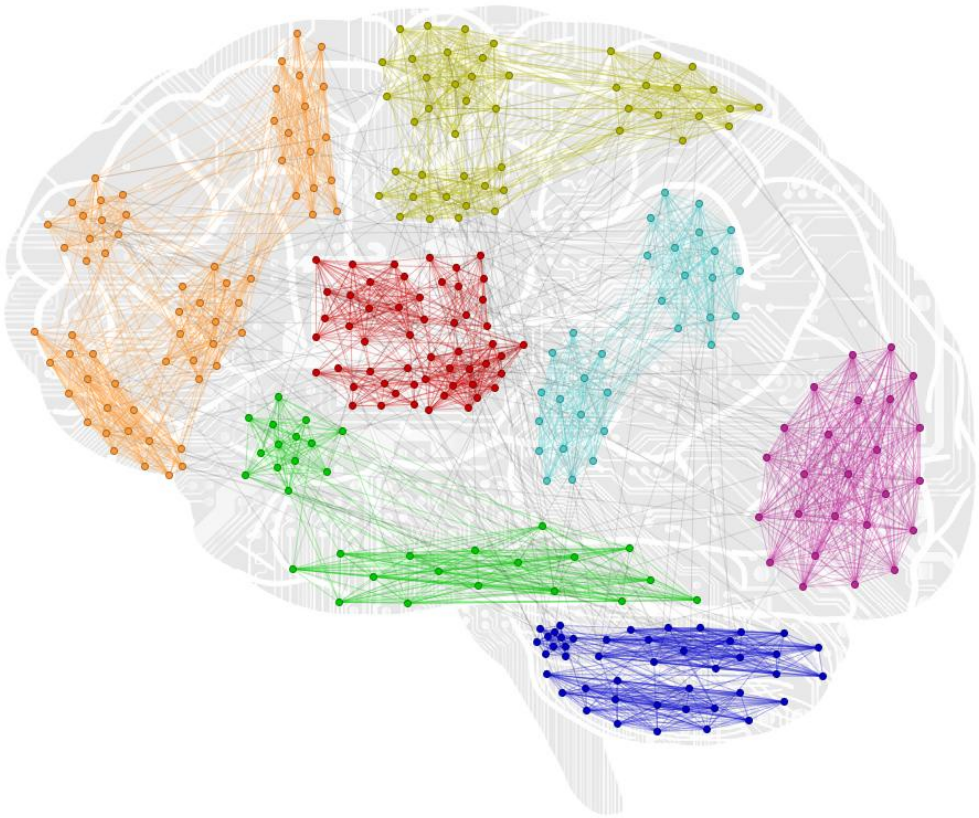
CHALMERS
UNIVERSITY OF TECHNOLOGY



Detection of Multi-Level Hierarchies in Multi-View Cancer Networks

with Applications to Glioblastoma Multiforme

Master's thesis in Complex Adaptive Systems

PHILIPP ARNDT

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018

Master's thesis 2018

Detection of Multi-Level Hierarchies in Multi-View Cancer Networks

with Applications to Glioblastoma Multiforme

PHILIPP ARNDT

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2018

Detection of Multi-Level Hierarchies in Multi-View Cancer Networks
with Applications to Glioblastoma Multiforme
PHILIPP ARNDT

Supervisor: Rebecka Jörnsten, Department of Mathematical Sciences
Examiner: Rebecka Jörnsten, Department of Mathematical Sciences

Master's Thesis 2018
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Artistic visualization of an unweighted network displaying multi-level hierarchical community structure, vaguely oriented on regions in the human brain that are responsible for different tasks.

