information

PIN . . . . . Protein interaction network

PPI . . . . . Protein–protein interaction

RNA . . . . . Ribonucleic acid

TDC . . . . . Temporal decay coupling

TF . . . . . Transcription factor

SBM . . . . . Stochastic Block Model

List of Symbols

N	Number of nodes in a network
M	Number of edges in a network
L	Number of layers in a multilayer network
T	Number of time-layers in a temporal multilayer network
$\mathbf{A}$	Adjacency matrix with elements A_{ij}
$\mathbf{Q}$	Modularity matrix with elements Q_{ij}
γ	Resolution parameter in a modularity function
ω	Inter-layer coupling in multilayer networks
$\mathbb{A}$	Supra-adjacency matrix
$\mathbf{C}$	Centrality matrix
$\mathbb{C}$	Supra-centrality matrix
α_i	Time-averaged centrality of node i in a temporal multilayer network
m_i	First-order mover score of node i in a temporal multilayer network
k_i	Degree of node i in a monolayer network
$\mathbf{k}_i$	Degree vector of node i in an edge-coloured multigraph
$\mathbf{P}_i$	Promiscuity vector of node i in an edge-coloured multigraph
$\mathbf{P}'_i$	Normalized promiscuity vector of node i in an edge-coloured multigraph
p_i	Promiscuity of node i in an edge-coloured multigraph

1 Introduction

Contents

1.1 Cellular Biology as a Complex System 1
1.1.1 Systems Biology 1
1.1.2 Complex Systems and Networks 2
1.2 Protein Interaction Networks 5
1.2.1 Challenges in Protein Interaction Analysis 9
1.3 Thesis Overview 11

1.1 Cellular Biology as a Complex System

1.1.1 Systems Biology

Cells are the fundamental building blocks of life and consist of millions of molecules [190, 403]. *Molecular biology* is the study of these molecules, their interactions, and the implications for cells and large-scale organisms [470].

A reductionist method of examing biological systems (e.g., cells) is to dissect them into constituent parts (e.g., molecules) and investigate their characteristics (e.g.,

physical and chemical properties or biological functions [464]). The focus of molecular biology has been to investigate single molecules in detail. While this approach gives valuable insights into the function of cells, it has limitations [340]. Most notably, molecules interact with each other to fulfil their biological functions [180]. For example, molecules participate in metabolic pathways, signalling pathways, or other 'functional modules'. Such functional modules are sets of molecules that underlie biological functions. *Systems biology* aims to understand higher levels of organisation, for example, in cells. Such holistic approaches do not exclusively examine the constituent parts; they also explicitly consider large-scale interaction structure between them [13]. As disease often arises from dysfunctionality inside cells, systems biology may also have implications for medical treatment and drug discovery [486]. In this thesis, we focus on proteins, one particular type of molecule in cells (see Section 2.1), and investigate their function from a systemic perspective.

1.1.2 Complex Systems and Networks

Many real-world systems — including cells — are *complex systems* [47, 442]. Here, a system is a set of units that form an integrated whole. It is not straightforward to derive a clear definition that unites all complex systems observable in nature, society, and technology [245]. They do, however, often exhibit common characteristics:

1. They consist of a large number of interacting elements.

2. They exhibit *emergence*; this refers to a collective behaviour that is difficult to

anticipate from the behaviour of the elements.

3. The collective behaviour is not driven by a central controller.

As the behaviour of an entire system is difficult to anticipate from the behaviour of the parts, one calls it *complex*. While the concept of *emergence* is (at least) as old as Aristotle[1](384–322 BC), the availability of data and the development of mathematical and computational methods has led to increased attention in the recent decades [327].

In biology, complex systems occur over a large range of sizes and spatial scales [422]. In an ant colony, thousands of individual insects collaborate and exhibit self-organisation with animals fulfilling distinct tasks [118]. Species interact with each other in ecological niches — for example, through predator–prey interactions [469]. A human brain consists of billions of neurons that connect to each other via synapses [186, 453]. Patterns on shells and on animals's fur may form through an interplay of diffusion and chemical reactions [27, 460]. Complex biological systems also arise at different timescales. Chemical reactions can take from femtoseconds to hours [388, 410]. Organisms adapt to evolutionary pressures over millions of years [483]. Additionally, complex systems often interact with each other. For example, evolutionary pressure may lead to the natural selection of certain fur patterns or preserve the structure of proteins with important biological functions. The interaction between complex systems of different temporal and spatial scales can be crucial for their behaviour. In some cases, however, it is fruitful to discuss them independently of one another.

[1] 'The totality is not, as it were, a mere heap, but the whole is something besides the parts.' Aristotle *Metaphysics* Book VIII, 1045a.8–10 [389].

There are plenty of examples of complex systems that are not predominantly biological[2], e.g., climate, river basins, cities, economics, the internet, and ultimately the entire universe [330]. In this thesis, we focus on biological systems at a molecular level. Some of the tools and methods that we develop, however, are also applicable for studying other complex systems (e.g., 'promiscuity' in Chapter 3 and node-weighted community detection in Chapter 5).

One aim of complex-systems science is to understand how emergence occurs for a given system. Another (ambitious) aim is to find common principles that govern different complex systems. Many complex systems — whether human-made, such as power-grids, or natural, such as brains — fulfil certain functions. Scholars try to gain understanding of the extent to which the structure of the complex system facilitates these functions.

To help understand a system, it is often useful to construct a *mathematical model*[3], which is a simplified mathematical representation of a real-world system. It is important that one does not implement all features of a real system in its model but only essential ones, which can be different for different research questions about the same system. Many different mathematical objects can be useful representations of real-world systems. For example, one can model the growth of bacteria using differential equations [322], the spread of tumours using 'agent-based' models [18], and a neuron as a set of coupled differential equations (such as the Hodgkin–Huxley model) [191]. For many

[2]Many of these systems might be influenced by living organisms, but their foremost features are not of a biological nature.

[3]A *biological model* is an organism of a non-human species that is studied to better understand the biological processes [144].

real-world systems, there exist multiple mathematical models, potentially with different scopes (i.e., situations to which the model is applicable). A *Poisson process*, for example, is another model of neuronal activity [436]. These different models often represent different *levels of abstraction* (i.e., different amounts of detail concerning the (biological) processes that govern a system). One challenge for researchers is the selection of appropriate models for a (biological) system [168].

Networks — in their simplest form *graphs* — are one way to represent complex systems [330]. In Section 2.3, we give a precise mathematical definition; loosely speaking a graph is a collection of points and pairwise connections between them. Leonard Euler introduced them in 1753 in his work on the *Seven Bridges of Königsberg* [140] and they attracted attention in the 20$^{\text{th}}$ century from a wide range of disciplines [330, 440], including the physical sciences (e.g., [136]), sociology (e.g., [476]), and climatology (e.g., [125, 433]). In biology, networks have been used to describe brains (e.g., [430]), diseases (e.g., [28]), ecological interactions (e.g., [210]), cellular dynamics (e.g., [306]), metabolism (e.g., [13, 298]), social networks of animals (e.g., [241]), and many other phenomena. In this thesis, we focus on *protein interaction networks* (PINs), which are representations of the interactions of proteins in cells.

1.2 Protein Interaction Networks

In this section, we discuss the relevance PINs for systems biology, medicine, and pharmacology. In Section 2.2, we give a detailed description of the construction

and analysis of PINs.

As discussed in Section 1.1, cells consist of millions of molecules of many different types. These work together to form higher entities — cells and multicellular organisms — so cells are examples of complex systems [284]. *Biological pathways* are interactions and chemical reactions between the molecules in a cell [222]. Proteins are active macromolecules that are crucial in these pathways. Prominent examples of proteins are *enzymes*, which catalyse chemical reactions. Proteins often act in combination with each other — for example, in the form of complexes that consist of multiple proteins. We give more background information on proteins in Section 2.1.

We can represent a set of protein–protein interactions (PPIs) in an organism as a network. These PINs can help to disentangle the complex and multimodal function of proteins in cellular systems [497]. Here we review some applications of PINs.

Reacting to external and internal stimuli is crucial for a cell's survival [209]. PINs process biological information and organise cellular response. To enable information processing, many PINs have common organisational principles in the form of *motifs*[4] [87, 305, 411, 434]. These are small subnetworks that occur more often in a network than expected from random null models. In PINs, for example, feedback loops tend to be overrepresented [414, 495]. Such motifs can stabilise the abundance of a protein in an organism [467] and examining the presence of network motifs can also give insights into the way that a cell processes information.

[4] A similar concept to motifs are *graphlets* [372].

Evolution has shaped cellular organisation, and has also shaped PINs [296]. Some PPIs or motifs are preserved by evolution and thus are present in similar forms across species [261]. One can thus compare PINs of different species to detect their evolutionary similarity [12]. One can use PINs to predict interactions between proteins [263] or functional orthology (i.e, genes in different species that evolved from a common ancestral gene [418]).

PINs can also give insight into the function of individual proteins. For some proteins, we know (some of) their biological functions. Researchers collect such information in databases such as the *Gene Ontology Database* (see Section 2.9). For other proteins, that information is not available. One can use PINs to predict the function of proteins [412]. The common principle of such predictive tasks is that proteins that interact with one another are more likely to have similar functions than proteins that do not interact. Different computational approaches have used this 'functional assortativity' to predict the function of proteins — for example, by looking at the function of a protein's neighbours (e.g., [85, 407]), by separating a PIN into functional homogenous subnetworks (e.g., [323]), and by examining groups of proteins (e.g., [180, 363]).

Malfunction in PINs can manifest in disease or cell death [395, 398]. PINs can therefore give inights into diseased organisms (e.g., [78, 205, 207, 473]). The two main areas where PINs can contribute are. (1) the identification of disease-related proteins or sets of proteins and (2) the study of network properties that change in disease. Proteins associated with cancer have, on average, twice as many interaction

partners than other proteins [220]. One can use such information to predict breast-cancer outcomes [452]. In general, 'influential' proteins in PINs (see Section 2.5) tend to be associated with diseases [488].

As PINs are often perturbed in disease, they can potentially also help in drug development [98, 197]. Most medical treatments focus on one component (e.g., a protein) in a malfunctioning molecular pathway [7]. Network-based approaches have the potential to identify drug targets based on protein's position in a PIN [486]. PINs help to predict the interaction between multiple drugs [203], and examining PINs may also help the development of multi-target drugs and combinatorial therapies [97].

Network science provides a large selection of mathematical methods to examine PINs [135, 330]. In this thesis, we focus on the examination of properties in the form of centrality measures and communities in PINs.

Centrality Measures

Centrality measures are designed to identify crucial actors or sets of potentially important actors in a network [52]. These tools are used, for example, to identify *essential* proteins in a PIN [181, 216]. Without these essential proteins an organism is not able to survive. We discuss centrality measures in Section 2.5.

Community Detection

Community-detection methods are designed to find clusters that consist of internally densely connected nodes with sparse external connections [148, 369]. As biological

function is modular [80, 180], such methods can be used to identify sets of proteins with similar functions [262, 363]. We discuss community-detection methods in Section 2.6.

1.2.1 Challenges in Protein Interaction Analysis

Despite the success of PINs and their analysis, it is important to note that they are simplified representations of cells and the complex interactions between proteins [409]. The notion that there is one network of pairwise interactions in cells that organises biological function is a vast oversimplification [475]. Mathematical models that are a more realistic, but still abstract, representation of PINs might help to incorporate data from multiple experimental sources. This could provide additional insight into the biological processes. Different generalisations of networks are potentially suited for biological analysis; we explore some of them in this thesis.

In an organism, there is not just one PIN, because PINs are condition-specific [374]. For example, a PIN may change during disease, over time, or due to an external stimulus [152]. In this thesis, we focus on two settings for condition-specific PINs: PINs that change over time (*temporal PINs*) and *tissue-specific PINs*.

At the moment, only in limited cases is it possible to measure *in vivo* a PIN in different cell types or at different time points. There are, however, ways to construct condition-specific PINs by combining a PIN with data from other experiments, such as gene-expression measurements [84, 309]. Studies that combine different types of biological data are often called *integrative biology*. Such approaches have the potential to counteract some biases or limitations of single experimental techniques [84, 275, 309].

In this thesis, we use two generalisations of ordinary graphs, *multilayer networks* (MLNs) and *node-weighted networks*, to construct and analyse context-specific PINs. We discuss them in Sections 2.7 and 5.2, respectively.

Tissue-specific PINs

A PIN describes interactions that can occur between proteins and other molecules. It is known, however, that the presence of proteins differs between cells and tissue types [54, 400]. 'Housekeeping proteins' are present in all tissues and commonly believed to fulfil essential functions in cells. Tissue-specific proteins are present only in some tissues and are thus expected to have specific functions. One can measure the presence of proteins indirectly by gene expression (see Section 2.1). Thus far, analysis of PINs in combination with tissue-specific gene-expression data has focused on one of two approaches. Most commonly, one examines a PIN independently of the gene expression and compares inights from both data sets (e.g., [279]). Alternatively, one can construct tissue-specific PINs by deleting proteins that are not present in a certain tissue (e.g., [271]). One can also analyse PINs in a tissue-specific way by incorporating a hierarchy of tissue similarity into the analysis (e.g., [506]).

In this thesis, we explore two approaches to examine tissue-specific PINs. In Chapter 3, we use tissue-specific gene-regulatory data to construct two tissue-specific PINs as an MLN. In Chapter 5, we construct tissue-specific PINs by integrating a PIN with tissue-specific gene-expression measurements as node-weighted networks.

Temporal PINs

It is believed that the modular organisation of PINs allows cells to adapt to stimuli and change their behaviour over time [174]. One can try to assess different roles proteins might have in this organisation (e.g., so-called 'party' and 'date' hubs [174]). While there has been discussion as to whether discrete classifications of roles is fruitful [3, 238], the general dynamic nature of PINs has not been disputed. One approach to investigate temporal PINs is to proceed similarly to tissue-specific PINs. First, one deletes proteins that are not present at a certain time point to construct a PIN for this time point [416]. Second, one repeats this for every time point of interest and compares, for example, structural properties of a PIN at different points in time. This approach, however, does not take into account the temporal succession of the events. In Chapter 4, we use MLNs to construct temporal PINs from temporal gene-expression profiles and a time-independent PIN. In such temporal PINs, one can examine the change of an interaction network over time as one mathematical object. Specifically, we investigate centrality of proteins over time and reconfiguration of modular structure in temporal PINs.

1.3 Thesis Overview

In this thesis, we explore three different network generalisations for the investigation of PINs (see schematic presentations in Fig. 1.1). In Chapter 2, we discuss relevant background information about proteins, PINs, and network science. In Chapter 3, we introduce 'promiscuity' as a measure of the variability of a node's importance across

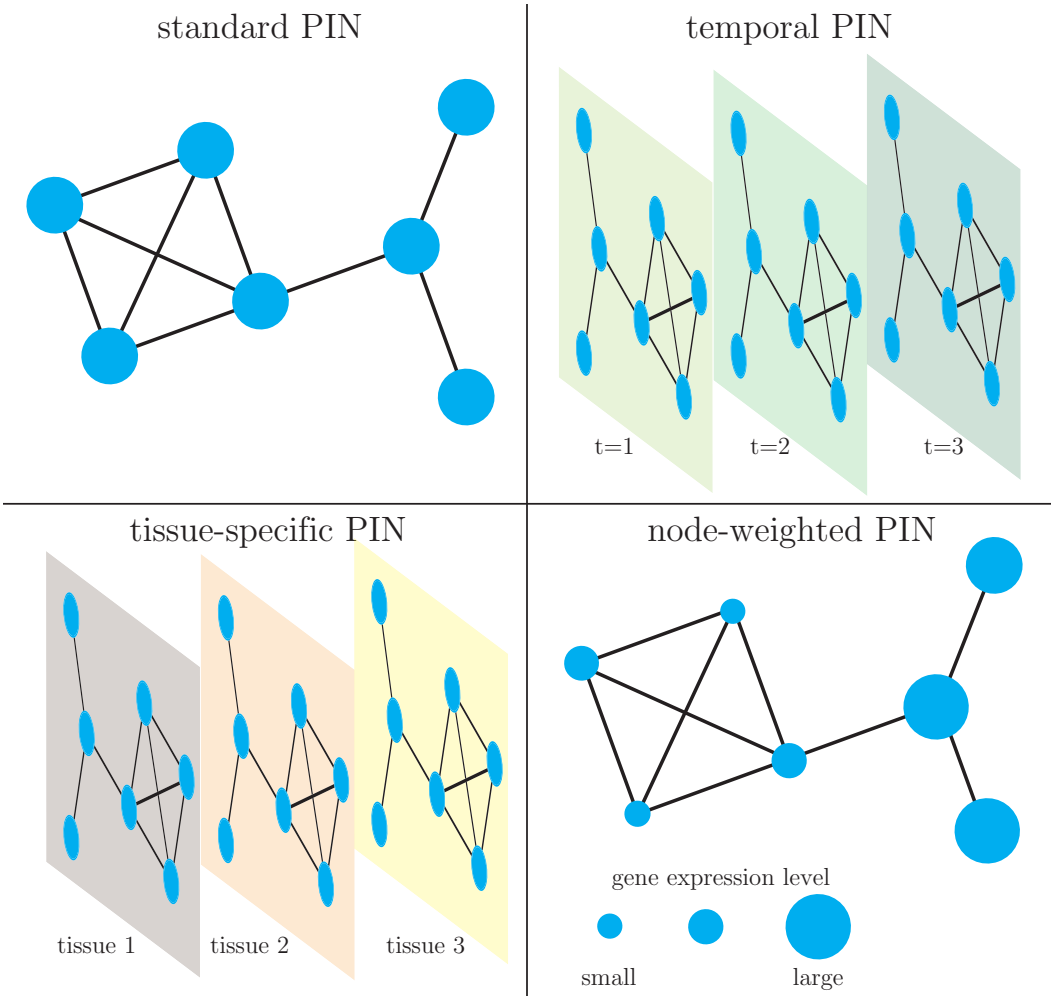


Figure 1.1: Schematic representations of the network generalisations employed in this thesis. We use MLNs to represent temporal and tissue-specific PINs and node-weighted networks to combine a PIN with gene expression data.

layers. We calculate node promiscuities in two tissue-specific protein–DNA interaction networks. In Chapter 4, we demonstrate that one can combine a PIN with temporal gene-expression profiles to construct a temporal PIN. We use an eigenvector-based centrality measure to examine the change of centrality in a temporal PIN of the yeast cell cycle. In Chapter 5, we examine modular structure in node-weighted networks. We then construct tissue-specific PINs as node-weighted networks and detect modular

structure in them. We conclude in Chapter 6.

In Appendix A, we discuss the community-detection algorithm that we employ in this thesis. In Appendix B, we investigate modular structure in a temporal PIN. In Appendix C, we give details about a parameter-identification algorithm that we use in this thesis. In Appendix D, we compare a stochastic block model with the node-weighted modularity that we introduce. In Appendix E, we discuss the construction of some networks that we investigate in this thesis.

2
Foundations

Contents

2.1 Proteins 17
2.1.1 Gene Expression, Protein Biosynthesis, and Protein Abundance 21
2.1.2 Measurement of Gene Expression 25
2.2 Protein Interaction Networks 27
2.2.1 Measurement of Protein–Protein Interactions 29
2.2.2 Genetic Interactions 30
2.2.3 Protein-Interaction Databases 31
2.2.4 Reliability and Coverage of Protein Interaction Networks . 32
2.3 Networks 34
2.4 Synthetic Network Models 40
2.4.1 Erdős–Rényi Model 40
2.4.2 Configuration Model 40
2.4.3 Stochastic Block Model 41
2.4.4 Regular Ring Lattice 42
2.5 Centrality Measures in Networks 42
2.5.1 Eigenvector-based Centralities 44
2.6 Community Detection in Networks 45
2.6.1 Modularity 46
2.7 Multilayer Networks 51
2.7.1 Monolayer Networks 53
2.7.2 Edge-coloured Multigraphs 54
2.7.3 Multiplex Networks 54
2.7.4 Temporal Networks 56
2.7.5 Community Detection in Multilayer Networks 58

2.8 Community Detection in Protein Interaction Networks . 59
2.8.1 Community Detection in Monolayer Protein Interaction Networks 59
2.8.2 Community Detection in Multilayer Protein Interaction Networks 61
2.9 Gene Ontology Enrichment 62
2.9.1 Gene Ontology Data 62
2.9.2 Hypergeometric Test 64

In this chapter, we give background information about the concepts and methods that we use in the thesis. In Section 2.1, we discuss proteins and their importance for cells. In Section 2.2, we describe the construction of PINs. In Section 2.3, we give background information about networks. In Section 2.4, we describe some synthetic network models. In Sections 2.5 and 2.6, we discuss centrality measures and community-detection methods for networks, respectively. In Section 2.7, we describe several multilayer networks. In Section 2.8, we discuss community-detection in PINs. In Section 2.9, we describe how one can compare results from community-detection methods with gene-ontology data.

2.1 Proteins

Proteins are long polymer chains that are present in all living cells and have biological functions [102]. The monomers of proteins are the *amino acids*, of which there are twenty-two types[1]. Amino acids have distinct chemical properties (e.g., the electrical charge of their side chains). Amino acids may attach to another by covalent bonds to form *polypeptides*. Proteins consist of at least one polypeptide. The length of proteins ranges from dozens to ten thousands[2] [444] and their median length in humans is approximately 400 amino acids [62]. The order of the amino acids in these proteins, called *amino acid sequence*, is important for the formation of a protein's three-dimensional structure and largely determines its biological function.

The formation of a protein's three-dimensional structure is called *protein folding* [158]. Proteins fold because it is energetically favorable; a mixture of hydrophobic interactions, intramolecular hydrogen bonds, and van der Waal forces hold the structure together [120]. While a protein's amino acid sequence mainly determines its structure, its environment (e.g., pH value, temperature, and the presence of other molecules) may also influence it [185]. Protein structures are not fixed and a protein may have multiple metastable confirmations that can be crucial for its biological function [196].

For some proteins, one can determine structure experimentally. *X-ray crystallography* is the most successful method of structure determination, but others (e.g.,

[1] Only twenty of them are present in the so-called 'standard genetic code' [17]. The other two, selenocysteine and pyrrolysine, can be synthesised under special circumstances.

[2] The largest known protein in humans is *titin*. It is a structural component of muscular tissue and consists of 30 000 amino acids [343].

nuclear magnetic resonance spectroscopy and cryo-electron microscopy) also exist [419]. As of August 2018, the Protein Data Bank (PDB) stores the experimentally determined structures of 133 464 proteins and 120 564 of them are X-ray structures [38]. Experimental determination of protein structures lags behind the identification of new protein sequences [397]. Researchers therefore attempt to infer a protein's three-dimensional structure from its amino acid sequence. This computational protein structure prediction has improved in the last decades [319].

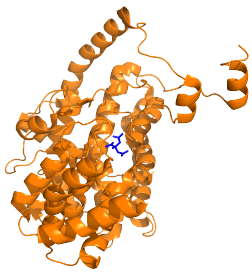
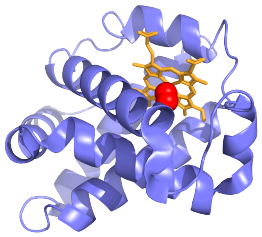
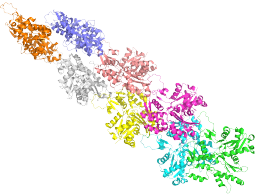
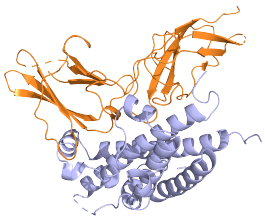
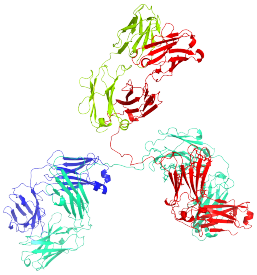
Function	Description	Example		
		Name [PDB ID]	Description	Illustration
Enzymes	catalyse chemical reactions that take place in cells	Citrate synthase [1CTS]	catalyses the first step in citric acid cycle to release energy	
Transport and storage proteins	bind and carry atoms and small molecules within cells and throughout the body	Haemoglobin [1HHO]	transports oxygen in the blood of vertebrates	
Structural components	provide structure and support for cells; on a larger scale, they also allow the body to move	Actin filaments [6BNO]	form the cytoskeleton	
Messenger proteins	transmit signals to coordinate biological processes between different cells, tissues, and organs	Growth hormone (purple) and receptor (orange) [1A22]	stimulate cell growth	
Antibodies	bind to specific foreign particles to help protect the body	Immunoglobulin [1IGT]	neutralise viruses and bacteria.	

Table 2.1: Examples of proteins and their functions. We use PyMol [117] to illustrate them and show the proteins using 'cartoon representations'. We present proteins at different spatial scales.

A protein's three-dimensional structure is important because it determines a protein's ability to interact with other molecules [11]. These interaction partners, called *ligands*, can be one of a variety of biological molecules, including proteins, deoxyribonucleic acid and Ribonucleic acid (DNA and RNA; molecules that carry genetic information), lipids, ions, or small molecules. Such interactions can be stable or transient. Importantly, bindings between proteins and their ligands are very *specific*. A protein binds only to a subset of the molecules that it encounters. Nevertheless, a protein can interact with many different molecules. For example, actin, the most abundant protein in most eukaryotic cells [124], can bind to itself, to the protein profilin, small molecules as latrunculin, and many other molecules. When actin binds to itself, it forms *microfilaments*. These constitute one part of the cytoskeleton, which gives cells their shape, allows cell movement, and is crucial for cell division [19, 204] (see the illustration of actin filaments in Table 2.1.) In Fig. 2.1, we illustrate the interaction between actin and profilin. By binding to actin, profilin regulates the formation of the cytoskeleton [68]. Latrunculin can bind to actin monomers and prevents them from polymerizing [57]. While this binding is toxic for most organisms, in small doses, these molecules may be beneficial as an antimetastatic drug [131]. Interaction between proteins and other molecules form the basis of PINs, which we discuss in more detail in Section 2.2.

A protein's biological function is strongly affected by its ability to interact with other molecules. We can characterise proteins by the type of function(s) they fulfil in

an organism. For example, they can be enzymes, hormones, structural components, transporters, or antibodies. In Table 2.1, we give some examples of the various functions that proteins fulfil in organisms.

Proteins often do not have a single biological function; instead they serve a combination of many different functions. The *chromosomal passenger complex*, for example, regulates many processes during cell division, and its location inside the cell changes over time [69, 156].

While proteins have specific biological functions, many biological functions need multiple proteins to work in concert to coordinate cellular processes [11]. More than 200 proteins, for example, coordinate the microtubule cytoskeleton, which itself is crucial for different biological processes, including cell division and intracellular transport [19]. Interaction between proteins often facilitate such coordination and therefore play a fundamental part in cellular organisation.

2.1.1 Gene Expression, Protein Biosynthesis, and Protein Abundance

DNA molecules hold biological information for the construction of proteins in the form of *genes* [11]. Not all genes, however, can produce proteins. The process by which a cell produces a 'gene product'[3] is called *gene expression*. Cells express genes and produce proteins through a process called *protein biosynthesis* [259]. This process consists of two major stages: *transcription* and *translation*.

[3]Other examples of non-protein gene products are transfer RNA, small nuclear RNA, micro RNA, and silencing RNA.

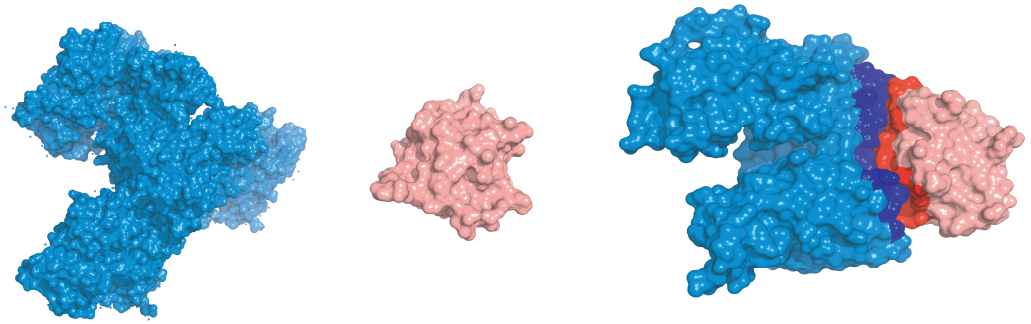


Figure 2.1: Actin–profilin complex in *Bos taurus*. Actin (blue) and profilin (red) form a complex. Profilin regulates the cytoskeleton and actin has a crucial role in its formation [68]. The proteins interact with each other through a combination of ionic, polar, and hydrophobic interactions [405]. We use PyMol [117] for illustration and show the proteins using 'surface representations'. In these, such surface is the manifold that would be traced out by water molecules in contact with a protein at all possible positions. We highlight the interaction interface of the two proteins in darker colours. The structures of actin, profilin, and the actin–profilin complex in *Bos taurus* (cattle) are from the PDB with IDs 2OAN [251], 1PNE [74], and 1HLU [83], respectively.

Transcription

Transcription occurs in the cell nucleus, where DNA and the enzyme RNA polymerase are present. RNA polymerase copies the information from a particular segment of the DNA (i.e., a gene) into a *messenger ribonucleic acid* (mRNA) molecule.

The transcription itself occurs in several steps of which initiation, elongation, and termination are the major ones [11, 26]. During *initiation*, RNA polymerase binds to a specific DNA sequence called *promoter* and locally unwinds the double-stranded DNA by breaking the hydrogen bonds between the strands. Subsequently, the RNA polymerase matches the single-stranded DNA sequence with a complementary sequence of nucleotide triphosphates. This *elongation* phase is the step-wise addition of nucleotides to the RNA chain. The RNA polymerase moves down the DNA double-helix

and rejoins the strands while growing the RNA chain. When the RNA polymerase reaches a *transcription terminator* at the end of a gene, the RNA is released.

An such obtained RNA molecule is called *precursor mRNA* (pre-mRNA). It undergoes *post-transcriptional modifications*[4] to become mRNA. The most notable modification is *splicing*, which is the removal of *introns* (non-coding regions of the RNA), and modifications of both ends of an RNA molecule. For many genes, some *exons* (coding regions of the RNA) might be removed, too. This is called *alternative splicing* and enables a cell to produce different mRNA molecules from a single gene [366]. In eukaryotes, most mRNAs exit the nucleus through the nuclear pores into the cytoplasm, where the protein synthesis continues.

Transcription factors (TFs) are proteins that up- or down-regulate the rate of transcription of specific genes by alternating the rate of any of these sub steps. They regulate by binding to specific parts of the DNA [252]. This allows cells to react to external stimuli, such as heat [317] or hormones [413]. Another regulation of gene expression is DNA methylation, which is the addition of a methyl group to a DNA segment [219]. It is commonly called 'gene silencing', and the organisation of methylation is crucial for embryonic development [73].

Translation

Translation is the synthesis of proteins from mRNA molecules. It occurs in the cytoplasm, where ribosomes (which are complexes of proteins and structural RNAs), are

[4]Some of these modifications are also co-transcriptional, i.e., they may also occur during transcription.

present. The ribosome decodes the mRNA to produce a polypeptide chain that becomes the protein. An mRNA's nucleotide sequence determines the amino acid sequence, where triplets of nucleotides (*codons*) encode amino acids. Similar to transcription, translation consists of an initiation, elongation, and termination phase[5]. These referer to the assembly of a ribosome around a specific start codon, the stepwise assembly of a polypeptide, and its release, respectively. *Translational control* is the regulation of these steps by altering their rate, where initiation is likely the limiting one [188]. Cellular levels of *initiation factors* affect the rate of this first step of translation [213]. The localisation of specific mRNA molecules to a particular subcellular regions can also influence translation rates [250]. Other translation steps can also be affected, so-called *release factors*, for example, catalyse translational termination [119].

Post-translational modifications modify polypeptides and potentially alter the biological function(s) of a final gene product. Among them are *phosphorylation* (i.e., the attachment of a phosphoryl group to the side chain of an amino acid), the linkage to other proteins (e.g., *ubiquitylation*), or *protein splicing* (i.e, removal of protein segments).

Protein Localisation & Degradation

The cellular levels of proteins are not exclusively determined by the rate of synthesis (i.e., the combined rate of transcription and translation) but also by the rate of *degradation* [91]. In this process, proteins are dismantled (e.g., by the organelle

[5]The release of a ribosome from an mRNA can be seen as a additional step called *ribosome recycling*.

lysosome or the protein proteasome) into their constituent amino acids. This removes damaged proteins and allows the cell to regulate other molecular pathways, for example, as response to external stimuli. The rate of degradation depends, among others, on the protein, the state of an organism, and external effects.

Eukaryotic cells are divided into *subcellular compartments* (e.g., nucleolus, mitochondria, endoplasmic reticulum). *Protein localisation* refers to the accumulation of a certain protein at a given site or compartment. The subcellular localisation of proteins allows them to fulfil their biological function in these compartments. To achieve this, proteins are translocated across biological membranes, for example, with the help of so-called *chaperone* proteins[6] [478].

In this thesis, we focus on the abundance of mRNA molecules as an indicator of the 'activity' of a given protein. As discussed in this section, this is an simplistic model of the underlying biology, as the functional proteins might have been altered by post-translational modifications, might be have been translocated, or might degrade quickly. A protein with a long half-life, for example, can have a high abundance, despite a small expression at a given point in time. A more complete (but also more complex) model of cellular biology would consider such effects.

2.1.2 Measurement of Gene Expression

As the regulation of gene expression is a fundamental mechanism in the organisation of cellular function, measuring gene expression across cells, tissues, space, and time

[6]Not all chaperone proteins are involved with protein transport. In general, molecular chaperones are proteins that influence the folding of proteins [178].

is crucial for understanding multicellular organisms. The gene products — proteins, most prominently — perform cellular functions, so measuring gene expression can reveal what functions are performed in a cell.

There exist techniques to directly measure the abundance of proteins in tissues; these include mass spectrometric approaches [164] and protein arrays [300]. The measurement of mRNA abundance, however, is easier and more affordable; and it is thus often preferable. The Pearson correlation between mRNA abundance and protein abundance varies from 0.4 to 0.8 [164, 406].

In this thesis, we use exclusively mRNA-abundance data. Different experimental techniques exist to measure the expression of genes in form of mRNA abundance. For many years, *Northern blotting* was the standard experimental technique for the detection and quantification of specific RNA levels [16].[7] In the last decades, however, microarrays and RNA sequencing became more widely used.

Microarrays

A *microarray* is a collection of single-stranded DNA probes that are bound to a glass slide and are complimentary to mRNA molecules in the sample [358]. These complimentary mRNA molecules anneal the single-stranded DNA probes by forming hydrogen bonds. One then washes unbound mRNA off and can visualise the amount of bound RNA to each probe, for example, if they are fluorescence labelled and activated by a laser. As each probe is specific for a given gene, one may obtain an estimate of mRNA

[7] It is named in analogy to Southern blotting, a method for the detection of specific DNA sequences [426].

levels in the sample[8]. As microarrays allow the measurement of expression levels of thousands of genes such data can also be used to construct gene-coexpression networks.

RNA Sequencing

Advances in massively parallel sequencing allowed the development of *RNA Sequencing* [292]. In these experiments, one breaks mRNA molecules into small fragments and sequences them. One obtains quantitative expression levels for genes by aligning them to the genome.

2.2 Protein Interaction Networks

As discussed in Section 2.1, proteins interact with other molecules and these interactions fulfil specific biological functions. One important type of interaction are those between proteins. We understand these *protein–protein interactions* (PPIs) as physical contacts with molecular docking between proteins that occur in a cell. This docking is the formation of a joint molecular interface that connects two or more proteins. Proteins can bind to other proteins in various ways [11]. The three main interaction types are surface–string interaction, helix–helix interaction, and surface–surface interaction. The last is the most common and occurs if the surfaces of two proteins match precisely. Such interactions can be very strong, because a large number of weak interactions form between the two binding proteins. Most often, a combination of ionic, polar, and hydrophobic interactions holds the proteins together [493]. In Fig. 2.1, we show

[8]Further analysis steps (e.g., normalisation of the raw signal) are often necessary.

actin and profilin, which form a complex through surface–surface interaction. One can classify PPIs depending on their strength, the interaction partner, and whether they are *obligate* [339].

One often distinguishes multiprotein complexes into obligate and non-obligate complexes [339]. In the former, the interacting proteins are not found on their own in the cell. In the latter, a complex' constituent proteins can also be present individually.

PPIs occur between either identical or distinct proteins. One calls the former *homooligomers* and the latter *heterooligomers*. The actin filament in Table 2.1 consists exclusively of actin proteins and is thus a homooligomer. The actin–profilin complex in Fig. 2.1 is a heterooligomer because it consists of two different proteins.

One can attempt to classify PPIs depending on their lifetime into *permanent* and *transient* interactions [339, 417]. The former creates stable complexes that consist of multiple proteins and fulfil specific biological function(s). Enzymes that consist of multiple subunits, each of which is a protein, often form stable interactions. Transient interactions are reversible and often short-lived with lifetimes of seconds [364]. Many proteins undergo conformational changes (i.e., a change in three-dimensional structure) upon transient interaction [364]. This change may allow an interaction with another ligand, and it thus triggers a *signalling cascade*. One example is the transient interaction of G-proteins[9] with membrane-bound G-protein-coupled receptors. This interaction activates a target in the plasma membrane and it thus causes a

[9]G-proteins are so-called because they bind the guanine nucleotides **G**DP and **G**TP.

cascade of other signalling events.

Many PPIs do not belong to a distinct type [339]. Rather, a continuum exists between non-obligate and obligate and between transient and stable interactions. Physiological conditions and external influences (e.g., temperature and the presence of other proteins) can change the stability of these complexes [351]. Additionally, drugs can influence an organism by inhibiting the interaction between proteins [25, 312].

2.2.1 Measurement of Protein–Protein Interactions

Many different experimental techniques can detect interactions between proteins [337, 417]. High-throughput techniques allow the screening of a large number of proteins at the same time, whereas low-throughput techniques only screen pairs of proteins. We also distinguish between *in vivo* and *in vitro* methods. The former are methods that measure PPIs in a living organism. The latter are methods that measure the PPIs outside of their normal biological context.

PPIs can be experimentally validated either biochemically (including affinity purification technology [373, 383], x-ray crystallography [224, 236], far-Western blotting [487], pull down [281], enzymatic [421]), via microscopy (fluorescence resonance energy transfer [458], or imaging [277]), or *in vivo* (yeast two-hybrid screens [145], protein complementation assay [354], genetic interactions). In this thesis, we use data from the BioGRID database [77, 432], which is an aggregation of many different of such experimental methods (see Fig. 2.2). In general, we distinguish between methods that detect actual physical interaction or proximity between proteins and methods

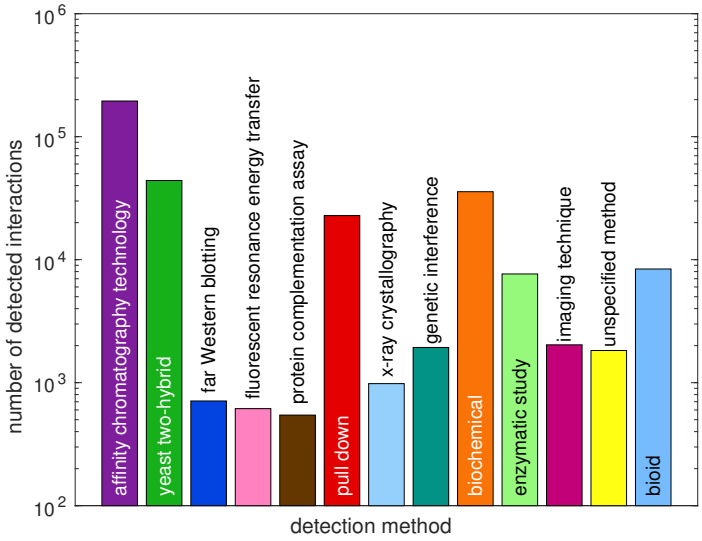


Figure 2.2: Experimental methods in the BioGRID **database.** We show the number of PPIs in the BioGRID database for thirteen experimental methods. Affinity chromatography technology, which includes TAP, detects most PPIs.

that infer them indirectly.

2.2.2 Genetic Interactions

One can measure *genetic interactions*, which are functional relationships between genes or their products (e.g., proteins) [94, 289]. These interactions can be PPI but, for example, also protein–DNA and RNA–DNA interactions [378]. Genetic approaches, however, cannot distinguish between direct and indirect interactions. An experiment might indicate an interaction between two genes that is mediated by some other protein.

To examine genetic interactions, one usually creates *double mutants*. In these cells, two genes are disabled due to synthetic mutations. One then investigates the effect of these mutations. An extreme negative genetic interaction occurs when two mutations, neither lethal individually, cause cell death when occuring together. There

exist several experimental technqiues (most notably *synthetic genetic arrays* [342, 455]) to measure such interactions. The detection of genetic interactions was first developed for yeast but has been used in other organisms [93]. The application to human cells, however, is challenging due to the enormous number of combinations ($\sim$ 200 million) and the rarity of strong genetic interactions [198, 254].

2.2.3 Protein-Interaction Databases

As different experimental techniques for detecting PPIs have different biases, it is important to integrate results from many different types of experiments [454]. Such information is available from protein-interaction databases. There exist many different databases of PPIs. Among the most popular are the Biological General Repository for Interaction Datasets BioGRID [432], the Human Protein Reference Database HPRD [230], the Homologous Interactions HINT database [103], and the Search Tool for the Retrieval of Interacting Genes/Proteins STRING [449]. For reviews of some databases, see [376, 417]. As we show in Table 2.2, the databases vary in the number N of proteins and the number M of pairwise interactions.

In this thesis, we use data exclusively from the BioGRID database [77, 432]. It is freely available, regularly updated, and has interaction data for 66 organisms. The interactions are an aggregation from publications that describe low-throughput experiments. As of August 2018, it has interactions from 65 816 publications. Experts curate publications to identify PPIs and add them to the database, which has both, genetic and physical PPIs. We use BioGRID because it is the largest available PPI

database	version	number N of proteins	number M of interactions	reference
BioGRID*	3.4.163	23 098	334 684	[432]
HPRD	Release 9	30 047	41 327	[230]
HIPPIE	2.1	16 792	340 629	[81]
STRING *	10.5	18 838	11 353 057	[450]
HINT *	4	16 221	170 804	[103]

Table 2.2: Size of PPI databases. We give the sizes of some PPI databases in the most recent versions (status as of August 2018). We indicate multiorganism databases with an asterisk (*) and for them, we give only the number N of human proteins and the number M of interactions between them.

database with only experimental interactions.[10]

2.2.4 Reliability and Coverage of Protein Interaction Networks

The detection of PPIs with high-throughput techniques such as Y2H assays or TAP is not always reliable [115, 202]. We distinguish two types of errors: detecting interactions that are not present (false positives) and missing existing interactions (false negatives). In 2006, Hart et al. estimated that only half of all human PPIs are known [177]. Huang and Bader (2008) estimated the false-positive rate of Y2H screens for yeast to be 9.9 % and the false-negative rate to be 51 %.

Another way to create large-scale PINs is to curate published literature, taking advantage of the fact that low-throughput techniques that individual papers describe are more reliable [384]. The error rates for such literature-curated protein interaction data sets are often smaller than large-scale experiments. The estimates of the error rates range from 2 % to 35 % [99, 394].

[10] The STRING database is larger but includes computational predicted interactions (e.g., based on similar genomic context).

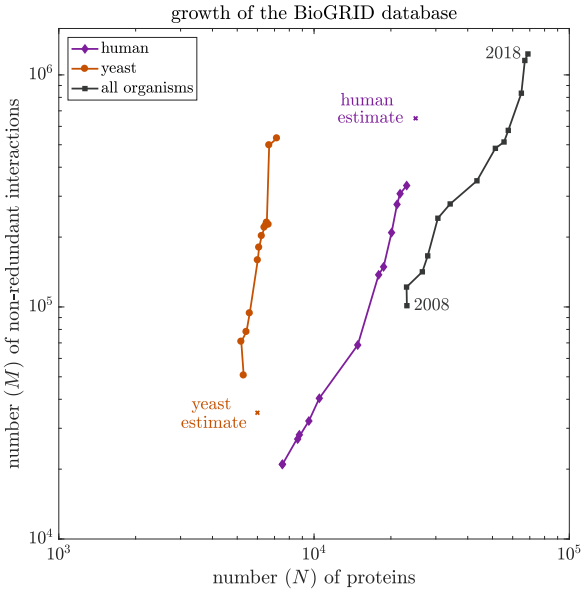


Figure 2.3: Growth of the BioGRID database. We show the number N of unique proteins and the number M of non-redundant interactions in the BioGRID database for *Homo sapiens* (human; purple diamonds), *Saccharomyces cerevisiae* (yeast; orange disks), and all organisms (black squares). Each marker indicates the number in July each year from 2008 to 2018. While the databases tend to increase in size, there are exceptions. We also indicate estimated interactome sizes for human and yeast as squares [443]. For yeast, note that the known interactome exceeds its estimate, whereas for human the current data is smaller than its estimate.

In recent years, the reliability and coverage of PINs has improved [81, 450]. The BioGRID database, for example, has experienced an increase in the number of curated interactions by 30 % from 2015 to 2017 [77] (see Fig. 2.3). The PINs are, however, most likely not complete. An estimate of the size of the human PIN is approximately 25 000 proteins with 700 000 interactions [443], and the most recent BioGRID database has only 20 914 proteins and 365 547 interactions.

Some databases, such as STRING [450] and HIPPIE [81], quantify the reliability of an interaction with a confidence score. One can use these scores to threshold PINs by keeping only PPIs that have at least a certain reliability level [55]. This approach

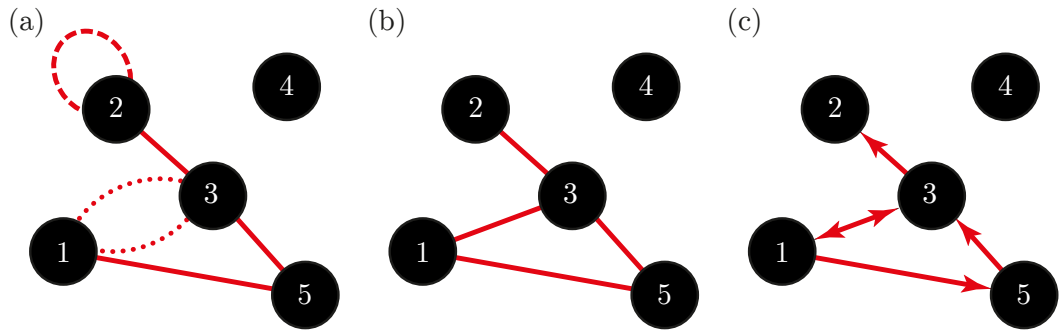


Figure 2.4: A network consisting of $N = 5$ nodes. (a) This graph consists of $N = 5$ nodes (black disks) that are connected to each other with $M' = 6$ edges (red lines). We show self-loops as dashed arc and multiedges as dotted arcs. (b) A simple graph has no self-loops and no multiedges. The number of edges is $M = 4$. The neighbourhood of node 1 is $\mathcal{N}(1) = \{3, 5\}$. (c) In a directed graph, edges have a direction. The graph in (b) is the underlying undirected graph of (c).

tends to decrease the false-positive rate but increase the false-negative rate. These scores also tend not to be comparable across databases [55].

In this thesis, we construct PINs from the literature-curated database BioGRID. We refrain from filtering these networks so that we do not further increase the false-negative rates. When interpreting our results, however, it is important to be aware that the networks are error-prone and do not cover all proteins.

2.3 Networks

One often represents PPI data as a network. In this representation the nodes represent proteins and the edges between nodes indicate an interaction between them. We can use methods from network science, which we discuss in this section, to analyse PINs.

A *graph* is an ordered pair $G = (V, E)$ composed of a set V of nodes (i.e., vertices) that are connected pairwise via edges (i.e., links) from the set $E \subset V \times V$ [4, 330].

A node and an edge are called *incident*, if the node is one of the two nodes the edge connects. The size of the graphs is given by the number of nodes $N = |V|$. The number of edges is $M = |E|$. Nodes that are adjacent to each other via an edge are called *neighbours*, and the set of all neighbours of a node i is its *neighbourhood* $\mathcal{N}(i)$. A graph without nodes that are adjacent to themselves (called 'self-loops'), and without multiple connections between a pair of nodes (called 'multiedges') is called a *simple graph*. The most common linear-algebraic representation of a graph is the *adjacency matrix* $\mathbf{A}$,which is a binary $N \times N$ matrix with elements

$$A_{ij} = \begin{cases} 1\,, & \text{if nodes } i \text{ and } j \text{ are adjacent}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{2.1}$$

The adjacency matrix for the simple graph in Fig. 2.4b is the 5×5 matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ \mathbf{1} & \mathbf{1} & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & 0 & \mathbf{1} & 0 & 0 \end{pmatrix}.$$

The edge list for the same network is

$$E = \{(1,3),(1,5),(2,3),(3,5)\}\,.$$

A network's edges can also have a weight $w_{ij} \in \mathbb{R}$. In this case, the entry A_{ij} is the weight w_{ij} of edge (i,j). As an edge connects two nodes without a sense of direction, the matrix is symmetric (i.e., $A_{ij} = A_{ji}$).

In an undirected network, the *degree* k_i of a node i is the number of edges incident to i. One calculates the degree from the adjacency matrix with the sum

$$k_i = \sum_{j=1}^{N} A_{ij} \,. \tag{2.2}$$

When summing the degrees of all nodes, we obtain twice the number of edges, because each edge is connected to exactly two nodes:

$$\sum_{i \in V} k_i = 2M \,. \tag{2.3}$$

The mean degree is then

$$\langle k \rangle = \frac{\sum_{i \in V} k_i}{N} = \frac{2M}{N} \,. \tag{2.4}$$

To compare the number M of edges present in a network with the maximum possible number, we compute the *density* of a network:

$$\rho = \frac{2M}{N(N-1)} \in [0, 1] \,, \tag{2.5}$$

where $N(N-1)/2$ is the maximum number of undirected edges in a network with N nodes.

In Fig. 2.5, we show the degree distribution of a human PIN. As in many real-world networks, it has a broad degree distribution (i.e., many nodes have a degree much higher than the mean degree).

node i	1	2	3	4	5	$\sum_{i \in V} k_i$	$\langle k_i \rangle$
degree k_i	2	1	3	0	2	8	8/5

Table 2.3: Node degrees for network shown in Fig. 2.4b. The sum over all node degrees is equal to twice the number of edges of the graph. The mean degree is $\langle k \rangle = 8/5$.

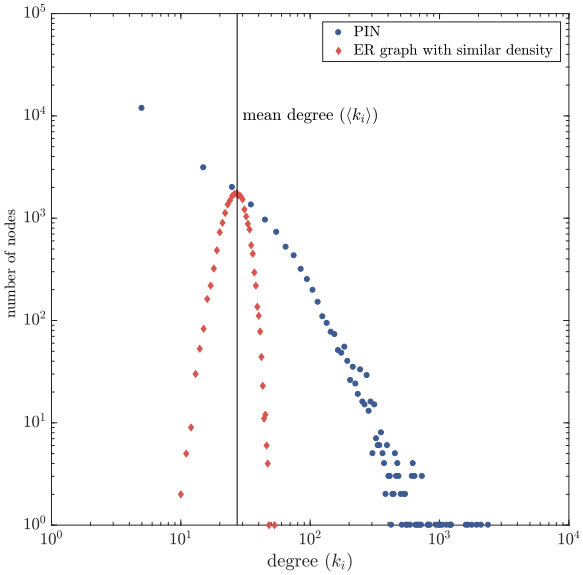


Figure 2.5: Degree distribution of a human protein interaction network and an Erdős–Rényi network. We show the degree distribution of a PIN as blue disks and the degree distribution of an ER graph (see Section 2.4) as red diamonds. We constructed the PIN from the BioGRID database (version 3.4.163) [432]. There are 5 322 nodes with degree $k_i = 1$. The protein ubiquitin has the largest degree ($k_i = 2\,387$) and thus is adjacent to almost 10 % of all proteins. We construct the ER graph using the $G(N, p)$ model (see Section 2.4). Its connection probability p is the density $\rho_{\text{PIN}} \approx 0.0012$ of the PIN. The vertical line indicates the mean degree $\langle k_i \rangle$ of the PIN, which is approximately the same for the ER graph.

Directed Networks

In a *directed network*, the edges have a direction (i.e., the edge (i, j) indicates the presence of an edge only from i to j). The adjacency matrix for the directed graph shown in Fig. 2.4c is the asymmetric 5×5 matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \end{pmatrix}.$$

In a directed network, a node i has an in-degree $k_i^{(\text{in})}$ and an out-degree $k_i^{(\text{out})}$.

These are, respectively, the number of inbound and outbound edges:

$$k_i^{(\text{in})} = \sum_{j=1}^{n} A_{ij}\,, \quad \text{and} \quad k_i^{(\text{out})} = \sum_{j=1}^{n} A_{ji}\,. \tag{2.6}$$

For a directed graph $G = (V, E)$, we can construct an *underlying graph* $\tilde{G} = (V, \tilde{E})$, which is an undirected graph with $(i, j) \in E \to (i, j) \in \tilde{E}$. In Fig. 2.4b, we show the underlying graph of Fig. 2.4c.

Subgraphs

The graph $G' = (V, E)$ is a *subgraph* of the graph $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. We then write $G' \subseteq G$.

Paths, Components, and Geodesic Distance

A *walk* in a graph $G = (V, E)$ is an alternating sequence

$$(i_0, e_1, i_1, e_2, \ldots, e_l, i_l)\,, \tag{2.7}$$

of nodes and edges that begins and ends with a node. For each $l' \in \{1, 2, \ldots, l\}$, the edge $e_{l'}$ connects the nodes $i_{l'-1}$ and $i_{l'}$. Node i_0 is called the *initial node*, and node i_l is called the *final node*. The number $l \in \mathbb{N}$ is the length of the walk. A *path* in G is a walk in which all its nodes are distinct.

node i	1	2	3	4	5	$\sum_{i \in V} k_i$
in-degree $k_i^{(\text{in})}$	1	1	2	0	1	5
out-degree $k_i^{(\text{out})}$	2	0	2	0	1	5

Table 2.4: Nodes degrees for the directed network in Fig. 2.4c. The sum over all node in-degrees is the same as the sum over all node out-degrees.

An undirected graph $G = (V, E)$ is *connected* if for every pair of distinct nodes $i,\ j \in V$, there exists a path from i to j.

A *connected component* H of an undirected graph G is a subgraph $H \subseteq G$ that is connected. Each graph G has a unique collection of connected components $H_1 \ldots, H_c$. The number c of connected components is uniquely determined by G.

In directed graphs, we distinguish between *weak* and *strong* connectedness. A directed graph is *weakly connected* if its underlying graph $\tilde{G}$ is connected. A directed graph is *strongly connected* if for every pair of distinct nodes $i,\ j \in V$, there exists a path from i to j.

A *weakly (strongly) connected component* of a directed graph G is a subgraph $H \subseteq G$ that is weakly (strongly) connected. In Fig. 2.4c, we show a directed graph with two weakly connected components and three strongly connected components.

The *geodesic distance* $d(i, j)$ from node i to node j is the length of a shortest path from i to j. In directed graphs, the distance $d(i, j)$ does not necessarily coincide with $d(j, i)$. In the directed graph in Fig. 2.4c, for example, $d(3, 5) = 2 \neq d(5, 3) = 1$. The *diameter* D of a network is the length of the longest shortest path between all pairs of nodes. That is

$$D = \max_i \max_j d(i, j)\,. \tag{2.8}$$

2.4 Synthetic Network Models

In addition to the investigation of networks that one constructs from data, it can be beneficial to study synthetic networks, as this allows testing of hypotheses for them. There exists a large number of such *generative models* [330].

2.4.1 Erdős–Rényi Model

One of the most simple network models is the *Erdős–Rényi* (ER) model [11]. There exist two variants, the $G(N, p)$ model and the $G(N, M)$ model, both creating networks of N nodes. The $G(N, p)$ model produces networks in which there is an edge between each distinct pair of nodes with independent probability p, whereas the $G(N, M)$ model yields a network with exactly M edges, chosen uniformly at random from the collection of all graphs which have N nodes and M edges. In the former, the number of edges is a random variable with expected value $\binom{N}{2}p$. The distribution of the degree k_i of node i is binomial

$$P(k_i = k) = \binom{N-1}{k} p^k (1-p)^{N-1} \,. \tag{2.9}$$

2.4.2 Configuration Model

The binomial degree distribution of ER graphs is atypical of empirical networks. For example, real networks often have heavy-tailed distributions [424] (see, for example, degree distribution of a human PIN in Fig. 2.5). Using a *configuration model*, we

[11] It is also sometimes called a 'random network', which is ambiguous because any network model that is not entirely deterministic is *random* by definition.

construct a random network with N nodes and a given degree sequence $\vec{k} = \{k_i\}$, where k_i is the degree of node i [151]. There are two primary methods to create networks with this configuration model: randomisation and stub matching [226].

2.4.3 Stochastic Block Model

A *stochastic block model* (SBM) is a random graph model that can have communities or other mesoscale structures [1, 476]. It produces graphs with N nodes, each of which belongs to exactly one of r communities. We denote this community assignment by $g_i \in \{1, \ldots, r\}$. The connection probability between each pair (i, j) of nodes is $p_{ij} = \Omega_{g_i, g_j}$. The $r \times r$ matrix Ω is called the *edge-probability matrix* and is symmetric for undirected graphs. Despite its simplicity, an SBM can create structures that are similar to empirical networks. This includes modular [116], core–periphery [260, 386], and multipartite structures [328].

Various generalisations of the SBM exist, including a variant with nodes that belong to multiple modules [9], ones with nodes that have labels [183, 200, 331], ones with weighted edges [8]. Another extension is networks with nodes that are embedded in a continuous latent space (specifically, a metric space) and connect them with a probability that depends on their distance [192, 333, 375].

2.4.4 Regular Ring Lattice

The *regular ring lattice* is a *regular* network, i.e., all nodes have the same degree. The adjacency matrix of a regular ring lattice with N nodes with degree $k \in [2, N-1]$ is

$$\mathbf{A}_{ij} = \begin{cases} 1\,, & \text{if } |i-j| \mod \left(N-1-\frac{K}{2}\right) \leq \frac{K}{2}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{2.10}$$

2.5 Centrality Measures in Networks

In many networks, it is desirable to quantify the importance of nodes [330].[12] There are many possible notions of importance in a network and, accordingly, many definitions of centrality measures [44]. In this section, we discuss some of them and focus on *eigenvector-based centralities* in Subsection 2.5.1.

The simplest centrality measure is the degree k_i of node i. As we discussed in Section 2.3, this is the number of edges incident to node i. Although the degree is a simple measure of centrality, it can give valuable insights. In PINs, for example, nodes with large degrees tend to represent proteins that are essential for an organism's survival [133, 216].

Another centrality measure is the *closeness centrality*, which measures the mean distance from a node to other nodes. Suppose that d_{st} is the length of a geodesic path from node s to node t. One version of closeness centrality C_i of node i is then

[12] There exist also measures of centrality for edges or other substructures (e.g. [113]). We omit them here and mean by centrality exclusively the centrality of nodes.

$$C_i = \frac{N}{\sum_{j \neq i} d_{ij}} \,. \tag{2.11}$$

Geodesic betweenness centrality measures the extent to which a node lies on shortest paths between other nodes in a network. The geodesic betweenness of node i is

$$B_i = \sum_{s \neq i} \sum_{t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \,, \tag{2.12}$$

where σ_{st} is the total number of geodesic paths from source node s to target node t and $\sigma_{st}(i)$ is the number of those paths that pass through node i. Joy et al. (2005) claimed that in a PIN of yeast, nodes with large geodesic betweennesses tend to be evolutionary preserved[13] [223]. Hwang et al. (2008) used geodesic betweenness to identify genes associated with asthma [206].

This list of centrality measures is very far from exhaustive, and researchers have defined many centrality measures for numerous applications [330]. For investigations of PINs, this includes measures that take into account gene-expression information [308]. A type of subgraph centrality is reported to be correlated positively with lethality in PINs [139]. Despite the evidence of correlations between various types of centrality and 'essentiality' in PINs, the prediction of whether a protein is essential or not in a PIN has high error rates [355]. One reason for this may be insufficient PIN data, as discussed in Subsection 2.2.4.

[13] In the investigation of genetic data, a 'conserved sequence' is a sequence that occurs similarly in different species. As such sequences are altered only lightly by evolution, it is typically assumed that they have a special importance [437].

centrality	centrality matrix $\mathbf{C}$	reference
eigenvector centrality	$\mathbf{A}$	[51]
hub	$\mathbf{A}\mathbf{A}^\intercal$	[233]
authority	$\mathbf{A}^\intercal\mathbf{A}$	[233]
PageRank	$(1-\psi)\mathbf{A} + (\psi/N)\mathbf{J}_N$	[348]

Table 2.5: Examples of eigenvector-based centralities and their respective centrality matrices as a function of the adjacency matrix A. PageRank, in its most common form, has a free parameter ψ, which is the probability of not following a direct edge but instead entering any node at random with uniform probability. The matrix $\mathbf{J}_N = \mathbf{1}^\intercal\mathbf{1}$ is the $N \times N$ matrix with all elements equal to 1.

2.5.1 Eigenvector-based Centralities

One class of centrality measures are *eigenvector-based centralities* [50, 51, 160]. For these, the nodes' centralities are given by the entries of the dominant eigenvector of a *centrality matrix*.[14] A network's centrality matrix $\mathbf{C}$ can be one of various functions of the networks adjacency matrix $\mathbf{A}$ [138]. We give some examples of such eigenvector-based centralities and their respective centrality matrices in Table 2.5.

For a network, an eigenvector-based centrality of node i is the ith element of the leading eigenvector $\mathbf{x}$ of the centrality matrix $\mathbf{C}$. This is given by the equation

$$\mathbf{C}\mathbf{x} = \lambda\mathbf{x}\,, \tag{2.13}$$

where λ is the largest eigenvalue of $\mathbf{C}$. Different numerical methods enable an efficient calculation of eigenvalues. Among them are *power iteration* and *implicit restarted Arnoldi methods* [307, 425]. The latter is used by the MATLAB function `eigs`, which returns the largest eigenvector.[15]

[14] There exist also centrality measures that involve the calculation of eigenvectors but the centralities are not elements in the eigenvector themeselves, e.g. Katz centrality [227]. In this thesis, we call 'eigenvector-based centralities' only those that eigenvectors of a centrality matrix.

[15] Computing only the largest eigenvector is considerably faster than computing all eigenvectors

Among the most prominent of these eigenvector-based centrality measures are hub and authority scores [233] and PageRank [348]. In PINs, eigenvector-based centralities can be used for studying proteins that are associated with diseases [347].

2.6 Community Detection in Networks

Thus far, we have examined networks either from a local or a global perspective. The former is the calculation of information concerning the nodes (e.g., centralities of nodes). The latter is the investigation of properties concerning a whole network, such as the diameter of a network, degree distributions, or the relationship between degree and betweenness centrality). Another perspective is to examine so-called 'mesoscale structures'. These are structures in a network that lie at levels between local and global scales [64]. Examples of such mesoscale structures are communities, core–periphery structures [53, 387], and multipartite organisation [137]. In PINs, mesoscale structures may coincide with functional modules or multiprotein complexes [428]. Investigating mesoscale structures can thus provide insights into the function of PINs.

A particular type of mesoscale feature are *communities*. Intuitively, communities are sets of nodes that are more densely connected among each other than to communities of other nodes [148, 150, 369]. The precise definition of a community, however, can depend on the network, the application, and the detection method [314].

Community detection is the task of finding communities in a network [148, 150, 369]. Many different methods have been developed to detect such communities. One

with the `eig` function.

of the oldest is Min Bisection, which is the division of nodes into two equal-sized groups, with as few edges as possible between them. In this thesis, we use *modularity maximisation* and discuss it in detail in Subsections 2.6.1. Other methods include maximum-likelihood estimation of SBMs [9, 116, 361], information-theoretical approaches [391], iterative removal of edges [159], random walks on networks [169], entries of the *Fiedler vector*[16] [328], and *communicability*-based clustering [134]. For an extensive (but not exhaustive) comparative review of different methods, see [248].

The detection of communities in real-world networks has led to insights in a wide variety of disciplines. These include social networks, materials, mobility patterns, and political interaction and voting [148, 150, 369]. In Section 2.8, we discuss the application of community-detection methods to PINs.

In this thesis, we exclusively consider non-overlapping communities (i.e., each node belongs to exactly one community). There also exist methods to detect overlapping communities, in which a node can belong to multiple communities [9, 150, 350, 369].

2.6.1 Modularity

We can use *modularity maximisation* to partition the set of nodes of a network into communities. A *partition* $\mathcal{G} = \{G^{(1)}, \ldots, G^{(n_{\text{com}})}\}$ is a collection of n_{com} sets such that $V = \bigcup_{s=1}^{n_{\text{com}}} G^{(s)}$ and $G^{(s)} \cap G^{(t)} = \emptyset$ for all s and t. That is, each node belongs to exactly one set G_s. We call these sets *communities*[17]. For convenience, we denote the group

[16]The Fiedler vector is the eigenvector associated with the second-smallest eigenvalue of a graph's combinatorial *Laplacian matrix* $\mathbf{L}$, which has elements $L_{ij} = k_i\delta_{i,j} - A_{ij}$ [143].

[17]Another common name is *module*. In this thesis, we use 'module' for a functional set (e.g., groups of proteins with common biological functions).

assignment of node i with g_i in the vector $\mathbf{g} \in \mathbb{N}^N$. We usually label the communities with integers $\{1, 2, \ldots, n_{\text{com}}\}$, where n_{com} is the number of communities. As with most clustering algorithms, however, the labelling of the clusters is arbitrary; and switching the labels of two clusters does not change a partition [435]. Usually, we do not know the number n_{com} of communities; for the modularity-based approach that we use, we do not specify it before applying a community-detection algorithm.[18]

Modularity

$$Q(\mathbf{g}) = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij})\, \delta(g_i, g_j)\,, \tag{2.14}$$

is a quality index for a partition $\mathbf{g}$ of a network with adjacency matrix $\mathbf{A}$ into communities. It is the sum of the weight of intracommunity edges minus the expected weight of intracommunity edges under a random *null model*. The Kronecker-delta $\delta(g_i, g_j)$ is 1 if nodes i and j belong to the same module ($g_i = g_j$) and 0 otherwise.

The element P_{ij} of the null-model matrix $\mathbf{P}$ is the expected connection strength between nodes i and j under a null model. There are numerous choices for this null model. There exist null models for networks with spatial embeddedness or that are otherwise influenced by space [141, 396] or correlation networks [34, 285], and for many other situations [150]. We use the widely-used Newman–Girvan null model $P_{ij} = k_i k_j/(2m)$ [332]. This is the expected connection strength if one rewires edges uniformly at random while preserving the expected strength distribution. In this

[18] Some algorithms, however, need the number n_{com} of communities as an input. This includes k-means clustering of a communicability matrix, maximum-likelihood estimation of a stochastic block model, and many others [150].

case, modularity Q is

$$Q(\mathbf{g}) = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(g_i, g_j) \,. \tag{2.15}$$

One of the shortcomings of modularity is that it fails to detect communities smaller than a certain size [149]. To examine communities at different scales, one can define a multi-resolution variant of modularity [380]

$$Q(\mathbf{g}, \gamma) = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(g_i, g_j) \,. \tag{2.16}$$

The resolution parameter γ allows the detection of modular organisation at multiple scales. For large values of γ, one finds many small modules, by contrast, communities increase in size as $\gamma \to 0$. In Fig. 2.6, we illustrate four modularity-based partitions of a PIN of the *human immunodeficiency virus* (HIV).

Thus far, we have explained modularity (see Eqn. 2.16), which gives one way to measure the quality of a partition $\mathbf{g}$ in a network with adjacency matrix $\mathbf{A}$. In community detection, we do not know the partition *a priori*. We rather want to detect communities in a network. We can detect communities by finding a partition with maximal modularity. One calls the task of finding such partitions the *modularity-maximisation problem* [34]. Ideally, we would like to test all possible partitions and choose one with maximal modularity. The number of possible partitions of a network with N nodes is the *Bell number*

$$B_N = \sum_{k=0}^{N+1} \binom{N-1}{k} B_k \,, \tag{2.17}$$

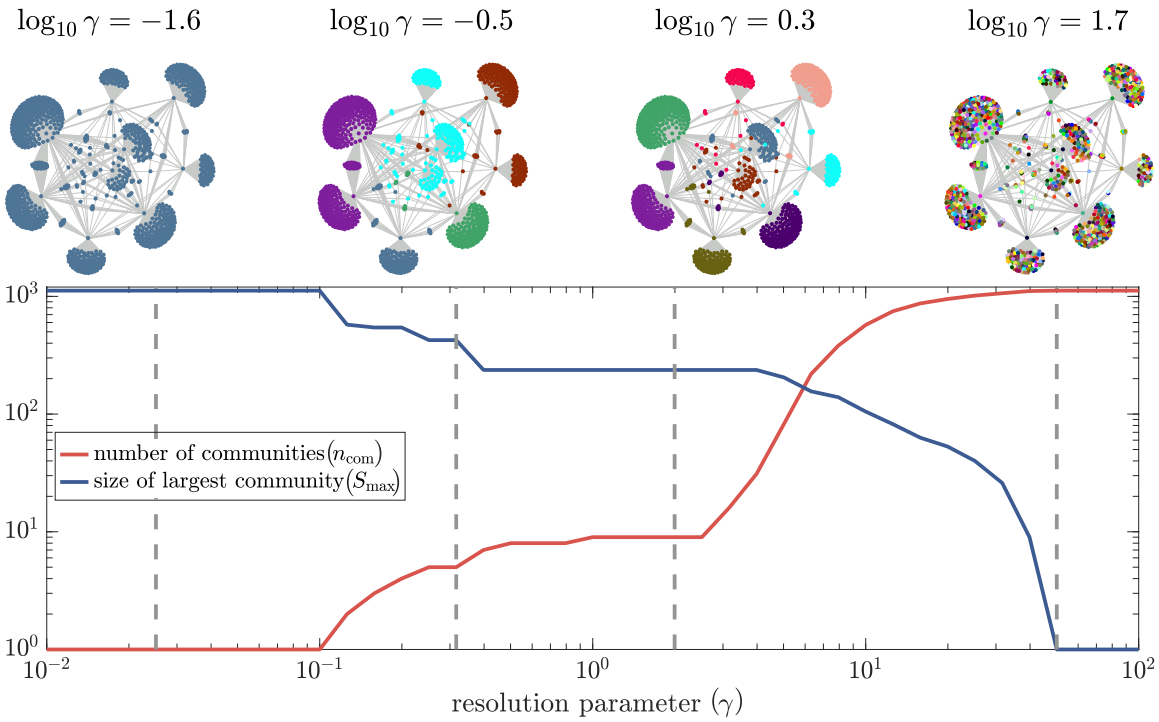


Figure 2.6: Community structure in a PIN of HIV. We show the communities in the HIV PIN for the resolution parameter $\log_{10}\gamma \in \{-1.6, -0.5, 0.3, 1.7\}$. For small γ, all nodes belong to the same community; for large γ, each node belongs to its own community. The number n_{com} of communities increases with γ, and the size S_{max} of the largest community decreases with γ. The network consists of $N = 1\,119$ nodes and $M = 1\,970$ edges; we constructed it using the BIOGRID database version 3.4.161 [77]. Note the logarithmic scale of both axes.

which grows very rapidly with N. For networks with $N = 7$ and $N = 8$ nodes, for example, there exist 877 and 4 140 possible partitions, respectively (see Table 2.6). Finding a partition that yields maximal modularity by comparing all partitions is thus not possible for almost all networks.

Finding a partition that maximises modularity is NP-hard; that is, there exists no correct polynomial-time algorithm to solve this problem for every instance (unless P = NP) [58]. Therefore, one uses heuristic algorithms to find approximate solutions to the problem. Among them are simulated annealing, extremal optimisation, and

spectral partitioning [58].

In this thesis, we use the GenLouvain implementation (version 2.1.1) [217] of a generalised Louvain algorithm [45] for Matlab to find an approximate solution to the modularity-maximisation problem. The GenLouvain algorithm returns a partition $\mathbf{g}$. For a detailed discussion of this computational method, see Appendix A. Empirical observations suggest that the run-time complexity of the Louvain algorithm for graphs is $\mathcal{O}(N \log\langle k\rangle)$ [45].

Although modularity-maximisation algorithms are used widely for the study of networks, they come with some drawbacks: modularity maximisation prefers partitions with communities of a similar size [249, 329], it is not always a consistent estimator[19] of partitions [43], and the modularity function often has multiple near-optimal partitions [163]. For all community-detection algorithms, including modularity-maximisation, there exists a detectability threshold [116, 324, 500]

[19] An estimator of a parameter is called *consistent* if it converges in probability to the true value of this parameter.

N	0	1	2	3	4	5	6	7	8	9	10
B_N	1	1	2	5	15	52	203	877	4140	21147	115975

Table 2.6: The number of possible partitions for a network with N nodes is the Bell number. The Bell number B_N grows rapidly with the size N of the network. Even for small networks, it is thus not feasible to check all possible partitions to find one with maximal modularity.

2.7 Multilayer Networks

As we discussed in Chapter 1, graphs are mathematical models of complex systems. These graphs describe pairwise interactions between a system's entities. In many systems, these interactions are more complicated, for example, they change over time or there are different types of interactions [42, 101, 232]. One can use *multilayer networks* to describe such interaction patterns. In this thesis, we use a general mathematical description of multilayer networks from [232]. From now on, we call 'normal' networks 'monolayer networks' if it is necessary to distinguish between them and multilayer networks.

A *multilayer network* (MLN) is a quadruplet $M = (V_M, E_M, V, \mathcal{L})$, whose components we discuss now.[20] The network consists of $N = |V|$ *physical nodes* and $L = |\mathcal{L}|$ *layers*, which are composed of d different *aspects* (e.g., 'tissue type' and 'time'), each of which has a set of elementary layers L_a (e.g., 'muscle' and 'nervous' for tissue type and '1','2', and '3' for time). We construct the sequence $\mathcal{L} = L_1 \times \cdots \times L_d$ of sets of layers as all possible combinations of these elementary layers by using a Cartesian product[21]. The physical nodes are represented in the layers as *states nodes* $V_M \subseteq V \times \mathcal{L}$, which can also be described as *node-layer tuples*: The entry $i\mathbf{a} \in V_M$ indicates whether node $i \in V$ exists on a specific layer $\mathbf{a} \in \mathcal{L}$, with $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$. The edges $E_M \subseteq V_M \times V_M$

[20] There is also the notion of 'multi-layer neural networks', which are computational models with multiple processing layers and can be used for so-called 'deep learning' [256, 447]. In this thesis, we use 'multilayer network' in the standard way from network science, rather than in this other way.

[21] For two sets A and B, their *Cartesian product* $A \times B$ is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$.

connect these state nodes in a pairwise way. We differentiate between *intralayer edges* $E_L = \{(i\mathbf{a}, j\mathbf{b}) \in E_M : \mathbf{a} = \mathbf{b}\}$ and *interlayer edges* $E_C = \{(i\mathbf{a}, j\mathbf{b}) \in E_M : \mathbf{a} \neq \mathbf{b}\}$. Edges can have a weight, which are given by a *weight function* $w : E_M \to \mathbb{R}$.[22] In the case of an unweighted network, we define $w(e) = 1$ for all $e \in E_M$. In analogy to an adjacency matrix for a monolayer network we can represent a multilayer network as an *adjacency tensor*

$$\mathbb{A}_{j\mathbf{b}}^{i\mathbf{a}} = \begin{cases} w((i\mathbf{a}, j\mathbf{b}))\,, & \text{if } (i\mathbf{a}, j\mathbf{b}) \in E_M\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{2.18}$$

We can use the state nodes V_M and the edges E_M of a multilayer network M to construct an *underlying graph* $G_M = (V_M, E_M)$, which we can interpret as a labelled graph by using the information from V and $\mathcal{L}$. The labels in this graph reflect the node-layer identity of the state nodes in the associated multilayer network. The adjacency matrix of this underlying graph is the multilayer network's *supra-adjacency matrix*. Specifically, one can 'flatten' an adjacency tensor into a supra-adjcacency matrix [96, 162, 423]. This is the unfolding of a 4^{th}-order tensor $\mathbb{A}_{j\mathbf{b}}^{i\mathbf{a}} \in \mathbb{R}^{N\times N\times L\times L}$ into a square matrix $\mathbf{A}^{\text{supra}} \in \mathbb{R}^{NL\times NL}$, i.e., a 2^{nd}-order tensor. The supra-adjacency matrix is

$$\mathbf{A}_{i\mathbf{a},j\mathbf{b}}^{(\text{supra})} = \begin{cases} w((i\mathbf{a}, j\mathbf{b}))\,, & \text{if } (i\mathbf{a}, j\mathbf{b}) \in E_M\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{2.19}$$

Such a flattening can be fruitful for many applications, including the study of

[22] As for a weighted monolayer networks, the range of this weight function is not necessarily restricted to real numbers $\mathbb{R}$. For our work, however, this range is sufficient.

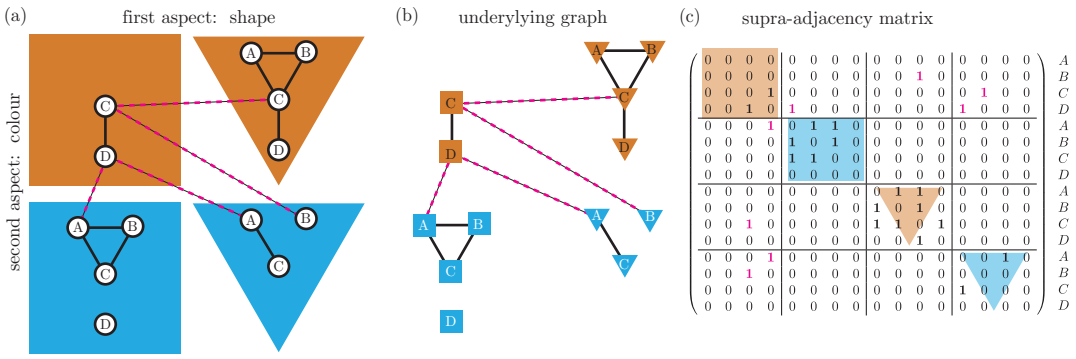


Figure 2.7: An example of the most general type of multilayer network that we use in this thesis. (a)An MLN $M = (V_M, E_M, V, \mathcal{L})$, as discussed in this thesis and introduced in [232]. The network M has a total of four physical nodes, so $V = \{A, B, C, D\}$, and two aspects. For illustrative purposes, we choose 'colour' and 'shape' as our aspects; their elementary layer sets are $L_1 = \{\text{orange}, \text{blue}\}$ and $L_2 = \{\square, \triangledown\}$. There are thus a total of four layers: $\mathcal{L} = \{(\text{orange}, \square), (\text{blue}, \square), (\text{orange}, \triangledown), (\text{blue}, \triangledown)\}$, as indicated by the layers' appearances. Each layer includes a subset of the physical-node set V as indicated by the set of node-layer tuples $V_M \subseteq V \times L_1 \times L_2$.[23] Edges connect some of the nodes in a pairwise manner, both within and across layers. We show intra-layer edges as solid lines and inter-layer edges as dashed lines. (b) The underlying graph $G_M = (V_M, E_M)$ of the MLN M. We show inter-layer and intra-layer edges using the same convention as in (a). We give the labels, which indicate the node-layer identity of each state node, as the appearance (colour and shape) of the nodes V_M. (c) The adjacency matrix of this underlying graph is the multilayer network's supra-adjacency matrix. Intra-layer edges are represented by **1** entries in the matrices along the diagonal. We show inter-layer edges **1** in magenta.

spreading processes [162] and the definition of multilayer generalisations of centrality measures [108]. It does, however, also remove information about the aspects and thus is a simplification of the full adjacency tensor.

After summarising this general framework of multilayer networks, we now discuss some special cases, which we use in the remainder of this thesis.

2.7.1 Monolayer Networks

A *monolayer* (also called monoplex) network is a 'normal' graph, as defined in Section 2.3. It is a special case of a MLN M outlined above with the number d

of aspects being zero. In this case, the set V_M of state nodes and the set V of physical nodes are identical. Similarly, if every aspect has a single elementary layer, the number of layers is $L = 1$, which also can be represented as monolayer network.

In these cases, the adjacency matrix (see Eqn. (2.1)) and the adjacency tensor (see Eqn. (2.18)) are equivalent. Note also that the underlying graph G_M represents all connectivity information.

2.7.2 Edge-coloured Multigraphs

An *edge-colored multigraph* is an $(L+1)$-tuple $M = (V, E_1, E_2, \ldots, E_L)$. The set V of nodes is the same as in a single-layer network, and each of the edge sets $E_i \subset V \times V$ connects pairs of nodes [336, 357]. An edge-coloured multigraph is a special case of a multilayer network without interlayer edges, such that $E_M = E_L$. In this edge-coloured multigraph, each edge colour corresponds to a layer in a multilayer network. In Fig. 2.8a, we show an edge-coloured multigraph with three colours (yellow, orange, and red) and three physical nodes (1, 2, and 3).

2.7.3 Multiplex Networks

Multiplex networks or *multirelational networks* are networks with different types of interactions. They are therefore similar to the edge-coloured multigraphs. The difference

[23] For this example, the set of node-layer tuples is

$$\begin{aligned} V_M = \{&(C, \text{orange}, \square), (D, \text{orange}, \square), \\ &(A, \text{blue}, \square), (B, \text{blue}, \square), (C, \text{blue}, \square), (D, \text{blue}, \square), \\ &(A, \text{orange}, \triangledown), (B, \text{orange}, \triangledown), (C, \text{orange}, \triangledown), (D, \text{orange}, \triangledown), \\ &(A, \text{blue}, \triangledown), (B, \text{blue}, \triangledown), (C, \text{blue}, \triangledown)\} \subseteq V \times L_1 \times L_2 \,. \end{aligned}$$

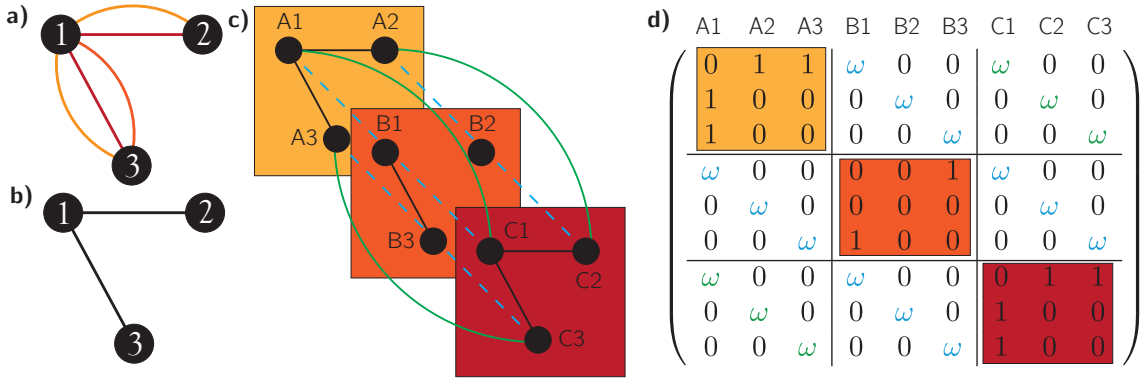


Figure 2.8: A few different types of multilayer networks. (a) Edge-coloured multigraph with three different types of interactions. (b) We can construct a monolayer network by deleting parallel edges. In (c), we show a MLN that consists of three layers. We show the corresponding supra-adjacency matrix in (d). Each layer's unweighted adjacency matrix is positioned along the diagonal in a block-matrix form. In the most commonly-used multilayer representation of a temporal multilayer network, nodes representing the same gene (or in a general case some other entity) are connected in adjacent layers (blue, dashed edges). In a categorical MLN, all layers connect with interlayer edges, so the green, curved edges can also be present. In the supra-adjacency matrix, all those inter-layer edges are usually chosen to have a weight which is the coupling parameter ω.

is that multiplex networks also have interlayer edges. Commonly, they connect nodes that represent the same physical entity, and we call them *coupling edges*. Thus, the set of interlayer edges is $E_C = \{(i\mathbf{a}, j\mathbf{b}) \in E_M : i = j\}$. It can be fruitful to assign these interlayer edges weights $\omega((i\mathbf{a}, j\mathbf{b}))$, which one calls *interlayer coupling*. In principle, this weight can be different for each pair of nodes. For simplicity, one often chooses this coupling as ω for all nodes.

In Fig. 2.8c, we show a MLN, that consists of three physical nodes (1,2, and 3), three layers, and nine state nodes (A1, A2, ..., C2, C3). The intra-layer edges are given by the edge-coloured multigraph. The inter-layer edges (dashed and solid lines) connect state nodes that represent the same physical node (e.g., A1 with B1 and C1).

In Fig. 2.8d, we show a corresponding supra-adjacency matrix.

2.7.4 Temporal Networks

There are various ways to represent temporal networks (i.e., networks whose nodes, edges, interaction structure, and/or interaction weights changes over time) [194, 294]. Two ways to represent them are *event-based* representations and *snapshot* representations. In this thesis, we exclusively use the latter but discuss both to make the difference clear.

Event-based Representation

We consider an event-based temporal network with the node set V as a time-ordered list of I events:

$$\mathcal{T}_{\text{event}} = \{(u_i, v_i, t_i) : i = 1, 2, \ldots, I \text{ with } u_i, v_i \in V\}\,, \tag{2.20}$$

where (u_i,v_i) is the node pair that interacts at the ith event[24], which occurs at time $t_i \in [0, t_{\max}]$. This representation is similar to an edge list in time-independent networks. For each edge (i.e., event), there is an associated time; a pair of nodes can occur in multiple events. One example of real-world data in this format is time-stamped messaging between individuals (e.g., via e-mail or on an online social network).

[24] A more general description, which includes the *duration* Δt of an event, is also possible [71]. For our applications, however, this is not necessary, so we omit it. Phone calls, with a finite duration Δt, are an application for which it can be relevant to consider contact duration.

Snapshot Representation

A snapshot representation of a temporal network is a discrete-time sequence of graphs

$$\mathcal{T}_{\text{snapshot}} = \{G_1, G_2, \ldots, G_T\}\,, \tag{2.21}$$

where T is the number of graphs (i.e., the number of discrete time points for which an interaction is measured). Each of the graphs $G_t = (V, E_t)$, with $E_t \subset V \times V$, represents the connectivity at time t. We assume that the node set V is the same for each snapshot graph G_t; it is possible to define a more general variant, in which nodes can appear and disappear. This snapshot representation can be written algebraically in a node-by-node-by-time ($N \times N \times T$) adjacency tensor in which the nonzero elements $A_{ij}^{(t)}$ indicate the presence and weight of an edge from node i to node j in time layer t. All temporal PINs that we discuss in this thesis are constructed from discrete-time-point data and thus are naturally given in a snapshot representation, in which each layer represents one time point.

Multilayer Network Representations of Temporal Networks

We can construct a temporal MLN $\mathcal{T} = (V_M, E_M, V, \mathcal{L})$ with a single aspect (time) from a temporal network in snapshot representation $\mathcal{T}_{\text{snapshot}}$, without loss of information in the following way: The set of physical nodes V is the same as for the snapshot graphs G_t. The temporal multilayer network $\mathcal{T}$ consists of T layers $\mathcal{L} = L_1, \ldots, L_T$ (one for each snapshot graph). We choose a variant in which every physical node is present in each of the layers; the node-layer couples are therefore given by $V_M = V \times \mathcal{L}$. The

edges between these state nodes are intralayer edges E_L and interlayer edges E_C. The former originate from the snapshot graphs $((i,t),(j,t)) \in E_L \leftrightarrow (i,j) \in E_t$. The latter, also called 'identity edges', link state nodes representing the same physical node in time-adjacent layers, so $((i,t),(i,t+1)) \in E_C$ for all $i \in V, t \in \{1,\dots,T-1\}$.[25]

At first glance, the creation of a MLN representation with inter-layer edges seems overly complicated in comparison with the simpler (and computationally less demanding), snapshot representation. For many applications (e.g., community detection and centrality computation), however, the introduction of inter-layer edges can be insightful, as we will discuss in Sections 4.5 and 2.8.

In Fig. 2.8c, we show a temporal MLN. The temporal MLN consists of three physical nodes (1,2, and 3), three layers, and nine state nodes (A1, A2, ..., C2, C3). The intra-layer edges are given by the edge-coloured multigraph. The inter-layer edges (solid lines) connect state nodes that correspond to the same physical node in adjacent time layers (e.g., A1 with B1). In Fig. 2.8d, we show a corresponding supra-adjacency matrix. Note, that in this temporal MLN, the green ω are zero, whereas in the multiplex MLN the are nonzero.

2.7.5 Community Detection in Multilayer Networks

Various community-detection methods have been generalised to MLNs. Examples include InfoMap [107], modularity [320], and maximum-likelihood estimations of SBMs [105, 362, 431]. In this thesis, we use a modularity-based approach [320] because

[25] One can also construct temporal networks in which not exclusively time-adjacent layers are connected.

it is readily available in the GenLouvain software [217] and it has been successfully applied to a wide variety of applications [31, 32, 34].

In temporal MLNs, detecting modular structure has given insights into financial correlations [34], brain connectivity [32], social networks [22], and several other fields. In multiplex MLNs, modular structure has been examined in examples such as social-interaction networks [105] and the interaction between bacteria in humans [431].

Mucha et al. (2010) derive a multilayer generalisation of modularity as

$$Q(\mathbf{g}) = \frac{1}{2\mu} \sum_{ijs} \left\{ \left(A_{ijs} - \gamma_s \frac{k_i k_j}{2m_s} \right) \delta(g_s, g_r) + \omega_{ist} \right\} \delta(g_{is}, g_{jr}) \,, \tag{2.22}$$

where A_{ijs} are the adjacency matrix entries A_{ij} for each layer s and μ is the total edge weight. The resolution parameter γ_s can be layer-dependent. We, however, choose them uniformly, i.e. $\gamma_s = \gamma$ for all layers s. The interlayer coupling parameter ω_{ist} gives the connection strength between layers. For temporal MLNs, we only couple time-adjacent layers, so $\omega_{ist} = 0$ for all $|s - t| > 1$. We treat all interlayer edges with the same coupling ω, which is called *homogenous interlayer coupling*. More general coupling procedures are possible but rarely used in practice [232].

2.8 Community Detection in Protein Interaction Networks

2.8.1 Community Detection in Monolayer Protein Interaction Networks

There exists evidence that the functional organisation of cells is modular [180]. It has been hypothesised, that such an organisation would be robust against external attacks

and that it would allow a cell to adapt its biological processes to external influences [14]. Two approaches that are suitable for examining such a modular organisation in cells are the investigation of gene-expression data and the examination of PINs. The former is the identification of genes that are often expressed with one another (i.e., they are 'co-expressed') [229, 404]. The latter is the identification of sets of proteins that tend to interact with one another and is one focus in this thesis.

One can use community-detection methods to identify modular structures in PINs, and researchers have applied many different methods to do so (e.g., [10, 65, 80, 127, 268, 283, 302, 363, 415].) It is hoped that such detected modules tend to consist of proteins with a similar biological function. Some communities in PINs have been associated with certain diseases [201, 321].

Functional modules in PINs can have many different scales, so multiresolution methods can reveal different levels of organisation [262]. Communities in PINs can help to identify the function of proteins [257] and proteins that are drug targets [310].

One can also identify the position of single proteins in a modular structure in a PIN [39, 170, 171, 234]. Such studies have given insight into the roles that different proteins play in PINs, such as connecting to proteins in the same community or connecting to nodes in different communities. The latter can be crucial in the communication between different functional modules.

2.8.2 Community Detection in Multilayer Protein Interaction Networks

In recent years, MLN have been explored as a tool to study PPI data [123]. For example, one can use MLNs to combine a PIN with drug–protein and drug–drug interations to try to predict side effects [505]. One can also use MLNs to encode interactions between different biological entities in the layers (e.g., genes and micro-RNAs[26] [187]). One can also examine a PIN as an edge-coloured multigraph in which each colour indicates a different experimental technique [288]. Somewhat surprisingly, Mangioni et al. assessed, however, that treating the layers in a human PIN independently — with no interlayer coupling — yields more biologically relevant modules than incorporating interlayer coupling. One can combine different modes of interactions (transcription-factor co-targeting, microRNA co-targeting, protein–protein interaction, and gene co-expression networks) in MLNs [67]. The authors reported that a community-detection algorithm helps to identify cancer-related genes.

In Appendix B, we use MLN to detect modular structures in a temporal PIN. Community detection in MLNs is only one way to detect community structure in temporal networks [390]. We choose this approach because it has been succesfully applied to several temporal networks (e.g., [32, 34, 320]).

Cellular systems are dynamic (i.e, they change in time) and respond to environmental signals [480, 490]. Functional modules, and thus also structural communities

[26]Micro-RNA are small non-coding RNA molecules that are involved in post-translational modification and gene silencing.

in PINs, are likely to have an important role in such response [112, 174]. The study of communities in PINs, however, so far has focussed on time-independent networks. Temporal clustering approaches can help to examine temporal changes in multiprotein complexes [345].

2.9 Gene Ontology Enrichment

2.9.1 Gene Ontology Data

The *Gene Ontology* (GO) project provides functional terms that describe genetic products, which include proteins [20, 89]. The ontology data base covers three domains: *cellular component*, *molecular function*, and *biological process*.

Each GO term belongs to exactly one of these three domains. In Table 2.7, we show *pyruvate metabolic process* as one example of a GO term. The terms in the GO data base are related to each other, because, for example, a given process is a subprocess of another. A 'cellular metabolic process', for example, is always also a 'metabolic process'. These relationships can be presented in a *directed acyclic graph*[27]. In Fig. 2.9, we show a subgraph of this graph. Parent terms in this graph are more general terms and a term can have multiple parents [382].

The GO annotations are created either by curators or automatically. The evidence used to create the annotation can either be experimental, computational, indirectly derived, or unknown. In this thesis, we use all available GO annotations, independent

[27] As the name suggests, a *directed acyclic graph* is a graph with directed edges and without any directed cycles.

of the evidence supporting their creation.

Accession	GO:0006090
Name	pyruvate metabolic process
Ontology	biological process
Alternate IDs	GO:0006087
Definition	The chemical reactions and pathways involving pyruvate, 2-oxopropanoate.
Synonyms	pyruvate dehydrogenase bypass RELATED pyruvate metabolism EXACT

Table 2.7: Example Gene Ontology Term. The GO term *pyruvate metabolic process* is part of the domain 'biological process'.

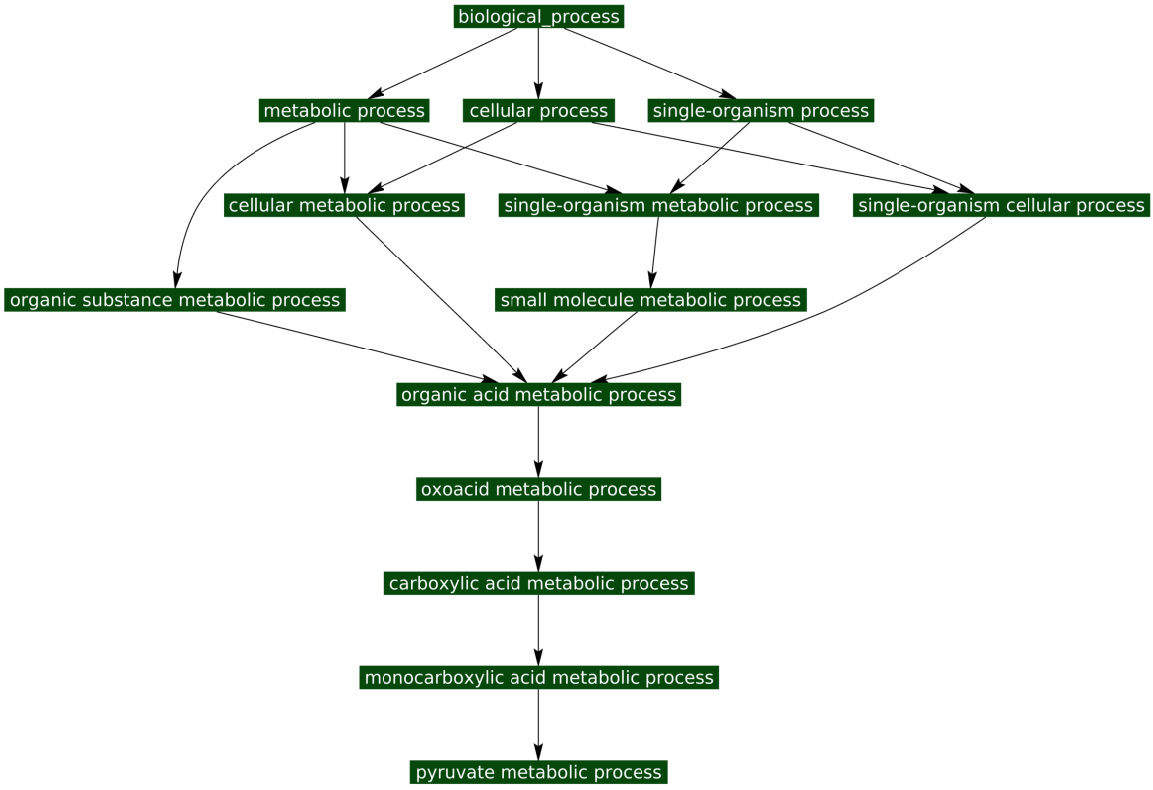


Figure 2.9: A subgraph of the domain 'biological process'. The relationships between ancestor terms of the term *pyruvate metabolic process* yields a directed acyclic graph. Every other term is reachable from the term *biological process*.

2.9.2 Hypergeometric Test

GO terms allow enrichment analysis of gene sets [304]. This is the detection of whether any GO terms are strongly associated with genes in a given set. It is necessary to be more sophisticated than counting only the number of occurrences, because some terms are associated with a large number of genes and an apparent over-representation can occur at random. One can also use GO-term enrichment to examine whether sets of gene products (e.g., proteins) are significantly enriched.

Using the *hypergeometric distribution* (HGDF) is the standard, but not the only possible, approach for identifying whether an individual GO term is enriched [128, 408].[28] The HGDF is a discrete probability distribution that gives the probability of k successes in n draws without replacement from a population of size N with K success states. In the notion of GO enrichment, a success is annotation with a given term and N is the size of the network. The probability mass function (PMF) is

$$P(X = k) = \frac{\binom{N}{K}\binom{N-K}{n-k}}{\binom{N}{n}} . \tag{2.23}$$

The probability that k or more genes from a target set (i.e., a community) of size B are associated with a given GO is then given by the hypergeometric tail

$$P(X \geq k) = \sum_{k'=k}^{\min(n,B)} \frac{\binom{N}{K}\binom{N-K}{n-k'}}{\binom{N}{n}} . \tag{2.24}$$

One calculates the hypergeometric p-value as the probability $P(X \geq k)$ of drawing at least k successes at random. If this probability is below a certain significance level, we

[28] For example, one can compare the functional similarity of proteins in a detected community [263].

reject the null hypothesis that the effect is detected at random. There is no clear way of choosing an appropriate significance level. In this thesis, we choose $\sigma = 0.05$.

There are multiple online tools that allow detection of gene enrichment. Among them are PANTHER [303, 304] and GORILLA [128]. In this thesis, we mainly use our own implementation in MATLAB, because it allows automated testing of a large number of sets of genes (or communities) in PINs. Comparing the results of different tools does not always yield identical results. The reasons for the discrepancies are the use of slightly different approaches. GORILLA compares against a background set of the tested genes, whereas PANTHER compares against the set of all genes for a given organism. GORILLA uses a hypergeometric probability distribution for testing, whereas PANTHER uses the Mann–Whitney test [290]. In our own implementation we test against a background set of all genes that are present (as proteins) in a PIN we examine. This ensures that we do not introduce additional biases from the selection of proteins that are present in a PIN.

While GO-term enrichment is used widely, it has limitations. Most notably, well-studied proteins tend to have more GO annotations than less-studied proteins; this is a so-called *examination bias* [282]. Due to this uneven distribution of functional annotations, GO-enrichment tools favour communities that consist of well-studied proteins with many annotations.