Typeset in LaTeX
Gothenburg, Sweden 2018

Detection of Multi-Level Hierarchies in Multi-View Cancer Networks
with Applications to Glioblastoma Multiforme
PHILIPP ARNDT
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Biological and other complex networks are generally believed to be hierarchically organized. High-throughput molecular sequencing technologies now make it possible to reconstruct large-scale biological networks based on different types of data at the genomic, transcriptomic, epigenomic, proteomic and metabolomic levels. A crucial task is to effectively integrate and analyze these different "views" of the data to gain a systems-level understanding of biological components, processes, and their functions. We here review recent advances in molecular data integration, multi-view learning and multi-level hierarchical community detection in big data networks. We then propose a novel method that integrates multiple views of similarities between data points into a single network via a diffusion process, and detects communities on multiple levels of hierarchy. On simulated data, we show that our approach is indeed able to capitalize on both common and complementary information contained in multiple views for the identification of an underlying multi-level hierarchical community structure. We apply our method to gene expression, copy number aberration and DNA methylation data from Glioblastoma Multiforme tumor samples to identify groups of genes that are highly co-regulated during disease progression. We verify that the resulting community structure is indeed representative of biological function by identifying various communities in which genes associated to known biological processes are highly overrepresented on statistically significant levels. We visualize the resulting network based on its multi-level hierarchical structure to allow for easy, intuitive exploration of the data.

Keywords: systems biology, cancer, data integration, multi-view data, network fusion, community detection, multi-level hierarchy, overrepresentation analysis, complex networks

Acknowledgements

I would like to thank my advisor Rebecka Jörnsten for guiding me through the worlds of Systems Biology, Biostatistics and Big Data throughout the last months. I am grateful to my best friend Eleftherios Filippakis for countless sauna sessions, discussing relevant topics and ideas in science and technology over some of the finest beers this planet has to offer. This project would also have been impossible without my beloved coffee maker, relentlessly working day and night shifts to fuel my inspiration.

I would like to thank my reviewers Jacob Söderström and Eleftherios Filippakis for many helpful comments on the manuscript.

Philipp Arndt, Gothenburg, May 2018

Contents

Abstract v

Acknowledgements vi

List of Figures x

List of Tables xii

List of Acronyms xiii

1 Introduction 1

1.1 Background . . . 1

1.1.1 Systems Biology . . . 2

1.1.1.1 Holism Versus Reductionism . . . 2

1.1.1.2 A Renewed Interest . . . 2

1.1.1.3 Prospects for Precision Medicine . . . 3

1.1.2 Molecular Cancer Research . . . 3

1.1.2.1 Cancer: A Deadly Disease . . . 3

1.1.2.2 Cancer Data Initiatives and Profiling Approaches . . . 3

1.1.2.3 Finding Meaning in the Data . . . 4

1.1.3 Big Data . . . 5

1.1.3.1 Volume . . . 5

1.1.3.2 Velocity . . . 6

1.1.3.3 Variety . . . 6

1.1.4 Machine Learning and Cluster Analysis . . . 6

1.1.4.1 Supervised Versus Unsupervised Learning . . . 7

1.1.4.2 Clustering Methods . . . 8

1.1.4.3 A Difficult, Subjective Task . . . 10

1.1.5 Network Modeling . . . 12

1.1.5.1 Terms and Definitions . . . 12

1.1.5.2 Community Detection . . . 14

1.1.5.3 Molecular Biological Networks . . . 16

1.1.6 Multi-View Data . . . 17

1.1.6.1 The Promise and the Challenge . . . 17

1.1.6.2 Multiple Views of Molecular Profiling Data . . . 17

1.1.6.3 Different Types of View Integration in Cancer Data . . . 19

1.2 A Desirable Framework for Integrative Exploratory Molecular Data Analysis . . . 20

2 Related Literature **22**
2.1 Unsupervised Multi-Omics Learning in Cancer Research . . . 22
2.1.1 Matrix Factorization Approaches . . . 23
2.1.2 Correlation and Covariance-Based Approaches . . . 24
2.1.3 Bayesian Approaches . . . 26
2.1.4 Network-Based Approaches . . . 27
2.1.5 Multiple Kernel and Multi-Step Approaches . . . 28
2.2 Multi-Level Hierarchical Community Detection in Large-Scale Complex Networks . . . 29
2.2.1 Modularity Optimization Approaches . . . 32
2.2.2 Methods Based on Information Theory . . . 33
2.2.3 Kernel Spectral Clustering Methods . . . 34
2.2.4 Techniques Based on Statistical Significance . . . 35