Multiple-Testing Correction

The hypergeometric test, as outlined above, allows an enrichment test for a single GO term against a set of genes. Usually, one wants to test whether or not a large number

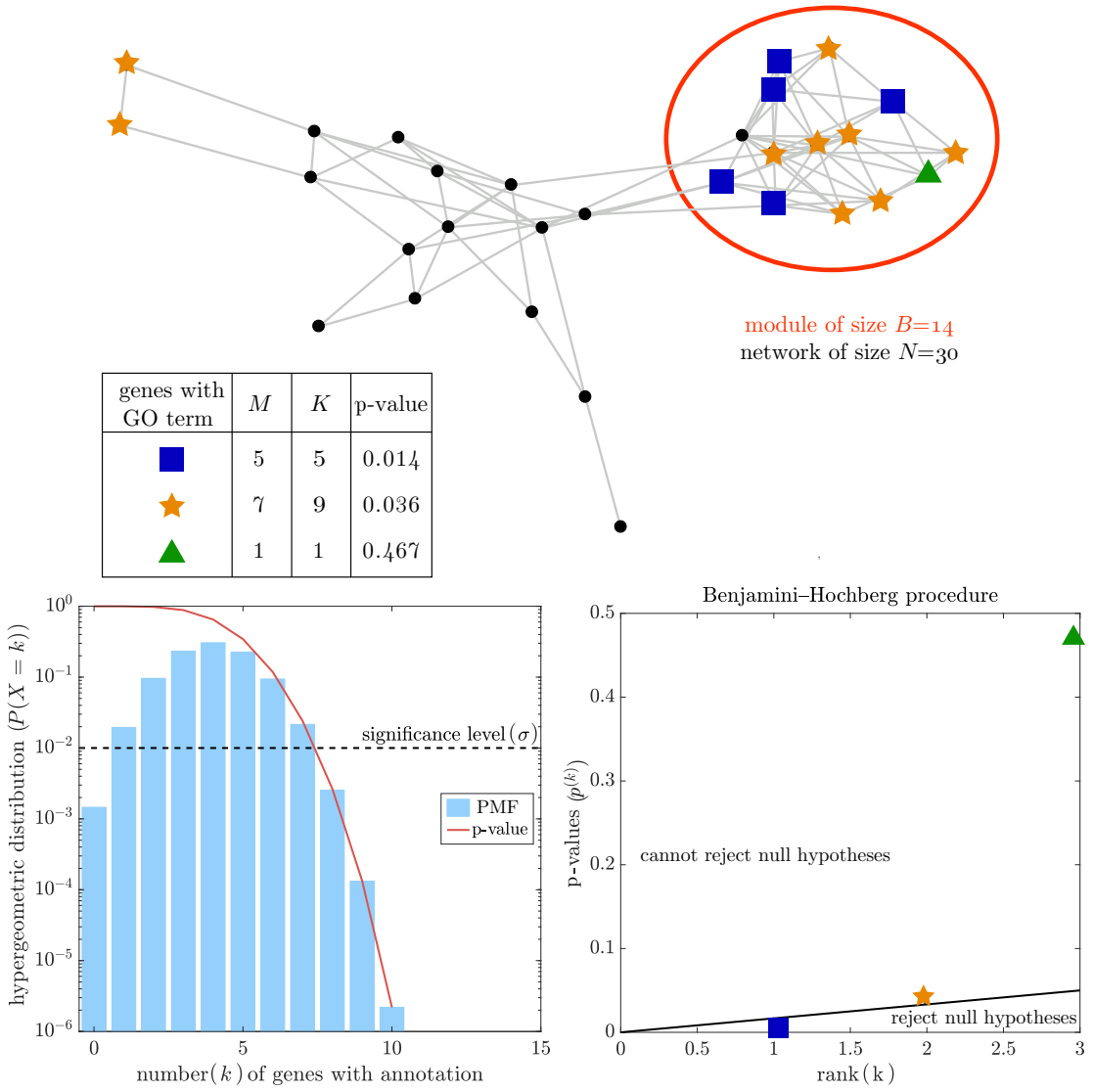


Figure 2.10: GO enrichment with a hypergeometric test and multiple-testing correction. The hypergeometric test for population size $N = 30$, number $S = 10$ of successes, and number $M = 12$ of draws. We test three different annotations and highlight proteins associated with a given annotations as blue squares, orange stars, and green triangles, respectively. In the module, we interpret the former two as significant, because their p-values are below the significance level of $\sigma = 0.05$. When we correct for multiple testing with the Benjamini–Hochberg procedure, however, only the blue-square annotation is significant.

of terms are enriched in a given set (e.g., a community). Testing repeatedly gives a high probability of false positives (i.e., detecting enriched terms that are only enriched at chance). Such *multiple testing* makes it necessary to apply a correction method.

The Bonferroni correction and the Benjamini–Hochberg procedure are two widely-used methods to correct for multiple testing [122]. We use the latter because the Bonferroni correction tends to produce a large number of false negatives. Bonferroni corrections control the family-wise error rate, which is the probability of producing one or more false positives, in a set of tests. By contrast, the Benjamini–Hochberg procedure instead controls the false-discovery rate [36], which is more appropriate for our investigation, as we usually do not mind having a small number of false positives for having a much higher number of true positives.

We use the Benjamini–Hochberg procedure as follows. We test n hypotheses — in our case, one hypothesis consists of the enrichment of a single GO term in a community — and the hypergeometric test yields p-values of $p^{(1)}, p^{(2)}, \ldots, p^{(n)}$. We sort the p-values in increasing order, such that $p^{(i)} \leq p^{(i+1)}$ for all i. Given a significance level σ, we calculate the largest s such that

$$p^{(s)} \leq \frac{s}{n}\sigma\,, \tag{2.25}$$

and we reject the null hypothesis for all $s' = 1, \ldots, s$. For GO enrichment, we declare these s terms to be enriched.

3

'Promiscuity' of Nodes in Multilayer Networks

Contents

3.1 Promiscuity of Nodes 71
3.1.1 Definition of Promiscuity 71
3.1.2 Proof that Promiscuity $p_i \in [0, 1]$ 78
3.1.3 Proof that Promiscuity $p_i = 1$ in Colour-regular Edge-coloured Multigraphs 79
3.2 Promiscuity p_i in Synthetic Networks 80
3.2.1 Network Composed of an Erdős–Rényi Layer and a Ring Lattice Layer 81
3.2.2 Network Composed of Two Erdős–Rényi Layers 85
3.2.3 Conclusion for Synthetic Two-Layer Networks 87
3.3 Promiscuity p_i in Synthetic Networks under Rewiring . . 88
3.3.1 Network Construction 88
3.3.2 Randomization 89
3.3.3 Promiscuity p_i under Randomization 90
3.4 Promiscuity p_i in Empirical Networks 96
3.5 Promiscuity in Tissue-specific Protein Interaction Networks 100
3.5.1 Tissue-specific Transcription Factor Regulation 101
3.5.2 Tissue-specific Gene Regulation 105
3.6 Conclusions 109

This chapter is based on a paper draft that is joint work with Gorka Zamora-López, Jonny Wray, Charlotte M. Deane, Jürgen Kurths, and Mason A. Porter.

As we discussed in Chapter 1, MLNs are a useful tool for integrating different types of biological data [507]. The development of structural measures for the analysis of these MLNs is an active field of research [33, 42, 110]. In this chapter, we examine the extent to which node importances vary across layers in edge-coloured multigraphs, a special case of MLNs that constitute the simplest representation of a multirelational network (see Fig. 3.1a). To do this, we introduce a quantity called *promiscuity*, which gives an information-theoretical measure of a node's distribution of neighbours across different layers in comparison with a (random) null model. Our approach is simple, but its yields valuable insights into the structure and function of multirelational networks. We find, for example, that different types of networks have rather different node-promiscuity distributions.

Similar to approaches for examining modular structures in monolayer [170, 171] and multilayer [33] networks, we examine promiscuity versus node degrees to create a two-dimensional depiction of the roles that nodes play in a multirelational network. In contrast to the *multilayer participation* approach from Battiston et al. (2014) we use normalized entropy as measure of variability, and we take into account the different densities of network layers. To avoid confusion, we note that Papadopoulos et al. (2016) defined the 'particle promiscuity' as the fraction of communities in which a node participates [353].

To demonstrate use on biological networks, we focus on the analysis of two regulation

interaction MLNs. In these, we represent tissue types as layers. Computing the promiscuities of nodes in these networks allows us to examine the variability of gene importance across tissues based on the number of their regulatory partners. We find that the vast majority of transcription factors (TFs) and genes are not cell-type specific. Some of them, however, have small promiscuities and their regulatory influence is thus tissue-specific.

The rest of this chapter is structured as follows. First, in Section 3.1, we motivate and define promiscuity in edge-coloured multigraphs. We then generalise the definition of promiscuity to directed and weighted networks. We prove that in all networks, the promiscuity $p_i \in [0, 1]$ and that in a certain type of synthetic networks, which we call *colour-regular*, all nodes have minimal promiscuity $p_i = 0$. Second, in Sections 3.2 and 3.3, we compute numerically and analytically the promiscuity in three types of synthetic networks. In some cases, we are able to derive explicit expressions; in others we obtain only upper bounds. Third, in Sections 3.4 and 3.5, we compute the promiscuity for networks constructed from real-world data, including two regulation networks. We conclude in Section 3.6.

3.1 Promiscuity of Nodes

3.1.1 Definition of Promiscuity

In this chapter, we discuss *edge-coloured multigraphs*. In Section 2.7, we showed that this is a special kind of MLN $\mathcal{M}$. Recall that an edge-coloured multigraph is an $(L+1)$-tuple $M = (V, E_1, E_2, \ldots, E_L)$, where V is the set of nodes and $N = |V|$. Each

of the l edge sets $E_i \subset V \times V$ consists of pairs of nodes. We refer to the edge sets as 'layers'. One way to interpret this definition is that each layer's edges have a unique colour. The network is a multigraph because it has parallel edges, so multiple edges can connect the same pair of nodes. *Multirelational network* [448] is another term to describe the same mathematical structure and it is especially used in the study of social networks. Unlike in many other MLNs, there are no interlayer edges.

In Fig. 3.1a, we show an edge-coloured multigraph with $N = 12$ nodes, $m = 17$ edges, and $L = 3$ layers. The nodes have different numbers of incident edges. Some nodes (specifically, nodes A, E, and G) have incident edges of only one colour, whereas others (nodes C and D) have incident edges of all three colours. We introduce a scalar quantity that we call *promiscuity* to quantify such variability, while taking into account that, in a random-graph model with uniformly random connection probabilities, a node is more likely to connect to a layer that includes a larger number of edges.

In a monolayer network, a node i's degree k_i is defined as the number of its neighbours; it is equal to the number of edges that are incident to it. For each node in an edge-coloured multigraph, we construct a degree vector

$$\mathbf{k}_i = \left(k_i^1, k_i^2, \ldots, k_i^L\right) , \tag{3.1}$$

where each element k_i^l is the number of neighbours that node i has in layer l.

For the example network in Fig. 3.1a, we show the degree vector of each node in Fig. 3.1b. We summarise this information using two scalar values; 'multilayer degree' and 'promiscuity'. Multilayer degree $K_i = \sum_{l=1}^{L} k_i^l$ of a node i is the total number

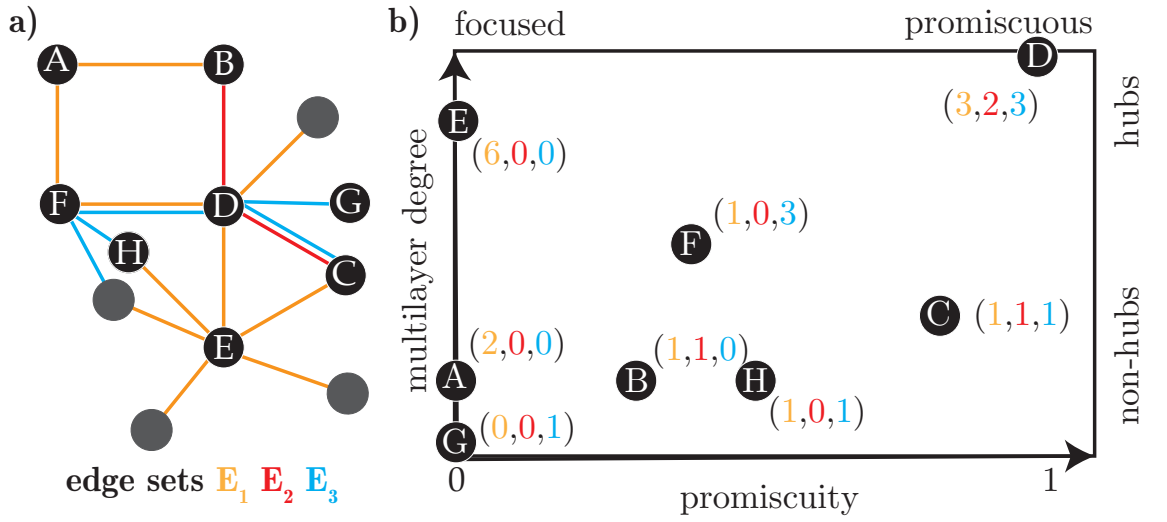


Figure 3.1: (a) Illustration of a MLN as an edge-coloured multigraph and (b) the role of some of the nodes in this network. We examine a node's role by calculating promiscuity–degree (p_i, K_i) coordinates, next to which we show its degree vector $\mathbf{k}_i$. Nodes with promiscuity $p_i = 0$ (specifically, A, E, and G) link to nodes in a single layer, and nodes (specifically, C and D) with almost maximal promiscuity (i.e., with $p_i \to 1$) link equally to all layers. Other nodes (specifically, B, F, and H) are dominated by edges of some layers, but not in all of them. This example network includes nodes that play various roles, ranging from promiscuous non-hubs to monolayer hubs (as well as other types). Unlabelled nodes in panel (a) have identical promiscuity–degree (p_i, K_i) coordinates as some of the labelled ones, so we do not show them in panel (b).

of its incident edges across all layers. The multilayer degree K_i is a measure of the aggregate importance of node i in a MLN, but it does not incorporate information about how its importance varies across layers.

To measure this variability in importance, we investigate the distribution of a node's degree among the different layers in a multirelational network. To do this, it is important to consider that the layers may have very different densities. Thus, we define each node's *promiscuity vector*

$$\mathbf{P}_i = \left(P_i^1, P_i^2, \ldots, P_i^L\right), \quad \text{with} \quad P_i^l = k_i^l / \langle k_l \rangle \,, \tag{3.2}$$

as the number of neighbours in each layer l divided by the layer's mean degree $\langle k_l \rangle = N^{-1} \left(\sum_{i \in N} k_i^l \right)$. The normalization by the mean degree represents a comparison with a randomization (i.e., a null model) that preserves the number $M_l = \sum_{i=1}^{N} k_{i,l}/2$ of edges in each layer l, but is otherwise maximally random. Specifically, $P_i^l = 0$ if node i has no connections in layer l, and $P_i^l = 1$ if it has as many connections as expected in the null model. When $P_i^l < 1$, there are fewer connections than expected; when $P_i^l > 1$, there are more connections than expected. We define the *promiscuity*

$$p_i = -\left(\sum_{l=1}^{L} P_i^l \ln P_i^l \right) \Big/ \ln L = H(\mathbf{P'}_i) \Big/ \ln L \tag{3.3}$$

of node i as the normalized Shannon entropy $H(\mathbf{P'_i})/\ln(L)$ of the normalized promiscuity vector $\mathbf{P'}_i$. The Shannon entropy of a normalized vector $\mathbf{x} = (x_1, \ldots, x_L)$ of L elements is $H(\mathbf{x}) = \sum_{l=1}^{L} -x_l \ln(x_l)$. If $x_l = 0$ for some l, the value of the corresponding summand $0 \ln(0)$ is equal to 0, which is consistent with the limit $\lim\limits_{x \to 0+} x \ln(x) = 0$. [1] The factor $\ln(L)^{-1}$ guarantees that promiscuity takes a value in $[0, 1]$. (See proof in Subsection 3.1.2.) The minimum promiscuity $p_i = 0$ indicates that node i has neighbors only in a single layer. Node i has maximum promiscuity $p_i = 1$ if it connects to nodes in each layer in equal proportion. In Subsection 3.1.3, we prove that all nodes

[1] We here show this one-sided limit with L'Hôpital's rule [111]

$$\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{\frac{\mathrm{d}f(x)}{\mathrm{d}x}}{\frac{\mathrm{d}g(x)}{\mathrm{d}x}} . \tag{3.4}$$

In our case, $c = 0^+$, $f(x) = \ln(x)$, and $g(x) = x^{-1}$. The first derivatives are $\frac{\mathrm{d}f(x)}{\mathrm{d}x} = x^{-1}$ and $\frac{\mathrm{d}g(x)}{\mathrm{d}x} = -x^{-2}$. Finally, we obtain

$$\lim_{x \to 0^+} x \ln(x) = \lim_{x \to 0^+} -x = 0 . \tag{3.5}$$

have maximal promiscuity $p_i = 1$ in a certain type of edge-coloured multigraph. The promiscuity of an isolated node (which has multilayer degree $K_i = 0$) is not defined, and in practice we remove such nodes before our computations.

One can examine node coordinates (p_i, K_i) in the parameter plane of values of promiscuity p_i and multilayer degree K_i to give insights into a node's role in a multirelational networks in a way that is similar to prior approaches for modular networks [170, 171, 234]. For the example network in Fig. 3.1a, we show these coordinates in Fig. 3.1b. Nodes A, E, and G are each connected only within a single layer (i.e., with edges of the same colour), so they have promiscuities of 0. We call such nodes 'focused' nodes, because they take part in only one layer of the MLN.

Nodes with promiscuity values $p_i \in (0, 1]$ connect to nodes in at least two layers, and larger values of p_i indicate that the distribution of attached edges is closer to that in the uniform null model. We refer to nodes with larger promiscuity values as 'more promiscuous'. Node D is the most promiscuous node in the network in Fig. 3.1a, because it connects to nodes in all layers with an almost uniform distribution.

Our example network also illustrates that there is not a precise relationship between multilayer degree K_i (which we show on the vertical axis of Fig. 3.1b) and promiscuity p_i. Instead, we observe a variety of (p_i, K_i) coordinates of nodes in this multirelational network, and we find it useful to distinguish promiscuous hubs, focused non-hubs, and nodes with intermediate roles between these extremes. Note that distinguishing a 'hub' node from a 'non-hub' is a difficult issue, and (as has been illustrated with

modular networks [3]) it is not necessarily desirable to use such discrete classes for nodes of different degrees. Following this, we refrain from partitioning the continuum of promiscuities into precise discrete classes.

Promiscuity p_i for multirelational networks that are both weighted and directed

In Subsection 3.1.1, we defined promiscuity p_i for each node i in undirected, unweighted networks. We now give the definition of in-promiscuity $p_i^{(\mathrm{in})}$ and out-promiscuity $p_i^{(\mathrm{out})}$ for each node i in directed, weighted networks.

Let W be the $N \times N$ adjacency matrix of an edge-weighted network with N nodes. The entry $W_{ij} \in \mathbb{R}$ is nonzero if there is an edge that starts from node i and ends at node j, and it is equal to 0 if there is no edge. Each node i has an in-strength of $s_i^{(\mathrm{in})} = \sum_{j=1}^{N} W_{ji}$ and an out-strength of $s_i^{(\mathrm{out})} = \sum_{j=1}^{N} W_{ij}$.

We represent a multirelational network using an edge-coloured multigraph, which consists of a 'stack' of L adjacency matrices $W^{(l)}$, where $l \in \{1, \ldots, L\}$ denotes the lth layer. Thus, $W^{(l)}$ encodes the edge information for layer l (i.e., relationship type l). The in-strength and out-strength vectors of node i are

$$\mathbf{s}_i^{(\mathrm{in})} = \left(s_{i,1}^{(\mathrm{in})}, s_{i,2}^{(\mathrm{in})}, \ldots, s_{i,L}^{(\mathrm{in})}\right) , \tag{3.6}$$

$$\mathbf{s}_i^{(\mathrm{out})} = \left(s_{i,1}^{(\mathrm{out})}, s_{i,2}^{(\mathrm{out})}, \ldots, s_{i,L}^{(\mathrm{out})}\right) , \tag{3.7}$$

where $s_{i,l}^{(\mathrm{in})} = \sum_{j=1}^{N} W_{ji}^{(l)}$ and $s_{i,l}^{(\mathrm{out})} = \sum_{j=1}^{N} W_{ij}^{(l)}$. As in an undirected and unweighted

multirelational network, we construct in-promiscuity and out-promiscuity vectors

$$\mathbf{P}_i^{(\text{in})} = \left(P_{i,1}^{(\text{in})}, P_{i,2}^{(\text{in})}, \ldots, P_{i,L}^{(\text{in})}\right) \text{ , with } P_{i,l}^{(\text{in})} = s_{i,l}^{(\text{in})} / \langle s_l \rangle \text{ ;} \tag{3.8}$$

$$\mathbf{P}_i^{(\text{out})} = \left(P_{i,1}^{(\text{out})}, P_{i,2}^{(\text{out})}, \ldots, P_{i,L}^{(\text{out})}\right) \text{ , with } P_{i,l}^{(\text{out})} = s_{i,l}^{(\text{out})} / \langle s_l \rangle \tag{3.9}$$

by scaling with each layer's mean strength $\langle s_l \rangle = \sum_{i=1} s_{i,l}^{(\text{in})} / N$. Note that we use the same normalization for both the in-promiscuity and out-promiscuity vectors, because the mean in-strength equals the mean out-strength in any directed network.

We then compute normalized promiscuity vectors

$$\mathbf{P'}_i^{(\text{in})} = \mathbf{P}_i^{(\text{in})} \Bigg/ \sum_{l=1}^{L} P_{i,1}^{(\text{in})} \text{ ,} \tag{3.10}$$

$$\mathbf{P'}_i^{(\text{out})} = \mathbf{P}_i^{(\text{out})} \Bigg/ \sum_{l=1}^{L} P_{i,1}^{(\text{out})} \text{ .} \tag{3.11}$$

Finally, we compute the normalized Shannon entropy of these normalized promiscuity vectors to obtain in-promiscuity and out-promiscuity:

$$p_i^{(\text{in})} = - \underbrace{\left(\sum_{l=1}^{L} P_{i,l}^{(\text{in})} \ln P_{i,l}^{(\text{in})} \right)}_{\text{Shannon Entropy} H(\mathbf{P}_i^{(\text{in})})} \Bigg/ \ln (L) \text{ ,} \tag{3.12}$$

$$p_i^{(\text{out})} = - \underbrace{\left(\sum_{l=1}^{L} P_{i,l}^{(\text{out})} \ln P_{i,l}^{(\text{out})} \right)}_{\text{Shannon Entropy} H(\mathbf{P}_i^{(\text{out})})} \Bigg/ \ln (L) \text{ .} \tag{3.13}$$

The normalization by $\ln (L)$ guarantees that both promiscuities, $p_i^{(\text{in})}$ and $p_i^{(\text{out})}$, have values $[0, 1]$. We provide the proof that promiscuity values lie in $[0, 1]$ in Subsection 3.1.2 for undirected and unweighted multirelational networks, but this is also true when we allow directions and/or weights.

3.1.2 Proof that Promiscuity $p_i \in [0,1]$

Promiscuity p_i is defined as the normalized Shannon entropy

$$p_i = \sum_{l=1}^{L} \frac{P_i'^l \ln(P_i'^l)}{\ln(L)} \tag{3.14}$$

of the normalized promiscuity vector $\mathbf{P}_i'$. As we will prove in this section, because the normalized entropy $H(\mathbf{P}_i')/\ln(L) \in [0,1]$, it is also true that $p_i \in [0,1]$. In this section, we prove this result explicitly for undirected, unweighted networks. However, the same result holds for the weighted and directed variants, but we omit it for simplicity.

We start by examining the maximum and minimum of the Shannon entropy $H(\mathbf{x})$ of a normalized vector $\mathbf{x} = (x_1, x_2, \ldots, x_L)$ with L elements. We exponentiate the entropy to obtain

$$\begin{aligned} \exp(H(\mathbf{x})) &= \exp\left(-\sum_{l=1}^{L} x_l \ln(x_l)\right) = \prod_{l=1}^{L} \exp\left(-x_l \ln(x_l)\right) \\ &= \prod_{l=1}^{L} \left(\frac{1}{x_l}\right)^{x_l} \leq \sum_{l=1}^{L} x_l \frac{1}{x_l} = L\,. \end{aligned} \tag{3.15}$$

The inequality arises from the following weighted inequality of arithmetic and geometric means [35, 176]:

$$\prod_{l=1}^{L} y_l^{w_l} \leq \sum_{l=1}^{L} w_l y_l\,, \tag{3.16}$$

which holds for $y_l \geq 0$, $w_l > 0$, $\sum_{l=1}^{L} w_l = 1$.

For the inequality in Eqn. (3.16), the equality occurs, if and only if $y_i = y_j$ for all $i, j = 1, 2 \ldots, L$. Thus, we conclude that entropy has an upper bound of $H_{\max} = \ln(L)$, which it achieves if and only if all $x_l = 1/L$. This maximum

possible entropy is exactly the normalization factor of the normalized entropy, so the maximum promiscuity is $p_i = 1$.

Now we show that the entropy $H(\mathbf{x})$ is minimized if one particular outcome x_l is 1 and all other entries are 0. For this case, we see that the entropy is 0 by calculating

$$H((1,0,\ldots,0)) = -1\ln(1) - (L-1)\times 0 = 0\,. \tag{3.17}$$

The entropy $H(x) = \sum_{l=1}^{L}(-x_l \ln(x_l))$ is nonnegative, because each of the summands $-x_l \ln(x_l)$ is nonnegative for $x_l \in [0,1]$, which holds for a normalized vector $|\mathbf{x}| = 1$. Thus, $H_{\min} = 0$ is indeed the minimum entropy and also the minimum normalized entropy, which we calculate by scaling by a factor of $\ln(L)^{-1}$.

We have shown that the range of the entropy H for normalized vectors is $[0, \ln(L)]$. Using the fact that $\mathbf{P}'_i$ is normalized, we can thus conclude that the promiscuity p_i must lie within the interval $[0,1]$. The result that the promiscuity is normalized also holds for directed and weighted promiscuity definitions, which we gave in Subsection 3.1.1, because they are also defined as normalized Shannon entropies of normalized vectors.

3.1.3 Proof that Promiscuity $p_i = 1$ in Colour-regular Edge-coloured Multigraphs

A graph $G = (V, E)$ is 'regular' if every node has the same number of neighbours, so if $k_i{=}k_j$ for all i, j [4]. We call an edge-coloured multigraph $M = (V, E_1, E_2, \ldots, E_L)$ *colour-regular* if for each edge set E_l, all nodes have the same number of attached edges. Thus, $k_i^l = k_j^l$ for all i, j for each layer $l \in \{1, \ldots, L\}$.

All nodes have the same degree k_i^l in each layer l, so every node has the same

degree vector $\mathbf{k}_i$. The mean degree of each layer is then $\langle k_l \rangle = k_i^l$. Accordingly, the promiscuity vector of each node is

$$\mathbf{P}_i = \left(\frac{k_i^1}{\langle k_1 \rangle}, \frac{k_i^2}{\langle k_2 \rangle}, \ldots, \frac{k_i^L}{\langle k_L \rangle} \right) = (1, 1, \ldots, 1) \,, \tag{3.18}$$

which is the all-ones vector $\mathbf{e}$. The normalized promiscuity vector is then

$$\mathbf{P}'_i = \frac{1}{L}\mathbf{e} \,. \tag{3.19}$$

As we showed in Subsection 3.1.2, this vector yields the maximum promiscuity $p_i = 1$. All nodes have the same promiscuity vector $\mathbf{P}_i$, so they all have the maximum promiscuity $p_i = 1$.

We confirmed the correctness of this result with calculations on synthetic colour-regular networks of various sizes N, numbers L of layers, and edge densities (results not shown). In Section 3.3, we analyse one example network and the influence of a particular randomisation procedure on the promiscuity.

3.2 Promiscuity p_i in Synthetic Networks

In this section, we calculate node promiscuities in two synthetic networks with $L = 2$ layers and derive analytical expressions for p_i. First, in Subsection 3.2.1, we analyse a network with one regular ring lattice layer and one layer with an ER graph in the $G(N, p)$ variant. Second, in Subsection 3.2.2, we analyse a network with both layers being ER graphs. We refer to the former as RL–ER network and to the latter as ER–ER network.

In general, the promiscuity is only a function of the degree vectors $\mathbf{k}_i$ and not of the network topology itself. Therefore, any randomisation of network topology that

preserves the degree vectors $\mathbf{k}_i$ does not change the numerical and analytical results for $p_i(K_i)$. Thus, the results in this section hold for all MLNs that have the same degree distributions in each layer as the ones we construct here. To be precise, these are a binomial distribution for the ER layer and a delta-distribution for the ring lattice network, which we discuss in more detail in Subsection 3.2.1.

3.2.1 Network Composed of an Erdős–Rényi Layer and a Ring Lattice Layer

In Subsection 3.1.3, we showed that an edge-coloured multigraph that consists of regular layers has the same promiscuity $p_i = 1$ for all nodes. We now construct a network that consists of an Erdős–Rényi (ER) graph in one layer and an r-regular ring lattice network, in which all nodes have the same degree r, in the other layer. We call them 'ER' layer and 'ring lattice' layer. In the ER layer, each of the $N(N-1)/2$ possible (undirected) edges exists with connection probability p; this network model is commonly called the $G(N,p)$ variant of the ER graph. The $G(N,m)$ variant is another ER graph model; see Section 2.4 for a discussion of both. The degree distribution in the ER layer follows a binomial distribution

$$P(k_i^1 = k) = \binom{N-1}{k} \rho^k (1-p)^{N-1-k} \,. \tag{3.20}$$

We construct a ring lattice layer by creating a ring lattice network with all nodes having degree $k_i^2 = r$. The degree distribution

$$P(k_i^2 = k) = \begin{cases} 1, & \text{for } k = r \,, \\ 0, & \text{otherwise} \,, \end{cases} \tag{3.21}$$

is thus a δ-function.

This particular network allows us to investigate promiscuity p_i as a function of the degree k_i^1 in the ER layer. First, we discuss our computational results, and we then derive an analytical expression for $p_i(K_i)$.

In Fig. 3.2, we show (p_i, K_i) for all nodes i in a single network with $N = 10,000$ nodes. The network consists of a ring lattice layer with $r = 4$, and an ER layer with connection probability $p = 10^{-3}$. Nodes with no edges in the ER layer have the minimum multilayer degree $K_i = r = 4$ and a promiscuity of $p_i = 0$. In contrast, nodes that connect with an edge in the ER layer have a nonzero promiscuity $p_i > 0$.

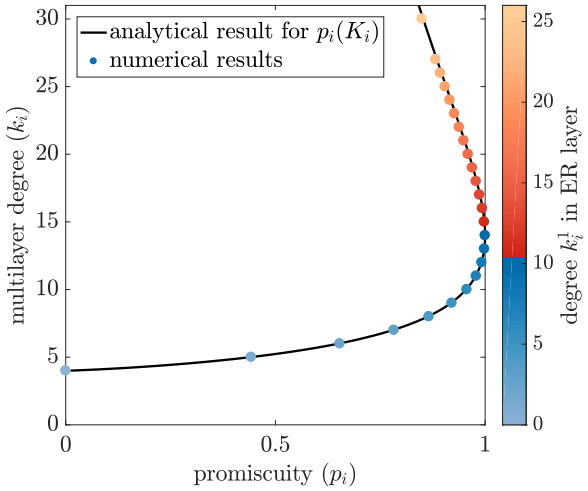


Figure 3.2: Promiscuity p_i and multilayer degree K_i for a synthetic network consisting of an ER layer and a ring lattice layer. Promiscuity p_i and multilayer degree K_i for a synthetic network consisting of an ER layer and a ring lattice layer.

For nodes with ER-layer degrees $k_i^1 < \langle k_1 \rangle$, the promiscuity p_i increases with increasing degree (see the blue-shaded disks). A node i reaches a maximum promiscuity of $p_i \approx 1$ when its ER-layer degree is close to the mean degree in this layer ($\langle k_1 \rangle \approx 9.9$).

For nodes with ER-layer degrees $k_i^1 > \langle k_1 \rangle$, the promiscuity p_i decreases with increasing degree (see the red-shaded disks).

We want to derive an analytical expression for promiscuity $p_i(K_i)$ with respect to the multilayer degree K_i. In this particular two-layer network, the promiscuity vector $\mathbf{P}_i$ is

$$\mathbf{P}_i = \left(\frac{k_i^1}{\langle k_1 \rangle}, \frac{k_i^2}{\langle k_2 \rangle} \right) \approx \left(\frac{k_i^1}{(N-1)p}, \frac{r}{r} \right) , \tag{3.22}$$

where we approximate the mean degree $\langle k_1 \rangle$ in the ER layer by $(N-1)p$, which is appropriate for large N. For the normalized promiscuity vector $\mathbf{P'}$, we thus calculate that

$$\mathbf{P}_i' = \left(\frac{x}{1+x}, \frac{1}{1+x} \right) , \quad \text{with} \quad x = \frac{k_{i,1}}{(N-1)p} \in [0, 1/p] . \tag{3.23}$$

After some algebra, we then obtain

$$p_i(x) = \frac{1}{\ln(2)} \left(\ln(1+x) - \frac{x \ln(x)}{1+x} \right) . \tag{3.24}$$

To replace x with multilayer degree K_i, we use

$$K_i = k_i^1 + k_i^2 = x(N-1)p + r \Rightarrow x = \frac{K_i - r}{(N-1)p} . \tag{3.25}$$

We substitute Eqn. (3.25) into Eqn. (3.24) and obtain

$$p_i(K_i) = \frac{1}{\ln(2)} \left(\ln\left(1 + \frac{K_i - r}{(N-1)p} \right) - \frac{\frac{K_i - r}{(N-1)p} \ln\left(\frac{K_i - r}{(N-1)p} \right)}{1 + \frac{K_i - r}{(N-1)p}} \right) , \tag{3.26}$$

which is indistinguishable from the numerical results in Fig. 3.2. Potential discrepancies, which differ by less than 0.5 % for our example network with $N = 10,000$ nodes, arise from the approximation $\langle k_1 \rangle \approx (N-1)p$.

We also want to examine analytically when the promiscuity p_i achieves its maximum as we vary x. To do this, we compute the first derivative of Eqn. (3.24) with respect to x to obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}p_i(x) = -\frac{\ln(x)}{(1+x^2)\ln(2)}\,. \tag{3.27}$$

Setting $\frac{\mathrm{d}}{\mathrm{d}x}p_i(x)$ to 0 yields a potential maximum promiscuity $p_{\max} = 1$ for $x = 1$. We verify that this a maximum by calculating the second derivative of Eqn. (3.24) with respect to x as

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}p_i(x) = -\frac{(1+x^2) - 2x^2\ln(x)}{x(1+x^2)\ln(2)}\,, \tag{3.28}$$

and evaluating it at $x = 1$ yields

$$\left.\frac{\mathrm{d}p_i(x)^2}{\mathrm{d}x^2}\right|_{x=1} = -\frac{1}{\ln(2)} < 0\,, \tag{3.29}$$

where $p_{\max} = 1$ is a local maximum.

Therefore, a node that has exactly the mean degree $k_i^1 = \langle k_1 \rangle$ in the ER layer has maximum promiscuity. This matches the numerical result in Fig. 3.2, for which promiscuity p_i has a maximum of $p_i \approx 1$ at $K_i = r + k_i^1 \approx 14$. In practice, however, degree is a discrete variable, so we may not exactly achieve the maximum $p_i = 1$.

As noted at the beginning of Section 3.2, the analytical expression Eqn. (3.26) for the promiscuity p_i as a function of K_i does not depend on the particular topology of both layers but exclusively on the degree vectors $\mathbf{k}_i$. We verify this on this network by running a randomisation algorithm: We select any two edges $(i,j) \subset E_l$ and $(i',j') \subset E_l$

uniform at random without replacement, delete them, and create new edges (i, j') and (i', j) in layer l. (If the creation of the new edges is not possible because either edge already exist, we restore the old edges (i, j) and (i', j').) This procedure randomises the topology of the layers but preserves the degree vectors $\mathbf{k}_i$ for all nodes i. For any given realisation of an edge-coloured multigraph, we see that (p_i, K_i) is indistinguishable from the result without randomisation in Fig. 3.2 (results not shown). This is only true, however, if we randomise the same realisation of the network. In the case of the creation of an entirely new MLN from the same model we might receive slightly different (p_i, K_i) coordinates as the degree distribution in the ER layer is non-deterministic.

3.2.2 Network Composed of Two Erdős–Rényi Layers

In this section, we discuss a network that consists of two ER layers, each of which is a $G(n, p)$ graph, as discussed in Subsection 3.2.1. The two layers have different densities, and we refer to the one with connection probability $p_1 = 10^{-3.2}$ as the 'sparse' layer and the one with connection probability $p_2 = 5 \times 10^{-3}$ as the 'dense' layer. Both degrees k_i^1 and k_i^2 are now random variables, and each of them follows a binomial distribution. Additionally, the two layers are independent of each other.

In Fig. 3.3, we show (p_i, K_i) for all $N = 10,000$ nodes i in a single two-layer ER network. Similar to the RL–ER network, the promiscuity spans the whole possible range of $[0, 1]$. We observe a larger number of (p_i, K_i) combinations in the ER–ER network in comparison to the RL–ER network. This occurs because both degrees, k_i^1 and k_i^2, are random variables. To understand the situation, we colour-coded the

variability in the degree k_i^2 in the dense layer and plotted nodes whose degree in the sparse layer is below the mean as circles and those with degree above the mean as triangles. We show the analytical result from Eqn. (3.26) as a solid black curve. We derived this formula for the RL–ER network. It is nevertheless valid in the ER–ER network for the subset of all nodes that fulfil the condition $p_2^l = 1$, as then Eqn. (3.22) is valid. This condition is fulfilled for nodes with mean degree $k_i^2 = \langle k_2 \rangle$ in the sparse layer. It separates blue-shaded nodes from red-shaded nodes.

We observe that for nodes with $k_i^1 < \langle k_1 \rangle$ (see the circles), an increase in degree reduces the promiscuity p_i. By contrast, for nodes with $k_i^1 > \langle k_1 \rangle$ (see the triangles), an increase in degree increases p_i. Nodes with degree exactly $\langle k_1 \rangle$ have a maximum promiscuity $p_i \approx 1$ at $k_i^2 = \langle k_2 \rangle$.

To understand this situation, we first discuss nodes that follow the solution (3.26) for the RL–ER network and thus have $k_i^2 = \langle k_2 \rangle$. Nodes with $k_i^1 = 0$ have promiscuity $p_i = 0$ and the analytical approximation agrees with this value at the mean degree $\langle k_2 \rangle \approx 50$. The degree k_i^2 is distributed binomially, so there is some variability, but all nodes have the minimum promiscuity $p_i = 0$. Nodes with $k_i^1 = 1$ have a larger promiscuity ($p_i \approx 0.6$). The variability in k_i^2 influences the promiscuity p_i: nodes with degree larger than the mean have smaller promiscuities, and those with degree smaller than the mean have larger promiscuities. Qualitatively, this behaviour is similar for larger degrees k_i^1 in the sparse layer (e.g., the case $k_i^1 = 2$). Thus, promiscuity p_i decreases with increasing k_i^2. This effect inverts for nodes with degree above the mean

(see the triangles): nodes with smaller k_i^2 also have a smaller promiscuity p_i.

Overall, most nodes in this two-layer ER network have a large promiscuity p_i. However, some of the nodes have the minimum promiscuity $p_i = 0$. The relation between degree K_i and promiscuity p_i is less clear in this network than for networks with a regular layer and an ER layer. For networks with a larger number L of layers, the relationship will be even more complicated, as the promiscuity p_i summarizes a degree vector of L elements rather than only 2.

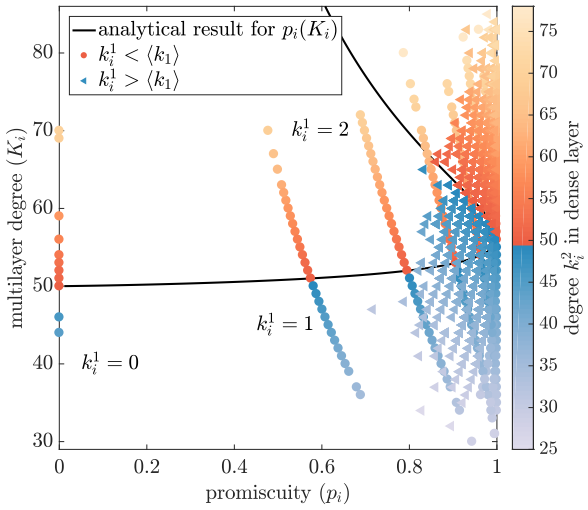


Figure 3.3: Promiscuity p_i and multilayer degree K_i for a synthetic two-layer network that consists of a sparse ER layer and a dense ER layer. We represent nodes with degree $k_i^1 < \langle k_1 \rangle$ in the sparse layer as circles and nodes in that layer with degree $k_i^1 > \langle k_1 \rangle$ as triangles. Colour indicates the degree k_i^2 in the dense layer. The solid black curve is the analytical result Eqn. (3.26) for a network that consists of an ER layer and a regular layer. Note that this curve separates blue-shaded nodes (with $k_i^2 < \langle k_2 \rangle$) from red-shaded nodes (with $k_i^2 \geq \langle k_2 \rangle$).

3.2.3 Conclusion for Synthetic Two-Layer Networks

In this section, we showed that synthetic two-layer networks can have nodes with promiscuity in the full domain of $[0, 1]$. In the case where one layer is regular, we

derived a closed-form expression of the function $p_i(K_i)$. For this analytical result, we showed that the maximum promiscuity $p_i \approx 1$ is achieved for nodes that connect to all layers in proportion to their density. We defined the promiscuity to show this behaviour. The calculations on synthetic networks also provide a sanity check for the behaviour of promiscuity. For the network consisting of two ER layers, we observe, that the relationship $p_i(K_i)$ is less clear than for networks with a regular layer and an ER layer. We found, however, that the analytical expression $p_i(K_i)$ from Eqn. (3.26) is valid for a subset of nodes, which fulfil $k_i^2 = \langle k_2 \rangle$.

3.3 Promiscuity p_i in Synthetic Networks under Rewiring

In this section, we show that rewiring MLNs can drastically change the promiscuity distribution $P(p_i)$. We illustrate this phenomenon with a specific example. We start with a network that, by construction, has $p_i = 0$ for all nodes. We then rewire a fraction β of its edges and compute the promiscuity distribution $P(p_i)$. We compare the numerically observed mean promiscuity $\langle p_i \rangle$ with an analytical expression.

3.3.1 Network Construction

In any edge-coloured multigraph in which each node has edges of only one colour, all nodes have the minimum promiscuity $p_i = 0$ by construction (see proof in Subsection 3.1.2). We construct such networks in the form of non-overlapping ring-lattice graphs as follows. For each layer, we construct a r-regular ring-lattice with

N_{ring} nodes (see Section 2.4 for a definition of the ring-lattice network). The union of these L ring-lattice networks into one network yields a graph of $N = L \times N_{\text{ring}}$ nodes. By keeping the layer information as edge colour, we construct an edge-coloured multigraph. In Fig. 3.4a, we show the adjacency matrix of such a network with $L = 5$ layers and $N_{\text{ring}} = 100$ nodes in each layer. This adjacency matrix is the direct sum of the adjacency matrices of the graphs in the individual layers.

3.3.2 Randomization

We randomize the MLN independently in each layer. For each layer l, we take a fraction of β of all existing edges $(i, j) \in E_l$, remove it, and then assign it new starting and terminal nodes i' and j'. We choose the new edges (i', j') uniformly at random from all non-edges $(i', j') \notin E_l$. The same edge can not be rewired twice because we choose the set of edges that we rewire once at the beginning. This is in contrast to an iterative randomisation. Note that this randomization does not preserve the multilayer degree distribution $P(K_i)$; it preserves only the total number of edges in each layer. In Figs. 3.4, we show the supra-adjacency matrices of the multilayer ring lattice after rewiring fractions $\beta = 0.05$ and $\beta = 0.95$ of its edges. The ring-lattice structure (mostly) disappears by the time that $\beta = 0.95$ and nodes are attached increasingly to edges from different layers.

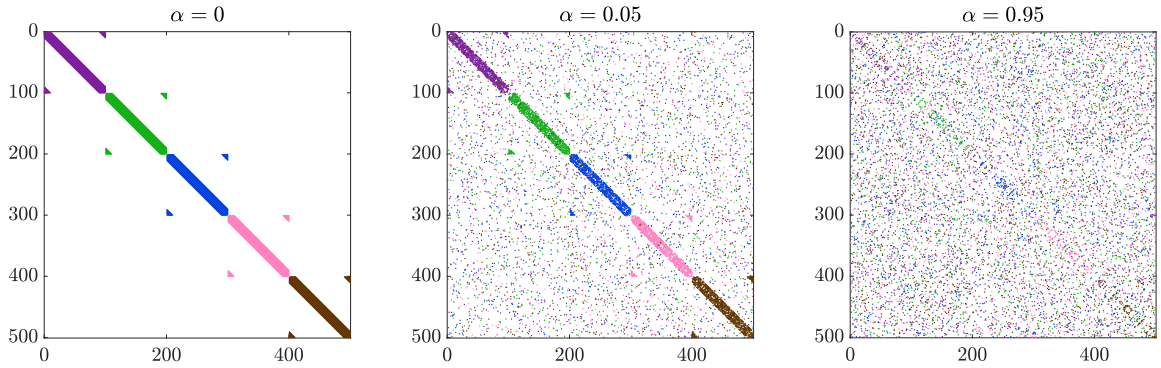


Figure 3.4: Supra-adjacency matrices of a ring-lattice MLN under randomization. (a) The network consists of $L = 5$ layers (which we indicate using distinct colours) with $N_{\text{ring}} = 100$ nodes in each layer. (Thus, there are a total of $N = 500$ nodes.) These nodes are connected as a ring lattice in which each node i has degree $k_i = 10$. (b) We rewire, uniformly at random, a fraction $\beta = 0.05$ of the edges in the original network. For details of the rewiring process, see Subsection 3.3.2. (c) We randomly rewire a fraction $\beta = 0.95$ of the edges in the original network.

3.3.3 Promiscuity p_i under Randomization

We investigate the change of the promiscuity distributions $P(p_i)$ of a network with uniform promiscuity $p_i = 0$, as introduced in Subsection 3.3.1, under the rewiring from Subsection 3.3.2. We construct multilayer ring-lattice networks with $N_{\text{ring}} = 1\,000$ nodes of degree $k_i = 100$ in each of the $L = 5$ layers. In Fig. 3.5a, we show the change of promiscuity p_i under randomization of a fraction $\beta \in [0, 1]$ of all edges. In the original network (i.e., with $\beta = 0$), each node i has a promiscuity of $p_i = 0$, as one can see in the green violin plot. In the violin plot, we smooth the distribution with a Gaussian kernel-density estimator with a smoothing bandwidth $w = 0.1$ [23, 193]. We show the mean promiscuity $\langle p_i \rangle = \sum_{i=1}^{N} p_i$ as a purple curve. It increases under randomization and approaches $\langle p_i \rangle \approx 1$ for $\beta = 1$, which indicates randomization of all edges. In this case the layers approximately resemble ER graphs. They do not exactly

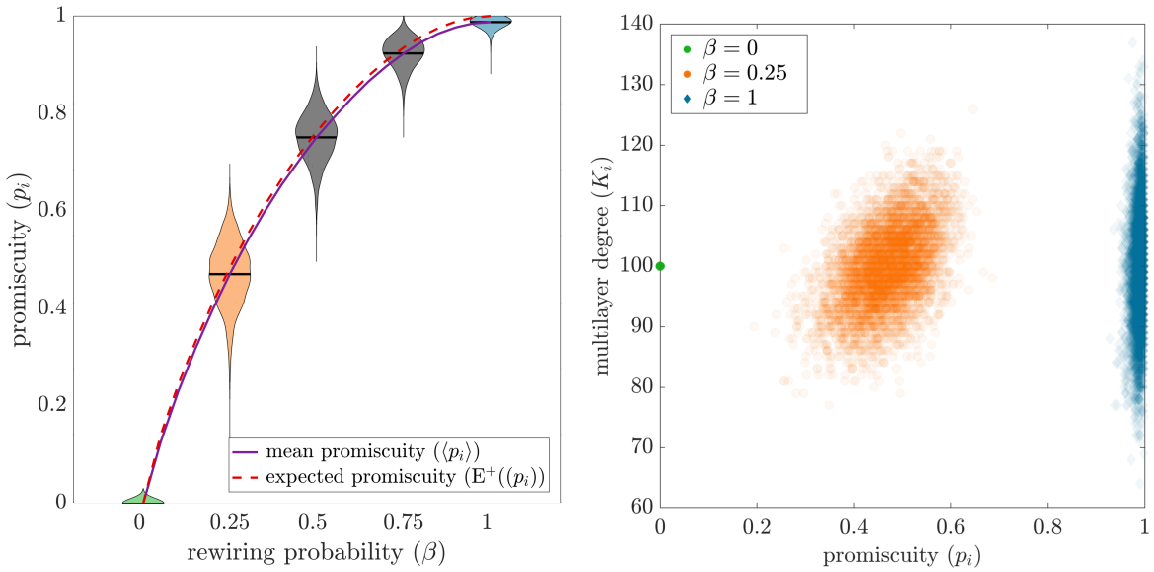


Figure 3.5: Promiscuity p_i of a ring-lattice MLN under randomization. (a) Violin plots of promiscuity distributions $P(p_i)$ of a ring-lattice network under randomization for rewiring parameter $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$. We show the numerically calculated mean promiscuity $\langle p_i \rangle$ versus β as a solid purple curve and the expected promiscuity $\mathrm{E}(p_i(\beta))$, calculated analytically using Eqn. (3.41), as a dashed red curve. (b) Promiscuity p_i and multilayer degree K_i for all nodes in randomized networks for rewiring parameter $\beta \in \{0, 0.25, 1\}$. We construct and rewire these networks in the same way as the ones that we described in Fig. 3.4, but now we use $N_{\mathrm{ring}} = 1,000$ nodes of degree $k_i = 100$ in each of the $L = 5$ layers. We make nodes translucent to make it easier to see node overlaps.

resemble ER as each edge is only randomised exactly once: Therefore, we know that the edge randomised last is with certainty not present; in an ER graph this is not the case. The promiscuity distribution $P(p_i)$ widens under randomization, as we show in orange for $\beta = 0.25$, and it narrows again towards $\beta = 1$ (see the blue points).

In Fig. 3.5b, we show the promiscuity p_i and multilayer degree K_i of all nodes for the the original network ($\beta = 0$; solid green disk) and two randomized versions (as orange circles for $\beta = 0.25$ and blue diamonds for $\beta = 1$). In the original network, all nodes have $(p_i, K_i) = (0, 100)$, by construction. For $\beta = 0.25$, the promiscuity p_i and multilayer degree K_i are spread out over a wide variety of values. For $\beta = 1$, the multilayer degree

is even wider, but the spread of the promiscuity is smaller, as discussed earlier.

We now derive an analytical expression of the mean promiscuity $\langle p_i \rangle(\beta)$. For this, we investigate the degree vector $\mathbf{k}_i$ of node i. Without rewiring, the degree of node i is k_i and all attached edges are exactly in one layer, by construction.

Without loss of generality, the degree vector of node i is then

$$\mathbf{k}_i(\beta = 0) = \left(k_i, \underbrace{0, \ldots, 0}_{L-1 \text{ elements}} \right) . \tag{3.30}$$

'Without loss of generality' refers to the fact that the nonzero element of $\mathbf{k}_i$ is not necessarily the first one. The promiscuity, however, is invariant under permutation of the degree vector, so we assume this particular order for notational simplicity.

To investigate the effect of rewiring on the degree vector $\mathbf{k}_i$ of node i, we distinguish between two random processes, which occur as consequence of the rewiring: the removal of edges connecting to i and the attachment of new edges to i. We consider both processes to be independent of each other and we compute the expectancy values $\mathrm{E}_{\text{removal}}(\Delta\mathbf{k}_i(\beta))$ and $\mathrm{E}_{\text{attach}}(\Delta\mathbf{k}_i(\beta))$ of the change in degree vector $\Delta\mathbf{k}_i(\beta)$ separately for the two processes.[2] We then combine the two results and obtain an expected value of the promiscuity vector under rewiring. Finally, we use Jensen's inequality to obtain an upper bound of the expected promiscuity.

We begin by treating the edge rewiring as a sequence of $n \in \mathbb{N}$ independent events with 'success' probability $q \in [0, 1]$. The notion of what a success entails is

[2] In reality there is small dependence of both processes because an edge (i, j) can not be reattached to the nodes i and j it was removed from. In the limit of large graphs $N \to \infty$, however, this effect is negligible.

different for removal and attachment processes and we will describe it below. We can use the binomial distribution

$$\Pr(X = s) = \binom{n}{s} q^s (1-q)^{n-s} \tag{3.31}$$

to find the probability of getting exactly s successful events. Keep in mind that for a random variable $X \sim \mathrm{B}(n, q)$ that follows the binomial distribution with parameters n and q, the expected value of X is

$$\mathrm{E}(X) = n \times q\,. \tag{3.32}$$

We now analyse the removal process. For the removal process, a 'success' is the removal of an edge of layer l to node i. Each of the k_i edges incident to i has a probability β of being removed. Each of these removal steps is independent of the others, so the number $-\Delta k_i^1$ of successful removals follows a binomial distribution $-\Delta k_i^1 \sim \mathrm{B}(k_i, \beta)$. By construction, each of these edges is initially in the same layer, so no edges in other layers can be removed; this gives $\Delta k_i^l = 0$ for all $l \neq 1$. Therefore, the expected value of the removal process is

$$\mathrm{E}_{\mathrm{removal}}(\Delta \mathbf{k}_i(\beta)) = \left(-k_i\beta,\ \underbrace{0, \ldots, 0}_{L-1 \text{ elements}} \right)\,. \tag{3.33}$$

To obtain an expression for the attachment of new edges to i, we investigate each layer independently. For the attachment process, a 'success' is the attachment of a new edge of layer l to node i. In each layer there are $k_i N_{\mathrm{ring}}/2$ edges. For an edge to potentially become attached to node i the edge has to become rewired first,

for which the probability is β. The probability of attaching such a rewired edge to node i is $1/N + 1/(N-1) \approx 2/N = 2/(L \cdot N_{\text{ring}})$. Accordingly, the number of edges that become attached to node i in layer l follows the binomial distribution $\text{B}(k_i N_{\text{ring}}/2, 2\beta/(L \cdot N_{\text{ring}}))$. The expected value of the number $\Delta k_i^l(\beta)$ of new edges of layer l to become attached to node i is then

$$\text{E}_{\text{attach}}(\Delta k_i^l(\beta)) = \beta \frac{k_i N_{\text{ring}}}{2} \frac{2}{L N_{\text{ring}}} = \frac{\beta k_i}{L} \,. \tag{3.34}$$

As this is identical for each layer, we obtain

$$\text{E}_{\text{attach}}(\Delta \mathbf{k}_i(\beta)) \approx \left(\frac{\beta k_i}{L}, \underbrace{\frac{\beta k_i}{L}, \ldots, \frac{\beta k_i}{L}}_{L-1 \text{ elements}} \right) , \tag{3.35}$$

as the expected change of the degree vector of node i due to the attachment process.

To obtain an expected value $\text{E}(\mathbf{k}_i(\beta))$ under the rewiring process, we combine the degree vector $\mathbf{k}_i(\beta = 0)$ without rewiring with the expected values of both processes:

$$\begin{aligned} \text{E}(\mathbf{k}_i(\beta)) &= \mathbf{k}_i(\beta = 0) + \text{E}_{\text{removal}}(\Delta \mathbf{k}_i(\beta)) + \text{E}_{\text{attach}}(\Delta \mathbf{k}_i(\beta)) \\ &= \left(k_i(1-\beta) + \frac{\beta k_i}{L}, \underbrace{\frac{\beta k_i}{L}, \ldots, \frac{\beta k_i}{L}}_{L-1 \text{ elements}} \right) . \end{aligned} \tag{3.36}$$

The removal process decreases the number of edges incident to i and the attachment increases it. The expected multilayer degree $\text{E}(K_i(\beta)) = k_i$, however, is unchanged under the rewiring. The mean degree in each layer is $k_i N_{\text{ring}}/(2N) = k_i/(2L)$, so the expected promiscuity vector of node i is

$$\text{E}(\mathbf{P}_i(\beta)) = \left(2L(1-\beta) + 2\beta, \underbrace{2\beta, \ldots, 2\beta}_{L-1 \text{ elements}} \right) , \tag{3.37}$$

and the expected normalized promiscuity vector of node i is

$$\mathrm{E}(\mathbf{P}'_i(\beta)) = \left(1-\beta+\beta/L, \underbrace{\beta/L, \ldots, \beta/L}_{L-1 \text{ elements}}\right) . \tag{3.38}$$

Using the definition of promiscuity, Eqn. (3.3), we obtain an expected value of the promiscuity of node i of

$$\mathrm{E}(p_i(\beta)) = -\frac{1}{\ln(L)}\mathrm{E}(\sum_{l=1}^{L}\left(f(P_i'^l(\beta)\right)), \tag{3.39}$$

with $f(x) = x\ln(x)$. The layers are independent of each other, so we can use $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ to obtain

$$\mathrm{E}(p_i(\beta)) = -\frac{1}{\ln(L)}\sum_{l=1}^{L}\left(\mathrm{E}(f(P_i'^l(\beta))\right) . \tag{3.40}$$

To further analyse the expected promiscuity, we apply Jensen's inequality, which requires $f(x)$ to be convex [215]. The function $f(x)$ is convex, because a twice differentiable function $f(x)$ of one variable x is convex on an interval (a, b) if and only if its second derivative is nonnegative there. Computing the second derivative yields $\mathrm{d}f(x)^2/\mathrm{d}x^2 = 1/x > 0$ for all $x \in [0, \infty)$. Thus, we can apply Jensen's inequality $f(\mathrm{E}(X)) \leq \mathrm{E}(f(X))$ to obtain an upper bound for the expected promiscuity of

$$\mathrm{E}(p_i(\beta)) \leq \mathrm{E}^+(p_i(\beta)) = \frac{1}{\ln(L)}\left(-(L-1)\frac{\beta}{L}\ln(\frac{\beta}{L}) - (1-\beta+\beta/L)\ln(1-\beta+\beta/L)\right) . \tag{3.41}$$

The case of complete rewiring, $\beta = 1$, yields an upper bound of the expected promiscuity $\mathrm{E}^+(p_i(\beta = 1)) = 1$. Without rewiring, $\beta = 0$, the terms $\frac{\beta}{L}\ln(\frac{\beta}{L})$ are taken to be zero, and we obtain $\mathrm{E}^+(p_i(\beta = 0)) = 0$.

We illustrate this upper bound $\mathrm{E}^+(p_i(\beta))$ as a dashed red curve in Fig. 3.5. The expected promiscuity's upper bound $\mathrm{E}^+(p_i(\beta))$ increases monotonously with β. It agrees well with the numerically calculated mean promiscuity $\langle p_i \rangle(\beta)$ and is an upper bound for it for all β. The larger the rewiring β, however, the larger the discrepancy between the numerical mean and the expected value's upper bound. This occurs, because with each rewiring step, the variance $n \times q(1-q)$ of the binomial distribution increases. This results in more nodes having degree vectors less similar than the expected one $\mathrm{E}(\mathbf{k}_i(\beta))$. Therefore, the expected promiscuity $\mathrm{E}^+(p_i(\beta))$ is less representative of the promiscuity values across the population of all nodes.

3.4 Promiscuity p_i in Empirical Networks

We illustrate how promiscuity can reveal insightful information about the organization of multirelational networks. We calculate promiscuity distributions for eleven empirical networks (which we construct from ten data sets; see Table 3.1) of different types: transportation, economic trade, social, and biological. We discuss the construction of 'human gene' and 'protein regulation' network in Section 3.5, where we analyse them also in more detail. For information about the remaining networks' construction see Appendix E.

Two of the networks, the World Trade Web (a trade network contains both exports and imports) and the advice aspect of a cognitive social structure (CSS; 'advice given' and 'advice received'), are directed networks, so we calculate both the in-degree and the out-degree promiscuities for the nodes in these examples. To give a coarse comparison,

network	nodes	layers	number nodes ($N = \|V\|$)	number of layers (L)	edge type	references
UK transportation	stops	modes of transportation	267,031	6	unweighted, undirected	[155]
London transportation	train stations	Underground, Overground, DLR	369	3	unweighted, undirected	[109]
air transportation	airports	airlines	3,182	540	unweighted, undirected	[356]
Aarhus	individuals	social interactions	61	5 0	unweighted, undirected.	[286]
international trade	countries	commodities	133	81	weighted, directed	[244]
cognitive social structure	individuals	individuals' perceptions	21	21	unweighted	[240]
Friendship	–	–	–	–	undirected	
Advice	–	–	–	–	directed	
protein regulation	transcription factors	cell types	538	41	unweighted, undirected	[326]
human brain	brain regions	individuals	998	5	weighted, undirected	[172]
human gene	genes	tissues	16,974	32	unweighted, undirected	[291]
protein–protein interactions	proteins	experiments	21,412	13	unweighted, undireced	[432]

Table 3.1: Description of the ten examined data sets and the real-world objects that are represented by the nodes, edges, and layers in the multirelational networks. We give the number $N = |V|$ of nodes, the number L of layers, edge type (unweighted or weighted, undirected or directed), and a citation to the associated data. The *cognitive social structure* data set consists of two multirelational networks: an undirected friendship network and a directed advice network, which we treat as two independent multirelational networks. The DLR layer in the *London transportation network* stands for 'Docklands Light Railway'.

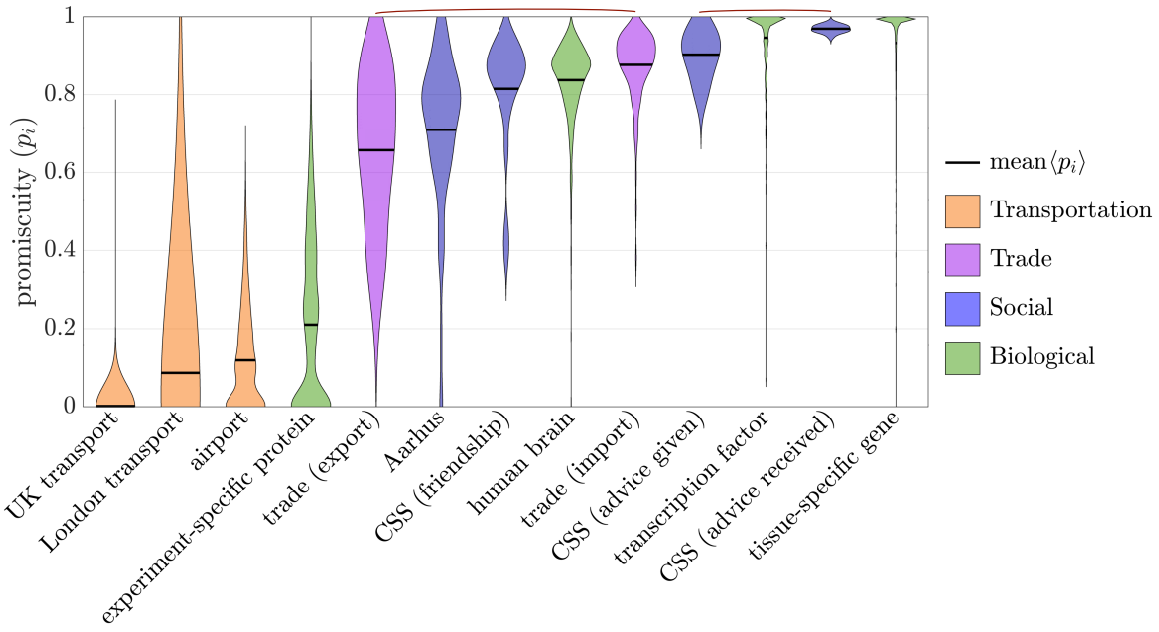


Figure 3.6: Violin plots [193] of the promiscuity distribution $P(p_i)$ across all nodes in eleven different networks. Two networks (trade and CSS advice) are directed networks, so we compute in-promiscuity and out-promiscuity in these cases. We use red arcs on top to associate distributions that arise from the same network. We estimate a continuous distribution for the promiscuity p_i with a normal-distribution kernel estimator. We indicate the mean promiscuity $\langle p_i \rangle$ for each network with a black horizontal line, and we use this value to sort the networks from small values ($\langle p \rangle_{\text{UK}} \approx 10^{-3}$) to large ones ($\langle p \rangle_{\text{gene}} \approx 0.93$). Colour indicates the type of data that we use for network construction.

we first illustrate that the promiscuity distribution differs across networks.

For each network, we calculate the promiscuity p_i of each node i and estimate a continuous distribution for the whole network by multiplication with a normal distribution kernel. (see Fig. 3.6). The violin plots [193] show the networks' normalized distribution of promiscuity p_i and mean promiscuity $\langle p_i \rangle$. The networks have very different mean promiscuities, ranging from $\langle p \rangle_{\text{UK}} \approx 10^{-3}$ (which is barely above the minimum value) to very large values, culminating with $\langle p \rangle_{\text{gene}} \approx 0.93$. Additionally, some distributions are tightly peaked (e.g., advice received in the CSS data set), and

others have a much broader distribution (e.g., trade exports). This illustrates that empirical multirelational networks can have very different promiscuity distributions. As we illustrated in Section 3.2, this is also true for synthetic networks.

The three transportation networks [109, 155, 356], which we show in orange, have broad promiscuity distributions, with many nodes with small promiscuity. Such small-promiscuity nodes are stations that incident predominantly by edges in a single layer. One can construe the large-promiscuity nodes as connector stations that facilitate changes between different modes of transportation. These transportation networks include both 0-promiscuity nodes (most of the nodes), which enable traffic along one mode of transportation, and intermediate-promiscuity nodes (which serve as connectors between different modes). The global air transportation network has many airports of minimum promiscuity $p_i = 0$ and a broad tail of larger-promiscuity nodes. These are international hub airports that are used by many different airlines.

The social networks that we examined have intermediate to large mean promiscuities $\langle p_i \rangle$. *Aarhus* consists of five kinds of online and offline relationships (Facebook, leisure, work, co-authorship, and lunch) between the employees of the Department of Computer Science at Aarhus University [286], and nodes span the whole promiscuity range. The other three promiscuity distributions come from the same *cognitive social structure* (CSS) network [240], which encodes the social structure of $N = 21$ people in a group. Each of the $L = 21$ layers represents the social structure as perceived by one person. The data includes both an undirected friendship network and a directed advice network.

The large mean promiscuity of the CSS suggests that different individuals perceive the degrees of nodes in the social structure as similar to each other. Two people have mean promiscuities much smaller than the mean in the CSS friendship network, as they estimate the number of their friends to be much larger than that which is reported by the other participants.

The import and export trade networks have rather different promiscuity distributions. The export trade network has a smaller mean and is distributed more broadly. Most countries have a large import promiscuity, indicating that they import a wide variety of goods.

The tissue-specific biological networks (tissue-specific gene [291] and transcription factor [326]) have large mean promiscuities $\langle p_i \rangle$ but broad distributions. This is also the case for the brain-connectivity data across individuals ('human brain') [172] but not for the experiment-specific protein [432] network. The latter suggests that the interactions of many proteins were detected using only a subset of available experimental methods.

3.5 Promiscuity in Tissue-specific Protein Interaction Networks

Genes can function in different ways in different tissues [165]. For example, genetic diseases can manifest in certain tissues, even though the same genome is present across all tissues [246]. This indicates that it is important to conduct tissue-specific analysis of genes and how they influence each other.

Transcription factors (TFs) are a specific type of proteins, which influence the

synthesis of other proteins from genetic information and may thereby exert a strong influence on the development of a cell [252]. It is known that some TFs are expressed ubiquitously across human tissues, whereas others are tissue-specific, either in their expression levels or whether they are expressed at all [465]. One can use such tissue-specific TFs as biomarkers of cell fate — e.g., as a prognostic tool for detecting breast cancer [272].

TFs fulfil their function in a complex regulatory network [270, 313, 481]. It is common belief that 'master regulators' are topologically central TFs in these networks [258, 385] and drive cell differentiation [75, 494].

Different tissues, however, are controlled by different TF networks [326]. Therefore, gene expression across a whole organism is controlled by multiple tissue-specific networks, which can be described as a MLN.

In this section, we examine two tissue-specific gene regulation networks. First, in Subsection 3.5.1, we examine the network of regulation between TFs across tissues. Second, in Subsection 3.5.2, we examine a network of regulation between genes (not exclusively those producing TFs) across tissues, which we create from a different data set.

3.5.1 Tissue-specific Transcription Factor Regulation

In this subsection, we examine the variability of the importance of TFs across different tissue types by computing the promiscuity p_i of nodes in such a MLN.

Network Construction

To construct a MLN we use data from [326]. Neph et al. (2012) [326] used DNase footprinting to determine whether a TF interacts with the transcriptional start site of other TF genes; this is called 'cross regulation'. Such cross-regulation was measured with *DNase footprinting* [60, 153] across $L = 41$ diverse tissue types. DNase footprinting allows the detection of protein–DNA interactions. Neph et al. measured the pairwise interactions, i.e., the cross-regulation, between $N = 538$ TFs. [3]

From this data we create an edge-coloured multigraph in which the nodes represent TFs and the layers the interactions in different tissues. Thus, the multigraph consists of $N = 538$ nodes and $L = 41$ layers. An undirected edge exists between nodes i and j in layer l if they influence each others expression in the tissue type represented by layer l.

[3] The paper [326] describes 475 sequence-specific TFs. Their publicly available data set, however, has 538 unique TFs.

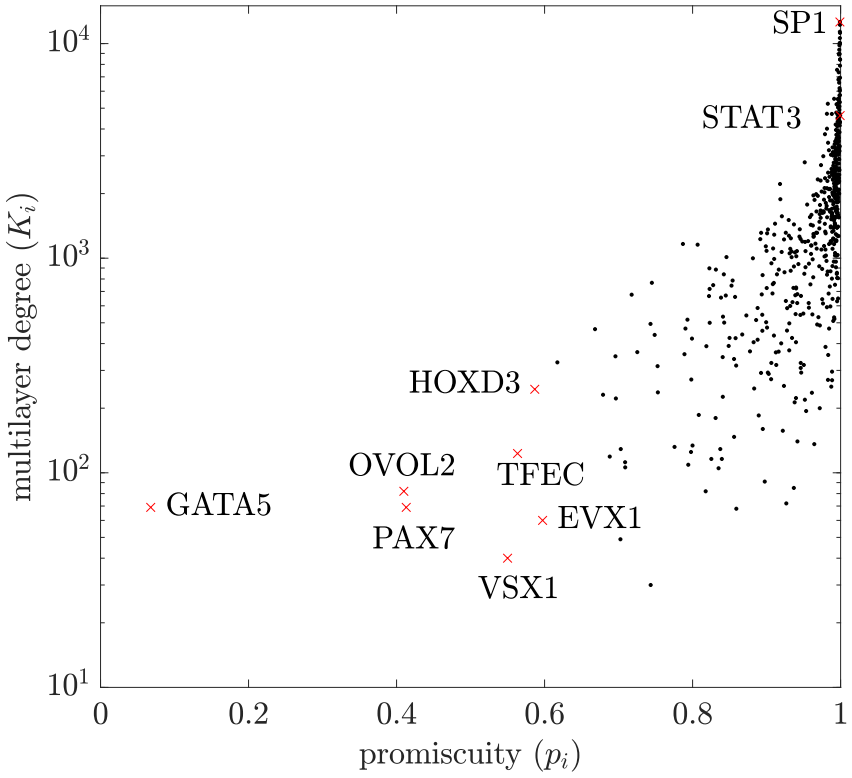


Figure 3.7: Multilayer organisation of transcription factors in human cells. We illustrate TFs at their degree–promiscuity (K_i, p_i) coordinates. The majority of TFs have relatively large promiscuity p_i and therefore interact in different cell types. By comparison, a small number of TFs (e.g., GATA5) have a smaller promiscuity and thus might be potential biomarkers. We highlight labelled TFs with a red cross.

Analysis

We calculate the promiscuities of nodes in the MLN to find TFs that are specific to certain cell types. TFs that are important in a small subset of tissues are candidates for biomarkers. As such computing the promiscuities of TF in a tissue specific regulatory network is a quantitative approach to detect biomarkers.

Almost all TF have large promiscuities p_i and thus are promiscuous and not tissue-specific (see Fig. 3.7). The TF with the largest multilayer degree $K_i = 12,611$ is 'Specificity Protein 1' (SP1). It also has a large promiscuity $p_i \approx 0.986$. SP1 is essential for the regulation of many genes involved in most cellular processes, including cell growth and differentiation [104, 266]. SP1 is expressed ubiquitously in human cells and its overexpression leads to apoptosis in cancer cells [86]. The largest promiscuity $p_i \approx 0.989$ is only a bit larger and reached by the TF 'Signal transducer and activator of transcription 3' (STAT3). This TF is ubiquitously expressed in human cells and also an oncogene [471].

The TF with the smallest promiscuity p_i is 'GATA Binding Protein 5' (GATA5)[4], which is important in embryonic development and a potential lung-tissue biomarker [491] and is not expressed ubiquitously in human cells (see, e.g., FANTOM5 data set in the Human Protein Atlas [367]). It is strongest expressed in testes and muscle cells. Another specialized TF is 'Paired box 7' (PAX7), which is known to be important for the specification of the neural crest [30].

[4]GATA transcription factors are a family of transcription factors, which are characterized by their ability to bind to the DNA sequence 'GATA' [235].

'Ovo like zinc finger 2' (OVOL2), which has the third smallest promiscuity $p_i \approx 0.5$, belongs to the OVOL gene family. OVOL genes are involved in epithelial development and differentiation [243, 264]. According to the FANTOM5 data set it is expressed in most but not all human tissue types.

In their original study, Neph et al. [326] report also some tissue-specific TFs. They identify some of the small-promiscuity nodes (e.g. GATA5) as tissue-specific TFs. They do not report other small-promiscuity TF (e.g., PAX7 or OVOL2). Neph et al. [326] hand-picked tissue-specific TFs from inspecting their data. In contrast, promiscuity is a summary statistic that allows a quantitative approach for finding possible tissue-specific TFs.

3.5.2 Tissue-specific Gene Regulation

In this subsection, we examine a tissue-specific gene regulatory network by investigating genes' promiscuities. In contrast to the data in Subsection 3.5.1, we do not investigate exclusively the regulation between TFs but also between normal genes.

Network Construction

We construct a multilayer gene regulation network by using data from [291]. Marbach et al. (2016) [291] inferred the regulatory network of $N = 16,974$ genes for $L = 32$ tissues. To do this, they integrated TF sequence motifs with promoter and enhancer activity data across cell types. Their pipeline consists of three steps: (1) measuring the activity of promoters and enhancers in different tissues, (2) identification which

TF binds to a given promoter or enhancer, (3) linking promoters and enhancers to gene products. The combination of these steps is therefore an indirect identification of physical TF–gene interaction in different tissues.

From these data we create an edge-coloured multigraph in which the nodes represent genes and the layers represent the interactions in different tissues. Thus, the multigraph consists of $N = 16,974$ nodes and $L = 32$ layers. An undirected edge exists between nodes i and j in layer l, if they regulate each other in the tissue type represented by layer l.

Analysis

The genes' promiscuities p_i span the entire range $[0, 1]$ (see Fig. 3.8a). However, the mean promiscuity $\langle p_i \rangle_{\text{all}} \approx 0.93$ is large, so most genes regulate many different tissues. In this data set, there is a positive correlation (the Spearman coefficient is $r_s \approx 0.64$) between multilayer degree K_i and promiscuity p_i. Thus, genes with more regulatory influence on other genes tend to exert such influence across different tissues. Only 152 genes have the minimum promiscuity $p_i = 0$ and thus regulate a single tissue. These include many transcription factors (e.g., GATA, DOBOX, and FOX genes), as well as genes of unknown function (e.g., C16ORF92 and C1ORF105). The transcription factors are therefore tissue-specific. The gene TNP2, which plays a key role in the replacement of histones in spermatids [311, 401], is the minimum-promiscuity gene with the largest multilayer degree (it has $K_i = 284$). It is exclusively expressed in testes and is therefore a potential biomarker for this tissue.

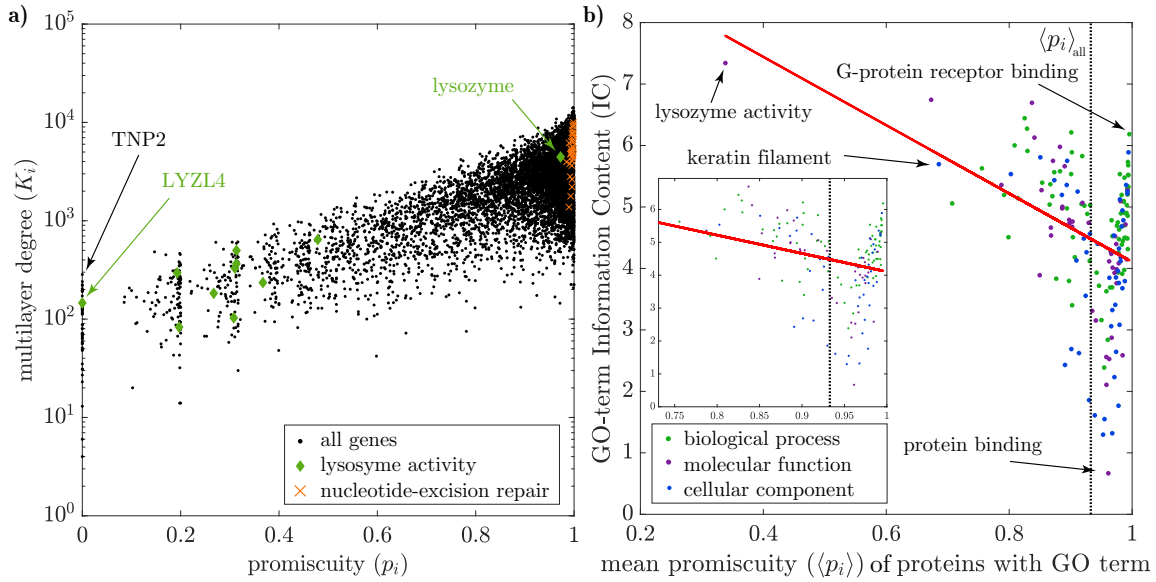


Figure 3.8: Tissue-specific organization of genes in human cells. (a) Genes in (p_i, K_i) space. Most genes have a large promiscuity p_i and thus regulate many different tissues. (b) For 167 GO annotations, genes with the given annotation have significantly different promiscuity distributions from those without the annotation. Of these GO terms, the ones with larger mean promiscuities $\langle p_i \rangle$ tend to have smaller Information Contents (ICs), so more genes have this annotation. The dotted vertical line indicates the mean promiscuity over all proteins $\langle p_i \rangle_{\text{all}}$.

We compare the promiscuities p_i of genes with annotation data from three gene ontologies (GOs) to examine whether certain biological processes, molecular functions, or cellular components have noticeable promiscuity distributions. (See in Section 2.9 for details about GO data and methodology). For each gene we compute promiscuity p_i, which is a continuous variable. Our aim is to detect whether GO terms are associated predominantly with genes of a certain promiscuity. For this, we use a non-parametric Kruskal–Wallis one-way analysis of variance . We choose a p-value of 0.01 to test whether genes with a given annotation are distributed differently than those without it. Out of $47,116$ terms, 167 pass a Bonferroni-corrected test, which corrects for multiple

comparison. We refer to these terms as 'enriched'. For each of these 167 enriched terms, proteins annotated with a given term have significantly different promiscuity distributions from proteins without this particular annotation. For each of these 167 terms, we show in Fig. 3.8b the mean promiscuity $\langle p_i \rangle$ of all genes with a given GO term. We compare this with the term's information content

$$\mathrm{IC} = -\ln\left(N_{\mathrm{annotated}}/N\right), \tag{3.42}$$

where $N_{\mathrm{annotated}}$ is the number of genes with a given annotation. IC is a measure of whether a term is used widely (small IC) or little used (large IC). IC and mean promiscuity $\langle p_i \rangle$ are negatively correlated with each other (p-value 10^{-7}). The Pearson correlation is $\rho_p \approx -0.37$ and we indicate the trend line in red in Fig. 3.8b.

We conclude from this that more widely-used terms tend to have larger mean promiscuity $\langle p_i \rangle$. We find that the GO term with the smallest IC is 'protein binding', which is associated with genes that have promiscuity larger than the mean $\langle p_i \rangle_{\mathrm{all}}$. The term with the smallest IC, 'lysozyme activity', also has the smallest mean promiscuity $\langle p_i \rangle \approx 0.4$; and it thus, in comparison with other terms, tends to be associated with tissue-specific genes. In Fig. 3.8a, we highlight all genes associated with 'lysozyme activity' as green diamonds. All but lysozyme have promiscuities $p_i < 0.5$. LYZL4, for which $p_i = 0$, belongs to a group of 'lysozyme-like' genes [502] that are active in human-host defence and are active exclusively in the male reproductive system, which matches their minimal promiscuity.

Some cellular components also exhibit promiscuity-enrichment. 'Keratin filament',

for example, has a small promiscuity; it is found only in a small subset of human cell types, including hair and epidermis. The cellular components with small IC are used widely (e.g., 'cytoplasm' and 'nucleus'), and their large promiscuities likely arises because all eukaryotic cells have cytoplasm and all (with a small number of exceptions) have a nucleus. Such widely-used terms tend to have a mean promiscuity $\langle p_i \rangle$ close to the overall mean promiscuity $\langle p_i \rangle_{\text{all}} \approx 0.93$, because they are associated with most genes. The terms with largest mean promiscuity $\langle p_i \rangle$, however, tend to have a larger IC. These terms are associated with specific functions, such as 'G-protein receptor binding' or 'nucleotide excision repair', that are essential for many different tissue types [253, 379].

In this analysis of a tissue-specific gene regulatory MLN, we find that only a small number of genes have the minimal promiscuity $p_i = 0$. These genes are tissue-specific regulators. Most genes, by contrast, have promiscuities close to the maximum $p_i \approx 1$ and thus their influence varies to lesser extent across tissues. We find that some GO terms tend to be associated with a genes with certain promiscuity p_i. For these enriched terms, we compare the mean promiscuities of associated genes with the IC. We find a negative correlation between them, so terms that are used more widely tend to be associated with genes that are more influential across tissues.

3.6 Conclusions

In this chapter, we introduced promiscuity p_i as a measure of the variability of a node's degree k_i across the layers in a MLN in comparison with a uniform null model. We

found that real-world networks and synthetic networks can have a variety of different promiscuity distributions. We then proved that the promiscuity $p_i \in [0, 1]$.

We also illustrated that there exist a particular type of synthetic networks, called colour-regular multigraphs, which have maximal promiscuity $p_i = 1$ for all nodes i. By contrast, a network with all nodes of minimal promiscuity $p = 0$ is a union of completely separated layers. We can interpret the promiscuity distribution in a network as a way to probe the possibility of interactions between layers. In the case of transportation networks, for example, nodes with nonzero promiscuities serve as transit nodes between layers, which represent different modes of transportation. This has potentially important implications for network function and dynamical process on them. For example, spreading processes can be influenced strongly by the structure of a MLN [106]. In contrast to earlier approaches to analyse the interaction between layers, e.g., the *local overlap* [41] or *multilayer participation* [46], the promiscuity takes the density of network layers into account.

We investigated the promiscuity in different types of synthetic networks. For two-layer networks consisting of one regular layer and one ER-layer, we derived an explicit formula $p_i(K_i)$, which agrees well with numerical computations. We created networks with $p_i = 0$ and observed the effect of a rewiring procedure that preserves the number of edges in each layer, on the promiscuity. We then derived an upper bound for the expected mean promiscuity in these networks. This result matches our numerical computations well but the discrepancy between both increases

with increasing randomisation.

In our examination of two networks of tissue-specific gene regulation, we calculated that most genes have large promiscuities and are thus influential in many different tissues. However, we also found some genes of small promiscuity, which indicates a tissue-specific function; these are candidates for tissue-specific biomarkers. Some of theses genes are indeed producing transcription factors with tissue-specific functions and abundance, but the function of others is unknown. Thus, calculating promiscuities may be helpful for discovering tissue-specific biomarkers. By comparing promiscuities with GO annotation data, we found that some biological functions are associated with certain promiscuity values. Biological functions that are associated with many genes tend to be associated with large promiscuity values.

We note that a direct comparison of node promiscuities between the two tissue-specific MLNs is not fruitful because the two networks have different numbers of nodes and layers. Furthermore, the tissue types are different. We do, however, observe the overall trend — most nodes have high promiscuities — in both networks.

Future Directions

The computation of promiscuity as a variability of a node's importance can be generalised to other node measures than the degree k_i. It is possible, e.g., to compute promiscuities of betweenness or eigenvector centrality. We note that the first-order mover score m_i of the authority centrality, which we used in Chapter B, resembles such an analysis of eigenvector-promiscuity, as it is a measure of variability of a node's

eigenvector centrality across layers. A crucial difference, however, is that the first-order mover score takes a temporal order of the layers into account, whereas in an edge-coloured multigraph there is no temporal ordering. Therefore, the promiscuity, as introduced in this chapter, neglects any particular ordering of layers.

In this study, we compared the degree of each node with a uniformly random null model to obtain the promiscuity vector. This is the simplest possible null model. Additional information, e.g., node positions in a physical space [29] or other nodal attributes [331] could be used to define more sophisticated null models. This could improve the identification of nodes that have extraordinary roles in a multirelational network.

It is known that not all TF–DNA interactions are functional [269]. Spivakov [429] argues that it would be fruitful to consider the 'potency' of such regulatory information. Such information — if experimentally available across tissues— would improve the computation of promiscuity of TF across tissues because it allows the construction of weighted tissue-specific regulatory MLNs.

4

Eigenvector Centrality in Temporal Protein Interaction Networks

Contents

4.1 Introduction 114
4.2 The Yeast Cell Cycle 116
4.3 Construction of Temporal Protein Interaction Networks . 118
4.4 Data 122
4.5 Eigenvector-based Centrality in Temporal Networks . . . 123
4.5.1 Inter-Layer Coupling of Centrality Matrices 124
4.5.2 Conditional Centrality for Temporal Multilayer Networks . 124
4.5.3 Singular Perturbation in the Strong-Coupling Limit 125
4.6 Results 127
4.6.1 Centrality in the Strong-Coupling Regime 127
4.6.2 Centrality Trajectories during the Yeast Cell Cycle 132
4.7 Discussion 136
4.7.1 Future Directions 140

4.1 Introduction

Recall that we can calculate centrality measures to quantify the importance of nodes in a network (see Section 2.5). For example, they can be used to identify 'essential' proteins in PINs [216]; these are proteins whose absence is lethal for an organism. One may test with mutagenesis experiments (also called 'gene knockout studies') the influence of making a gene inoperative, which results in the absence of proteins produced from this gene [485]. Jeong et al. (2001) performed the first computational study of the relationship between centrality and essential proteins. For this, they computed changes in the network's diameter after deletion of certain nodes. Proteins whose deletion increases the network's diameter drastically tend to be 'essential', Since this first analysis, various other centrality measures have been proposed to identify essential or otherwise influential proteins; these include betweenness [499], closeness [2], *star centrality* [468], and DiffSLC, a weighted combination of eigenvector centrality and gene co-expression data [308]. Recently, Ashtiani et al. (2018) reviewed 27 common centrality measures and found that combinations of multiple measures are more successful at identifying essential proteins than single measures [21]. Centrality in PINs can give insights beyond the lethality of single deletions. For example, Alvarez-Ponce et al. (2017) found a positive Spearman correlation between the centrality of proteins in PINs and the evolutionary conservation of their structure [15].

One can also investigate link-centralities in PINs. In such computations, centralities indicate the importance of PPIs and not proteins. Agarwal et al. (2010) found

Pearson and Spearman correlations between the link-centrality and the functional similarity of the interacting proteins [3].

These studies suggest that analysing centrality of proteins in PINs may provide information about their biological functions. However, all of these approaches examined PINs in a non-state-specific form. Usually, one constructs such static PINs with all known proteins and measured interactions between them present.[1] As PINs are inherently dynamic, we hypothesise that centralities of proteins can also change over time. A change in centrality may perhaps reflect the change or fluctuation of a protein's importance during different stages of a dynamic process. This changing centrality may also influence lethality properties of protein-deletion, as a protein that is not central in an aggregated network can nevertheless be essential at a given point in time.

In this chapter, we describe methods to investigate changes in proteins' eigenvector centralities over time. We use an eigenvector-based temporal centrality, as described by Taylor et al. (2017) [451], for the the analysis of a temporal PIN. This allows us to investigate variations in a proteins' importance in a PIN over time. As a case study, we examine a temporal PIN of the cell cycle of the budding yeast *Saccharomyces cerevisiae*. We focus on the yeast cell cycle for two reasons, because it is a well-understood biological process and so allows us to validate our novel approach. Second, for the given data, which covers multiple cell cycles, we expect oscillatory change of centrality for some proteins. Such a behaviour has not been detected with the

[1] As discussed in Subsection 2.2.4, not all interactions are known and some of the detected interactions may be non-existent or condition-specific.

eigenvector-based centrality in temporal networks and therefore we want to validate whether this method is able to detect such oscillations.

The rest of this chapter is structured as follows. In Section 4.2, we discuss the importance of the yeast cell cycle. In Section 4.3, we describe a procedure to construct temporal PINs. In Section 4.5, we describe eigenvector-based centrality in temporal networks [451]. In Subsection 4.6.1, we examine this notion of centrality in the strong-coupling regime. In Subsection 4.6.2, we discuss centrality trajectories for selected proteins. We conclude in Subsection 4.7.

4.2 The Yeast Cell Cycle

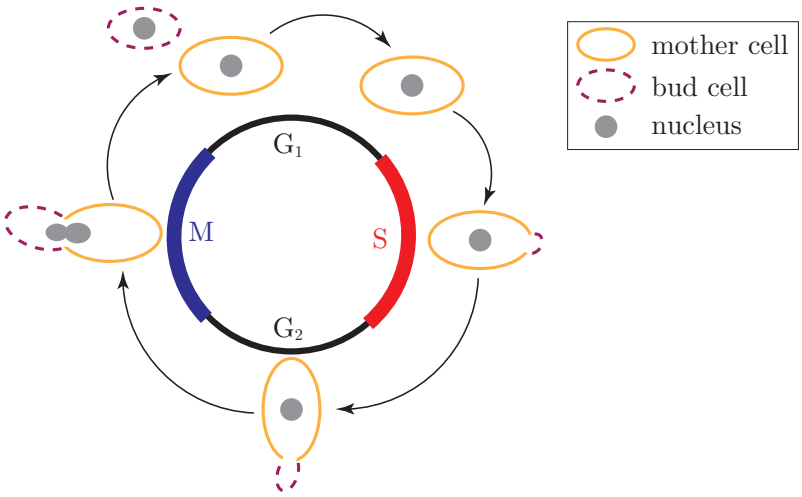


Figure 4.1: The mitotic cell cycle in yeast. During the mitotic cell cycle a yeast mother cell (solid ellipse) creates a bud (dashed ellipse) that develops into a daughter cell with the same genetic material in its nucleus (grey disk) as the mother cell. The cell cycle consists of four phases: 'Gap 1' (G_1), 'Synthesis' (S), 'Gap 2' (G_2), and 'Mitosis' (M).

Saccharomyces cerevisiae (*S. cerevisiae*) is a single-celled eukaryote and belongs to the biological kingdom *fungus* [126, 142]. Under good environmental conditions,

its cells double in number roughly every 100 minutes [189]. Like all fungi, yeast may reproduce sexually and asexually. The former is more common and a lack of nitrogen may trigger the latter [218]. *S. cerevisiae* grows asexually by *budding*[2], in which a 'mother' cell develops a small bud, which develops into a new organism with identical genetic material. To do this, yeast cells undergo a *mitotic cell cycle* (see Fig. 4.1).

A mitotic cell cycle is the succession of events that lead to the division of one mother cell into two cells with the same genetic information [316]. Between one cell division and the next, all essential parts of the cell — most notably, genetic material in the form of DNA — must be duplicated. The cell cycle consists of four phases: 'Gap 1' (G_1), 'Synthesis' (S), 'Gap 2' (G_2), and 'Mitosis' (M). In the S-phase, a cell replicates the DNA; in the M-phase, the mother cell divides into two cells. During the G-phases, the cells grow physically larger and produces proteins and organelles that are necessary for later steps. Checkpoints occur at the transition from G_1 to S and G_2 to M-phase [40]..

Each cycle consists of two regimes: a nonrespiratory regime without oxygen consumption and a respiratory one in which yeast consumes oxygen [459]. This results in a change of dissolved oxygen concentration (see Fig. 4.2). One therefore can measure the cell-cycle progression by measuring the oxygen concentration in the environment of the yeast organism.

The regulation of the succession of cell cycle phases is important for the survival of an organism. There are two families of proteins that are crucial for such regulation:

[2]This contrast with *fission*, during which the mother cell develops into two identical daughter cells. Fission yeast *Schizosaccharomyces pombe* reproduces asexually by fission.

cyclins and *cyclin-dependent kinase* (CDK) enzymes. CDKs are key regulators of the cell cycle [315]. While there is some difference in the precise regulation mechanism among eukaryotes, many features are conserved, and yeast is often used as a model system to study cell-cycle regulation [338]. In yeast, 'Cell Division Cycle 28' (CDC28) is the only CDK enzyme, but other organisms have multiple types of CDK enzymes.

Many of the approximately 6 000 genes in yeast are cell-cycle-dependent; in other words, they are transcribed at higher levels at some points of the cell cycle than at others. In Fig. 4.2, we illustrate the expression profiles (i.e., the amount of mRNA for each gene over time; see Subsection 2.1.2) of 100 genes [459]. The estimates of the number of these cell-cycle-dependent genes vary from around 800 to more than 3 000 [427, 459]. The iteration of multiple cell cycles leads to oscillatory changes in gene expression [76].

4.3 Construction of Temporal Protein Interaction Networks

Below we outline a procedure to construct a temporal PIN as a multilayer network from temporal gene expression profiles and a static PIN[3]. The static PIN has N nodes, which represent the proteins present in the organism. Its adjacency matrix $\mathbf{A}$ is a $N \times N$ matrix. The procedure we outline works, in principle, with weighted networks, if the edge weights quantify the experimental evidence for a PPI [55]. In our application, however, the PIN is unweighted.

We assume that there is a bijection between genes and proteins, such that each

[3]While we discuss 'gene expression profiles', the procedure is also suited for other 'omics' data (e.g., protein abundance; see Subsection 2.1.2).

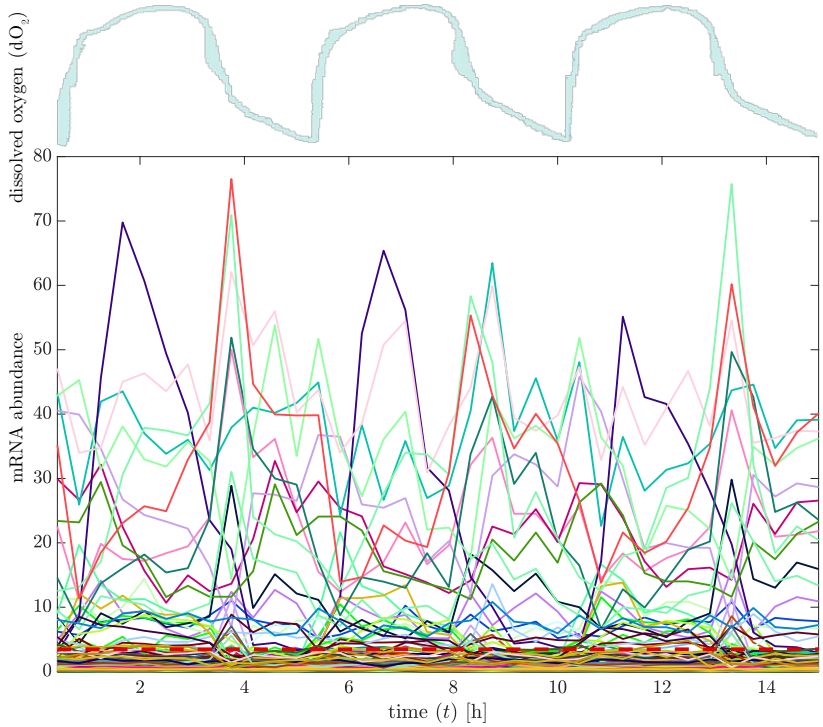


Figure 4.2: Gene-expression profiles during the yeast cell cycle. Each of the 100 curves is the expression of a single gene during three cell cycles. The dashed horizontal line indicates the threshold that we choose for the construction of the temporal PIN (see Subsection 4.3). The upper panel gives the concentration of dissolved oxygen dO_2 during three cell cycles. Yeast cells tend to enter the cell cycle when the oxygen increases. All data are from Tu et al. (2005).

protein is the product of exactly one gene and *vice versa*.[4] Each of these N genes has a temporal expression profile $E_{i,t}$ that gives the gene expression level of gene i at a discrete time t. We can represent the gene expression profiles of all N proteins as a $N \times T$ matrix $\mathbf{E}$. We binarise this matrix by thresholding it:

$$E_{i,t}^{\text{binary}} = \begin{cases} 0\,, & \text{if } E_{i,t} \leq E_0\,, \\ 1\,, & \text{if } E_{i,t} > E_0\,. \end{cases} \tag{4.1}$$

[4]This is a simplification of the actual biological processes. However, it is a common assumption because protein-interaction measurements usually do not allow the distinction between different *isomers* (i.e, multiple proteins that are the product of the same gene) [409].

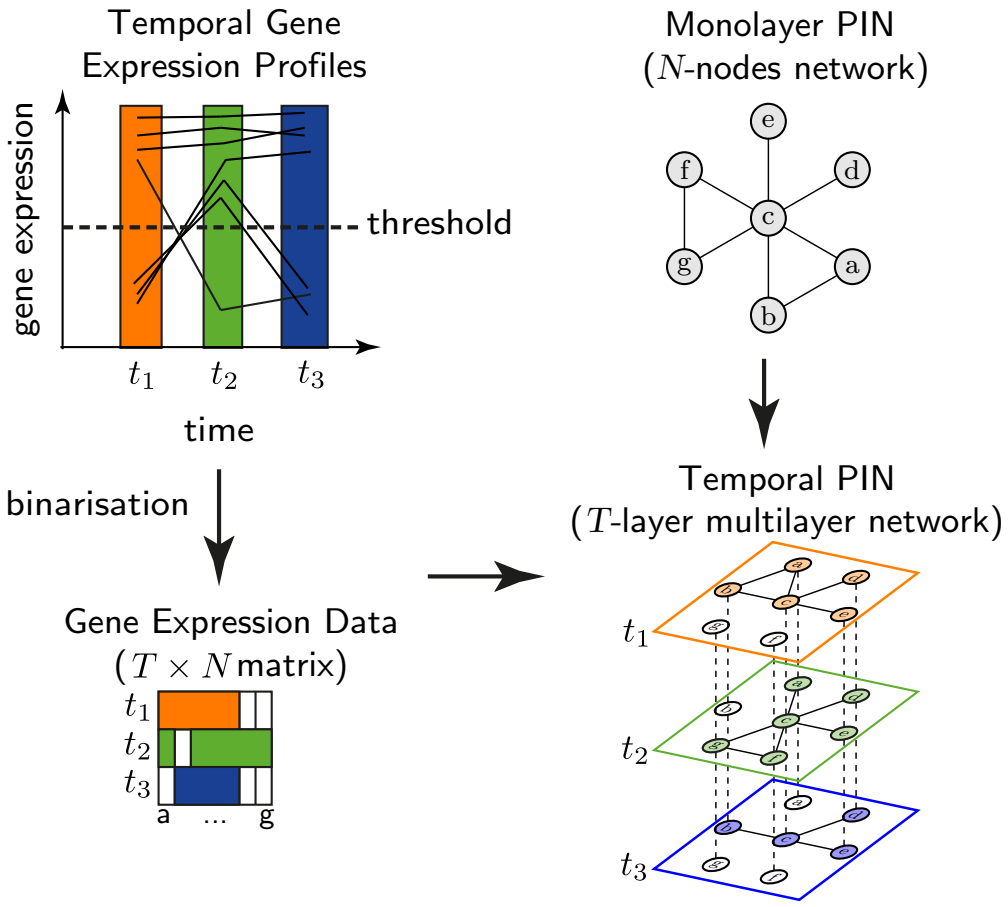


Figure 4.3: Construction of temporal PINs from temporal gene expression profiles and a static PIN. We binarise the N temporal gene expression profiles by thresholding them and obtain a $T \times N$ gene expression matrix. We construct a temporal PIN by combining a static PIN with temporal gene expression data in the following way. We create T copies of the PIN, each of which is one layer in a multilayer network and represents a time point. We then delete all connections of proteins transcribed by genes that are not expressed at a given time point. We show these 'ghost nodes' (e.g., nodes f and g in the layer representing t_1) colourless. Finally, we connect nodes that represent the same protein with interlayer edges if they are in adjacent temporal layers.

Thus, for each time point t, the matrix $E_{i,t}^{\text{binary}}$ indicates whether protein i is present ($E_{i,t}^{\text{binary}} = 1$) or not ($E_{i,t}^{\text{binary}} = 0$). There is no *a priori* 'optimal' choice of the threshold E_0. As a default choice, however, we choose the mean expression $\langle E \rangle = \left(\sum_{i,t} E_{i,t}\right)/(NT)$.

We construct a temporal PIN $\mathcal{T}$ in the snapshot representation, which is a sequence $\mathcal{T} = \{G_1, G_2, \dots G_T\}$ of graphs, where each graph G_t represents the connectivity at time t (see Subsection 2.7.4). First, we create T copies of the PIN, each of which is one snapshot. We then delete edges that are incident to nodes that represent unexpressed genes. Suppose that $\mathbf{A}$ is the adjacency matrix of a static PIN. The elements of the adjacency matrix $\mathbf{A}^{(t)}$ of G_t are then

$$A_{ij}^{(t)} = \begin{cases} A_{ij}\,, & \text{if } E_{i,t}^{\text{binary}} = 1\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{4.2}$$

The temporal network $\mathcal{T}$ includes only intralayer edges. As we outlined in Section 2.7, there are different possibilities for constructing interlayer edges in a temporal network. The most common one in the literature is to construct interlayer edges as 'identity edges' that connect nodes in time-adjacent layers if they represent the same gene [232].

This construction procedure assumes implicitly that every gene node i is present at every time layer t. It is tempting to remove such 'ghost nodes' that do not have intra-layer edges and connect only to their counterparts via inter-layer edges. In Fig. 4.3, the ghost nodes are opaque. For some applications, e.g., the construction of multilayer modularity matrices (see Section 2.8), removing ghost nodes is possible. The eigenvector-based temporal centrality that we use in this chapter, requires that all layers have the same number of nodes. Therefore we do not remove ghost nodes.

4.4 Data

We use two publicly available data sets to construct a temporal PIN of the yeast cell cycle, temporal RNA-abundance levels [459] and a time-independent PIN [77].

Tu et al. (2005) measured the RNA abundance in the yeast strain CEN.PK[5] with oligonucleotide arrays[6] in 25-minute intervals over 5 hours, which covers three consecutive cell cycles. We discuss oligonucleotide arrays in Subsection 2.1.2. They found 6 275 distinct mRNA molecules of genes that encode proteins. In Fig. 4.2, we show the expression profiles (i.e., the RNA-abundance over time) for some proteins. We chose this data because it is publicly available and widely used in other publications, which makes a comparison of results easier.

We map the mRNA molecules to the associated proteins and construct a static PIN from PPI data from BioGRID (version 3.4.158) [77]. This is not possible if the mRNA molecule is a non-coding RNA (e.g., tRNA).

We use the procedure that we discussed in Section 4.3 to construct a temporal PIN. As binarisation threshold P_0, we choose the mean expression $\langle P \rangle \approx 3.5$. We choose this intermediate value to balance between the two extremes: a threshold of $P_0 = 0$ would lead to all nodes being present in all layers, and a threshold of $P_0 = \infty$ would result in no nodes present in all layers. We do remove proteins that are never

[5] CEN.PK is a laboratory strain and popular in systems-biology studies [463].

[6] Oligonucleotide arrays do have some methodological disadvantages [237]. Specifically, the hybridization efficiency of probe sets might be different for different genes. In principle, RNA-sequencing data would overcome these problems. As RNA sequencing is a newer experimental technique, however, time series data of good quality is not available.

construction step	number of entities	number of removed entities
expression profiles	6 275	
RNA not associated with a protein		2 196
RNA thresholding		925
isolated nodes		388
static network	2 766	
temporal multilayer network	$36 \times 2\,766 = 99\,576$	

Table 4.1: The number of entities in the yeast cell cycle data. We show the number of entities (mRNA molecules or proteins) that are present in the original data (expression profiles) and removed at subsequent steps in the temporal PIN creation process. The static network consists of 2 766 nodes and the temporal multilayer network of 99 576 nodes.

expressed above the threshold P_0 and proteins without interaction partners (isolated nodes). Table 4.1 details the number of entities (mRNA molecuels or proteins) that are removed at each of these construction steps. The final temporal network consists of 2 766 physical nodes, each representing a protein, in 36 layers, which results in 99 576 state nodes. The temporal difference between adjacent layers is 25 minutes.

4.5 Eigenvector-based Centrality in Temporal Networks

In this section, we present eigenvector-based centralities in temporal networks, as discussed by Taylor et al. (2017) [451]. This is an extension of eigenvector-based centralities for (non-temporal) monolayer networks, which we discussed in Subsection 2.5.1. While we use exclusively the authority score, which is one particular choice of eigenvector-based centrality, we discuss the computation in a general way that allows the computation of any kind of eigenvector-based centrality in temporal networks.

4.5.1 Inter-Layer Coupling of Centrality Matrices

Suppose that we have a temporal network in its snapshot representation (see Subsection 2.7.4). That is, the adjacency matrix at time t is given by $\mathbf{A}^{(t)}$ and $A_{ij}^{(t)}$ encodes the presence and weight of the edge from node i to node j. Recall that in a monolayer network, the centrality matrix $\mathbf{C}$ is a function of the adjacency matrix $\mathbf{A}$ and one obtains an eigenvector-based centrality by computing the leading eigenvector of $\mathbf{C}$. Analogously, we can compute the centrality matrix $\mathbf{C}^{(t)}$ at time t as a function of the adjacency matrix $\mathbf{A}^{(t)}$ at time t. We couple these centrality matrices with inter-layer couplings of strength ω to obtain a *supra-centrality matrix*

$$\mathbb{C}(\varepsilon) = \begin{bmatrix} \varepsilon\mathbf{C}^{(1)} & \mathbf{I} & 0 & \dots \\ \mathbf{I} & \varepsilon\mathbf{C}^{(2)} & \mathbf{I} & \ddots \\ 0 & \mathbf{I} & \varepsilon\mathbf{C}^{(3)} & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix} . \tag{4.3}$$

For diagonal and uniform coupling, a supra-centrality matrix, as a supra-adjacency matrix (see Section 2.7), is a $NT \times NT$ matrix. A supra-centrality matrix is a function of the scaling factor $\varepsilon = 1/\omega$. This notation is convenient for singular-perturbation analysis in the strong-coupling limit $\varepsilon \to 0^+$ (i.e., $\omega \to \infty$), which we discuss in Subsection 4.5.3.

4.5.2 Conditional Centrality for Temporal Multilayer Networks

We study the dominant eigenvector $\mathbf{v}(\varepsilon)$ of the supra-centrality matrix $\mathbb{C}(\varepsilon)$. This is

$$\mathbb{C}(\varepsilon)\mathbf{v}(\varepsilon) = \lambda_{\max}(\varepsilon)\mathbf{v}(\varepsilon) , \tag{4.4}$$

where $\lambda_{\max}(\varepsilon)$ is the largest positive eigenvalue of $\mathbb{C}(\varepsilon)$. The eigenvector $\mathbf{v}(\varepsilon)$ has NT elements, each of which is the centrality of a node-layer pair (i,t). We can rewrite this vector into an $N \times T$ matrix $\mathbf{W}$, whose entries

$$W_{it}(\varepsilon) = v_{N(t-1)+i}(\varepsilon)\,, \tag{4.5}$$

indicate the centrality of node i at time t. Often, we want to investigate the importance of a physical node i, which represents a protein, relative to other physical nodes in layer t. To do this, we compute the *conditional centrality*

$$Z_{it} = \frac{W_{it}}{\sum_i W_{it}}\,. \tag{4.6}$$

We thereby can investigate a *centrality trajectory* as an ordered sequence $\{Z_{it}\}$ of a physical node i over time.

4.5.3 Singular Perturbation in the Strong-Coupling Limit

Taylor et al. (2017) provided a procedure for computing a 'time-averaged centrality' in the limit $\varepsilon \to 0^+$. In this limit, the conditional node-layer centrality of every physical node is constant across the time layers (i. e., $Z_{it} = Z_{it'}$ for all times t and t').

Taylor et al. (2017) demonstrated that one can obtain these time-averaged centralities as the elements of the leading eigenvector of the $N \times N$ matrix $\mathbf{X}^{(1)}$. These elements are

$$X_{ij}^{(1)} = \gamma_1^{-1} \sum_t C_{ij}^{(t)} \sin^2\left(\frac{\pi t}{T+1}\right)\,, \tag{4.7}$$

where $\gamma_1 = \sum_{t=1}^{T} \sin^2(\pi t/(T+1))$. One thus obtains the time-averaged centralities $\{\alpha_i\}$ as elements of the vector α by solving the eigenvector equation

$$\mathbf{X}^{(1)}\alpha = \lambda_1 \alpha \,. \tag{4.8}$$

They also calculated *first-order mover scores*, which are a measurement of the variation over time of each physical node's centrality trajectory. To do this, they constructed a matrix $\mathbf{X}^{(2)}$ with elements

$$X_{ij}^{(2)} = \mathbb{U}_i^{\intercal}\mathbb{P}^{\intercal}\mathbb{G}\mathbb{L}_0^{\dagger}\mathbb{P}\mathbb{U}_j \,, \tag{4.9}$$

where $\mathbb{L}_0^{\dagger} = (\lambda_0\mathbf{I} - \mathbf{A})^{\dagger} \bigotimes \mathbf{I}$ denotes the Moore–Penrose pseudoinverse of $\mathbb{L}_0$, the matrix $\mathbb{G} = \mathrm{diag}[\mathbf{C}^{(1)}, \ldots, \mathbf{C}^{(T)}]$, and $\mathbb{P}$ is the stride-permutation matrix with entries $P_{kl} = 1$ for $l = \lceil k/N \rceil + T[(k-1) \mod N]$ and $P_{kl} = 0$ otherwise [161]. The ceiling function $\lceil x \rceil$ maps x to the least integer greater than or equal to x (i.e., $\lceil x \rceil = \min\{n \in \mathbb{Z} | n \geq x\}$). The vector $\mathbb{U}_i$ has T blocks with N elements each and so a total of NT elements.The ith block is $\mathbf{u}$, which is the leading eigenvector of the adjacency matrix $\mathbf{A}$, and all other elements are 0. One then solves for β using the equation

$$\left(\mathbf{X}^{(1)} - \lambda_1\mathbf{I}\right)\beta = (\lambda_2\mathbf{I} - \mathbf{X}^{(2)})\alpha \,, \tag{4.10}$$

where $\lambda_2 = \alpha^{\intercal}\mathbf{X}^{(2)}\alpha$. One then obtains the first-order-mover scores $\{m_i\}$ by solving

$$m_i^2 = \beta_i^2 + \sum_{t=1}^{T} \left([\mathbb{L}_0^{\dagger}\mathbb{G}\mathbf{v_0}]_{i+t(N-1)}\right)^2 \,, \tag{4.11}$$

where $\mathbf{v_o} = \sum_j \alpha_j \mathbb{PU}_j$.

This framework for eigenvector-based centrality in temporal networks allows the investigation of different centrality matrices $\mathbf{C}$. We summarise some prominent choices in Table 2.5. For the investigation of the yeast temporal PIN we use the authority score, which we calculate as the leading eigenvalue of the centrality matrix $\mathbf{C} = \mathbf{A}\mathbf{A}^\intercal$, where $\mathbf{A}$ is the adjacency matrix and $\mathbf{A}^\intercal$ its transpose. Choices other than $\mathbf{A}\mathbf{A}^\intercal$ for the centrality matrix are possible. One can, for example, calculate a temporal variant of the PageRank score.

4.6 Results

4.6.1 Centrality in the Strong-Coupling Regime

We begin by calculating the temporal authority centrality of proteins in the strong-coupling regime (see Subsection 4.5.3). We summarise the centrality trajectory of each node in terms of two scalar values: time-averaged authority centrality α_i, which is a measure of a protein's overall centrality; and first-order-mover score m_i, which is an indicator of the magnitude of the variability in centrality.

We list the proteins that have the largest time-averaged authority centralities α_i in the left part of Table 4.2; all central proteins (according to this measure) are ones with important functions for the yeast organism. Not all are essential, however (essential proteins are in bold). The most central proteins are *ribosome-associated molecular*

	Time-Averaged Centralities			First-Order-Mover Scores		
Rank	Gene	α_i	m_i	Gene	α_i	m_i
1	SSB2	0.1544	3.9723	PAT1	0.0415	132.8032
2	**ACT1**	0.1525	5.1916	**CDC11**	0.0240	88.3696
3	**RPN11**	0.1476	4.9006	**SSC1**	0.0495	81.291
4	**NAB2**	0.1368	25.8395	**CDC28**	0.0106	75.0849
5	SBP1	0.1276	3.3581	**RPB3**	0.0067	74.1525
6	HSP82	0.1163	2.5787	BRE5	0.0391	68.1169
7	RPN10	0.1092	4.1812	**RVB1**	0.0720	66.8207
8	GIS2	0.1002	2.1381	BEM2	0.0279	64.8161
9	TOM1	0.0998	43.36	DHH1	0.0487	63.2704
10	RPN1	0.0992	5.1957	**STT3**	0.0353	58.2226
11	**RSP5**	0.0985	19.8528	UMP1	0.0136	57.6226
12	**RPT5**	0.0919	4.5613	**CDC34**	0.0313	57.2429
13	**HYP2**	0.0881	1.5251	STE24	0.0352	57.1430
14	HSC82	0.0844	1.8852	CCR4	0.0583	56.8237
15	YDJ1	0.0814	2.3955	**SUP35**	0.0469	56.0839
16	**CMD1**	0.0812	1.8014	KSP1	0.0186	54.1306
17	**CDC48**	0.0808	2.3380	SYS1	0.0279	51.4935
18	**RPT6**	0.0800	2.1521	APC11	0.0321	51.3182
19	RPN4	0.0790	15.8850	**TAF5**	0.0150	49.9215
20	RVS161	0.0785	2.5643	SEC28	0.0291	49.5234
21	CSG2	0.0768	1.8814	**RPB7**	0.0201	48.6788
22	**RPT4**	0.0767	1.06390	SIF2	0.0127	46.4559
23	GET2	0.0764	9.35124	TPS2	0.0408	46.0501
24	**RPN6**	0.0762	2.1065	SLA2	0.0457	45.6933
25	RVS167	0.0753	1.554	**TIF11**	0.0162	45.6790

Table 4.2: Top time-averaged centralities α_i and first-order-mover scores m_i in the temporal PIN during the yeast cell cycle. Genes in bold are essential, according to the Saccharomyces Genome Database (SGDB) [82].

chaperone (SSB2), *actin* (ACT1), *ubiquitin carboxyl-terminal hydrolase* (RPN11), and *nuclear polyadenylated RNA-binding protein* (NAB2). Although detecting such central proteins is interesting, the time-averaged centrality α_i, by definition, does not provide insight into the temporal variability of a node's importance. We thus calculate the first-order-mover scores m_i that are a measure of the variability of nodes' importance in

the limit $\varepsilon \to 0^+$. It does not allow conclusions about the direction of the importance change. We summarise the proteins with largest scores in the right part of Table 4.2. The four proteins with the largest first-order-mover scores are *DNA topoisomerase 2-associated protein* (PAT1), *cell division control protein 11* (CDC11), *heat shock protein* (SSC1), and *cyclin-dependent kinase 1* (CDC28). All these proteins are known to control multiple processes in cell cycle procession and cellular growth: CDC28, the catalytic subunit of the main yeast cyclin-dependent kinease, drives progress through the cell cycle and is essential for the start of the cell cycle [301, 445]. PAT1 is an important element in developmental control [489]. SSC1 is a component of the endonuclease I-Sce that is important for polypeptidechain cleavage during the cell cycle [179].

The two measures, time-averaged authority α_i and first-order-mover score m_i, identify high ranks for different proteins. Thus, proteins that are important for an organism averaged over time do not necessarily change their centrality drastically. Although this is not surprising in this context, qualitatively different results were obtained on computations of a temporal network of the United States Ph.D. exchange in mathematics, where high-centrality universities correlate with high-mover-score universities [451].

In Fig. 4.4, we show the time-averaged centrality α_i and first-mover score m_i of all proteins. The two measures are positively correlated (Pearson correlation $\rho_{\text{Pearson}} \approx 0.20$ and Spearman correlation $\rho_{\text{Spearman}} \approx 0.08$), so proteins that are central tend to change their centrality more than low-centrality proteins. Many proteins,

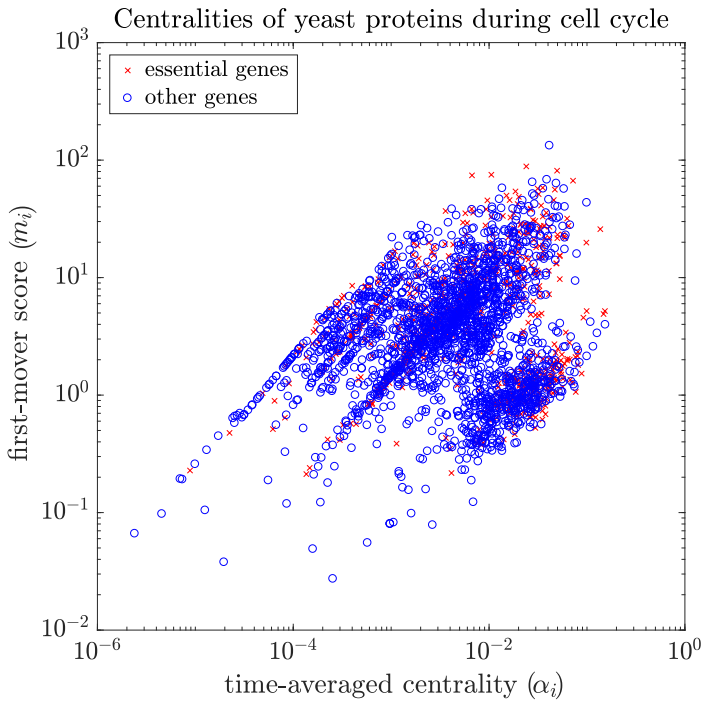


Figure 4.4: Time-averaged centralities α_i and first-mover scores m_i for proteins during the yeast cell cycle. We show essential genes as red crosses and all other genes as blue circles.

however, depart from this trend.

We distinguish between proteins that are products of essential genes, as listed in the Saccharomyces Genome Database [82] (shown as red crosses), and all other proteins (shown as blue circles). Of the 2,766 proteins, 679 are classified as essential. The essential genes have a mean time-averaged centrality of $\langle \alpha_i \rangle_{\text{essential}} \approx 0.0159$, which is approximately 50 % larger than the $\langle \alpha_i \rangle_{\text{other}} \approx 0.0096$ of all other genes. Essential genes also have a larger mean first-mover-score than other genes with $\langle m_i \rangle_{\text{essential}} \approx 8.0851$ versus $\langle m_i \rangle_{\text{other}} \approx 5.6451$. A two-sample t-test rejects the null hypothesis that essential and other genes are sampled from the same population; the p-values are 9×10^{-21} (time-averaged centrality) and 5×10^{-10} (first-mover score).

The above computations indicate that essential genes tend to have a higher time-averaged centrality α_i and first-mover score m_i than non-essential genes. The former is known from the analysis of the static representations of PINs [181, 216]. The latter suggests that genes can be essential because they are crucial during a certain stage of the cell cycle but not at all times.

An example of this is *septin* CDC11, which has the second largest mover score $m_i \approx 88$. Similar to actin (ACT1), septins form complexes and are a part of the cytoskeleton [318]. CDC11 is a mitotic septin, which localises to the bud neck before cell division and forms filaments and controls the mitotic cell cycle [231, 477]. The formation of these filaments is essential for cell survival [299]. In contrast to actin, which is an essential part of the cytoskeleton and has a large time-averaged centrality but a smaller mover score $m_i \approx 5$, the presence of CDC11 is only necessary during mitosis. This indicates that a gene can be essential for yeast due to its centrality during a specific phase of the cell cycle rather than its time-averaged centrality.

The difference in centrality between essential and non-essential genes is statistically significant. But the relationship between 'essentiality' and temporal centrality is not straightforward. For example, the top-ranked genes from both measures, SSB2 and PAT1, are each non-essential (see Table 4.2). SSB2 is a molecular chaperone that binds to the ribosome and assists with cotranslational folding [360]. Many proteins are expected to fold cotranslationally [114, 214, 255], so SSB2 may have a crucial function in the translational process. Its large time-averaged centrality $\alpha_i \approx 0.15$ is

potentially an indicator for this. Its deletion is non-lethal, however, as some of its functions can be covered by the structurally similar SSA and SSB1 proteins [325]. Nevertheless, such a deletion comes with drawbacks, as mutant SSB1–SSB2 strains grow slowly, contain a small number of translating ribosomes, and are hypersensitive to several inhibitors of protein synthesis.

4.6.2 Centrality Trajectories during the Yeast Cell Cycle

We now examine the authority centrality trajectories of proteins during the yeast cell cycle. For this, we plot the *centrality trajectory* as the conditional authority centralities (see Subsection 4.5.2) in each layer. The centrality trajectories (4.6) depend on the choice of the inverse coupling ε. Therefore, we first investigate the centrality trajectory of a single protein to try to identify an appropriate choice of ε; using that value ε, we then examine the centrality trajectory of other proteins.

Influence of the Inverse Coupling ε on Centrality Trajectories

In Fig. 4.5, we show the authority centrality trajectories of the protein expressed by the SLA2 gene[7], which has rank 1,947 and 24 for time-averaged centrality and first-mover score, respectively. It is not extremely important averaged over time, but its importance changes strongly. This is consistent with experimental results in which the deletion of this gene is non-lethal but results in cell-cycle delay [154]. SLA2 is one of many proteins organising actin filaments, specifically for bipolar bud site selection [371]. Therefore, we expect its centrality to be large in the early S

[7]The name stands for 'Synthetic Lethal with ABP1' [195].

phase and small otherwise. For illustration, we chose this protein as its centrality trajectory is strongly driven by the cell cycle. We show the results of more proteins in Subsection 3.6. We study the conditional node-layer centralities versus time t for $\varepsilon = \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. Recall that the conditional node-layer centrality indicates the centrality of node-layer pair (i, t) with respect to all node-layer pairs (j, t) at time t, as discussed in Subsection 4.5.2. Thus, it represents the importance of a protein in comparison with all proteins in the same time-layer.

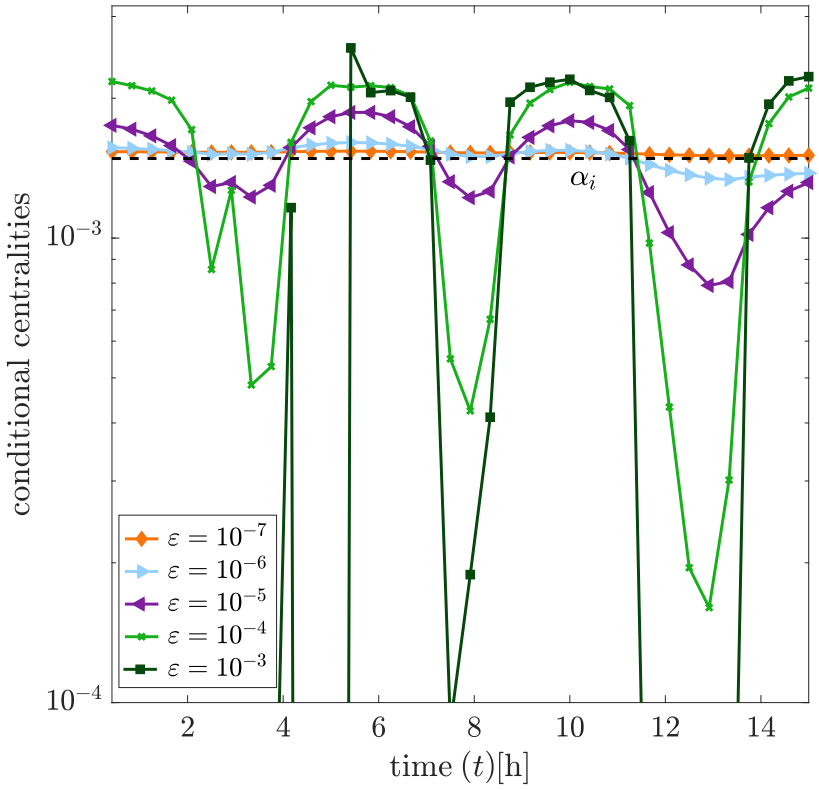


Figure 4.5: Centrality trajectories of SLA2 during yeast cell cycles for several values of the coupling parameter. We choose coupling parameter $\varepsilon = \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. We indicate the normalised time-averaged authority $\alpha_i \approx 0.0015$ with the horizontal dashed line. Note the logarithmic scale on the vertical axis.

We also show the time-averaged authority α_i, which is normalised such that

$\sum_{i=1}^{N} \alpha_i = 1$, as a horizontal dashed line. This time-independent value is the conditional node-layer centrality in the limit $\varepsilon \to 0^+$, which closely matches the numerical result for $\varepsilon = 10^{-7}$. For small ε, the trajectory generally varies slowly over time; however, for larger ε there is larger temporal variation from t to $t + 1$.

There is no *a priori* best choice for ε. A strong temporal change of conditional centralities may be desirable for investigations of some temporal properties but not for others. In the following Subsection 3.6, we focus on $\varepsilon = 10^{-5}$ because the visual inspection indicates a temporal oscillation. Smaller values (e.g., $\varepsilon = 10^{-3}$) tend to give a more volatile behaviour that we do not intend for our application. We note that other choices are possible and $\varepsilon = 10^{-4}$ gives qualitatively similar behaviour (results not shown).

Centrality Trajectories

In Fig. 4.6, we show the authority-centrality trajectories of the three top-ranked and three bottom-ranked proteins according to time-averaged centrality and first-order-mover score. As expected, the top centrality-ranked proteins (SSB2, ACT1, and RPN11) have large conditional centralities. Their conditional centralities also do not vary strongly, which is in accordance with the observation that they have small first-order-mover scores $m_i < 6$ (see Table 4.2).

The bottom-ranked proteins according to centrality (FOL1, YFR020W, and SEY1) have conditional centralities smaller than 10^{-6} for most time points. As they are only expressed at $t \leq 2\,\text{hours}$ (SEY1) or $t \geq 12\,\text{hours}$ (FOL1 and YFR020W)

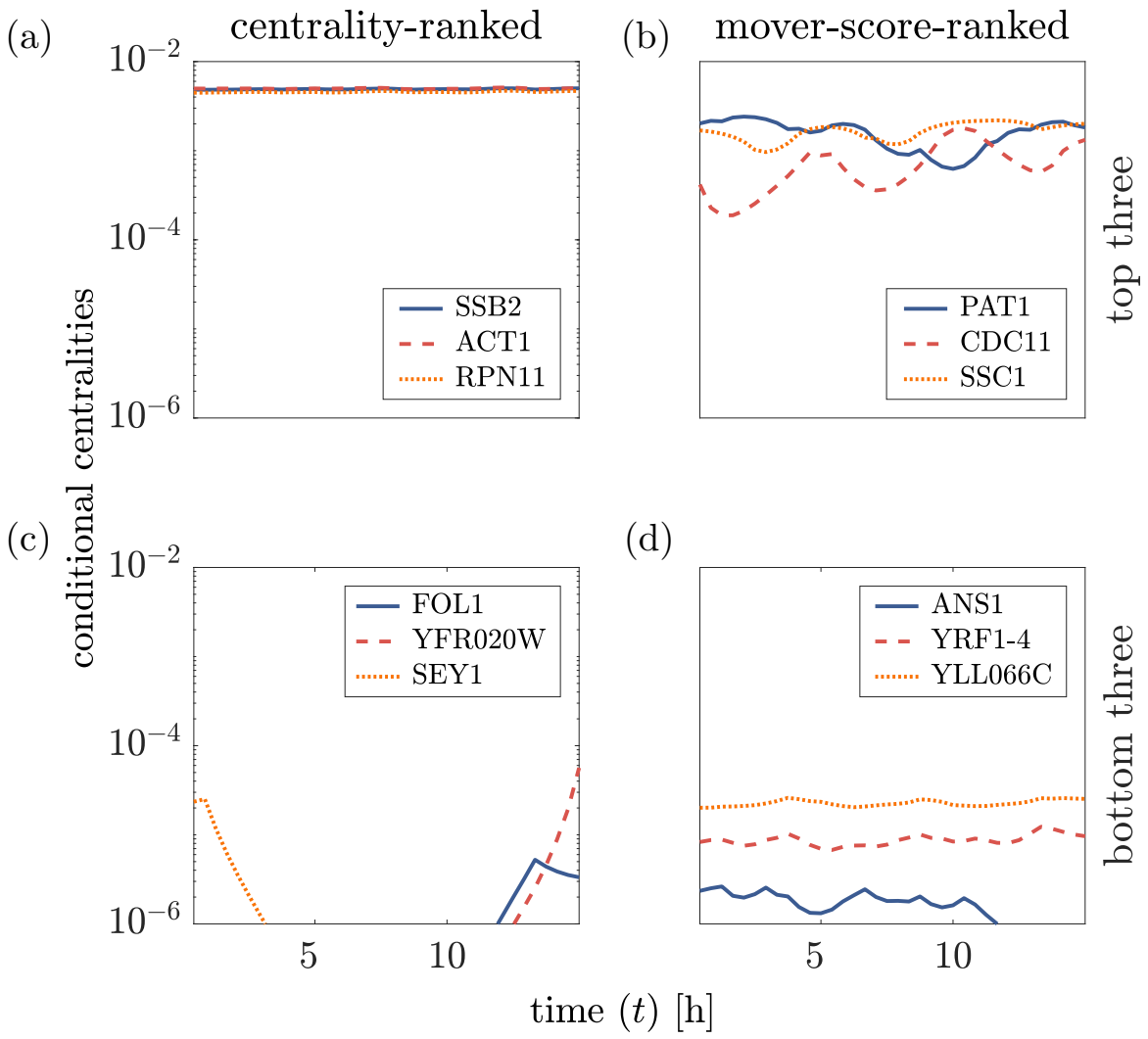


Figure 4.6: Centrality trajectories during yeast cell cycles for the top three and bottom three proteins according to time-averaged centrality α_i and mover score m_i. All trajectories are for $\varepsilon = 10^{-5}$. Note the logarithmic scale of the vertical axes and that we do not show centralities smaller than 10^{-6}.

and have small conditional centralities at these points, the overall time-averaged centralities α_i are small.

We expect the top-ranked proteins according to mover-score (PAT1, CDC11, and SSC1) to change their centrality over time. For these three proteins, we observe the change in centrality as fluctuations. CDC11, which is a protein with crucial importance

during the cell cycle and bud formation [147], has a periodic-seeming oscillation with local maxima at 5 hours and 10 hours. For PAT1, a protein required for the chromosome transmission during cell division [472], the presence of oscillations is less clear. Some of its local minima are at times when CDC11 has local maxima, indicating that both proteins are important at different times during the cell cycle.

The bottom mover-score-ranked proteins ANS1, YRF1-4, and YLL066C have small time-averaged centralities α_i. ANS1 is not expressed for the time-layers $t > 12\,\mathrm{h}$ but nevertheless has a small mover score m_i because even when expressed its centrality is very small around 10^{-5}. The proteins YRF1-4 and YLL066C are expressed at all time points and show only small fluctuations in their centrality trajectories.

4.7 Discussion

This case study of temporal centrality calculation in the temporal PIN of the yeast cell cycle illustrates its applicability for the analysis of biological data. For example, we found that essential genes tend to have larger time-averaged centralities α_i and first-mover scores m_i than other genes. The examination also gives insights into practical considerations (e.g., the choice of ε) concerning temporal centrality in PINs. We found in the strong-coupling regime ($\varepsilon \to 0^+$) that the centrality trajectories vary slowly with time, so they stay very close to the time-averaged centrality α_i.[8] As ε

[8]For some nodes, there is a small (less than 5 %) discrepancy between time-averaged centrality α_i and the centrality trajectory for small values ε (e.g., $\varepsilon = 10^{-7}$; results not shown). Numerical inaccuracies in eigenvector calculations are a potential cause for this discrepancy. Another possible cause is that small finite values of epsilon are still different from $\varepsilon \to 0^+$.

increases, we observe that conditional centrality begins to vary more from one time-layer to the next. For the protein SLA2, multiple choices of ε in this regime indicate an oscillation and thus give qualitatively similar results. However, the magnitude of the observed fluctuations vary strongly. We thus suggest that a detailed analysis of the strength of these oscillations might not be fruitful. We rather focus on their existence because not all proteins show such a behaviour. In the end, we decide on an intermediate value of $\varepsilon = 10^{-5}$ for further analysis because it shows the oscillatory behaviour but seems less sensitive to outliers.

We note that the computation of temporal centrality in the strong-coupling limit, which does not require one to choose ε — rather $\varepsilon \to 0$ is implied by the computation — gave biological insights that are consistent with existing understanding of the yeast cell cycle: The proteins with large time-averaged centrality α_i (SSB2 and ACT1) have known crucial biological functions, so they are important at all times[9]. By contrast, proteins with large first-mover-scores m_i (CDC11 and CDC28) have important biological functions in some phases of the mitotic cell cycle; so their centralities change. As the yeast cells undergo multiple cell cycles we expect these changes to manifest as approximately periodic fluctuations.

The first-mover-score is only an indication of such fluctuation and could also hint at other temporal changes (e.g., a linear increasing trend or an abrupt increase in one time layer). To directly observe the nature of the temporal change, one has to investigate the

[9]The associated genes could be considered a temporal variant of 'housekeeping genes'. Housekeeping genes are genes that are expressed in all cell types [504].

centrality trajectories. In our computations, we found that some proteins have a roughly constant centrality, others are only active at a small number of time points, and some have an oscillatory behaviour. Investigating centrality trajectories allows one to compare temporal change of multiple proteins. Different proteins have local maxima of centrality at different time points, because they are active at different phases of the cell cycle.

As hypothesised by Yu et al. (2007), in general, more central proteins also tend to fluctuate stronger in their centrality [499]. The most central proteins, however, are not the ones with the largest first-mover scores; rather they are proteins with crucial biological functions and are thus present at all time points. A systematic study of to what extent nodes with large time-averaged centralities intrinsically have larger first-mover-scores, would strengthen this finding.

We found that essential proteins tend to have larger time-averaged centralities and first-mover scores. Based on previous studies of PINs [21, 216], the result for time-averaged centralities is unsurprising, but the first-mover scores give novel insights. It demonstrates that the presence of proteins, such as CDC11 and CDC28, during certain phases of a cell cycle is crucial even though static centrality calculations have not suggested that they are important. Although these insights are not surprising from a gene-expression perspective, such time-dependence has thus far been ignored in the analysis of PIN.

Conversely, a protein with strongly changing expression levels can only have a large first-mover score, if it has also a large static centrality. The cyclin CLN3,

for example, is only expressed in the G_1 stage to promote the transition to the S phase. Its static centrality, however, is small and we observe relatively small time-averaged centralities and first-mover scores. A further investigation of the relationship between the centralities in temporal and static PINs would be fruitful to understand this phenomenon.

Our analysis gives a novel perspective for studying temporal gene expression data integrated with a PIN to evaluate the importance of proteins and how those importances change over time. In this case study, we saw that our results are consistent with existing biological understanding of the cell cycle. We expect that this approach will also give useful insights on other applications. Such analysis is potentially useful for identifying proteins that change their behaviour during the progression of a disease, as a result of medication, ageing, or the development of animal cells (e.g., the model organism *Drosophila melanogaster*).

As we constructed the temporal PIN with gene-expression data, it comes with the associated caveats. While gene expression gives some (limited) information about the rate of protein synthesis, it is not necessarily directly connected with protein abundance. As we discussed in Subsection 2.1.1, the rate of degradation is another factor that strongly influences the presence of proteins. The direct measurement of protein abundance would overcome this problem.

4.7.1 Future Directions

Although our investigation of centrality in a temporal PIN of the yeast cell cycle gave reasonable biological insights, there are several shortcomings in this approach that are desirable to address.

First, we construct temporal PINs binary; that is, a protein is either present at a given time point or not. Thresholding gene expression data is a straightforward way to construct the temporal PIN, but a more sophisticated method (e.g., involving edge-weighted networks) can potentially give a more nuanced picture of temporal changes of centralities. It would, for example, allow to represent a weak (but present) interaction between proteins of small abundance. The temporal eigenvector-based centrality framework, as introduced by Taylor et al. (2017) allows edge weights [451]. However, there is not a canonical way to integrate gene-expression information as edge weights and as such, so it is important to consider other approaches.[10]

Second, the employed multilayer approach for the network construction uses discrete time points as layers. The gene expression during the cell cycle was measured at discrete time points, so this was appropriate for this application. This will often be the case for gene-expression measurements. However, recent advances in centrality analysis for continuous temporal networks [6] may be more appropriate for biological data with continuous time measurements. For example, in single-cell expression data, one

[10]In fact, we undertook such efforts and the results gave very 'unstable' centrality trajectories, in the sense that, even for small ε, the centrality trajectories fluctuated strongly. This does not seem appropriate for the present application, and it may be a consequence of the broad distribution of gene-expression measurements or high level of measurement error. In principle, it may be possible to use some sort of preprocessing to ameliorate such problems.

can estimate a continuous *pseudotime* [381]. Pseudotime is an unobserved (latent) dimension that is an indicator of cells' progress through a dynamic process (e.g., an organism's cell cycle). A temporal centrality analysis may help reveal the influence of individual proteins over time for this type of data.

Third, one might consider some methodological extensions of the eigenvector-based temporal centrality. For example, one might couple different proteins with different coupling parameters ε or couple temporal layers that are not equidistant in time..

Fourth, we took mRNA-abundance data from oligonucleotide assays to create a temporal PIN. The advances of RNA sequencing might allow the application of this approach to RNA-seq data which measures the mRNA abundance directly and thus has less methodological problems concerning normalisation.

5

Node-weighted Protein Interaction Networks

Contents

5.1 Motivation 144
5.2 Model Specification 147
5.2.1 Asymptotic Discussion 152
5.3 Synthetic Examples 154
5.4 Application to Tissue RNA-Abundance Data 159
5.4.1 Choice of the Resolution Parameter γ 160
5.4.2 Comparison with Null Models 162
5.4.3 Comparison across Tissues 166
5.4.4 Comparison of Detected Partitions with Gene Ontology Annotations 168
5.5 Conclusions 173

This chapter is based on a manuscript that is joint work with my supervisors Jonny Wray, Charlotte M. Deane, and Mason A. Porter.

5.1 Motivation

PINs are mathematical representations of interactions between proteins. The notion that there is one human PIN is a simplification, as discussed in Chapter 1. Proteins and their interactions are context-specific: they vary between tissues, over time, and between healthy and diseased states [475]. In particular, two proteins that are reported to interact with each other might not be present in the same tissue. Therefore, human PINs vary across tissues and tissue-specific gene expression influences the function of proteins [132]. As the proteins in structural modules in PINs often have common biological functions [262, 428], one might assume that these structural modules also differ across tissues. In this chapter, we examine this by constructing tissue-specific PINs and detecting community structure in them.

Databases (e.g., BioGrid [432]) aggregate protein interaction data from different experimental conditions, tissues, cell types, and across time. Therefore, they are intrinsically not context-specific. Furthermore, many experimental techniques are *in vitro* methods and are inherently not physiological [376] (see Subsection 2.2.1). This makes it challenging to use a database of PPIs to obtain context-specific insights into PINs. In this chapter, we aim to solve this problem by integrating a general PIN with context-specific data. In particular, we use *RNA-abundance measurements*, i.e., measurements of the level of messenger ribonucleic acid molecules (mRNA) in cells [474]. mRNA molecules carry information of DNA to ribosomes, which are the sites of protein synthesis in the cell (see Subsection 2.1.1 for more information about

protein biosynthesis). While there is some variability between mRNA abundance and protein abundance, they are correlated with each other [276]. Such mRNA abundance data is available for many context-specific situations, e.g., over time [346] or in different cancer cell lines [265]. We focus on tissue-specific mRNA abundance [461].

Different approaches have been used to represent and analyse context-specific PINs [496]; we discussed some of them in Subsection 1.2.1. A common approach is thresholding PINs [79, 239] which generates a context specific PIN from a general PIN by keeping only proteins that are present above a certain threshold. We used this approach to generate the temporal PIN of *S. cerevisiae* in Chapter 4. This approach, however, neglects the nuances of gene expression, as it makes a binary selection: a node is either kept in the network or it is not. In this chapter, we instead use RNA abundance levels without binarization to create mathematical representations of tissue-specific PINs.

We integrate mRNA abundance data with a PIN in the form of *node-weighted networks*, which we introduce in detail in Section 5.2. By incorporating mRNA abundance as node weights, we aim to obtain insights into the structure and function of PINs of tissues[1].

We then detect community structure in the tissue-specific PINs. In Fig 5.1, we show an example network, whose community structure changes under consideration of the node weights.

[1]In this investigation, we assume that there is injective function from the set of genes to the set of proteins. Thus, each protein and its associated RNA is the exclusive product of exactly one gene. This is a simplification of the protein biosynthesis process (See Subsection 2.1.1). However, it is a common assumption because experimental techniques that are able to distinguish interactions between protein variants are still in early development [92].

Node-weighted networks resemble so-called 'annotated networks' [200, 331]. Annotated networks are networks with metadata. Metadata is additional information about the nodes (or edges) in a network that are not used to create the network. Often, such information is used to validate detected partitions by comparing them with each other. In a social network, for example, we can compare detected partitions with information about the individuals. Traud et al. (2012) find that Facebook friendship networks in American universities often form predominantly in the same dormitory residence [457]. GO annotation data are a form of metadata and the comparison of network partitions with them, as presented in Section 2.9, is one approach to validate detected partitions in PINs.

Such metadata, however, does not always correlate with the communities in networks (e.g., if the community structure and the metadata capture different features of a network [359]). Recent approaches allow the consideration of such metadata for community detection itself [200, 331]. This raises the question of why we do not use these approaches for the detection of communities in tissue-specific PINs. The reason is that in the cases of annotated networks the metadata guides the community detection (i.e., some alignment between network structure and metadata is presumed). For tissue-specific PINs this is not necessarily the case.[2] Two genes can have a high expression in a tissue without belonging to the same structural or functional module.

The rest of this chapter is organised as follows. In Section 5.2, we introduce a

[2]In a large scale analysis across tissues, however, this might be the case. In fact, such information can be used to create *gene co-expression networks* [441]

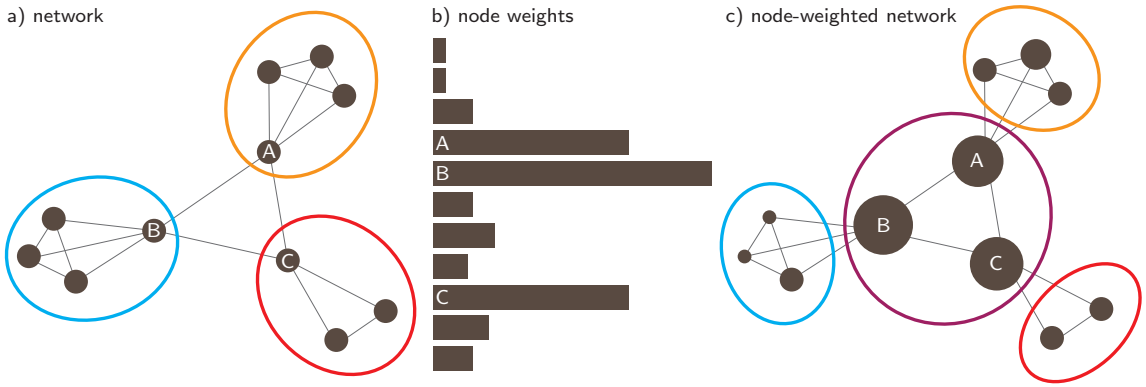


Figure 5.1: We can combine a network (a) with the node weights (b) to obtain a node-weighted network (c). In this schema, the original network has only three modules and the node-weighted network has an additional one. The three central nodes (A, B, and C) have large weights and thus form a module that was not obtained before, while the three other modules decreased in size.

node-weighted modularity function and analytically discuss some of its properties. In Section 5.3, we illustrate the community detection with node-weighted modularity on a synthetic network with overlapping community structure. In Section 5.4, we apply the node-weighted community detection to tissue-specific PINs. In Section 5.5, we conclude.

5.2 Model Specification

Although there is considerable research on networks with edge weights [330], node-weighted networks were introduced only recently [184]. Heitzig et al. (2012) [184] discuss network measures that take into account node weights. They investigated spatially embedded climate networks. In these, nodes represent measurement locations of observables (e.g. precipitation). As these locations are distributed unevenly on the surface of the earth, it is relevant to weight nodes according to surface-area size. These techniques have given insights into moisture transport in South America [501]

and trade balance [479]. Most importantly, this approach avoids systematic biases created by a larger node density at the poles in comparison with equatorial areas. In the context of PINs, we may consider the abundance of proteins as node weights. We then take these node weights into account when examining the network. Proteins with larger abundance have a larger weight in the network measures than proteins with small abundance. We develop our methodology for the detection of community structure in tissue-specific networks. One can, however, also apply the methodology to other node-weighted networks.

A *node-weighted network* is an ordered triple $G = (V, \mathcal{W}, E)$ where there is a set V of nodes that are connected pairwise via edges from the set $E \subset V \times V$. The *node weights* $\mathcal{W}$ are a map $\mathcal{W} : V \to \mathbb{R}_{\geq 0}$ and we denote the weight of node i as w_i[3].

It is insightful to generalise network properties to such node-weighted networks [184]. We focus on the detection of community structure in such node-weighted networks. In contrast to the approach in Heitzig et al., we include a free parameter s that allows us to scale to which extent we consider the node weights. For community detection in node-weighted networks, we define a node-weighted variant of the modularity function (see Section 2.6). Recall, that the modularity quality function for a network without node weights is

$$Q = \frac{1}{2m} \sum_{ij} Q_{ij} \delta(g_i, g_j) \text{ with } Q_{ij} = \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) . \tag{5.1}$$

The matrix elements Q_{ij} are positive if the connection between nodes i and j is

[3]Here we choose the node weights to be real and nonnegative. This is not necessarily the case in all applications.

larger than expected under the Newman–Girvan null model and negative if they are smaller than expected.[4] In this chapter, we refer to this modularity without node weights as 'ordinary' modularity.

Intuitively, we want to incorporate the node weights w_i into the modularity function in a way that changes the contribution of all terms $\{Q_{i1}, \dots Q_{in}\}$ associated with node i as a function of its weight w_i. Nodes with a larger weight should have a stronger contribution to the modularity function than nodes with a small weight. To do this, we define the *node-weighted modularity* as

$$\begin{aligned} Q^{\mathrm{nw}} &= \frac{1}{2m} \sum_{ij} W(w_i, w_j; s) Q_{ij} \\ &= \frac{1}{2m} \sum_{ij} W(w_i, w_j; s) \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(g_i, g_j) \,. \end{aligned} \tag{5.2}$$

The *weight function* $W(w_i, w_j; s)$ scales the weight of a particular element of the modularity matrix Q_{ij} and is a function of the associated node weights w_i and w_j. Different functional relationships $W(w_i, w_j; s)$ are reasonable. We use a function with sigmoidal shape, such that increasing the weight w_i of node i results also in larger values for the weight function $W(w_i, w_j; s)$. At the same time there is a saturation, i.e., increasing a large weight w_i has only small influence. Artificial neural networks often use similar functions [221]. In this thesis, we use a bivariate logistic function

$$W(w_1, w_2; s) = \left((1 + \exp\left(-2s \cdot w_1 + s\right))(1 + \exp\left(-2s \cdot w_2 + s\right)) \right)^{-1} \,. \tag{5.3}$$

[4] There are other null models, such as ones that incorporate spatial information [396]. See Section 2.6 for a literature review.

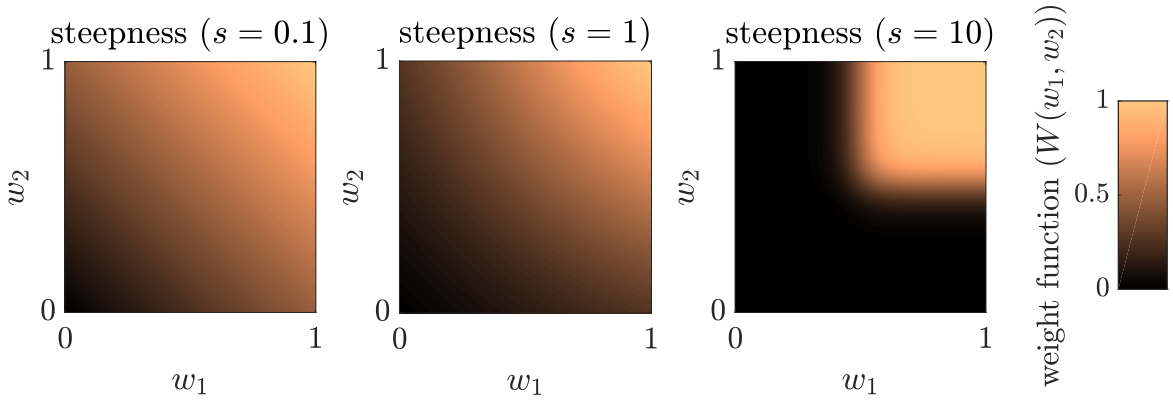


Figure 5.2: Weight function $W(w_1, w_2)$ for three different values of the steepness parameter $s \in \{0.1, 1, 10\}$. A small steepness s leads to a shallow slope with increasing w_i and a large steepness s leads to a steep increase. Note that $W(0, 0) = 0$ and $W(1, 1) = 1$ for all choices of $s > 0$.

The *steepness parameter* $s > 0$ tunes the consideration of the node weights for the community detection. For small choices of s the weight function W is 'flat', while large s leads to a steep increase (see Fig. 5.2). We choose this particular function $W(w_1, w_2; s)$ for three reasons. First, $W(w_1, w_2; s)$ is monotonic increasing in w_1 and w_2. This ensures that increasing the weight w_i of node i also increases its contribution to the node-weighted modularity (5.2). Second, the weight function approaches $W(0, 0; s) = 0$ and $W(1, 1; s) = 1$ for all steepness parameters $s > 0$. Nodes with weight $w_i = 0$ do not contribute to the modularity function and so $Q^{\text{nw}} \leq Q$. Third, the weight function is symmetric, so $W(w_1, w_2; s) = W(w_2, w_1; s)$.

Although definition (5.2) seems superficially similar to the community detection in edge-weighted networks, it is fundamentally different. Most crucially, our definition puts a weight of $W(w_i, w_j; s)$ not exclusively on edges $(i, j) \in E$ but also on modularity matrix elements $Q_{i'j'}$ of non-edges $(i', j') \notin E$.

The combination of both types of annotations — the detection of node-weighted community structure in a network with edge weights — is possible but beyond the scope of this chapter. Such an approach could be fruitful for examining tissue-specific PINs with edge-weights whose weights represent interaction strengths.

In Appendix D, we compare the definition of the node-weighted modularity (5.2) with the maximum likelihood (ML) function of a SBM. Both approaches are equivalent under specific conditions. Therefore, we can understand the node-weighted modularity not only as a multiplication of the elements of an ordinary modularity matrix with the weighting function (5.3) but also as the weighting of edges (and non-edges) in the ML function of an SBM.

Optimisation of Node-weighted Modularity

To detect communities in a node-weighted network we aim to find group assignments $\mathbf{g}$, which maximise the node-weighted modularity function (5.2). Optimization of a modularity function is NP-hard [58]. Numerical approaches, however, can give approximate solutions. We employ the GenLouvain community-detection algorithm [217], because it can be used with any quality function specified in terms of a modularity matrix, so we can use it with the node-weighted modularity matrix Q_{ij}^{nw} (for details about the GenLouvain algorithm see Appendix A).

The computation of a node-weighted modularity matrix is algorithmically simple: First, we define a normal modularity matrix $\mathbf{Q}$ with elements Q_{ij}, as outlined in Section 2.6. Second, we compute the weight function $W(w_i, w_j)$ for each pair of nodes

(i, j). Third, we compute each element of the node-weighted modularity

$$Q_{ij}^{\text{nw}} = W(w_i, w_j) \times Q_{ij} \,. \tag{5.4}$$

Passing this modularity to the GenLouvain algorithm then returns a partition $\mathbf{g}$ of the node-weighted network.

5.2.1 Asymptotic Discussion

We now examine analytically how the choice of the steepness parameter s influences the node weighted modularity Q^{nw}. Specifically, we investigate two limiting cases $s = 0$ and $s \to \infty$ and the case of uniform node weights.

For minimum steepness $s = 0$, the weight function $W(w_i, w_j; s = 0) = 1/4$ is identical for all pairs of nodes, independent of their node weights w_i and w_j. Thus, the node-weighted modularity Q^{nw} is a rescaled version of the original modularity Q, with $Q^{\text{nw}} = 4Q$. For any partition $\mathbf{g}$, the node-weighted modularity is therefore a quarter of the modularity. Let $\mathbf{g}^\star$ denote a partition that yields a global maximum of the modularity function. It is then also a global maximum of the node-weighted modularity function because the rescaling does not change the position $\mathbf{g}^\star$ of the maximum. A 'perfect' modularity maximisation algorithm, i.e., one that always returns one of the best partitions, would then return partition $\mathbf{g}^\star$ for both cases. Modularity maximisation algorithms, however, are approximate, thus we do not expect two runs to return *exactly* the same partition.

For infinite steepness $s \to \infty$, the weight function yields

$$\lim_{s\to\infty} W(w_1, w_2) = \begin{cases} 1\,, & \text{if } w_1 > 1/2 \text{ and } w_2 > 1/2 \\ 0 & \text{otherwise}\,. \end{cases} \tag{5.5}$$

Thus, the weight function approaches a two-variable Heaviside function. The node-weighted modularity Q^{nw} is then

$$\lim_{s\to\infty} Q^{\text{nw}} = \frac{1}{2m} \sum_{\substack{ij \\ \{w_i>1/2\} \\ \{w_j>1/2\}}} \left(A_{ij} - \frac{k_i k_j}{2m}\right) \delta(g_i, g_j)\,. \tag{5.6}$$

The node-weighted modularity (5.6) is a function of the group assignment g_i of nodes that have a node weight $w_i > 1/2$. The group assignments g_i of other nodes do not contribute to the modularity value.

Node-weighted Network with Identical Node Weights

For a node-weighted network $G = (V, \mathcal{W}, E)$ with uniform weights $\mathcal{W} : V \to c$ with $c \in \mathbb{R}_+$, the node-weighted modularity (Eqn. (5.2)) simplifies to

$$\begin{aligned} Q^{\text{nw}} &= \frac{C}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m}\right) \delta(g_i, g_j) \;\text{ with } C = W(c, c; s) = \text{Const.} \\ &= CQ\,. \end{aligned} \tag{5.7}$$

For $c > 0$, this is a rescaled version of the original modularity Q, detected community structure does not depend on the node weights w_i. Varying the steepness parameter s does not change the obtained community structure.

If all nodes have a zero weight $w_i = 0$, the node-weighted modularity (Eqn. (5.2)) is $Q^{\text{nw}} = 0$. One thus cannot detect community structure, because any partition $\mathbf{g}$ yields a modularity value of $Q^{\text{nw}} = 0$.

5.3 Synthetic Examples

We now demonstrate that the consideration of node weights can change detected community structure in node-weighted networks. To do this, we construct a synthetic node-weighted network and use the GENLOUVAIN algorithm to maximise the node-weighted modularity function Q^{nw}.

Network Construction

We construct a synthetic node-weighted network $G = (V, \mathcal{W}, E)$ with $|V| = n = 500$ nodes. To construct the edges, we use a SBM with two overlapping modules of equal size (See Subsection 2.4.3 for a definition of the SBM.) The connection probabilities are $p_{\text{in}} = 0.5$ inside a module and $p_{\text{ex}} = 0.05$ between modules. The two modules of size 250 overlap each other in a set of 50 nodes. These nodes belong to both modules and thus connect to both of them with connection probability p_{in}. In Fig. 5.3a we show one realisation of such a network with the two modules in red and blue and the overlap nodes in purple. The nodes $\{i = 1, \dots, 250\}$ belong to the red module and the nodes $\{i = 251, \dots, 500\}$ to the blue modules. The overlap nodes are therefore $\{i = 226, \dots, 275\}$.

We determine node weights $\mathcal{W}$ in a way that gives a larger weight to the nodes in the red module and the purple overlap. To achieve this, we choose $w_i = 1$ for all $\{i = 1, \dots, 275\}$, and $w_j = w_{\text{blue}}$ for all other nodes $\{j = 276, \dots, 500\}$, which are in the blue module.

Community Detection in Synthetic Node-weighted Networks

We use the GENLOUVAIN algorithm for node-weighted modularity maximisation to obtain community structure. First, we illustrate two example cases. Second, we examine the influence of the steepness s and the weights w_{blue} in the blue module on the detected community structure.

In Fig. 5.3b, we show detected partitions for $w_{\text{blue}} = 0.2$ and two choices for the steepness $s = 0$ and $s = 10$. As we derived in Subsection 5.2.1, in the case $s = 0$ the node-weighted modularity is a rescaled version of the modularity without node weights. Therefore we should receive a partition that is similar to the case of neglected node weights. For both choices of the steepness s, we successfully detect two modules, which we illustrate in blue and red. Of special interest is the module assignment of the 50 nodes in the overlap because they connect with equal probability with both modules. For both steepness values, some of these nodes belong to the red and some of them to the blue module. For $s = 0$, we observe that these nodes are distributed approximately equally between the red and blue module. For $s = 10$, however, almost all overlap nodes belong to the red module. To understand this, we investigate the weight function $W(w_1, w_2)$ for the overlap nodes. All of them have a weight w_1. The values w_2, however, are different for connections with the two modules: we weigh the connections with the red module with $W(1, 1; s = 10) \approx 1$ and those with the blue module with $W(1, w_{\text{blue}} = 0.2; s = 10) \approx 0.0025$. This illustrates that the red module 'pulls' the overlap nodes to its side. The blue module is still detected, even though all

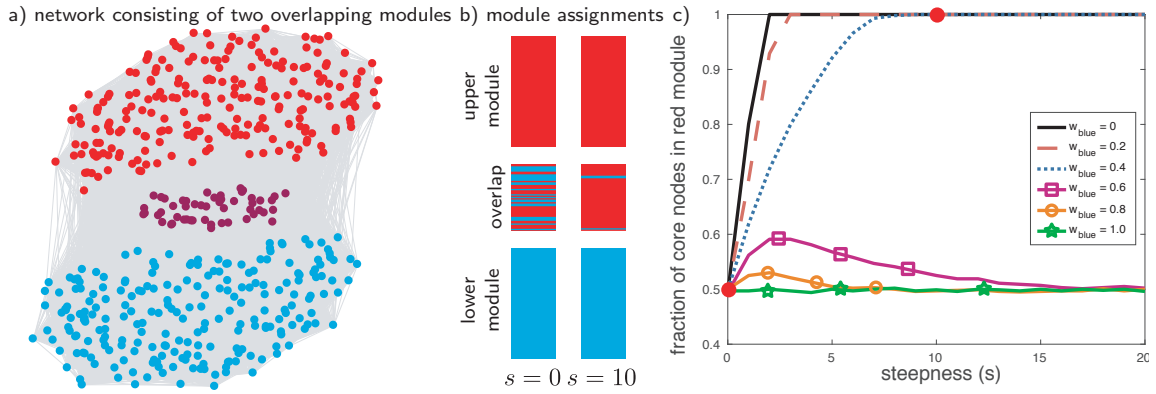


Figure 5.3: Community detection in a synthetic node-weighted network. (a) A network with $N = 500$ nodes in two modules (red and blue) with an overlap of 50 nodes, which we show in purple. (b) For $w_{\text{blue}} = 0.2$ and $s = 0$, we successfully detect the two modules and the overlapping nodes are distributed equally between them. With $s = 10$, the red module, which has larger node weights, dominates the modularity function (5.2) and so almost all nodes in the overlap are assigned to the red module. (c) We show the fraction f of nodes in the overlap that are assigned to the red module for different weights w_{blue} and steepness parameter s. The red dots indicate the parameter combinations we used in panel (b).

its nodes have small weights, so their weight function $W(w_{\text{blue}}, w_{\text{blue}}; s = 10) \approx 6 \times 10^{-6}$ is small. Only the nodes in the overlap have different module assignments from the the $s = 0$ case. This computation illustrates qualitatively that the consideration of node weights can change the detected community structure.

We now investigate the change of detected community structure quantitatively and vary the steepness parameter s and the weights w_{blue} in the blue module. For this, we compute

$$f = \frac{\text{overlap nodes in red module}}{\text{size overlap}} \in [0, 1]\,. \tag{5.8}$$

This fraction f is a measure of the extent to which the nodes in the overlap belong to the red module. In Fig. 5.3c, we show $f(s)$ as a function of the steepness s for $w_{\text{blue}} \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For all parameter choices, the fraction $f \gtrapprox 0.5$. A

fraction of $f = 0.5$ indicates that half of the nodes belong to the red and half of the nodes belong to the blue module. For minimum steepness $s = 0$, all curves coincide at $f \approx 0.5$, as the node-weights are ignored and this setting is identical to ordinary modularity maximisation.

For $w_{\text{blue}} = 1$, the weight of the red nodes, the fraction $f \approx 0.5$ is almost constant. As the node weights are identical for all nodes, varying the steepness s only scales the modularity function; we so obtain a similar community structure as with a normal modularity function for all choices of the steepness s. This agrees with our theoretical considerations using uniform node weights in Subsection 5.2.1.

When $w_{\text{blue}} < 1$, we observe two distinct behaviours. For $w_{\text{blue}} \leq 0.5$ the curves increase monotonic and approach $f = 1$ as $s \to \infty$. Thus, all nodes in the overlapping module belong to the red community, which has larger node weights. Smaller weights w_{blue} for the blue nodes lead to $f(s)$ having a steeper slope. For $w_{\text{blue}} > 0.5$, the fraction $f(s)$ increases with s, reaches a maximum, and asymptotically approaches $f = 1/2$. Thus, the nodes are distributed evenly between red and blue modules.

We can understand the two different asymptotic limits, $f = 0.5$ for $w_{\text{blue}} > 0.5$ and $f = 1$ for $w_i \leq 0.5$, from our theoretical considerations, in Subsection 5.2.1. For $s \to \infty$, the module assignment of nodes with weights $w_{\text{blue}} \leq 0.5$ do not contribute to the node-weighted modularity function. Therefore, all nodes in the overlap belong to the red module and $f \to 1$. For weights $w_{\text{blue}} > 0.5$, however, the weight function $W \approx 1$ and all nodes contribute equally to the node-weighted modularity function,

so the node-weighted modularity function is the same as in an unweighed network. The modularity maximisation algorithm returns therefore a partition similar to using modularity without node-weights.

With $w_{\text{blue}} > 0.5$ for the blue nodes, we observe a maximum in the curves $f(s)$.

This occurs because the weight function $W(w_{\text{red}} = 1, w_{\text{blue}}; s)$ between modules grows faster with steepness s than the weights $W(w_{\text{blue}}, w_{\text{blue}}; s)$ inside of the blue modules. Specifically, the weights between the modules are

$$W(w_{\text{red}} = 1, w_{\text{blue}}; s) = \frac{1}{(1 + \exp(-s))(1 + \exp(-2s \cdot w_{\text{blue}} + s))}, \tag{5.9}$$

and the weights inside the blue module are

$$W(w_{\text{blue}}, w_{\text{blue}}; s) = \frac{1}{(1 + \exp(-2s \cdot w_{\text{blue}} + s))^2}. \tag{5.10}$$

Their difference $\Delta W(S) = W(1, w_{\text{blue}}; s) - W(w_{\text{blue}}, w_{\text{blue}}; s)$ gives the weight between modules in comparison to the weight inside of the blue module. It is equal to

$$\Delta W(s) = \frac{1}{(1 + \exp(-2s \cdot w_{\text{blue}} + s))} \left(\frac{1}{(1 + \exp(-s))} - \frac{1}{(1 + \exp(-2s \cdot w_{\text{blue}} + s))} \right). \tag{5.11}$$

Evaluating the maxima of this function numerically yields $s \approx 1.96$ for $w_{\text{blue}} = 0.8$ and $s \approx 2.98$ for $w_{\text{blue}} = 0.6$, which agree with the observed maxima of $f(s)$ in Fig. 5.3c. Note that $\lim_{s\to\infty} \Delta W = 0$, which supports our asymptotic discussion and our numerical results that the overlap is distributed equally among the two modules and so $f \approx 1/2$.

Our examination of this synthetic node-weighted network indicates that node weights can influence detected community structure. Specifically, the overlap between

modules, thus nodes that belong to multiple modules, might change module assignment when considering node-weights. In the setting of PINs, modules overlap because proteins can have multiple functions that are organised in structural modules [274]. Depending on the context (e.g., gene expression in a certain tissue type) the proteins' might change the module they belong to and thus also their biological function.

5.4 Application to Tissue RNA-Abundance Data

Now we combine a PIN of human proteins with tissue-specific gene-expression data to construct node-weighted PINs. We use RNA-abundance data from the *Human Protein Atlas* [367, 461]. Uhlén et al. (2015) performed mRNA sequencing on Illumina HiSeq2000 and Illumina HiSeq2500 high-throughput sequencing system and they computed abundance measures using Kallisto v0.42.4. [59]. For each mRNA, the abundance is measured 'Transcript Per Million' (TPM). If there are multiple genes encoding the same protein a gene's mRNA abundance is the sum of the mRNA abundance of all its protein-coding transcripts.

The data set contains mRNA abundance of 13 290 protein-encoding genes across different tissue types. As an example, we focus on the data from two tissues, 'adipose tissue' and 'suprarenal glands'. While both tissues may produce hormones that influence each other [225] and have certain 'housekeeping' functions that are mutual for all human tissues[5], we expect them to also have tissue-specific biological functions [129]. Adipose

[5]The *Human Protein Atlas* concludes that 7 319 out of 19 613 protein-encoding genes are expressed in all tissues and are therefore potential housekeeping genes [367]

tissue (i.e. body fat) is a connective tissue whose main function is the storage of energy and the supply of free fatty acids [393]. The suprarenal glands produce hormones, that help in the regulation of metabolism, blood pressure, and stress response [72].

We use the mRNA abundance data to create node weights $w_i \in [0, 1]$. Specifically, suppose that x_i is the mRNA abundance associated with protein i. The weight of node i, which represents this protein, is then

$$w_i = \frac{\log_2 x_i}{\log_2 x_{\max}}\,, \tag{5.12}$$

where $x_{\max}$ is the maximum RNA abundance across all protein-encoding genes. For the weight function $W(w_i, w_j; s)$ we use (5.3).

We use the *Homo sapiens* BioGRID data base (version 3.4.146) to create an undirected PIN from these 13 290 proteins [77, 432]. After removing proteins for which no interaction is known, we obtain a network of $n = 10\,394$ proteins. In total, we find $m = 109\,713$ pairwise interactions between them, so the network has a density of $\rho \approx 0.002$. The node set V and edge set E are identical across the two tissues. Only the weights differ, as the mRNA abundance differs across tissues.

5.4.1 Choice of the Resolution Parameter γ

As discussed in Section 2.6, the choice of resolution parameter γ has a strong influence on the detected community structure. Smaller resolution parameters typically yield a smaller number of larger modules. Larger resolution parameters yield a larger number of smaller modules. In the study of networks in general and PINs specifically,

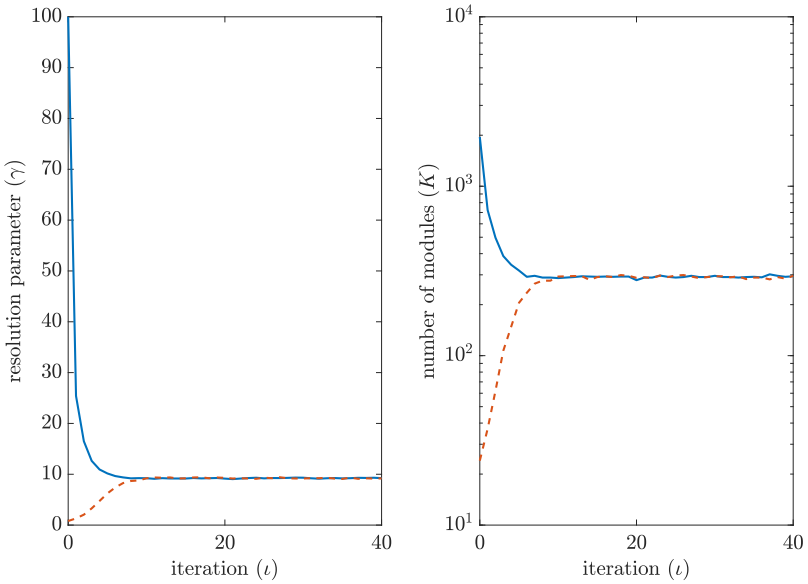


Figure 5.4: To fix the resolution parameter γ for the node-weighted community detection, we use an iterative algorithm for finding an 'optimal' choice of γ. We start the procedure at two drastically different resolution parameters $\gamma_0 \in \{1, 100\}$. The former returns a partition into 24 modules and the latter returns a partition into almost 2 000 modules. For each iteration ι we give the updated resolution parameter γ in panel (a) and the number of detected modules n_c in panel (b). We indicate this output with dashed red and solid blue curves for the two different initial conditions. For both starting values γ_0, the procedure converges to $\gamma \approx 9.1$, which returns a partition into 287 modules.

there is not one correct choice of γ. Instead, different choices may reveal different organisational principles [262]. In our investigation, however, we want to focus on the influence of the steepness parameter s, so we fix the resolution parameter γ. We use an iterative procedure that converges to a value of γ that 'optimally' describes the detected community structure in a network. This procedure was introduced in [329]. It is a special case of the algorithm by Pamfil et al. [352], which we describe in Appendix C.

We show the results of this iterative estimation of the resolution parameter γ in Fig. 5.4. We use two different initial resolution parameter values: $\gamma_0 \in \{1, 100\}$. We use them in the node-weighted modularity (5.2) and use GenLouvain to obtain

partitions. The former yields a partition into 24 modules and the latter yields a partition into almost 2000 modules. After each iteration, we obtain a partition with n_c modules. We use this partition to estimate a new resolution parameter γ. For both initial conditions, the iterative procedure converges in approximately twenty steps to a resolution parameter of $\gamma \approx 9.1$. The number n_c of detected modules fluctuates slightly around 287, as the GenLouvain algorithm is non-deterministic.

The convergence of the iterative algorithm does not guarantee that the obtained community structure is biologically relevant. Instead, it arises from the theoretical consideration that for this value the network closely resembles a particular SBM. Nevertheless, we choose this resolution parameter $\gamma = 9.1$ for future analysis. In comparison with social networks [329], this is a high γ value, which indicates that we find small and densely connected modules.

5.4.2 Comparison with Null Models

In this subsection, we discuss the influence of the steepness parameter s on the detected community structure in the tissue-specific PIN of adipose tissue. When analysing such partitions, it is important to consider that modularity-maximisation algorithms always return a network partition. A comparison with null models allows us to examine whether an obtained partition represents meaningful community structures in a network or whether it might have arisen at random. Similar approaches helped identify relevant temporal scales in temporal multilayer networks [31] and incorporate specific features of the network structure (e.g., signed edges [456], space embeddedness [141, 396],

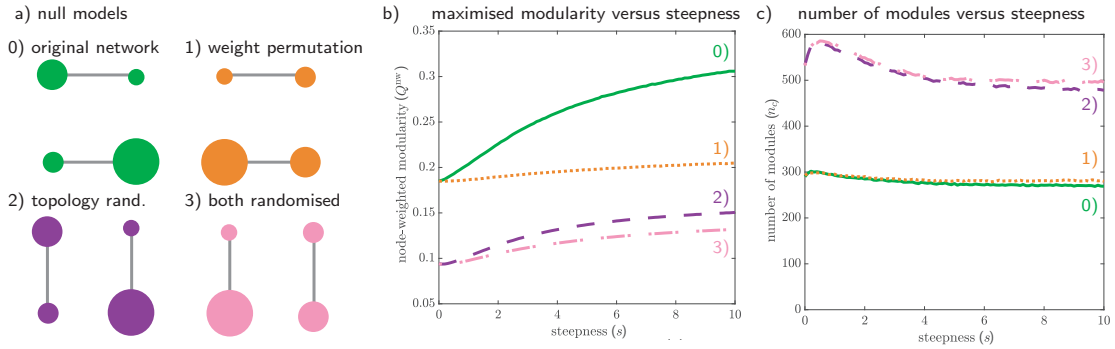


Figure 5.5: We compare the detected community structure in the PIN with three null models. (a) The null models are randomised versions of the original network (1; in green). They are node-weighted networks with either permuted weights (2; in orange), randomised network topology (3; in purple), or both (4; in pink). b) The node-weighted modularity Q^{nw} increases with the steepness s. c) The number n_c of detected modules is fairly constant for (1) and (2). It first increases and then decreases for (3) and (4). For steepness $s \gtrapprox 4$ it is also fairly constant for (3) and (4). We calculate the mean of all curves over 20 realisations of null-model creation and GenLouvain modularity optimisation.

and many more [150]).

In Fig. 5.5a, we show a schematic network, consisting of four nodes, and the three null models, which we construct from the original network. In green (see subpanel 1), we show a scheme representing the original data, which is the node-weighted network $G = (V, W, E)$. We construct the null models as follows.

Null Model 1: Weight Permutation

We obtain a *permuted-weights* null model $G = (V, W', E)$ by a random permutation W' of node weights.[6] We generate such a permutation uniformly at random; that is, each of the $n!$ permutations W' is equally likely to appear.

[6] A permutation is a rearrangement of an ordered set S into an order S'.

Null Model 2: Topology Randomisation

We construct a *randomised-network* null model $G = (V, W, E')$ by randomising the edges, while preserving the degree k_i for each node i.

We obtain a new edge set E' by using an iterative rewiring procedure on the original edge set E [293]. This procedure rewires one edge at each step. Uniformly at random, we choose one edge $(i, j) \in E_l$, remove it, and then assign it new starting and terminal nodes (i', j'). We choose the new edge (i', j') uniformly at random from all non-edges $(i', j') \notin E_l$. We repeat this procedure $m = |E| = |E'|$ times. This procedure preserves the degree k_i for each node i.

Null Model 3: Weight Permutation and Topology Randomisation

We obtain a *both-randomised* null model $G = (V, W', E')$ by combining null models (1) and (2); in other words, the new weights W' are a random permutation of W and the new edge set E' is a rewired version of E and the degree k_i of each node i is preserved.

We now examine the community structures that we detect in a node-weighted network in comparison with the ones we obtain from the three null models. For the modularity definition (5.1), we still use the Newman–Girvan null model.

All calculations concern the tissue-specific PIN of adipose tissue with $n = 13\,394$ proteins. In Fig. 5.5b, we show the node-weighted modularity $Q^{\mathrm{nw}}(s)$ as a function of the steepness parameter $s \in [0, 10]$. For minimum steepness $s = 0$, the original network and

the permuted-weights null model have the same node-weighted modularity. This occurs because, for $s = 0$, the node-weighted modularity reduces to the ordinary modularity and neglects node weights. Therefore, a permutation of edge weights has no influence on the node-weighted modularity function. Similarly, the *randomised-network* null model and the *both-randomised* null model have the same node-weighted modularity for $s = 0$.

For steepness $s > 0$, the node-weighted modularity is largest for the original data, second-largest for the *permuted-weights* null model, third-largest for the *randomised-network* null model, and smallest for the *both-randomised* null model. That the permutation of node weights decreases the modularity indicates a correlation between network topology and node weights (i.e., RNA abundance). Such a correlation has been identified directly in *Saccharomyces cerevisiae* [167], *Arabidopsis thaliana* [482], and *Homo sapiens* [173]. We observe that randomising both — network topology and node weights — results in the smallest modularity. This indicates that information from both — PIN structure and RNA abundance — contributes relevant information to the detection of community structure.

In Fig. 5.5c, we show the number n_c of modules, which we detect by maximising the node-weighted modularity $Q^{\mathrm{nw}}(s)$. As the the modularity, for $s = 0$, the number n_c of modules does not change if we permute the weights. Therefore, the original network and the permuted-weights null model have the same number n_c of modules. The same is true for the randomised-network null model and the both-randomised null model. Both null models with randomised topology (see purple and pink curves),

have almost twice the number of modules than the networks with the original topology. Randomising the node weights w_i results in a small change of the number n_c of modules. For all steepness values s, the original network has the smallest number n_c of modules. We find that varying the steepness has only a minor influence on n_c.

5.4.3 Comparison across Tissues

Thus far, we focussed on the detection of community structure in the node-weighted PINs of a single tissue. We now examine the partitions that we detect in two tissue-specific PINs. To compare the partitions, we compute normalised mutual information (NMI) between two partitions S_1 and S_2:

$$\mathrm{NMI}(S_1, S_2) = \frac{H(S_1) - H(S_1|T_1)}{S_1}\,, \tag{5.13}$$

where $H(x)$ is the Shannon entropy as discussed in 3.1. The $\mathrm{NMI}(S_1, S_2) = 1$ if the two partitions are identical and $\mathrm{NMI}(S_1, S_2) = 0$ if they provide no information about each other [267].

The detection of community structure with the GenLouvain algorithm is stochastic, — different realisations can return different partitions. We perform 20 realisations of each computation. This allows us to compare the NMI between different realisations in the same tissue with the NMI of partitions from different tissues.

In Fig. 5.6, we show the NMI of the partitions of two tissues and $r = 20$ realisations versus the steepness parameter $s \in [0, 10]$. There are three curves: the NMI of partitions of adopise tissue (red curve), the NMI of partitions of suprarenal glands (blue curve), and

the NMI of partitions across tissues (purple curve). At a minimum steepness $s = 0$, all curves coincide, because the node weights are neglected in the node-weighted modularity function (5.2). At this value, the NMI is approximately 0.78. It is smaller than 1 because repeating the community-detection algorithm does not yield the same partition.

For steepness $s > 0$, we find that the NMI is larger inside each tissue than across tissues. Thus, the partitions that we detect by maximising node-weighted modularity (5.2) vary more across tissues than within the same tissue. With increasing steepness, we observe an increase of the NMI within each tissue but a decrease across tissues. For both tissues, the internal NMI first increases with s, reaches a maximum at $s \approx 4$ and then plateaus at NMI ≈ 0.81. Therefore, at intermediate steepness $s \approx 4$, repeated modularity maximisations return more similar partitions, in comparison with the ordinary modularity.

The decrease of cross-tissue NMI indicates that the detected community structures become increasingly different in the two tissues as the node weights are considered more strongly in node-weighted modularity (5.2). Overall, this trend — the decreasing NMI across tissues and an initial increase of NMI with steepness inside each tissue — indicates that we detect partitions that are different in the two tissues. These partitions might consist of structural modules that are tissue-specific.

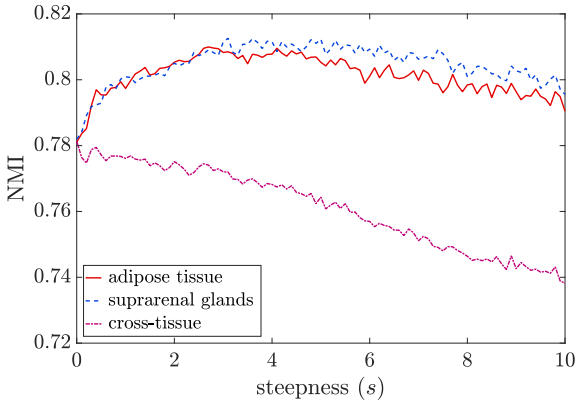


Figure 5.6: The NMI of community structure in two tissues and across tissues. The mean NMI, averaged over twenty realisations, for two tissues (red and blue curves) first increases with the steepness parameter s, reaches a maximum at $s \approx 4$ and then plateaus. Comparing partitions across tissues (purple curve) gives lower NMI and decreases with increasing steepness s.

5.4.4 Comparison of Detected Partitions with Gene Ontology Annotations

Thus far, we have looked at community structure in the node-weighted PINs of two tissues. Comparing the detected partitions indicated that different tissues have similar but not identical community structures. Now we examine whether modules in these tissue-specific PINs also have tissue-specific biological functions by comparing them with GO annotations.

We use the hypergeometric test (see Section 2.9), to compare the detected partitions with GO annotations. We use the Benjamini–Hochberg procedure to control the false-discovery rate at level $\sigma = 0.05$.

We find that approximately 62 % of all detected modules, across the two tissues and different values for the steepness s, are enriched for at least one annotation. This is a similar value to that reported by other authors [262, 365]. Lewis et al., for example,

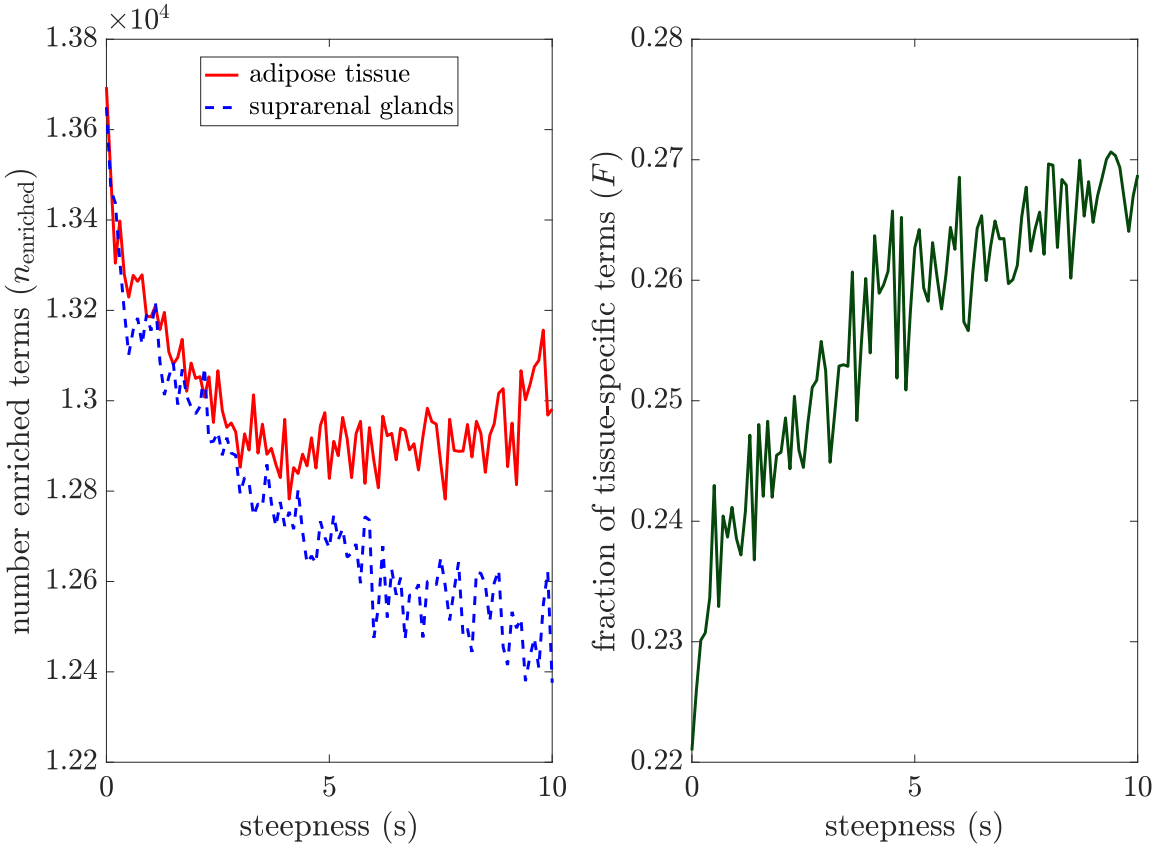


Figure 5.7: The number n_{enriched} of enriched GO terms as a function of steepness s. (a) The number n_{enriched} of enriched GO terms of detected partitions varies with the steepness parameter s. We show results for two tissues (red and blue line). Overall, the number n_{enriched} of enriched terms decreases with s. (b) The fraction F of tissue-specific terms, however, increases with steepness s.

report that approximately 300 out of approximately 600 modules are enriched (see Fig. 2 in [262]). The precise value, however, depends on the number of communities and tends to be higher for a smaller number of modules. We do not investigate this here as we fixed the resolution parameter γ.

Of particular interest is whether GO enrichment of the detected partitions changes with the steepness s. The set $\{t_1\}$ of GO terms that are enriched for a given partition is the union of enriched terms of all modules in this partition (i.e., $\{t_1\} = \bigcup_i^{n_c} \{t_l\}$,

Rank	Adipose Tissue	Adrenal Gland
1	negative regulation of lipid storage	m-xylene catabolic process
2	regulation of sequestering of triglyceride	(1->6)-beta-D-glucan metabolic process
3	methylammonium transport	regulation of single stranded viral RNA replication
4	positive regulation of sequestering of triglyceride	corticosteroid hormone secretion
5	regulation of spermatid nuclear differentiation	detection of electrical stimulus
6	complement component C3a binding'	o-xylene catabolic process

Table 5.1: Top enriched GO terms for adipose tissue and adrenal gland. We show the six most enriched GO terms for both tissues for partitions detected through node-weighted modularity optimisation with a steepness of $s = 9$.

where $\{t_l\}$ are all enriched terms of modules l). In Fig. 5.7a, we examine the number n_{enriched} of enriched terms for each partition as a function of the steepness s. The curves are the average values over 20 realizations. For $s = 0$, we find $n_{\text{enriched}} \approx 13\,700$ enriched terms. As the steepness $s = 0$ ignores node weights and both networks have the same edge set, it is (approximately) the same for both tissues. For both tissues, we observe that the number n_{enriched} of enriched terms decreases with the steepness s. When we consider the RNA abundance data in the form of node-weights w_i the number of biological enriched functions in the PIN decreases.

To investigate GO enrichment further, we examine whether the enriched GO terms are *tissue-specific*. In this context, we define tissue-specific as a GO term that is enriched in one tissue both not the other. We now investigate what percentage of the enriched terms are tissue-specific. Specifically, let $\{t_1\}$ and $\{t_2\}$ be the sets of enriched terms of adipose tissue and suprarenal glands, respectively. We then compute the fraction of tissue-specific GO terms:

$$F = 1 - \frac{\{t_1\} \cap \{t_2\}}{\{t_1\} \cup \{t_2\}}\,, \tag{5.14}$$

where $\cap$ and $\cup$ are the intersection and union, respectively. In Fig. 5.7b, we show $F(s)$ over 20 realisations. Overall, the fraction F of tissue-specific terms is approximately 0.2. Note that even for $s = 0$ we identify approximately 0.2 of all enriched terms as 'tissue-specific', despite ignoring the node weights. This occurs because different runs of the GenLouvain algorithm with the same modularity function can return different partitions, as discussed in Subsection 5.4.3. We observe an increasing $F(s)$, which indicates that more of the enriched terms are tissue-specific for larger steepness s.

In Table 5.1, we give the top enriched GO terms for both tissues. A manual inspection of the enriched GO terms indicates that some of them indeed seem tissue-specific. For the adipose tissue, for example, we find 'negative regulation of lipid storage' and 'regulation of sequestering of triglyceride' enriched. Triglycerides are the main components of human adipose tissue [247] and for the adrenal gland we find 'corticosteroid hormone secretion' enriched. These hormones are a class of steroid hormones that are products of adrenal cells [88]. We also find enriched terms that do not seem appropriate for these tissues, including 'detection of electrical stimulus involved in sensory perception'. It is possible that we erroneously identified modules with such biological functions. Potentially, some of the detected modules are enriched for a biological function even if this function is not relevant for the tissue.

To complement these manual inspections with a quantitative analysis of the enriched GO terms we investigate their hierarchy level. As discussed in Section 2.9, the GO terms are organised in a hierarchical way in three directed acyclic graphs [20]. On

top of these hierarchies are the terms 'biological function', 'cellular component' and 'molecular function'. Intuitively, the GO terms farther away from these top terms give rise to more specific annotations for proteins, whereas the higher-level terms indicate more general annotations. We compute for each term t the hierarchy level h_t as the length of the shortest path to the top term in its hierarchy. We obtain for each term an integer hierarchy level $h_t \in [0, 9]$ and smaller h_t indicate more general GO terms. We now examine whether the mean hierarchy

$$\langle h \rangle = \frac{\sum_{t=1}^{n_{\text{enriched}}} h_t}{n_{\text{enriched}}} \tag{5.15}$$

of enriched terms changes with the steepness s. In Fig. 5.8, we show the curves $\langle h \rangle(s)$ as a mean over twenty realisations. The solid red curve indicates the adipose tissue and the dashed blue curve indicates the suprarenal gland. For both tissues, we observe that the mean hierarchy $\langle h \rangle$ increases when we consider the node weights. In absolute values, however, the change is less than 1 %. The increase in mean hierarchy $\langle h \rangle$ could hint at the detection of modules with more specific terms when we consider the node weights.

We conclude that the number of enriched terms decreases when considering RNA abundance data as node weights in PINs. However, incorporation of RNA abundance data increases the detection of modules with terms that are enriched in one tissue but not the other. Such terms are candidates of tissue-specific functions. When analysing the hierarchy level of these terms we find that they tend to have a larger hierarchy level and thus are more specific. This trend, however, is small and not conclusive.

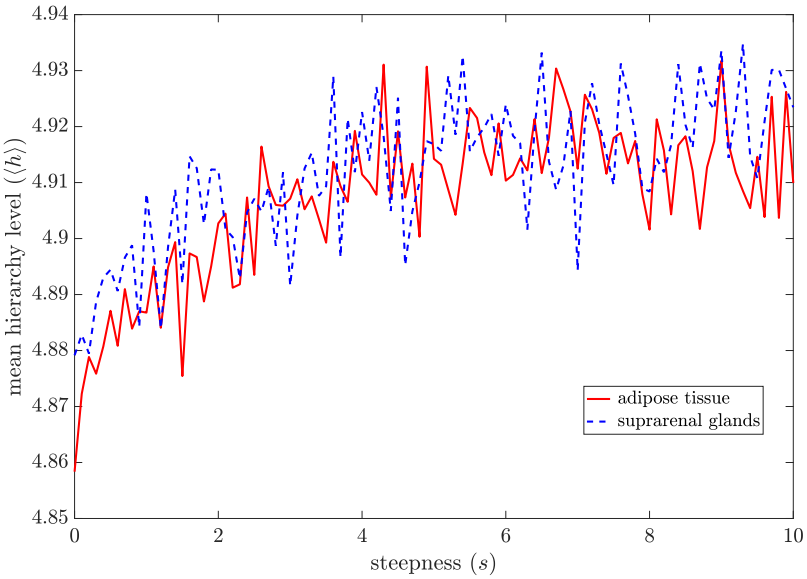


Figure 5.8: The mean hierarchy level $\langle h \rangle$ of all terms that are enriched for two tissue-specific PINs as a function of the steepness s. We show adipose tissue as solid red curve and suprarenal glands as dashed blue curve. For both tissues, the mean hierarchy level increases slightly with steepness s. This indicates that the enriched terms might become more 'specific' when we consider RNA-abundance as node weights.

5.5 Conclusions

In this chapter, we introduced a node-weighted variant Q^{nw} of the modularity function Q for the detection of community structure in PINs. This approach allows us to examine protein interaction networks together with protein abundance levels. It is relevant to develop such a methodology for data integration, as proteins that interact with each other are not always present at the same time in a given cell or tissue. Therefore, node-weighted networks are one way to construct tissue-specific PINs.

For a synthetic network with a planted overlapping community structure, we showed that the detected community structure changes considerably when node weights are taken into account. Specifically, nodes that structurally belong to multiple modules

join the module with large weights. In a PIN, these nodes might represent proteins that are involved in multiple biological functions and potentially change context-specifically.

We then examined biological data in the form of a PIN and RNA-abundance information for two different tissues. We compared maximising the node-weighted modularity Q^{nw} in such a tissue-specific network with maximising the node-weighted modularity Q^{nw} in three null models. We found that a permutation of network structure, as well as node weights, decreases the node-weighted modularity Q^{nw}. The number of detected modules stays fairly constant when changing the influence of the node weights. We then compared detected partitions in two tissue-specific PINs. For each tissue separately, the NMI over multiple realisation increases, if we consider RNA-abundance. By contrast, the NMI decreases across tissues. This indicates that we detect community structures that are more tissue-specific when we consider the RNA-abundance as node weights.

To complement these findings, we compare the detected modules with GO annotations. While the number of enriched biological functions decreases when considering RNA-abundance, the fraction of tissue-specific terms increases. This indicates that we indeed find modules that are structurally and functionally tissue-specific.

While these are promising results, we should also be aware that this is not a conclusive validation of our approach. It would be fruitful, for example, to test our approach of integrating protein interaction networks with mRNA abundance by testing its performance to predict communities or pathways involved in diseases. If successful,

the identified proteins could be perturbed with gene-knock-out studies to test whether they do indeed cause certain malfunctions in an organism.

Future Work

Our general definition of node-weighted modularity (5.2) does not depend on the specific weighting function (5.3) that we choose. Different functional forms likely change detected community structures. Potentially, different weighting functions are appropriate for different applications of node-weighted networks.

We focussed on the investigation of node-weighted PINs for only two tissues (adipose tissue and suprarenal glands). A further, large-scale investigation may give insights into community structure of PINs across tissues. This would potentially reveal structural modules that are present in certain types of tissues, e.g., muscle, but not in others.

One can extend the notion of a node-weighted modularity function to the modularity function for multilayer networks (see Section 2.8). This would allow the investigation of tissue-specific temporal PINs.

In our investigation, we identified 'reasonable' tissue-specific GO terms manually. Ideally, we would do a statistical analysis to determine whether the terms we detect are indeed tissue-specific, for example, as a ROC curve. This is not possible because there is no large-scale information whether a GO term is specific to a given tissue. Potentially, an automated literature search could help to create such data [70]. Alternatively, one can use mRNA-abundance measurements to obtain tissue-specific GO terms [503]. One has to be careful with such an investigation because we use the same data for

the construction of the tissue-specific PIN.

6
Conclusions

Examining protein–protein interactions is important for improving understanding of the biological processes in cells and the human body. One can use protein interaction networks (PINs) to represent interactions between proteins. Analysing PINs with the tools of network science, such as community detection or centrality measures, can give insights into the biological organisation in cells.

PINs, in the form of networks, are mathematical models of the complex processes in cells. The advantage of a network representation is that it allows the application of a wide range of analytical and computational tools. We must be aware, however, that the underlying biological processes are more complex than any representation as a network. Mathematical structures other than networks can incorporate more information about the biology and may therefore give novel insights into PINs. In this thesis, we investigated some generalised network structures and evaluated the extent to

which they are suitable for the investigation of PINs. We focused foremost on multilayer networks (MLNs), which are one way to encode protein interaction in multiple tissues, across time, or from different experiments in one mathematical structure. We also investigated node-weighted networks as another way to construct tissue-specific PINs.

In Chapter 3, we introduced *promiscuity* as a measure of the variability of node importance across layers. We investigated real-world networks from a wide variety of applications and showed that they can have very different promiscuity distributions. For some synthetic network models, we were able to derive exact or approximate analytic expressions. By examining PINs from multiple tissues, we found that a small number of transcription factors can be used as biomarkers. We compare proteins' promiscuities with gene-ontology information and found that some biological functions tend to be associated with large promiscuity values.

In Chapter 4, we demonstrated that MLNs can be used to represent temporal PINs. As a case study, we investigated a temporal PIN during the yeast cell cycle with an eigenvector-based centrality. We found that essential proteins tend to have larger time-averaged centralities and larger changes in centrality than non-essential proteins. We then examined proteins with the largest centralities or the largest changes in centrality. We found that the former have crucial roles in cell biology (e.g., actin as structural component) and the latter are important in certain phases of the cell cycle.

Proteins that can interact with each other are not necessarily present in the same cell at the same point in time. In Chapter 5, we investigated node-weighted

networks to integrate protein-interaction information with measurements of protein abundance. We focused on the detection of modular structures in such node-weighted networks and introduced a generalisation of the modularity quality function that takes account node weights. In a synthetic network, we showed that node weights can have significant impact on detected modular structure. Using gene-expression information of different tissue types, we showed that detected modular structure varies across tissue types. Analysing GO enrichment in the modular structure suggests that we detected modules with tissue-specific functions.

The work in this thesis illustrates promising approaches for the examination of protein-interaction data and gene-expression data in an integrated manner. One important caveat is that, for such integrated data analysis, we needed to weigh the extent to which we incorporate the different data sets. In the study of node-weighted PINs, for example, we had to choose a steepness parameter s as an indication of the influence of RNA-abundance data on the community detection algorithm. Extreme cases of such free parameters may indicate that it is permissible to neglect one of the data sets. For the minimal steepness $s = 0$, for example, we recover modularity maximisation in an unweighted network. It is typically unclear how to choose such parameters. For this application, we choose such parameters by comparing our algorithmically-obtained partitions with GO data. The development of systematic methods to infer the values of such parameters is an important research direction (e.g., [329, 334, 352]). One option is to use cross-validation approaches, in which one separates a network into a

training and a test set. Such approaches have been successful for maximum-likelihood estimations of stochastic block models (SBMs; e.g., [100, 462]). The comparison of node-weighted modularity with a node-weighted SBM, which we present in Appendix D, may provide a starting point for a cross-validation approach.

We examined tissue-specific and temporal PINs as MLNs. There are different biological processes involved with protein regulation. One may model such processes as MLNs. Most notably, MLNs can perhaps be useful for capturing alternative splicing [61, 228, 295] or post-translational modifications [48]. Such processes enable cells to produce multiple proteins from one gene. In such an MLN representation, one layer can encode gene-expression data, and a second layer can encode protein-abundance data. Interlayer edges indicate which proteins are a product of which gene.

Other mathematical extensions of graphs may help for exploring the complex functions of proteins. Proteins interact with each other not only in pairs but also in higher orders, e.g., in triples or more. Multi-protein complexes can consist of more than two proteins. It is thus compelling to examine pairwise interactions differently than multi-protein complexes. To represent this information in a PIN, one can use hypergraphs or simplices [37, 287, 344]. The development of tools to analyse such mathematical structures is an active field of research [402, 438, 439], and these tools might provide novel insights (e.g., whether multi-protein complexes tend to be more central than pairwise interactions in a PIN).

Overall, the study of protein interactions as generalised networks enables one to

investigate temporal change and tissue-specificity in biological processes. As such, MLNs and node-weighted networks are promising tools for an analysis that integrates protein interaction data with other biological data.

A

The GenLouvain Community Detection Algorithm

A.1 The Original Louvain Algorithm

The *Louvain Algorithm* (LA) is a method to detect communities in networks by modularity optimisation [45]. As introduced in Chapter 2, the modularity function Q is defined as

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(g_i, g_j) \,. \tag{A.1}$$

The community detection problem is then the uncovering of a group assignment g_i for each node i that maximises this modularity function. The LA is a heuristic method for this optimisation problem. It consists of two stages, 'community reassignments' and 'coarse graining', which are repeated iteratively.

The 'community reassignment' stage starts by assigning a different community to each node. Then we consider for each node i the possibility to join the community

of any of its neighbours $j \in \mathcal{N}(i)$. Node i is placed in the module that maximises the gain in modularity by the reassignment, but only if it improves modularity. It is efficient to not compute the whole modularity with eqn. (A.1) but rather its change. If a node i is moved from a community consisting only of itself into a community C, the modularity change ΔQ is given by

$$\Delta Q = \left(\frac{\sum_{\text{in}} + k_{i,\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}} + k_{i,\text{in}}}{2m} \right)^2 \right) - \left(\frac{\sum_{\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right) , \qquad \text{(A.2)}$$

where $\sum_{\text{in}}$ is the sum of the weights of the links inside community C, $\sum_{\text{tot}}$ is the sum of the weights of the links incident to nodes in community C, k_i is the sum of the weights of the links incident to node i, $k_{i,\text{in}}$ is the sum of the weights of the links from i to nodes in community C and m is the sum of the weights of all the links in the network. If a node i moves from module C_1 to module C_2 we can compute the modularity change ΔQ by first moving the nodes from module C_1 into its own module and then join it with module C_2, thus computing eqn. (A.2) twice. The process of moving a random node into one of its neighbours communities is repeated until none of these moves can improve modularity, such that a local maximum of modularity is reached.

In the 'coarse graining' stage a new network is created. Its nodes are the communities found during the first phase. Weights between these are calculated as the sum of the weights of the links between the two modules. This new network is returned to the first stage for a new 'community reassignment.'

The iterative application of these two stages, called a 'pass', become faster because the number of nodes decreases with each coarse graining step. Passes are repeated

until there are no more changes, ideally reflecting a global maximum of the modularity Q. Usually only the final partition is returned, but together with earlier steps it gives a hierarchical modular structure.

LA allows a fast detection of communities in large networks and is widely used but comes with some restrictions. As with most community detection methods, it often fails to detect very small communities in large networks, something which is known as 'resolution limit' [149].

A.2 The GenLouvain Algorithm

GenLouvain is a 'generalized Louvain method for community detection' which is implemented in MATLAB [217]. We used the most recent 2.1 version. The GenLouvain algorithm uses different approaches for modularity optimisation, one of them being the original Louvain algorithm as outlined above. The user can define a modularity matrix $\mathbf{Q}$. This allows the incorporation of resolutions parameter γ, construction of modularity functions for MLN, and node-weighted modularity, as used in this thesis.

B

Community Structure in a Temporal Protein Interaction Network

Contents

B.1 Introduction 187
B.2 Data and Preprocessing 188
B.2.1 Multilayer Network Construction 189
B.3 Community Detection 190
B.3.1 Parameter Choice 192
B.4 Conclusions 196

B.1 Introduction

PINs allow the representation and analysis of biological processes in cells. As cells are dynamic and adaptive, these processes change over time. One form of adaptive regula-

This appendix is joint work with A. Roxana Pamfil and my supervisors Jonny Wray, Charlotte M. Deane, and Mason A. Porter.

tion is the change of gene expression, which may occur at very different time scales [280]: responses to environmental signals take minutes [498], the cell cycle length in yeast is 90 minutes [90], and developmental changes take days in *C. elegans* [446] and years in humans [63]. Despite their many differences, the presence of proteins changes during all these processes [466]. It is likely that protein abundance influences PPIs [182, 212].

There is evidence that biological function in cells is modular [39] and that these modules change over time [174]. In this chapter, we detect modular structure in a temporal PIN. This enables us to identify functional modules and also observe their temporal change.

As a case study we investigate a PIN of leukocytes (white blood cells) in human subjects receiving an inflammatory stimulus [66]. Inflammation is an attempt of the human immune system to protect a host from infections or other harmful stimuli. As such, inflammation plays a crucial role in many diseases, e.g., metabolic disorders [199], heart disease [175], cancer progression [95], and liver failure [242].

B.2 Data and Preprocessing

In this chapter, we examine a temporal network that is constructed from temporal gene-expression data and a monolayer PIN. The temporal network was provided in a snapshot representation (.i.e, one graph for each time point) by *e-Therapeutics*[1]. The temporal network was constructed from gene-expression data in human leukocytes

[1]My industrial supervisor Jonny Wray is an employee of the pharmaceutical company *e-Therapeutics*, who sponsored my DPhil project.

under two experimental conditions and at four time points [66]. Gene expression in whole blood leukocytes was determined with oligonucleotide arrays at 2, 4, 6, and 9 hours for eight human volunteers. Subjects were intravenously administered with either 'NIH Clinical Center Reference Endotoxin' (CC-RE-Lot 2) at a dose of 2 ng/kg body weight (treatment group; $n = 4$, one female and three males) or 0.9 % sodium chloride (placebo group; $n = 4$, one female and three males) over a 5-minute period. For details about the method of endotoxin administration see [146].

Researchers at *e-Therapeutics* used Robust Multi-array Average [211] for cross chip normalization and Limma [420] for differential expression calculations. These values are then used as node weights in a network extraction approach [121]. Specifically, *e-Therapeutics* used the Heinz implementation with the CPLEX solver [130]. The underlying PIN used in the network reconstruction is an in-house network from an integration of multiple sources, among them iRefindex [377], HIPPIE [399], TRANSFAC [484], and Pazar [368]. This yields a temporal PIN with $T = 4$ snapshot graphs with $N = 7928$ nodes each. Each network indicates the active parts of the underlying PIN in the treatment group in comparison with the placebo group at one time point.

B.2.1 Multilayer Network Construction

We construct a temporal MLN from a temporal network in a snapshot representation, as provided by *e-Therapeutics*. To do this, we construct interlayer edges between nodes representing the same protein in time-adjacent layers, i.e., diagonal and ordinal

coupling (for details see Subsection 2.7.4). This yields a MLN with $3 \times 7928 = 23784$ interlayer edges, each of strength ω.

B.3 Community Detection

We use the GenLouvain algorithm to maximise the multilayer modularity function (Eqn. 2.22) with homogeneous interlayer coupling of temporally adjacent layers, i.e., $\omega_{ist} = \omega$ for all $|s - t| = 1$. When applying the modularity-maximisation algorithm, we have to choose two parameters, resolution parameter γ and interlayer coupling ω. We obtain community structure across layers, i.e., each node-layer pair belongs to exactly one community.

In Fig. B.1, we show detected modular structures for four parameter choices as *alluvial diagrams* [297, 392]. In these alluvial diagrams, each block represents a community at one time point. The height of each block gives the size of the community, i.e., the number of nodes in this community for this point in time. The 'stream fields' between the blocks represent nodes that change their community from one time point to the next.

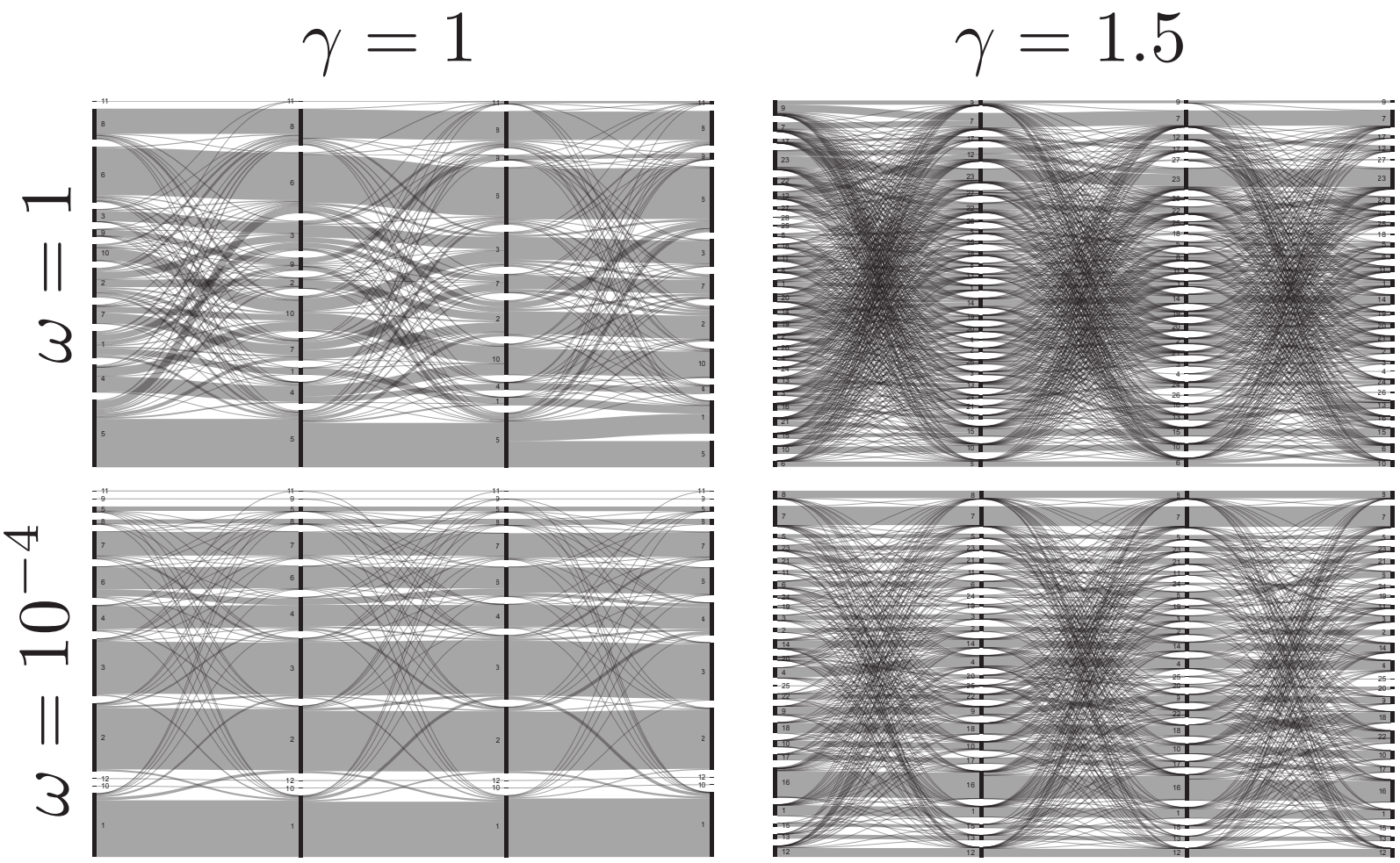


Figure B.1: Alluvial diagrams of modular structure in a temporal PIN. We show detected modular structures for four choices of the parameters γ and ω. We created the visualisations with RAWGraphs [297].

As the temporal network consists of four time layers, the alluvial diagrams have four time points. All four partitions change over time but we observe some differences between the partitions. For parameters ($\gamma = 1, \omega = 10^{-4}$), there are $n_{\text{com}} = 12$ communities and only a small number of nodes change their community assignment over time. For the same resolution factor $\gamma = 1$ and interlayer coupling $\omega = 1$, we detect $n_{\text{com}} = 11$ communities. The increased interlayer coupling results in an increased change of community assignment over time. For both choices of interlayer coupling ($\omega = 10^{-4}$ and $\omega = 1$), a resolution parameter $\gamma = 1.5$ yields partitions with a larger number of communities ($n_{\text{com}} = 28$ and $n_{\text{com}} = 25$, respectively) than a resolution parameter of $\gamma = 1.0$.

These examples of detected modular structure indicate that modularity-maximisation for MLN may yield partitions with different numbers of communities and different extents of temporal change. For our investigation of temporal PINs, it is not *a priori* known how to choose these parameters.

B.3.1 Parameter Choice

As our choice of (γ, ω) parameters influences the modular structures we obtain, we would like to choose them in a statistically-grounded way. Pamfil et al. (2018) introduced an iterative procedure to choose (γ, ω) parameters for community-detection in MLNs [352]. They discuss this procedure for temporal and multiplex MLNs. We use the variant for temporal MLNs (see Appendix C). This procedure uses a relation between multilayer modularity and a maximum-likelihood estimation of a temporal stochastic block model.

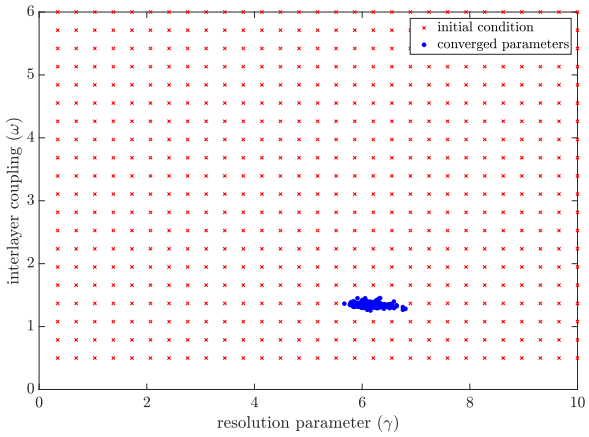


Figure B.2: Convergence of community-detection parameters γ and ω. We use an iterative algorithm for a selection of parameters for the community detection [352] (see Appendix C). We show 560 initial conditions (γ_0, ω_0) as red crosses and the final, converged parameters (γ, ω) as blue disks. All of these realisations converge to values close to $(\gamma \approx 6.15, \omega \approx 1.34)$.

Aim of this procedure is to select parameters that best explain the observed network (in a maximum-likelihood sense). The iterative algorithm works well for several multilayer benchmark networks. There exists, however, no guarantee that this procedure identifies parameters that are associated with a maximum likelihood. For this algorithm, one has to choose initial values of γ and ω.

We ran 560 trials of the iterative algorithm using initial values of γ and ω that were distributed on a grid in the intervals $[0.5, 10]$ and $[0.5, 6]$, respectively (see Fig. B.2). For all these starting conditions the procedure converges to parameters $(\gamma \approx 6.15, \omega \approx 1.34)$ with standard deviation of approximately $(0.18, 0.02)$. The mean normalised mutual information between all partitions that we obtain is 0.64. As the NMI is less than 1, there is some variability in the partitions we obtain.

As an example, we investigated one of the obtained partitions. The detected

community structure has 173 modules. Each of them spans all four time layers. Overall, there is little temporal change in their modular structure (91 % of nodes do not change their module assignment over time). Due to the large number of modules, a visualisation of the temporal modular structure as alluvial diagram is not fruitful and we omit it.

We calculate GO-term enrichment as discussed in Section 2.9: We employ hypergeometric tests with a significance level of $\sigma = 0.05$. We use the Benjamini–Hochberg procedure to correct for multiple testing as we test 8156 GO terms. Out of the 173 communities, 111 have at least one enriched biological function. Many of the enriched functions are relevant for the cellular response to inflammation, e.g., 'Wnt signaling pathway' [157], 'transmembrane receptor protein tyrosine kinase activity' [349], and 'SUMO ligase activity' [273]. This suggests that some of the detected communities in the temporal PIN are functional modules of inflammation.

We now examine the change of community structure and biological function. For this, we calculate the size S_1 of each community in the first layer and the size S_4 in the last layer. In Fig. B.3, we show the size S_1 in the first layer and the change $S_4 - S_1$ in community size from the first to the fourth layer. Overall, most of the 173 communities stay similar in size but the size of some communities increases whereas that of others decreases.[2] One community increases in size from $S_1 = 16$ nodes to $S_4 = 51$ nodes. For the first two layers, the community has no enriched biological function. In the last two layers, the community has the enriched function 'apoptotic

[2]The one implies the other, as each protein belongs to exactly one community.

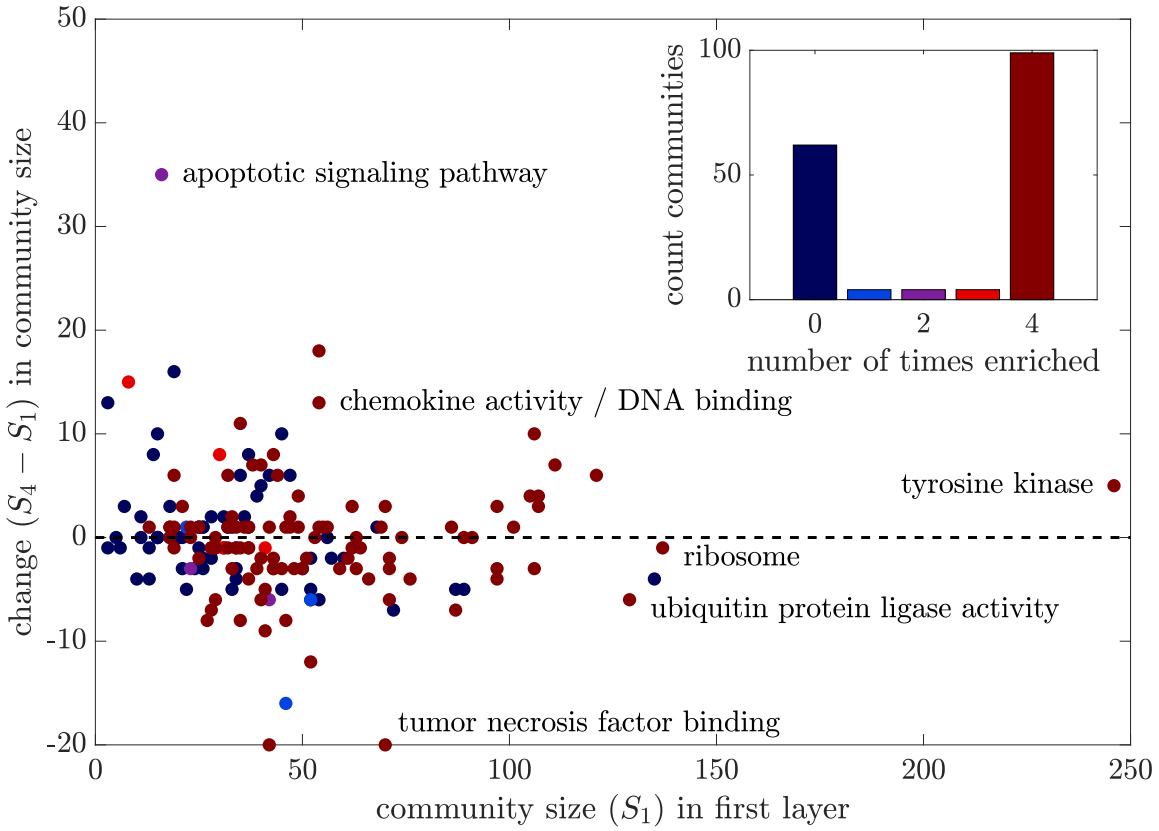


Figure B.3: Change in size of communities. We show the size S_1 in the first layer and the change in size $S_4 - S_1$ for the detected communities and colour-code the number of times each community is enriched (see inlay). Most communities are enriched at all four times (dark red) but others are never enriched (dark blue). We label selected communities with their most-enriched biological function, e.g., 'apoptotic signaling pathway'.

signaling pathway'. *Apoptosis* is a form of cell death responsible for the deletion of damaged cells [49]. Apoptosis of pro-inflammatory cells is a mechanism which terminates inflammatory response [166, 492]. Another community changes its enriched biological function from 'chemokine activity' to 'DNA binding'. Chemokines are crucial for the inflamamtory response in multicellular organisms and recruit immune-response cells to the site of infection [24, 370]. The community with the function 'tumor necrosis factor receptor binding' shrinks in size but the function is enriched at all four time

points. Tumor necrosis factors are crucial proteins for the organisation of inflammatory response [278] and can trigger inflammation [56, 208].

Overall, only a small number of communities change their function over time (see inlay in Fig. B.3). This is in accordance with the fact that most of them also do not change structurally, as less than 10 % of nodes change their module assignment.

We have to be aware that we investigated only one modular structure. As we discussed, however, the iterative algorithm may yield different community structures for each realisation. We therefore calculated GO-enrichment for all 580 community structures. We find 1321 GO-terms that are enriched in all partitions for at least one community. This indicates that we indeed obtain community structures that represent some modular functional organisation.

B.4 Conclusions

In this section, we examined modular structure in a human temporal PIN during the response to an inflammatory stimulus. We constructed a temporal PIN as a MLN.

We illustrated that the choice of resolution parameter γ and interlayer coupling ω can strongly influence the modular structure that we obtain from a modularity-maximisation algorithm.

We then employed an iterative algorithm to choose the parameters γ and ω in a statistically-grounded way [352]. Applying these parameters yields modular structures that change only little over time. Examining biological functions in detected modules

with GO enrichment indicated that most of them have biological functions. As the structural change was small the functional adaption was also small. We do, however, find some biological functions that become enriched exclusively at a subset of time stages.

Future Directions

It could be beneficial to investigate whether other temporal PINs have a more prominent change in community structure over time. We examined modular structure in two other temporal PINs: a developmental PIN in the roundworm *C. elegans* and a PIN during the yeast cell cycle (as described in Chapter 4). For both networks, the algorithm converged to parameter values that yield modular structure with almost no temporal change (results not shown).

The specification of the parameters γ and ω in a statistically-grounded way was fruitful and we obtained many communities that are functionally enriched. Results from static PINs indicate, however, that there is not a single scale of community structure in PINs [262]. By comparing partitions obtained from different parameters with GO-annotation data, we could potentially examine a multiscale temporal organisation of PINs and its change through time.

C

Relating Modularity Maximization and Stochastic Block Models in Multilayer Networks

In this appendix, we relate modularity maximization and stochastic block models (SBM) in multilayer networks (MLNs). All results are not original work but taken from [352] and [329]. While this method is applicable to temporal and multiplex MLNs, we exclusively discuss temporal MLNs. Relating modularity maximisation and a maximum-likelihood estimation of a SBM allows the design of a interative algorithm for the selection of the resolution parameter γ and the interlayer coupling ω.

Such an iterative algorithm was first developed for monolayer networks [329] and then adapted for MLNs [352]. Pamfil et al. (2018) present a version in which the resolution parameter γ and the coupling parameter ω are the same for all layers (but generalisations from this are possible). The algorithm works by repeatedly using a

modularity-maximisation algorithm to yield partitions and updating the parameters.

Algorithm 1, is this iterative algorithm, as derived in [352]. To use this algorithm, we have to estimate SBM parameters from partitions that we obtain from maximising a multilayer modularity. Specifically, one can estimate SBM parameters by calculating

$$\theta_{\text{in}} \approx \frac{\sum_{t=1}^{T} 2m_{\text{in}}^{t}}{\sum_{t=1}^{T} \frac{1}{2m_t} \sum_r \left(\kappa_r^t\right)^2}\,, \tag{C.1}$$

and

$$\theta_{\text{out}} \approx \frac{\sum_{t=1}^{T} 2m_{\text{out}}^{t}}{\sum_{t=1}^{T} \left[2m_t - \frac{1}{2m_t} \sum_r \left(\kappa_r^t\right)^2\right]}\,, \tag{C.2}$$

$$p \approx \frac{\frac{\text{Pers}(\mathbf{g})}{N(T-1)} - \frac{1}{K}}{1 - \frac{1}{K}} \tag{C.3}$$

where Pers($\mathbf{g}$) denote the number of instances in which a node belongs to the same community in consecutive layers.

The equations for updating γ and ω are

$$\gamma = \frac{\theta_{\text{in}} - \theta_{\text{out}}}{\log\theta_{\text{in}} - \log\theta_{\text{out}}}\,, \tag{C.4}$$

$$\omega = \frac{1}{\log\theta_{\text{in}} - \log\theta_{\text{out}}} \log\left(1 + \frac{p}{1-p}K\right)\,. \tag{C.5}$$

respectively.

Algorithm 1 Iterative algorithm for performing modularity maximization and estimating resolution and interlayer-coupling parameters in a multilayer network. From [352].

function IterativeModularityMaximization($\mathbf{A}$)
  initialize $\gamma = \gamma^{(0)}$ and $\omega = \omega^{(0)}$
  while not converged **do**
    $\mathbf{g} \leftarrow$ MaximizeModularity($\mathbf{A}, \gamma, \omega$) ▷ Community detection (e.g., using GenLouvain [217])
    $\theta_{\text{in}}, \theta_{\text{out}}, p, K \leftarrow$ EstimateSBMParameters($\mathbf{A}, \mathbf{g}$) ▷ Use detected communities to estimate SBM parameters
    $\gamma \leftarrow$ UpdateGamma($\theta_{\text{in}}, \theta_{\text{out}}$) ▷ Update γ using Eqn. (C.4)
    $\omega \leftarrow$ UpdateOmega($\theta_{\text{in}}, \theta_{\text{out}}, p, K$) ▷ Update ω using Eqn. (C.5)
  return γ, ω ▷ Optimal modularity parameters
  return g ▷ Detected communities using optimal parameters

D

Comparison of Node-weighted Modularity Function with Weighted Likelihood Stochastic Block Model

In Chapter 5, we introduced node-weighted modularity in an *ad hoc* way. We now motivate node-weighted modularity (5.2) from a probabilistic point of view. To do this, we introduce the *weighted likelihood stochastic block model*, an extension of the SBM, and show that there is an equivalence between both approaches.

Weighted Likelihood

Consider some data $x_1, x_2, \ldots, x_{n'}$, where each $x_{i'}$ is independent and identically distributed, from a parametric model $f(x_{i'}|\theta)$, where θ is a set of parameters. The likelihood of the data given a set of parameters θ is the joint probability function

$$L(x_{i'}|\theta) := \prod_{i'=1}^{n'} f(x_{i'}|\theta)\,. \tag{D.1}$$

One then maximises (D.1) under variation of θ to find a *maximum-likelihood estimate* (ML estimate) of θ. Alternatively, one might use Bayesian approaches to find a full posterior distribution of θ.

In some cases, we want to weight the data differently. Given scalar weights $w_{i'}$ for each data point the *weighted likelihood function* [5, 335] is given by

$$\widetilde{L}(x_{i'}|\theta) := \prod_{i'=1}^{n'} f(x_{i'}|\theta)^{w_{i'}} . \tag{D.2}$$

Weighted Likelihood SBM

We now follow the discussion in [329], though we generalise it slightly, to show that maximisation of such a weighted likelihood on an SBM is equivalent to the node-weighted modularity for a planted-partition model for a particular choice for the resolution parameter γ. Note that this does not show a general equivalence of the two models.

As we discussed in Subsection 2.4.3, in an SBM the probability that an edge (i, j) exists depends on the group memberships g_i and g_j of both nodes. Often, one studies this is a Poisson distribution, where $\omega_{s,r}$ is the expected number of edges between groups s and r [329]. The likelihood of the standard SBM is then

$$P(A|\Omega, \mathbf{g}) = \prod_{i=1}^{n} \frac{(\frac{1}{2}\omega_{g_i,g_j})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\omega_{g_i,g_j}/2\right) \prod_{i<j} \frac{\omega_{g_i,g_j}^{A_{ii}}}{(A_{ii})!} \exp\left(-\omega_{g_i,g_j}\right) . \tag{D.3}$$

Each of the multiplicands represents one element (i, j) in the adjacency matrix of the observed network. We now weigh each of these elements in the likelihood function (D.1) by the weight $W(w_i, w_j)$ that is a function of the node weights w_i and w_j. The functional form of the weight function $W(w_i, w_j)$ is not relevant for the following

derivation. In Chapter 5, we used a two-dimensional sigmoidal function (5.3).

Introducing the weighted version in accordance with eqn. (D.2) gives

$$P(A|\Omega,\mathbf{g}) = \prod_{i=1}^{n}\left[\frac{(\frac{1}{2}\omega_{g_i,g_j})^{A_{ii}/2}}{(A_{ii}/2)!}\exp\left(-\omega_{g_i,g_j}/2\right)\right]^{W(w_i,w_i))}\prod_{i<j}^{n}\left[\frac{\omega_{g_i,g_j}^{A_{ij}}}{(A_{ij})!}\exp\left(-\omega_{g_i,g_j}\right)\right]^{W(w_i,w_j))}. \tag{D.4}$$

Taking natural logarithm yields the log-likelihood

$$\log P(A|\Omega,\mathbf{g}) = \sum_{i=1}^{n} W(w_i,w_i))\left[\frac{A_{ii}}{2}\log\frac{1}{2}\omega_{g_i,g_j} - \log\left((A_{ii}/2)!\right) - \omega_{g_i,g_j}/2\right] \tag{D.5}$$

$$+\sum_{i<j}^{n} W(w_i,w_j)\left[A_{ij}\log\omega_{g_i,g_j} - \log\left(A_{ij}\right)! - \omega_{g_i,g_j}\right]. \tag{D.6}$$

Ignoring all constant terms (which we show in blue) simplifies the log-likelihood to

$$\log P(A|\Omega,\mathbf{g}) = \sum_{i,j=1}^{n} W(w_i,w_j)[A_{ij}\log\omega_{g_i,g_j} - \omega_{g_i,g_j}]\,. \tag{D.7}$$

In the degree-corrected version where we replace ω_{g_i,g_j} by $k_ik_j\omega_{g_i,g_j}/(2m)$ to

$$\log P(A|\Omega,\mathbf{g}) = \sum_{i,j=1}^{n} W(w_i,w_j)[A_{ij}\log\omega_{g_i,g_j} - \frac{k_ik_j\omega_{g_i,g_j}}{2m}]\,. \tag{D.8}$$

We now discuss the *planted-partition model*, a special case of the SBM. It consists of modules that are connected with $\omega_{g_i,g_j} = \omega_{\text{in}}$ internally and with ω_{out} across modules. We can then write

$$\begin{aligned}\omega_{g_i,g_j} &\implies (\omega_{\text{in}} - \omega_{\text{out}})\delta_{g_i,g_j} + \omega_{\text{out}}\,,\\ \log\omega_{g_i,g_j} &\implies (\log\omega_{\text{in}} - \log\omega_{\text{out}})\delta_{g_i,g_j} + \log\omega_{\text{out}}\,.\end{aligned} \tag{D.9}$$

Using (D.9) in the degree-corrected SBM likelihood (D.8) and ignoring constants we obtain

$$\log P(A|\Omega,\mathbf{g}) = \sum_{i,j=1}^{n} W(w_i,w_j)\left[A_{ij} - \frac{k_ik_j}{2m}\frac{\omega_{\text{in}}-\omega_{\text{out}}}{\log\omega_{\text{in}}-\log\omega_{\text{out}}}\omega_{g_i,g_j}\right]\delta(g_i,g_j)\,. \tag{D.10}$$

This is exactly the definition of the node-weighted modularity (5.2) if we choose the resolution parameter

$$\gamma = \frac{\omega_{\text{in}} - \omega_{\text{out}}}{\log \omega_{\text{in}} - \log \omega_{\text{out}}} \,. \tag{D.11}$$

This indicates that the heuristically defined node-weighted modularity function in eqn. (5.2) can be understood as the weighing of a node in the ML estimator of a SBM. This equivalence only holds for a planted partition model and a specific choice of γ. It nevertheless supports our heuristic definition from a probabilistic point of view, as this illustrates that we weigh elements in the ML function of a the planted-partition model.

E

Network Construction

In this appendix, we give details for the network construction of some data sets, which we analysed in this thesis.

International airline network.

We downloaded the worldwide airline data from `OpenFlights` [356]. It includes a total of $m = 34,230$ routes between $N = 3,182$ airports on $L = 540$ airlines around the world in January 2012. We use this data to create an undirected multilayer network.

Cognitive social structures.

The cognitive social structure (CSS) networks originate from [240]. The data consist of a group of $N = 21$ managers in a small high-tech firm and their perceptions of the pairwise social interactions among them. For the 'advice' network, each individual was asked 'Who would X go to for help or advice on work?', where 'X' is one member

of the group, and the questions was repeated with respect to each possible choice of 'X'. From the questionnaire data, we construct a directed multilayer network with $L = 21$ layers, where each layer represents the cognition of the social structure of one individual. Similarly, we construct an undirected 'friendship' CSS network from the same data set, but now using a question about friendships among the individuals.

Aarhus social network.

This multiplex social network consists of five types of online and offline relationships between the employees of the Department of Computer Science at Aarhus University [286]. Among the 142 employees, $N = 61$ individuals participated in the survey. The $L = 5$ layers are 'friendship on Facebook', 'repeated leisure activities', 'current working relationships', 'co-authorship of a publication', and 'regularly eating lunch together'. The data were obtained from a mixture of traditional survey-based methods and analysis of Facebook profiles.

United Kingdom transportation network.

The UK transportation network, from [155], consists of $L = 6$ layers that each represent one mode of transportation (rail, coach, air, ferry, underground, and bus). They connect a total of $N = 267,031$ stops in UK.

London transportation network.

The London transportation multiplex network consists of $L = 3$ layers, which each represent a type of public transportation (Underground, Overground, and the *Docklands*

Light Railway (DLR)) [109]. There are a total of $N = 369$ stations.

Structural human-brain network.

We used the structural brain connectivity data from [172] to construct a multilayer network for $L = 5$ individuals. The $N = 998$ nodes represent regions of interest, which are connected to each other via tracts of white matter.

Experiment-specific protein–protein interaction network.

We construct a multirelational network from the BIOGRID database (version 3.4.150) [432]. The nodes $N = 21,412$ nodes represent proteins, and the edges represent interactions between them. The database provides $L = 13$ different experimental methods for interaction measurements. This yields the network layers.

World Trade Web.

We extracted the 'World Trade Web' from the United Nations' *Comtrade Database* [244] to create an international trade network. We used the 2014 import and export data for 288 countries and neglected re-imports and re-exports. Additionally, import and export data sets usually do not concur. For example, the total export from Germany to the US is given as $123,248,179,842$ USD, whereas Germany has a reported export of goods with a value of $127,770,823,065$ USD. In such cases, we used the arithmetic mean of the two values. In cases in which one of the reported values is 0 or missing, we use the other given value for both values. We also only consider countries that report themselves, which removes 155 out of originally 288 countries. Therefore our multilayer

network has only $N = 133$ nodes. The Comtrade Database sorts the commodities into $L = 81$ different groups, such as 'Live animals' and 'Sugars and sugar confectionery' (see E.1). In total, the network has $m \approx 476,000$ pairwise trade values. For downloading the data, we used a modified version of `tradedownloader` [341].[1]

[1]The Comtrade Database application programming interface (API) does restrict the number of requests per hour. Users that exceed this limit are blocked. To prevent us becoming blocked we include an additional wait of one hour every 20 requests.

Live animals	Printed books, newspapers, pictures etc
Meat and edible meat offal	Silk
Fish, crustaceans, molluscs, aquatic invertebrates nes	Wool, animal hair, horsehair yarn and fabric thereof
Dairy products, eggs, honey, edible animal product nes	Cotton
Products of animal origin, nes	Vegetable textile fibres nes, paper yarn, woven fabric
Edible vegetables and certain roots and tubers	manmade staple fibres
Edible fruit, nuts, peel of citrus fruit, melons	Wadding, felt, nonwovens, yarns, twine, cordage, etc
Coffee, tea, mate and spices	Carpets and other textile floor coverings
Cereals	Special woven or tufted fabric, lace, tapestry etc
Milling products, malt, starches, inulin, wheat glute	Impregnated, coated or laminated textile fabric
Lac, gums, resins, vegetable saps and extracts nes	Articles of apparel, accessories, knit or crochet
Vegetable plaiting materials, vegetable products nes	Articles of apparel, accessories, not knit or crochet
Animal, vegetable fats and oils, cleavage products, et	Other made textile articles, sets, worn clothing etc
Meat, fish and seafood food preparations nes	Footwear, gaiters and the like, parts thereof
Sugars and sugar confectionery	Headgear and parts thereof
Cereal, flour, starch, milk preparations and products	Bird skin, feathers, artificial flowers, human hair
Vegetable, fruit, nut, etc food preparations	Stone, plaster, cement, asbestos, mica, etc articles
Miscellaneous edible preparations	Ceramic products
Beverages, spirits and vinegar	Glass and glassware
Residues, wastes of food industry, animal fodder	Pearls, precious stones, metals, coins, etc
Salt, sulphur, earth, stone, plaster, lime and cement	Articles of iron or steel
Ores, slag and ash	Copper and articles thereof
Mineral fuels, oils, distillation products, etc	Nickel and articles thereof
Inorganic chemicals, precious metal compound, isotope	Aluminium and articles thereof
Organic chemicals	Zinc and articles thereof
Fertilizers	Tin and articles thereof
Tanning, dyeing extracts, tannins, derivatives ,pigments etc	Other base metals, cermets, articles thereof
Essential oils, perfumes, cosmetics, toiletries	Tools, implements, cutlery, etc of base metal
Soaps, lubricants, waxes, candles, modelling pastes	Miscellaneous articles of base metal
Albuminoidal substances; modified starches; glues; enzymes	Electrical, electronic equipment
Photographic or cinematographic goods	Railway, tramway locomotives, rolling stock, equipment
Miscellaneous chemical products	Vehicles other than railway, tramway
Plastics and articles thereof	Aircraft, spacecraft, and parts thereof
Rubber and articles thereof	Ships, boats and other floating structures
Raw hides and skins (other than furskins) and leather	Clocks and watches and parts thereof
Furskins and artificial fur, manufactures thereof	Musical instruments, parts and accessories
Wood and articles of wood, wood charcoal	Arms and ammunition, parts and accessories thereof
Cork and articles of cork	Furniture, lighting, signs, prefabricated buildings
Manufactures of plaiting material, basketwork, etc.	Toys, games, sports requisites
Pulp of wood, fibrous cellulosic material, waste etc	Works of art, collectors pieces and antiques
Commodities not specified according to kind	

Table E.1: List (in no particular order) of the $L = 81$ commodities as provided in the COMTRADE data that we used to create the directed multilayer network of international trade. The label 'nes' stands for 'not elsewhere specified'. For example, 'Products of animal origin, nes ' indicates that it groups together animal products that do not belong into one of the other groups as 'Meat and edible meat offal'.

References

[1] Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86.

[2] Abedi, M. and Gheisari, Y. (2015). Nodes with high centrality in protein interaction networks are responsible for driving signaling pathways in diabetic nephropathy. *PeerJ*, 3:e1284.

[3] Agarwal, S., Deane, C. M., Porter, M. A., and Jones, N. S. (2010). Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Computational Biology*, 6(6):e1000817.

[4] Agnarsson, G. and Greenlaw, R. (2006). *Graph Theory: Modeling, Applications, and Algorithms*. Prentice-Hall, Inc.

[5] Agostinelli, C. and Greco, L. (2012). Weighted likelihood in Bayesian inference. In *46th Scientific Meeting of the Italian Statistical Society*.

[6] Ahmad, W., Porter, M. A., and Beguerisse-Díaz, M. (2018). Tie-decay temporal networks in continuous time and eigenvector-based centralities. *arXiv preprint arXiv:1805.00193*.

[7] Ahn, A. C., Tewari, M., Poon, C.-S., and Phillips, R. S. (2006). The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS Medicine*, 3(6):e208.

[8] Aicher, C., Jacobs, A. Z., and Clauset, A. (2014). Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248.

[9] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

[10] Airoldi, E. M., Blei, D. M., Fienberg, S. E., Xing, E. P., and Jaakkola, T. (2006). Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the International Biometrics Society Annual Meeting*, volume 15.

[11] Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2013). *Essential Cell Biology*. Garland Science.

[12] Ali, W., Rito, T., Reinert, G., Sun, F., and Deane, C. M. (2014). Alignment-free protein interaction network comparison. *Bioinformatics*, 30(17):i430–i437.

[13] Alm, E. and Arkin, A. P. (2003). Biological networks. *Current Opinion in Structural Biology*, 13(2):193–202.

[14] Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC press.

[15] Alvarez-Ponce, D., Feyertag, F., and Chakraborty, S. (2017). Position matters: Network centrality considerably impacts rates of protein evolution in the human protein–protein interaction network. *Genome Biology and Evolution*, 9(6):1742–1756.

[16] Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5350–5354.

[17] Ambrogelly, A., Palioura, S., and Söll, D. (2007). Natural expansion of the genetic code. *Nature Chemical Biology*, 3(1):29–35.

[18] An, G., Mi, Q., Dutta-Moscato, J., and Vodovotz, Y. (2009). Agent-based models in translational systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(2):159–171.

[19] Antrobus, R. and Wakefield, J. G. (2011). Isolation, identification, and validation of microtubule-associated proteins from drosophila embryos. In *Microtubule Dynamics*, pages 273–291. Springer.

[20] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25.

[21] Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H., Wolkenhauer, O., Mirzaie, M., and Jafari, M. (2018). A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology*, 12(1):80.

[22] Aslak, U., Rosvall, M., and Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. *Physical Review E*, 97(6):062312.

[23] Azzalini, A. and Bowman, A. W. (1997). Applied smoothing techniques for data analysis. *Oxford Statistical Science Series, Oxford*.

[24] Baggiolini, M. and Loetscher, P. (2000). Chemokines in inflammation and immunity. *Immunology Today*, 21(9):418–420.

[25] Bakail, M. and Ochsenbein, F. (2016). Targeting protein–protein interactions, a wide open field for drug design. *Comptes Rendus Chimie*, 19(1-2):19–27.

[26] Baker, T. A., Watson, J. D., Bell, S. P., Gann, A., Losick, M., and Levine, R. (2003). *Molecular biology of the gene*. Benjamin-Cummings Publishing Company.

[27] Ball, P. and Borley, N. R. (1999). *The self-made tapestry: Pattern formation in nature*, volume 198. Oxford University Press Oxford.

[28] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56.

[29] Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499(1-3):1–101.

[30] Basch, M. L., Bronner-Fraser, M., and García-Castro, M. I. (2006). Specification of the neural crest occurs during gastrulation and requires pax7. *Nature*, 441(7090):218–222.

[31] Bassett, D. S., Porter, M. A., Wymbs, N. F., Grafton, S. T., Carlson, J. M., and Mucha, P. J. (2013). Robust detection of dynamic community structure in networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(1):013142.

[32] Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., and Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7641–7646.

[33] Battiston, F., Nicosia, V., and Latora, V. (2014). Structural measures for multiplex networks. *Physical Review E*, 89(3):032804.

[34] Bazzi, M., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. (2016). Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Modeling & Simulation*, 14(1):1–41.

[35] Bellman, R. and Beckenbach, E. F. (1965). *Inequalities*. Springer.

[36] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

[37] Berge, C. (1973). *Graphs and Hypergraphs*. North-Holland Pub. Co.

[38] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *European Journal of Biochemistry*, 80(2):319–324.

[39] Bertin, N., Simonis, N., Dupuy, D., Cusick, M. E., Han, J.-D. J., Fraser, H. B., Roth, F. P., and Vidal, M. (2007). Confirmation of organized modularity in the yeast interactome. *PLoS Biology*, 5(6):e153.

[40] Bertoli, C., Skotheim, J. M., and De Bruin, R. A. (2013). Control of cell cycle transcription during G1 and S phases. *Nature Reviews Molecular Cell Biology*, 14(8):518.

[41] Bianconi, G. (2013). Statistical mechanics of multiplex networks: Entropy and overlap. *Physical Review E*, 87(6):062806.

[42] Bianconi, G. (2018). *Multilayer Networks: Structure and Function*. Oxford University Press.

[43] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, pages pnas–0907096106.

[44] Bloch, F., Jackson, M. O., and Tebaldi, P. (2017). Centrality measures in networks. *Available at SSRN 2749124*.

[45] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[46] Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., and Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122.

[47] Boccara, N. (2010). *Modeling Complex Systems*. Springer Science & Business Media.

[48] Bode, A. M. and Dong, Z. (2004). Post-translational modification of p53 in tumorigenesis. *Nature Reviews Cancer*, 4(10):793.

[49] Böhm, I. and Schild, H. (2003). Apoptosis: The complex scenario for a silent cell death. *Molecular Imaging & Biology*, 5(1):2–14.

[50] Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120.

[51] Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

[52] Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27(1):55–71.

[53] Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, 21(4):375–395.

[54] Bossi, A. and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5(1):260.

[55] Bozhilova, L. V., Whitmore, A. V., Wray, J., Reinert, G., and Deane, C. M. (2018). Measuring rank robustness in scored protein interaction networks. to be submitted to *Network Biology*.

[56] Bradley, J. (2008). TNF-mediated inflammatory disease. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 214(2):149–160.

[57] Braet, F., De Zanger, R., Jans, D., Spector, I., and Wisse, E. (1996). Microfilament-disrupting agent latrunculin a induces and increased number of fenestrae in rat liver sinusoidal endothelial cells: Comparison with cytochalasin b. *Hepatology*, 24(3):627–635.

[58] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2006). Maximizing modularity is hard. *arXiv preprint physics/0608255*.

[59] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525.

[60] Brenowitz, M., Senear, D. F., Shea, M. A., and Ackers, G. K. (1986). Quantitative DNase footprint titration: A method for studying protein-DNA interactions. *Methods in Enzymology*, 130:132–181.

[61] Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nature Genetics*, 30(1):29–30.

[62] Brocchieri, L. and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, 33(10):3390–3400.

[63] Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. *American Psychologist*, 32:513–531.

[64] Brugere, I., Gallagher, B., and Berger-Wolf, T. Y. (2018). Network structure inference, a survey: Motivations, methods, and applications. *ACM Computing Surveys (CSUR)*, 51(2):24.

[65] Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., et al. (2003). Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Research*, 31(9):2443–2450.

[66] Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., et al. (2005). A network-based analysis of systemic inflammation in humans. *Nature*, 437(7061):1032–1037.

[67] Cantini, L., Medico, E., Fortunato, S., and Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Scientific Reports*, 5:17386.

[68] Carlsson, L., Nyström, L.-E., Sundkvist, I., Markey, F., and Lindberg, U. (1977). Actin polymerizability is influenced by profilin, a low molecular weight protein in non-muscle cells. *Journal of Molecular Biology*, 115(3):465–483.

[69] Carmena, M., Wheelock, M., Funabiki, H., and Earnshaw, W. C. (2012). The chromosomal passenger complex (cpc): From easy rider to the godfather of mitosis. *Nature Reviews Molecular Cell Biology*, 13(12):789.

[70] Carvalhal, C., Deusdado, S., and Deusdado, L. (2012). Asap: An automated system for scientific literature search in pubmed using web agents. In *6th International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 73–78. Springer.

[71] Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-F., and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLOS ONE*, 5(7):e11596.

[72] Ceccato, F., Scaroni, C., and Boscaro, M. (2018). The adrenal glands. *Principles of Endocrinology and Hormone Action*, pages 387–421.

[73] Cedar, H. and Bergman, Y. (2012). Programming of DNA methylation patterns. *Annual Review of Biochemistry*, 81:97–117.

[74] Cedergren-Zeppezauer, E. S., Goonesekere, N. C., Rozycki, M. D., Myslik, J. C., Dauter, Z., Lindberg, U., and Schutt, C. E. (1994). Crystallization and structure determination of bovine profilin at 2.0 Angstrom resolution. *Journal of Molecular Biology*, 240(5):459–475.

[75] Chan, S. S.-K. and Kyba, M. (2013). What is a master regulator? *Journal of Stem Cell Research & Therapy*, 3(114).

[76] Chance, B., Estabrook, R. W., and Ghosh, A. (1964). Damped sinusoidal oscillations of cytoplasmic reduced pyridine nucleotide in yeast cells. *Proceedings of the National Academy of Sciences of the United States of America*, 51(6):1244–1251.

[77] Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379.

[78] Chautard, E., Thierry-Mieg, N., and Ricard-Blum, S. (2009). Interaction networks: from protein functions to drug discovery. A review. *Pathologie Biologie*, 57(4):324–333.

[79] Chen, G. and Wang, J. (2012). Identifying functional modules in tissue specific protein interaction network. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, pages 581–586. IEEE.

[80] Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290.

[81] Chen, J. Y., Pandey, R., and Nguyen, T. M. (2017). HAPPI-2: A comprehensive and high-quality map of human annotated and predicted protein interactions. *BMC Genomics*, 18(1):182.

[82] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., et al. (1998). SGD: Saccharomyces genome database. *Nucleic Acids Research*, 26(1):73–79.

[83] Chik, J. K., Lindberg, U., and Schutt, C. E. (1996). The structure of an open state of β-actin at 2.65 å resolution. *Journal of Molecular Biology*, 263(4):607–623.

[84] Cho, D.-Y., Kim, Y.-A., and Przytycka, T. M. (2012). Network biology approach to complex diseases. *PLoS Computational Biology*, 8(12):e1002820.

[85] Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics*, 22(13):1623–1630.

[86] Chuang, J.-Y., Wu, C.-H., Lai, M.-D., Chang, W.-C., and Hung, J.-J. (2009). Overexpression of Sp1 leads to p53-dependent apoptosis in cancer cells. *International Journal of Cancer*, 125(9):2066–2076.

[87] Ciriello, G. and Guerra, C. (2008). A review on models and algorithms for motif discovery in protein–protein interaction networks. *Briefings in Functional Genomics and Proteomics*, 7(2):147–156.

[88] Claman, H. N. (1972). Corticosteroids and lymphoid cells. *New England Journal of Medicine*, 287(8):388–397.

[89] Consortium, G. O. (2004). The Gene Ontology database and informatics resource. *Nucleic Acids Research*, 32(suppl_1):D258–D261.

[90] Cooper, G. M. and Ganem, D. (1997). The cell: A molecular approach. *Nature Medicine*, 3(9):1042–1042.

[91] Cooper, G. M. and Hausman, R. E. (2004). *The Cell: Molecular Approach*. Medicinska naklada.

[92] Corominas, R., Yang, X., Lin, G. N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S. A., et al. (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature Communications*, 5:3650.

[93] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. *Science*, 327(5964):425–431.

[94] Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306):aaf1420.

[95] Coussens, L. M. and Werb, Z. (2002). Inflammation and cancer. *Nature*, 420(6917):860.

[96] Cozzo, E., Kivelä, M., De Domenico, M., Solé, A., Arenas, A., Gómez, S., Porter, M. A., and Moreno, Y. (2013). Clustering coefficients in multiplex networks. *arXiv preprint arXiv:1307.6780*.

[97] Csermely, P., Agoston, V., and Pongor, S. (2005). The efficiency of multi-target drugs: the network approach might help drug design. *Trends in Pharmacological Sciences*, 26(4):178–182.

[98] Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & Therapeutics*, 138(3):333–408.

[99] Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., et al. (2008). Literature-curated protein interaction datasets. *Nature Methods*, 6(1):39.

[100] Dabbs, B. and Junker, B. (2016). Comparison of cross-validation methods for stochastic block models. *arXiv preprint arXiv:1605.03000*.

[101] D'Agostino, G. and Scala, A. (2014). *Networks of Networks: The Last Frontier of Complexity*, volume 340. Springer.

[102] Darnell, J. E., Lodish, H. F., Baltimore, D., et al. (1990). *Molecular Cell Biology*, volume 2. Scientific American Books New York.

[103] Das, J. and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92.

[104] Davie, J. R., He, S., Li, L., Sekhavat, A., Espino, P., Drobic, B., Dunn, K. L., Sun, J.-M., Chen, H. Y., Yu, J., et al. (2008). Nuclear organization and chromatin dynamics–SP1, SP3 and histone deacetylases. *Advances in Enzyme Regulation*, 48:189.

[105] De Bacco, C., Power, E. A., Larremore, D. B., and Moore, C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317.

[106] De Domenico, M., Granell, C., Porter, M. A., and Arenas, A. (2016). The physics of spreading processes in multilayer networks. *Nature Physics*, 12(10):901.

[107] De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015a). Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X*, 5(1):011027.

[108] De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., and Arenas, A. (2013). Mathematical formulation of multilayer networks. *Physical Review X*, 3(4):041022.

[109] De Domenico, M., Solé-Ribalta, A., Gómez, S., and Arenas, A. (2014). Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8351–8356.

[110] De Domenico, M., Solé-Ribalta, A., Omodei, E., Gómez, S., and Arenas, A. (2015b). Ranking in interconnected multilayer networks reveals versatile nodes. *Nature Communications*, 6:6868.

[111] de le Hospital, M. (1715). *Analyse des Infiniment Petits: Pour l'Intelligence des Lignes Courbes*. François Montalant.

[112] de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727.

[113] De Meo, P., Ferrara, E., Fiumara, G., and Ricciardello, A. (2012). A novel measure of edge centrality in social networks. *Knowledge-based Systems*, 30:136–150.

[114] Deane, C. M., Dong, M., Huard, F. P., Lance, B. K., and Wood, G. R. (2007). Cotranslational protein folding—Fact or fiction? *Bioinformatics*, 23(13):i142–i148.

[115] Deane, C. M., Salwiński, Ł., Xenarios, I., and Eisenberg, D. (2002). Protein interactions two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics*, 1(5):349–356.

[116] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701.

[117] DeLano, W. L. (2002). The pymol molecular graphics system. *http://www. pymol. org*.

[118] Detrain, C. and Deneubourg, J.-L. (2006). Self-organized structures in a superorganism: do ants "behave" like molecules? *Physics of Life Reviews*, 3(3):162–187.

[119] Dever, T. E. and Green, R. (2012). The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harbor Perspectives in Biology*, 4(7):a013706.

[120] Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155.

[121] Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein–protein interaction networks: An integrated exact approach. *Bioinformatics*, 24(13):i223–i231.

[122] Diz, A. P., Carvajal-Rodríguez, A., and Skibinski, D. O. (2011). Multiple hypothesis testing in proteomics: A strategy for experimental work. *Molecular & Cellular Proteomics*, 10(3):M110–004374.

[123] Domenico, M. D. (2018). Multilayer network modeling of integrated biological systems: Comment on "network science of biological systems at different scales: A review" by gosak et al. *Physics of Life Reviews*, 24:149 – 152.

[124] Dominguez, R. and Holmes, K. C. (2011). Actin structure and function. *Annual Review of Biophysics*, 40:169–186.

[125] Donges, J. F., Zou, Y., Marwan, N., and Kurths, J. (2009). The backbone of the climate network. *Europhysics Letters*, 87(4):48007.

[126] Duina, A. A., Miller, M. E., and Keeney, J. B. (2014). Budding yeast for budding geneticists: a primer on the saccharomyces cerevisiae model system. *Genetics*, 197(1):33–48.

[127] Dunn, R., Dudbridge, F., and Sanderson, C. M. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6(1):39.

[128] Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48.

[129] Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574.

[130] El-Kebir, M., Klau, G. W., and Engler, M. Heinz – Single species module discovery.

[131] El Sayed, K. A., Youssef, D. T., and Marchetti, D. (2006). Bioactive natural and semisynthetic latrunculins. *Journal of Natural Products*, 69(2):219–223.

[132] Emig, D. and Albrecht, M. (2011). Tissue-specific proteins and functional implications. *Journal of Proteome Research*, 10(4):1893–1903.

[133] Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40.

[134] Estrada, E. (2011). Community detection based on network communicability. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(1):016103.

[135] Estrada, E. (2012). *The Structure of Complex Networks: Theory and Applications*. Oxford University Press.

[136] Estrada, E. (2013). Graph and network theory in physics. *arXiv preprint arXiv:1302.4378*.

[137] Estrada, E. and Hatano, N. (2008). Communicability in complex networks. *Physical Review E*, 77(3):036111.

[138] Estrada, E. and Higham, D. J. (2010). Network properties revealed through matrix functions. *SIAM Review*, 52(4):696–714.

[139] Estrada, E. and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103.

[140] Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140.

[141] Expert, P., Evans, T. S., Blondel, V. D., and Lambiotte, R. (2011). Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19):7663–7668.

[142] Feldmann, H. (2011). *Yeast: Molecular and Cell Biology*. John Wiley & Sons.

[143] Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305.

[144] Fields, S. and Johnston, M. (2005). Whither model organism research? *Science*, 307(5717):1885–1886.

[145] Fields, S. and Song, O.-k. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245.

[146] Fong, Y., Marano, M. A., Moldawer, L. L., Wei, H., Calvano, S. E., Kenney, J. S., Allison, A. C., Cerami, A., Shires, G., and Lowry, S. (1990). The acute splanchnic and peripheral tissue metabolic response to endotoxin in humans. *The Journal of Clinical Investigation*, 85(6):1896–1904.

[147] Ford, S. K. and Pringle, J. R. (1991). Cellular morphogenesis in the *Saccharomyces cerevisiae* cell cycle: Localization of the CDC11 gene product and the timing of events at the budding site. *Genesis*, 12(4):281–292.

[148] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

[149] Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41.

[150] Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

[151] Fosdick, B. K., Larremore, D. B., Nishimura, J., and Ugander, J. (2018). Configuring random graph models with fixed degree sequences. *SIAM Review*, 60(2):315–355.

[152] Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics*, 29(3):150–159.

[153] Galas, D. J. and Schmitz, A. (1978). DNase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, 5(9):3157–3170.

[154] Gale, C. A., Leonard, M. D., Finley, K. R., Christensen, L., McClellan, M., Abbey, D., Kurischko, C., Bensen, E., Tzafrir, I., Kauffman, S., et al. (2009). Sla2 mutations cause swe1-mediated cell cycle phenotypes in *Candida albicans* and *Saccharomyces cerevisiae*. *Microbiology*, 155(12):3847–3859.

[155] Gallotti, R. and Barthelemy, M. (2015). The multilayer temporal network of public transport in great britain. *Scientific Data*, 2:140056.

[156] Gao, S., Giansanti, M. G., Buttrick, G. J., Ramasubramanyan, S., Auton, A., Gatti, M., and Wakefield, J. G. (2008). Australin: a chromosomal passenger protein required specifically for drosophila melanogaster male meiosis. *Journal of Cell Biology*, 180(3):521–535.

[157] George, S. J. (2008). Wnt pathway – A new role in regulation of inflammation. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 28:400–2.

[158] Gething, M.-J. and Sambrook, J. (1992). Protein folding in the cell. *Nature*, 355(6355):33.

[159] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

[160] Gleich, D. F. (2015). PageRank beyond the web. *SIAM Review*, 57(3):321–363.

[161] Golub, G. H. and Van Loan, C. F. (2012). *Matrix Computations*, volume 3. JHU Press.

[162] Gomez, S., Diaz-Guilera, A., Gomez-Gardenes, J., Perez-Vicente, C. J., Moreno, Y., and Arenas, A. (2013). Diffusion dynamics on multiplex networks. *Physical Review Letters*, 110(2):028701.

[163] Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106.

[164] Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, 4(9):117.

[165] Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569.

[166] Greenhalgh, D. G. (1998). The role of apoptosis in wound healing. *The International Journal of Biochemistry & Cell Biology*, 30(9):1019–1030.

[167] Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Research*, 29(17):3513–3519.

[168] Gross, E., Davis, B., Ho, K. L., Bates, D. J., and Harrington, H. A. (2016). Numerical algebraic geometry for model selection and its application to the life sciences. *Journal of The Royal Society Interface*, 13(123):20160256.

[169] Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM.

[170] Guimerá, R. and Amaral, L. A. N. (2005a). Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(02):P02001.

[171] Guimerá, R. and Amaral, L. A. N. (2005b). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

[172] Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., and Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, 6(7):e159.

[173] Hahn, A., Rahnenführer, J., Talwar, P., and Lengauer, T. (2005). Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6(1):112.

[174] Han, J.-D. J., Bertin, N., Tong, H., Goldberg, D. S., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88.

[175] Hansson, G. K. (2005). Inflammation, atherosclerosis, and coronary artery disease. *New England Journal of Medicine*, 352(16):1685–1695.

[176] Hardy, G. H., Littlewood, J. E., and Pólya, G. (1952). *Inequalities*. Cambridge University Press.

[177] Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120.

[178] Hartl, F. U. (1996). Molecular chaperones in cellular protein folding. *Nature*, 381(6583):571.

[179] Hartlerode, A., Odate, S., Shim, I., Brown, J., and Scully, R. (2011). Cell cycle-dependent induction of homologous recombination by a tightly regulated i-scei fusion protein. *PLOS ONE*, 6(3):e16501.

[180] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761supp):C47.

[181] He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genetics*, 2(6):e88.

[182] He, Y.-M. and Ma, B.-G. (2016). Abundance and temperature dependency of protein-protein interaction revealed by interface structure analysis and stability evolution. *Scientific Reports*, 6:26737.

[183] Heimlicher, S., Lelarge, M., and Massoulié, L. (2012). Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*.

[184] Heitzig, J., Donges, J. F., Zou, Y., Marwan, N., and Kurths, J. (2012). Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes. *The European Physical Journal B-Condensed Matter and Complex Systems*, 85(1):1–22.

[185] Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450(7172):964.

[186] Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3:31.

[187] Hernández-Lemus, E., Espinal-Enríquez, J., and de Anda-Jáuregui, G. (2018). Probabilistic multilayer networks. *arXiv preprint arXiv:1808.07857*.

[188] Hershey, J. W., Sonenberg, N., and Mathews, M. B. (2012). Principles of translational control: An overview. *Cold Spring Harbor Perspectives in Biology*, 4(12):a011528.

[189] Herskowitz, I. (1988). Life cycle of the budding yeast saccharomyces cerevisiae. *Microbiological Reviews*, 52(4):536.

[190] Ho, B., Baryshnikova, A., and Brown, G. W. (2018). Unification of protein abundance datasets yields a quantitative saccharomyces cerevisiae proteome. *Cell Systems*, 6(2):192–205.

[191] Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544.

[192] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

[193] Hoffmann, H. (2015). Simple violin plot using matlab default kernel density estimation. Matlab Central.

[194] Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125.

[195] Holtzman, D. A., Yang, S., and Drubin, D. G. (1993). Synthetic-lethal interactions identify two novel genes, SLA1 and SLA2, that control membrane cytoskeleton assembly in *Saccharomyces cerevisiae*. *The Journal of Cell Biology*, 122(3):635–644.

[196] Honeycutt, J. and Thirumalai, D. (1990). Metastability of the folded states of globular proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 87(9):3526–3529.

[197] Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682.

[198] Horlbeck, M. A., Xu, A., Wang, M., Bennett, N. K., Park, C. Y., Bogdanoff, D., Adamson, B., Chow, E. D., Kampmann, M., Peterson, T. R., et al. (2018). Mapping the genetic landscape of human cells. *Cell*, 174(4):953–967.

[199] Hotamisligil, G. S. (2006). Inflammation and metabolic disorders. *Nature*, 444(7121):860.

[200] Hric, D., Peixoto, T. P., and Fortunato, S. (2016). Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, 6(3):031038.

[201] Hu, K., Hu, J.-B., Xiang, J., Li, H.-J., Zhang, Y., Chen, S., and Yi, C.-H. (2017). Predicting disease-related genes by path-based similarity and community structure in protein-protein interaction network. *arXiv preprint arXiv:1707.06846*.

[202] Huang, H. and Bader, J. S. (2008). Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3):372–378.

[203] Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H., and Han, J.-D. J. (2013). Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Computational Biology*, 9(3):e1002998.

[204] Hughes, J. R., Meireles, A. M., Fisher, K. H., Garcia, A., Antrobus, P. R., Wainman, A., Zitzmann, N., Deane, C., Ohkura, H., and Wakefield, J. G. (2008). A microtubule interactome: Complexes with roles in cell cycle and mitosis. *PLoS Biology*, 6(4):e98.

[205] Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505.

[206] Hwang, S., Son, S.-W., Kim, S. C., Kim, Y. J., Jeong, H., and Lee, D. (2008). A protein interaction network associated with asthma. *Journal of Theoretical Biology*, 252(4):722–731.

[207] Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Research*, 18(4):644–652.

[208] Idriss, H. T. and Naismith, J. H. (2000). TNFα and the TNF receptor superfamily: Structure-function relationship (s). *Microscopy Research and Technique*, 50(3):184–195.

[209] Ingber, D. E. (2003). Tensegrity ii. how structural networks influence cellular information processing networks. *Journal of Cell Science*, 116(8):1397–1408.

[210] Ings, T. C., Montoya, J. M., Bascompte, J., Blüthgen, N., Brown, L., Dormann, C. F., Edwards, F., Figueroa, D., Jacob, U., Jones, J. I., et al. (2009). Ecological networks–beyond food webs. *Journal of Animal Ecology*, 78(1):253–269.

[211] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

[212] Ivanic, J., Yu, X., Wallqvist, A., and Reifman, J. (2009). Influence of protein abundance on high-throughput protein-protein interaction detection. *PLOS ONE*, 4(6):e5815.

[213] Jackson, R. J., Hellen, C. U., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, 11(2):113.

[214] Jacobs, W. M. and Shakhnovich, E. I. (2017). Evidence of evolutionary selection for cotranslational folding. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43):11434–11439.

[215] Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193.

[216] Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

[217] Jeub, L. G. S., Bazzi, M., Jutla, I. S., and Mucha, P. J. (2011-2014). A generalized Louvain method for community detection implemented in MATLAB, Version 2.1.

[218] Jones, E. W., Strathern, J. N., and Broach, J. R. (1981). *The Molecular Biology of the Yeast Saccharomyces, Life Cycle and Inheritance*. Cold Spring Harbor Laboratory.

[219] Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484.

[220] Jonsson, P. F. and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297.

[221] Jordan, M. I. et al. (1995). Why the logistic function? A tutorial discussion on probabilities and neural networks.

[222] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl 1):D428–D432.

[223] Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *BioMed Research International*, 2005(2):96–103.

[224] Juo, Z. S., Kassavetis, G. A., Wang, J., Geiduschek, E. P., and Sigler, P. B. (2003). Crystal structure of a transcription factor IIIB core interface ternary complex. *Nature*, 422(6931):534.

[225] Kargi, A. Y. and Iacobellis, G. (2014). Adipose tissue and adrenal glands: novel pathophysiological mechanisms and clinical applications. *International Journal of Endocrinology*, 2014.

[226] Karrer, B. and Newman, M. E. J. (2009). Random graph models for directed acyclic networks. *Physical Review E*, 80(4):046110.

[227] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

[228] Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. *Gene*, 514(1):1–30.

[229] Kerr, G., Ruskin, H. J., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, 38(3):283–293.

[230] Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2008). Human protein reference database—2009 update. *Nucleic Acids Research*, 37(suppl_1):D767–D772.

[231] Kim, H. B., Haarer, B. K., and Pringle, J. R. (1991). Cellular morphogenesis in the *Saccharomyces cerevisiae* cell cycle: Localization of the CDC3 gene product and the timing of events at the budding site. *The Journal of Cell Biology*, 112(4):535–544.

[232] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

[233] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery*, 46(5):604–632.

[234] Klimm, F., Borge-Holthoefer, J., Wessel, N., Kurths, J., and Zamora-López, G. (2014). Individual node's contribution to the mesoscale of complex networks. *New Journal of Physics*, 16(12):125006.

[235] Ko, L. and Engel, J. (1993). DNA-binding specificities of the GATA transcription factor family. *Molecular and Cellular Biology*, 13(7):4011–4022.

[236] Kobe, B., Guncar, G., Buchholz, R., Huber, T., Maco, B., Cowieson, N., Martin, J. L., Marfori, M., and Forwood, J. K. (2008). Crystallography and protein–protein interactions: Biological interfaces and crystal contacts. *Biochemical Society Transactions*, 36(6):1438–1441.

[237] Koltai, H. and Weingarten-Baror, C. (2008). Specificity of DNA microarray hybridization: Characterization, effectors and approaches for data correction. *Nucleic Acids Research*, 36(7):2395–2405.

[238] Komurov, K. and White, M. (2007). Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Molecular Systems Biology*, 3(1):110.

[239] Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2015). Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Research*, 44(D1):D536–D541.

[240] Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, 9(2):109–134.

[241] Krause, J., Lusseau, D., and James, R. (2009). Animal social networks: An introduction. *Behavioral Ecology and Sociobiology*, 63(7):967–973.

[242] Kuhla, A., Norden, J., Abshagen, K., Menger, M., and Vollmar, B. (2013). RAGE blockade and hepatic microcirculation in experimental endotoxaemic liver failure. *British Journal of Surgery*, 100(9):1229–1239.

[243] Kumar, A., Bhandari, A., Sinha, R., Sardar, P., Sushma, M., Goyal, P., Goswami, C., and Grapputo, A. (2012). Molecular phylogeny of OVOL genes illustrates a conserved C2H2 zinc finger domain coupled by hypervariable unstructured regions. *PLOS ONE*, 7(6):e39399.

[244] Labs, C. Comtrade – United Nations Commodity Trade Statistics Database.

[245] Ladyman, J., Lambert, J., and Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67.

[246] Lage, K., Hansen, N. T., Karlberg, E. O., Eklund, A. C., Roque, F. S., Donahoe, P. K., Szallasi, Z., Jensen, T. S., and Brunak, S. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20870–20875.

[247] Lampe, M. A., Burlingame, A., Whitney, J., Williams, M. L., Brown, B. E., Roitman, E., and Elias, P. M. (1983). Human stratum corneum lipids: Characterization and regional variations. *Journal of Lipid Research*, 24(2):120–130.

[248] Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5):056117.

[249] Lancichinetti, A. and Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E*, 84(6):066122.

[250] Lasko, P. (2012). mrna localization and translational control in drosophila oogenesis. *Cold Spring Harbor perspectives in biology*, page a012294.

[251] Lassing, I., Schmitzberger, F., Björnstedt, M., Holmgren, A., Nordlund, P., Schutt, C. E., and Lindberg, U. (2007). Molecular and structural basis for redox regulation of β-actin. *Journal of Molecular Biology*, 370(2):331–348.

[252] Latchman, D. S. (1997). Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, 29(12):1305–1312.

[253] Latimer, J. J., Nazir, T., Flowers, L. C., Forlenza, M. J., Beaudry-Rodgers, K., Kelly, C. M., Conte, J. A., Shestak, K., Kanbour-Shakir, A., and Grant, S. G. (2003). Unique tissue-specific level of DNA nucleotide excision repair in primary human mammary epithelial cultures. *Experimental Cell Research*, 291(1):111–121.

[254] Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013). Mapping genetic interactions in human cancer cells with rnai and multiparametric phenotyping. *Nature Methods*, 10(5):427.

[255] Law, E. C., de Oliveira, S. H., Kelm, S., Shi, J., and Deane, C. M. (2017). Investigating cotranslational folding in membrane proteins using fragment-based structure prediction. *Biophysical Journal*, 112(3):61a.

[256] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

[257] Lee, J. and Lee, J. (2013). Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction. *PLOS ONE*, 8(4):e60372.

[258] Lemus, E. H., López, K. B., Lemus, R., and Herrera, R. G. (2015). The role of master regulators in gene regulatory networks. *Papers in Physics*, 7:070011.

[259] Lengyel, P. and Söll, D. (1969). Mechanism of protein biosynthesis. *Bacteriological Reviews*, 33(2):264.

[260] Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th International Conference on World Wide Web*, pages 695–704. ACM.

[261] Lewis, A. (2011). *Communities and Homology in Protein-protein Interactions*. PhD thesis, Oxford University.

[262] Lewis, A. C., Jones, N. S., Porter, M. A., and Deane, C. M. (2010a). The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100.

[263] Lewis, A. C., Saeed, R., and Deane, C. M. (2010b). Predicting protein–protein interactions in the context of protein evolution. *Molecular BioSystems*, 6(1):55–64.

[264] Li, B., Mackay, D. R., Dai, Q., Li, T. W., Nair, M., Fallahi, M., Schonbaum, C. P., Fantes, J., Mahowald, A. P., Waterman, M. L., et al. (2002). The lef1/β-catenin complex activates movo1, a mouse homolog of drosophila ovo required for epidermal appendage differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6064–6069.

[265] Li, J.-R., Sun, C.-H., Li, W., Chao, R.-F., Huang, C.-C., Zhou, X. J., and Liu, C.-C. (2015). Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Research*, 44(D1):D944–D951.

[266] Li, L. and Davie, J. R. (2010). The role of sp1 and sp3 in normal and cancer cell biology. *Annals of Anatomy (the journal formerly known as 'Anatomischer Anzeiger')*, 192(5):275–283.

[267] Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154.

[268] Li, M., Wang, J., and Chen, J. (2008a). A graph-theoretic method for mining overlapping functional modules in protein interaction networks. In *International Symposium on Bioinformatics Research and Applications*, pages 208–219. Springer.

[269] Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C. L. L., et al. (2008b). Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLOS Biology*, 6(2):e27.

[270] Liang, Y., Tsoi, L. C., Xing, X., Beamer, M. A., Swindell, W. R., Sarkar, M. K., Berthier, C. C., Stuart, P. E., Harms, P. W., Nair, R. P., et al. (2017). A gene network regulated by the transcription factor VGLL3 as a promoter of sex-biased autoimmune diseases. *Nature Immunology*, 18(2):152.

[271] Lin, W.-h., Liu, W.-c., and Hwang, M.-j. (2009). Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. *BMC Systems Biology*, 3(1):32.

[272] Littlepage, L. E., Adler, A. S., Kouros-Mehr, H., Huang, G., Chou, J., Krig, S. R., Griffith, O. L., Korkola, J. E., Qu, K., Lawson, D. A., et al. (2012). The transcription factor ZNF217 is a prognostic biomarker and therapeutic target during breast cancer progression. *Cancer Discovery*, 2(7):638–651.

[273] Liu, B. and Shuai, K. (2008). Targeting the PIAS1 SUMO ligase pathway to control inflammation. *Trends in Pharmacological Sciences*, 29(10):505–509.

[274] Liu, C., Li, J., and Zhao, Y. (2010). Exploring hierarchical and overlapping modular structure in the yeast protein interaction network. In *BMC Genomics*, volume 11, page S17. BioMed Central.

[275] Liu, E. T. (2005). Systems biology, integrative biology, predictive biology. *Cell*, 121(4):505–506.

[276] Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3):535–550.

[277] Liu, Z., Xing, D., Su, Q. P., Zhu, Y., Zhang, J., Kong, X., Xue, B., Wang, S., Sun, H., Tao, Y., et al. (2014). Super-resolution imaging and tracking of protein–protein interactions in sub-diffraction cellular space. *Nature Communications*, 5:4443.

[278] Locksley, R. M., Killeen, N., and Lenardo, M. J. (2001). The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell*, 104(4):487–501.

[279] Lopes, T. J., Schaefer, M., Shoemaker, J., Matsuoka, Y., Fontaine, J.-F., Neumann, G., Andrade-Navarro, M. A., Kawaoka, Y., and Kitano, H. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, 27(17):2414–2421.

[280] López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8):583.

[281] Louche, A., Salcedo, S. P., and Bigot, S. (2017). Protein–protein interactions: Pull-down assays. In *Bacterial Protein Secretion Systems*, pages 247–255. Springer.

[282] Luecken, M., Page, M., Crosby, A., Mason, S., Reinert, G., and Deane, C. M. (2017). Commwalker: Correctly evaluating modules in molecular networks in light of annotation bias. *Bioinformatics*, 34(6):994–1000.

[283] Luo, F., Yang, Y., Chen, C.-F., Chang, R., Zhou, J., and Scheuermann, R. H. (2006). Modular organization of protein interaction networks. *Bioinformatics*, 23(2):207–214.

[284] Ma'ayan, A. (2017). Complex systems biology. *Journal of The Royal Society Interface*, 14(134):20170391.

[285] MacMahon, M. and Garlaschelli, D. (2013). Community detection for correlation matrices. *arXiv preprint arXiv:1311.1924*.

[286] Magnani, M., Micenkova, B., and Rossi, L. (2013). Combinatorial analysis of multiple networks. *arXiv preprint arXiv:1303.4986*.

[287] Malod-Dognin, N. and Przulj, N. (2018). Functional geometry of protein-protein interaction networks. *arXiv preprint arXiv:1804.04428*.

[288] Mangioni, G., Jurman, G., and De Domenico, M. (2018). Multilayer flows in molecular networks identify biological modules in the human proteome. *arXiv preprint arXiv:1801.10144*.

[289] Mani, R., Onge, R. P. S., Hartman, J. L., Giaever, G., and Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3461–3466.

[290] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

[291] Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, 13(4):366.

[292] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*.

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

[294] Masuda, N. and Lambiotte, R. (2016). *A Guidance to Temporal Networks*. World Scientific.

[295] Matlin, A. J., Clark, F., and Smith, C. W. (2005). Understanding alternative splicing: Towards a cellular code. *Nature Reviews Molecular cell biology*, 6(5):386–398.

[296] Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126.

[297] Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and Azzi, M. (2017). RAWGraphs: A visualisation platform to create open outputs. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, page 28. ACM.

[298] McCloskey, D., Palsson, B. Ø., and Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of escherichia coli. *Molecular Systems Biology*, 9(1):661.

[299] McMurray, M. A., Bertin, A., Garcia, G., Lam, L., Nogales, E., and Thorner, J. (2011). Septin filament formation is essential in budding yeast. *Developmental Cell*, 20(4):540–549.

[300] Melton, L. (2004). Protein arrays: proteomics in multiplex. *Nature*, 429(6987):101.

[301] Mendenhall, M. D. and Hodge, A. E. (1998). Regulation of CDC28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, 62(4):1191–1243.

[302] Mete, M., Tang, F., Xu, X., and Yuruk, N. (2008). A structural approach for finding functional modules from large biological networks. In *BMC Bioinformatics*, volume 9, page S19. BioMed Central.

[303] Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P. D. (2009). PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the GENE ONTOLOGY CONSORTIUM. *Nucleic Acids Research*, 38(suppl_1):D204–D210.

[304] Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene function analysis with the panther classification system. *Nature Protocols*, 8(8):1551–1566.

[305] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

[306] Mirollo, R. E. and Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics*, 50(6):1645–1662.

[307] Mises, R. and Pollaczek-Geiringer, H. (1929). Praktische Verfahren der Gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 9(1):58–77.

[308] Mistry, D., Wise, R. P., and Dickerson, J. A. (2017). Diffslc: A graph centrality method to detect essential proteins of a protein-protein interaction network. *PLOS ONE*, 12(11):e0187091.

[309] Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

[310] Mitsopoulos, C., Schierz, A. C., Workman, P., and Al-Lazikani, B. (2015). Distinctive behaviors of druggable proteins in cellular networks. *PLoS Computational Biology*, 11(12):e1004597.

[311] Miyagawa, Y., Nishimura, H., Tsujimura, A., Matsuoka, Y., Matsumiya, K., Okuyama, A., Nishimune, Y., and Tanaka, H. (2005). Single-nucleotide polymorphisms and mutation analyses of the TNP1 and TNP2 genes of fertile and infertile human male populations. *Journal of Andrology*, 26(6):779–786.

[312] Moellering, R. E., Cornejo, M., Davis, T. N., Del Bianco, C., Aster, J. C., Blacklow, S. C., Kung, A. L., Gilliland, D. G., Verdine, G. L., and Bradner, J. E. (2009). Direct inhibition of the notch transcription factor complex. *Nature*, 462(7270):182.

[313] Mol, M., Kosey, D., Boppana, R., and Singh, S. (2018). Transcription factor target gene network governs the logical abstraction analysis of the synthetic circuit in leishmaniasis. *Scientific Reports*, 8(1):3464.

[314] Moore, C. (2017). The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*.

[315] Morgan, D. O. (1995). Principles of cdk regulation. *Nature*, 374(6518):131.

[316] Morgan, D. O. (2007). *The Cell Cycle: Principles of Control*. New Science Press.

[317] Morimoto, R. I. et al. (1993). Cells in stress: transcriptional activation of heat shock genes. *Science*, 259:1409–1409.

[318] Mostowy, S. and Cossart, P. (2012). Septins: The fourth component of the cytoskeleton. *Nature Reviews Molecular Cell Biology*, 13(3):183.

[319] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (casp)—round xii. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15.

[320] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878.

[321] Muraro, D. and Simmons, A. (2016). An integrative analysis of gene expression and molecular interaction data to identify dys-regulated sub-networks in inflammatory bowel disease. *BMC Bioinformatics*, 17(1):42.

[322] Murray, J. D. (2002). *Mathematical Biology I. An Introduction*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer, New York, 3 edition.

[323] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(suppl_1):i302–i310.

[324] Nadakuditi, R. R. and Newman, M. E. J. (2012). Graph spectra and the detectability of community structure in networks. *Physical Review Letters*, 108(18):188701.

[325] Nelson, R. J., Ziegelhoffer, T., Nicolet, C., Werner-Washburne, M., and Craig, E. A. (1992). The translation machinery and 70 kd heat shock protein cooperate in protein synthesis. *Cell*, 71(1):97–105.

[326] Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286.

[327] Newman, M. E. (2011). Complex systems: A survey. *arXiv preprint arXiv:1112.1440*.

[328] Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104.

[329] Newman, M. E. J. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94(5):052315.

[330] Newman, M. E. J. (2018). *Networks: An Introduction (2nd Edition)*. Oxford University Press.

[331] Newman, M. E. J. and Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications*, 7:11863.

[332] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

[333] Newman, M. E. J. and Peixoto, T. P. (2015). Generalized communities in networks. *Physical Review Letters*, 115(8):088701.

[334] Newman, M. E. J. and Reinert, G. (2016). Estimating the number of communities in a network. *Physical Review Letters*, 117(7):078301.

[335] Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48.

[336] Ng, M. K.-P., Li, X., and Ye, Y. (2011). Multirank: co-ranking for objects and relations in multi-relational data. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1217–1225. ACM.

[337] Ngounou Wetie, A. G., Sokolowska, I., Woods, A. G., Roy, U., Loo, J. A., and Darie, C. C. (2013). Investigation of stable and transient protein–protein interactions: past, present, and future. *Proteomics*, 13(3-4):538–557.

[338] Nigg, E. A. (1995). Cyclin-dependent protein kinases: key regulators of the eukaryotic cell cycle. *Bioessays*, 17(6):471–480.

[339] Nooren, I. M. and Thornton, J. M. (2003). Diversity of protein–protein interactions. *The EMBO Journal*, 22(14):3486–3492.

[340] Nurse, P. (1997). Reductionism: The ends of understanding. *Nature*, 387(6634):657.

[341] O'Keeffe, E. Tradedownloader — Download trade data from Comtrade using the exposed Comtrade API. GitHub.

[342] Ooi, S. L., Shoemaker, D. D., and Boeke, J. D. (2003). DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nature Genetics*, 35(3):277.

[343] Opitz, C. A., Kulke, M., Leake, M. C., Neagoe, C., Hinssen, H., Hajjar, R. J., and Linke, W. A. (2003). Damped elastic recoil of the titin spring in myofibrils of human myocardium. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22):12688–12693.

[344] Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17.

[345] Ou-Yang, L., Dai, D.-Q., Li, X.-L., Wu, M., Zhang, X.-F., and Yang, P. (2014). Detecting temporal protein complexes from dynamic protein-protein interaction networks. *BMC Bioinformatics*, 15(1):335.

[346] Owens, N. D., Blitz, I. L., Lane, M. A., Patrushev, I., Overton, J. D., Gilchrist, M. J., Cho, K. W., and Khokha, M. K. (2016). Measuring absolute RNA copy numbers at high temporal resolution reveals transcriptome kinetics in development. *Cell Reports*, 14(3):632–647.

[347] Özgür, A., Vu, T., Erkan, G., and Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285.

[348] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab.

[349] Page, T. H., Smolinska, M., Gillespie, J., Urbaniak, A. M., and Foxwell, B. M. (2009). Tyrosine kinases and inflammatory signalling. *Current Molecular Medicine*, 9(1):69–85.

[350] Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

[351] Palleros, D. R., Shi, L., Reid, K. L., and Fink, A. L. (1994). HSP70–protein complexes. Complex stability and conformation of bound substrate protein. *Journal of Biological Chemistry*, 269(18):13107–13114.

[352] Pamfil, A. R., Howison, S. D., Lambiotte, R., and Porter, M. A. (2018). Relating modularity maximization and stochastic block models in multilayer networks. *arXiv preprint arXiv:1804.01964*.

[353] Papadopoulos, L., Puckett, J. G., Daniels, K. E., and Bassett, D. S. (2016). Evolution of network architecture in a granular material under compression. *Physical Review E*, 94(3):032908.

[354] Park, J., Back, J. H., Hahm, S.-H., Shim, H., Park, M. J., Ko, S.-I., Han, Y. S., et al. (2007). Bacterial beta-lactamase fragment complementation strategy can be used as a method for identifying interacting protein pairs. *Journal of Microbiology and Biotechnology*, 17(10):1607.

[355] Park, K. and Kim, D. (2009). Localized network centrality and essentiality in the yeast–protein interaction network. *Proteomics*, 9(22):5143–5154.

[356] Patokallio, J. (2012). OpenFlights: Airport, airline and route data.

[357] Pattison, P. (2012). Algebraic models for social networks. In *Computational Complexity*, pages 2925–2939. Springer.

[358] Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., and Fodor, S. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 91(11):5022–5026.

[359] Peel, L., Larremore, D. B., and Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548.

[360] Peisker, K., Chiabudini, M., and Rospert, S. (2010). The ribosome-bound HSP70 homolog SSBof *Saccharomyces cerevisiae*. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1803(6):662–672.

[361] Peixoto, T. P. (2014). Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1):012804.

[362] Peixoto, T. P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807.

[363] Pereira-Leal, J. B., Enright, A. J., and Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *PROTEINS: Structure, Function, and Bioinformatics*, 54(1):49–57.

[364] Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein-protein interactions: structural, functional, and network properties. *Structure*, 18(10):1233–1243.

[365] Pinkert, S., Schultz, J., and Reichardt, J. (2010). Protein interaction networks—more than mere modules. *PLoS Computational Biology*, 6(1):e1000659.

[366] Ponomarenko, E. A., Poverennaya, E. V., Ilgisonis, E. V., Pyatnitskiy, M. A., Kopylov, A. T., Zgoda, V. G., Lisitsa, A. V., and Archakov, A. I. (2016). The size of the human proteome: The width and depth. *International Journal of Analytical Chemistry*, 2016.

[367] Pontén, F., Jirström, K., and Uhlen, M. (2008). The human protein atlas—A tool for pathology. *The Journal of Pathology*, 216(4):387–393.

[368] Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M. I., Jiang, S., McCallum, A., Kirov, S., and Wasserman, W. W. (2008). The PAZAR database of gene regulatory information coupled to the orca toolkit for the study of regulatory sequences. *Nucleic Acids Research*, 37(suppl_1):D54–D60.

[369] Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 1164–1166.

[370] Proost, P., Wuyts, A., and Van Damme, J. (1996). The role of chemokines in inflammation. *International Journal of Clinical and Laboratory Research*, 26(4):211–223.

[371] Pruyne, D. and Bretscher, A. (2000). Polarization of cell growth in yeast. *Journal of Cell Science*, 113(4):571–585.

[372] Pržulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515.

[373] Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229.

[374] Rachlin, J., Cohen, D. D., Cantor, C., and Kasif, S. (2006). Biological context networks: a mosaic view of the interactome. *Molecular Systems Biology*, 2(1):66.

[375] Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919.

[376] Rao, V. S., Srinivas, K., Sujini, G., and Kumar, G. (2014). Protein–protein interaction detection: methods and analysis. *International Journal of Proteomics*, 2014.

[377] Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405.

[378] Reeve, E. C. (2014). *Encyclopedia of Genetics*. Routledge.

[379] Regard, J. B., Sato, I. T., and Coughlin, S. R. (2008). Anatomical profiling of g protein-coupled receptor expression. *Cell*, 135(3):561–571.

[380] Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1):016110.

[381] Reid, J. E. and Wernisch, L. (2016). Pseudotime estimation: Deconfounding single cell time series. *Bioinformatics*, 32(19):2973–2980.

[382] Rhee, S. Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–515.

[383] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(10):1030.

[384] Roberts, P. M. (2006). Mining literature for systems biology. *Briefings in Bioinformatics*, 7(4):399–406.

[385] Rodriguez-Caso, C., Medina, M. A., and Solé, R. V. (2005). Topology, tinkering and evolution of the human transcription factor network. *The FEBS Journal (Federation of European Biochemical Societies)*, 272(24):6423–6434.

[386] Rombach, M. P., Porter, M. A., Fowler, J. H., and Mucha, P. J. (2014). Core-periphery structure in networks. *SIAM Journal on Applied mathematics*, 74(1):167–190.

[387] Rombach, P., Porter, M. A., Fowler, J. H., and Mucha, P. J. (2017). Core-periphery structure in networks (revisited). *SIAM Review*, 59(3):619–646.

[388] Rosker, M. J., Rose, T. S., and Zewail, A. H. (1988). Femtosecond real-time dynamics of photofragment-trapping resonances on dissociative potential energy surfaces. *Chemical Physics Letters*, 146(3-4):175–179.

[389] Ross, W. D. (1948). *Aristotle's Metaphysics. A revised text with introduction and commentary*. Clarendon Press.

[390] Rossetti, G. and Cazabet, R. (2018). Community discovery in dynamic networks: A survey. *ACM Computing Surveys*, 51(2):35:1–35:37.

[391] Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the Untied States of America*, 105(4):1118–1123.

[392] Rosvall, M. and Bergstrom, C. T. (2010). Mapping change in large networks. *PLOS ONE*, 5(1):e8694.

[393] Sacks, H. S. and Fain, J. N. (2007). Human epicardial adipose tissue: A review. *American Heart Journal*, 153(6):907–917.

[394] Salwinski, L., Licata, L., Winter, A., Thorneycroft, D., Khadake, J., Ceol, A., Aryamontri, A. C., Oughtred, R., Livstone, M., Boucher, L., et al. (2009). Recurated protein interaction datasets. *Nature Methods*, 6(12):860.

[395] Sam, L., Liu, Y., Li, J., Friedman, C., and Lussier, Y. A. (2007). Discovery of protein interaction networks shared by diseases. In *Biocomputing 2007*, pages 76–87. World Scientific.

[396] Sarzynska, M., Leicht, E. A., Chowell, G., and Porter, M. A. (2015). Null models for community detection in spatially embedded, temporal networks. *Journal of Complex Networks*, 4(3):363–406.

[397] Saunders, R. and Deane, C. M. (2010). Protein structure prediction begins well but ends badly. *Proteins: Structure, Function, and Bioinformatics*, 78(5):1282–1290.

[398] Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218.

[399] Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLOS ONE*, 7(2):e31826.

[400] Schaefer, M. H., Lopes, T. J., Mah, N., Shoemaker, J. E., Matsuoka, Y., Fontaine, J.-F., Louis-Jeune, C., Eisfeld, A. J., Neumann, G., Perez-Iratxeta, C., et al. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Computational Biology*, 9(1):e1002860.

[401] Schlüter, G., Kremling, H., and Engel, W. (1992). The gene for human transition protein 2: Nucleotide sequence, assignment to the protamine gene cluster, and evidence for its low expression. *Genomics*, 14(2):377–383.

[402] Scholtes, I. (2017). When is a network a network?: Multi-order graphical model selection in pathways and temporal networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1037–1046. ACM.

[403] Schrodinger, E. (1967). *What is Life?: The Physical Aspect of the Living Cell and Mind and Matter; Mind and Matter*. Cambridge University Press.

[404] Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays—a technology review. *Nature Cell Biology*, 3(8):E190.

[405] Schutt, C. E., Myslik, J. C., Rozycki, M. D., Goonesekere, N. C., and Lindberg, U. (1993). The structure of crystalline profilin–β-actin. *Nature*, 365(6449):810.

[406] Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337.

[407] Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology*, 18(12):1257.

[408] Sealfon, R. S. G., Hibbs, M. A., Huttenhower, C., Myers, C. L., and Troyanskaya, O. G. (2006). Golem: An interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics*, 7(1):443.

[409] Sevimoglu, T. and Arga, K. Y. (2014). The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, 11(18):22–27.

[410] Shamir, M., Bar-On, Y., Phillips, R., and Milo, R. (2016). Snapshot: Timescales in cell biology. *Cell*, 164(6):1302–1302.

[411] Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427.

[412] Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, 3(1):88.

[413] Shaywitz, A. J. and Greenberg, M. E. (1999). Creb: A stimulus-induced transcription factor activated by a diverse array of extracellular signals. *Annual Review of Biochemistry*, 68(1):821–861.

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

[415] Shih, Y.-K. and Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping markov clustering. *Bioinformatics*, 28(18):i473–i479.

[416] Shinde, P. and Jalan, S. (2015). A multilayer protein-protein interaction network analysis of different life stages in *Caenorhabditis elegans*. *Europhysics Letters*, 112(5):58001.

[417] Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein–protein interactions. part i. experimental techniques and databases. *PLoS Computational Biology*, 3(3):e42.

[418] Singh, R., Xu, J., and Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):12763–12768.

[419] Slabinski, L., Jaroszewski, L., Rodrigues, A. P., Rychlewski, L., Wilson, I. A., Lesley, S. A., and Godzik, A. (2007). The challenge of protein structure determination—lessons from structural genomics. *Protein Science*, 16(11):2472–2482.

[420] Smyth, G. K., Ritchie, M., Thorne, N., Wettenhall, J., Shi, W., and Hu, Y. (2002). limma: Linear models for microarray and RNA-seq data user's guide.

[421] Söderberg, O., Gullberg, M., Jarvius, M., Ridderstråle, K., Leuchowius, K.-J., Jarvius, J., Wester, K., Hydbring, P., Bahram, F., Larsson, L.-G., et al. (2006). Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nature Methods*, 3(12):995.

[422] Solé, R. and Goodwin, B. (2000). How complexity pervades biology. *New York: Basic*.

[423] Sole-Ribalta, A., De Domenico, M., Kouvaris, N. E., Diaz-Guilera, A., Gomez, S., and Arenas, A. (2013). Spectral properties of the laplacian of multiplex networks. *Physical Review E*, 88(3):032807.

[424] Song, C., Havlin, S., and Makse, H. A. (2005). Self-similarity of complex networks. *Nature*, 433(7024):392.

[425] Sorensen, D. C. (1997). Implicitly restarted arnoldi/lanczos methods for large scale eigenvalue calculations. In *Parallel Numerical Algorithms*, pages 119–165. Springer.

[426] Southern, E. (2006). Southern blotting. *Nature Protocols*, 1(2):518.

[427] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297.

[428] Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12123–12128.

[429] Spivakov, M. (2014). Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*, 36(8):798–806.

[430] Sporns, O. (2010). *Networks of the Brain*. MIT press.

[431] Stanley, N., Shai, S., Taylor, D., and Mucha, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE Transactions on Network Science and Engineering*, 3(2):95–105.

[432] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539.

[433] Steffen, W., Rockström, J., Richardson, K., Lenton, T. M., Folke, C., Liverman, D., Summerhayes, C. P., Barnosky, A. D., Cornell, S. E., Crucifix, M., et al. (2018). Trajectories of the earth system in the anthropocene. *Proceedings of the National Academy of Sciences of the United States of America*, page 201810141.

[434] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., et al. (2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957–968.

[435] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

[436] Stevens, C. F. and Zador, A. M. (1996). When is an integrate-and-fire neuron like a Poisson neuron? In *Advances in Neural Information Processing Systems*, pages 103–109.

[437] Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W., and Hardison, R. (1999). Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Research*, 27(19):3899–3910.

[438] Stolz, B. J., Emerson, T., Nahkuri, S., Porter, M. A., and Harrington, H. A. (2018). Topological data analysis of task-based fMRI data from experiments on schizophrenia. *arXiv preprint arXiv:1801.10144*.

[439] Stolz, B. J., Harrington, H. A., and Porter, M. A. (2017). Persistent homology of time-dependent functional networks constructed from coupled time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4):047410.

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

[441] Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

[442] Stumpf, M., Balding, D. J., and Girolami, M. (2011). *Handbook of Statistical Systems Biology*. John Wiley & Sons.

[443] Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6959–6964.

[444] Su, M., Ling, Y., Yu, J., Wu, J., and Xiao, J. (2013). Small proteins: Untapped area of potential biological importance. *Frontiers in Genetics*, 4:286.

[445] Surana, U., Robitsch, H., Price, C., Schuster, T., Fitch, I., Futcher, A. B., and Nasmyth, K. (1991). The role of CDC28 and cyclins during mitosis in the budding yeast *Saccharomyces cerevisiae*. *Cell*, 65(1):145–161.

[446] Sutphin, G. L. and Kaeberlein, M. (2009). Measuring *Caenorhabditis elegans* life span on solid media. *Journal of Visualized Experiments*, 27(1152).

[447] Svozil, D., Kvasnicka, V., and Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems*, 39(1):43–62.

[448] Szell, M., Lambiotte, R., and Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13636–13641.

[449] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2014). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452.

[450] Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368.

[451] Taylor, D., Myers, S. A., Clauset, A., Porter, M. A., and Mucha, P. J. (2017). Eigenvector-based centrality measures for temporal networks. *Multiscale Modeling & Simulation*, 15(1):537–574.

[452] Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27(2):199–204.

[453] Telesford, Q. K., Simpson, S. L., Burdette, J. H., Hayasaka, S., and Laurienti, P. J. (2011). The brain as a complex system: using network science as a tool for understanding the brain. *Brain Connectivity*, 1(4):295–308.

[454] Titz, B., Schlesner, M., and Uetz, P. (2004). What do we learn from high-throughput protein interaction data? *Expert Review of Proteomics*, 1(1):111–121.

[455] Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368.

[456] Traag, V. A. and Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115.

[457] Traud, A. L., Mucha, P. J., and Porter, M. A. (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180.

[458] Truong, K. and Ikura, M. (2001). The use of fret imaging microscopy to detect protein–protein interactions and protein conformational changes in vivo. *Current Opinion in Structural Biology*, 11(5):573–578.

[459] Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, 310(5751):1152–1158.

[460] Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B*, 237(641):37–72.

[461] Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419.

[462] Vallès-Català, T., Peixoto, T. P., Sales-Pardo, M., and Guimerà, R. (2018). Consistencies and inconsistencies between model selection and link prediction in networks. *Physical Review E*, 97(6):062316.

[463] Van Dijken, J., Bauer, J., Brambilla, L., Duboc, P., Francois, J., Gancedo, C., Giuseppin, M., Heijnen, J., Hoare, M., Lange, H., et al. (2000). An interlaboratory comparison of physiological and genetic properties of four saccharomyces cerevisiae strains. *Enzyme and Microbial Technology*, 26(9-10):706–714.

[464] Van Regenmortel, M. H. (2004). Reductionism and complexity in molecular biology: Scientists now have the tools to unravel biological complexity and overcome the limitations of reductionism. *EMBO Reports (European Molecular Biology Organization)*, 5(11):1016–1020.

[465] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics*, 10(4):252.

[466] Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227.

[467] Vogelstein, B., Lane, D., and Levine, A. J. (2000). Surfing the p53 network. *Nature*, 408(6810):307.

[468] Vogiatzis, C. and Camur, M. C. (2017). Identification of essential proteins using induced stars in protein-protein interaction networks. *arXiv preprint arXiv:1708.00574*.

[469] Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically.

[470] Waddington, C. (1961). Molecular biology or ultrastructural biology? *Nature*, 190(4771):184–184.

[471] Wake, M. S. and Watson, C. J. (2015). STAT3 the oncogene—Still eluding therapy? *The FEBS Journal (Federation of European Biochemical Societies)*, 282(14):2600–2611.

[472] Wang, X., Watt, P. M., Louis, E. J., Borts, R. H., and Hickson, I. D. (1996). Pat1: a topoisomerase ii-associated protein required for faithful chromosome transmission in saccharomyces cerevisiae. *Nucleic Acids Research*, 24(23):4791–4797.

[473] Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology*, 30(2):159.

[474] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57.

[475] Washburn, M. P. (2016). There is no human interactome. *Genome Biology*, 17(1):48.

[476] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

[477] Weirich, C. S., Erzberger, J. P., and Barral, Y. (2008). The septin family of GTPases: Architecture and dynamics. *Nature Reviews Molecular Cell biology*, 9(6):478.

[478] Wickner, W. and Schekman, R. (2005). Protein translocation across biological membranes. *Science*, 310(5753):1452–1456.

[479] Wiedermann, M., Donges, J. F., Heitzig, J., and Kurths, J. (2013). Node-weighted interacting network measures improve the representation of real-world complex systems. *Europhysics Letters*, 102(2):28007.

[480] Wilkins, M. R. and Kummerfeld, S. K. (2008). Sticking together? falling apart? exploring the dynamics of the interactome. *Trends in Biochemical Sciences*, 33(5):195–200.

[481] Wilkinson, A. C., Nakauchi, H., and Göttgens, B. (2017). Mammalian transcription factor networks: Recent advances in interrogating biological complexity. *Cell Systems*, 5(4):319–331.

[482] Williams, E. J. and Bowles, D. J. (2004). Coexpression of neighboring genes in the genome of arabidopsis thaliana. *Genome Research*, 14(6):1060–1067.

[483] Wilson, A. C., Ochman, H., and Prager, E. M. (1987). Molecular time scale for evolution. *Trends in Genetics*, 3:241–247.

[484] Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: An integrated system for gene expression regulation. *Nucleic Acids Research*, 28(1):316–319.

[485] Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., et al. (1999). Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906.

[486] Wray, J. (2017). A complex system. *European Biopharmaceutical Review*, 28(32):28–32.

[487] Wu, Y., Li, Q., and Chen, X.-Z. (2007). Detecting protein–protein interactions by far Western blotting. *Nature Protocols*, 2(12):3278.

[488] Wuchty, S. (2014). Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19):7156–7160.

[489] Wyers, F., Minet, M., Dufour, M. E., Lacroute, F., et al. (2000). Deletion of the PAT1 gene affects translation initiation and suppresses a PAB1 gene deletion in yeast. *Molecular and Cellular Biology*, 20(10):3538–3549.

[490] Xiao, Q., Wang, J., Peng, X., and Wu, F.-X. (2013). Detecting protein complexes from active protein interaction networks constructed with dynamic gene expression profiles. *Proteome Science*, 11(1):S20.

[491] Xing, Y., Li, C., Li, A., Sridurongrit, S., Tiozzo, C., Bellusci, S., Borok, Z., Kaartinen, V., and Minoo, P. (2010). Signaling via Alk5 controls the ontogeny of lung clara cells. *Development*, 137(5):825–833.

[492] Xu, G. and Shi, Y. (2007). Apoptosis signaling pathways and lymphocyte homeostasis. *Cell Research*, 17(9):759.

[493] Yan, C., Wu, F., Jernigan, R. L., Dobbs, D., and Honavar, V. (2008). Characterization of protein–protein interfaces. *The Protein Journal*, 27(1):59–70.

[494] Yang, J., Mani, S. A., Donaher, J. L., Ramaswamy, S., Itzykson, R. A., Come, C., Savagner, P., Gitelman, I., Richardson, A., and Weinberg, R. A. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell*, 117(7):927–939.

[495] Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):5934–5939.

[496] Yeger-Lotem, E. and Sharan, R. (2015). Human protein interaction networks across tissues and diseases. *Frontiers in Genetics*, 6:257.

[497] Yook, S.-H., Oltvai, Z. N., and Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942.

[498] Yosef, N. and Regev, A. (2011). Impulse control: Temporal dynamics in gene transcription. *Cell*, 144(6):886–896.

[499] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4):e59.

[500] Zdeborová, L. and Krzakala, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552.

[501] Zemp, D., Wiedermann, M., Kurths, J., Rammig, A., and Donges, J. F. (2014). Node-weighted measures for complex networks with directed and weighted edges for studying continental moisture recycling. *Europhysics Letters*, 107(5):58005.

[502] Zhang, K., Gao, R., Zhang, H., Cai, X., Shen, C., Wu, C., Zhao, S., and Yu, L. (2005). Molecular cloning and characterization of three novel lysozyme-like genes, predominantly expressed in the male reproductive system of humans, belonging to the c-type lysozyme/alpha-lactalbumin family. *Biology of Reproduction*, 73(5):1064–1071.

[503] Zhu, J., Chen, G., Zhu, S., Li, S., Wen, Z., Li, B., Zheng, Y., and Shi, L. (2016). Identification of tissue-specific protein-coding and noncoding transcripts across 14 human tissues using rna-seq. *Scientific Reports*, 6:28400.

[504] Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. *Trends in Genetics*, 24(10):481–484.

[505] Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466.

[506] Zitnik, M. and Leskovec, J. (2017). Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198.

[507] Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91.