3 Theory **37**
3.1 Similarity Network Fusion . . . 37
3.1.1 Construction of Similarity Matrices . . . 37
3.1.2 Fusion of Similarities . . . 38
3.2 Multi-Level Hierarchical Kernel Spectral Clustering . . . 39
3.2.1 Predictive Kernel Spectral Clustering . . . 40
3.2.1.1 Representative Subset Selection . . . 40
3.2.1.2 Primal Formulation . . . 40
3.2.1.3 Dual Formulation . . . 41
3.2.1.4 Optimal Number of Clusters . . . 42
3.2.2 Muli-Level Hierarchy Detection . . . 44
3.2.2.1 Selection of Distance Thresholds . . . 44
3.2.2.2 Identification of Communities for the Whole Network 45
3.3 Order Statistics Local Optimization Method . . . 45
3.3.1 Statistical Significance of Communities . . . 45
3.3.1.1 Topological Relations . . . 46
3.3.1.2 Edge Weights . . . 47
3.3.1.3 The Combined Significance Score . . . 48
3.3.2 Single Community Analysis . . . 48
3.3.3 Network Analysis . . . 49
3.3.4 Hierarchical Structure . . . 50

4 Methods **52**
4.1 Simulation Study . . . 52
4.1.1 Fusion and Hierarchical Community Detection on Balanced Data with Symmetrical Hierarchical Structure . . . 53
4.1.2 Hierarchical Community Detection on Unbalanced Data with Asymmetrical Hierarchical Structure . . . 55
4.2 Data and Preprocessing . . . 59
4.3 View-Specific Similarity Estimation . . . 59

4.4 Network Fusion 60
4.5 Multi-Resolution Community Detection 62
4.6 Visualization of Multi-Resolution Networks 62
4.7 Gene Set Overrepresentation Analysis 64

5 Results **67**
5.1 Visual Exploration of the Hierarchical Community Structure 67
5.2 Network Statistics 71
5.3 Communities Related to Biological Function 72
5.4 A Potentially Important Community for Glioblastoma Multiforme . . 73

6 Discussion **77**
6.1 Future Work 77
6.2 Societal and Ethical Aspects 78
6.3 Conclusion 78

Bibliography **79**

List of Figures

1.1 Two examples of clustering based on distance in two-dimensional space. The left panel shows a relatively "easy" clustering task with clearly separated clusters, while the panel on the left illustrates a harder example where clusters seem to overlap. 9
1.2 An illustrative example which shows that clustering is not a well-defined task, but rather a subjective classification of data that depends on the research goal. 11
1.3 Visual 3-D representations of three examples of real world networks, plotted in Matlab. Left to right: a power grid, a network of email interactions, and a protein-protein interaction network. 12
1.4 Three equivalent ways of describing a network. From left to right: A visualization of nodes and edges, an edge list with weights, and an adjacency matrix. 13
1.5 An example of communities detected in an unweighted network. Here, node color indicates community membership. 14
1.6 An overview of different "omics" levels in systems biology, and the sequencing data types (or views) that they give rise to. 18

2.1 An illustration of how the precision matrix can often provide more useful information about the underlying system than the covariance matrix. Non-zero entries in the precision matrix of a chain harmonic oscillator represent direct interactions via a spring connection. 25
2.2 An example of communities detected in a large-scale social network using the Louvain algorithm, based on Wikipedia pages about famous public figures (nodes) and links between their pages (edges). The figure is reproduced from Biddulph [2012] under Creative Commons Attribution-ShareAlike 2.0 license (CC BY-SA 2.0). 31

4.1 The distributions from which entries in the different views were drawn, here illustrated for view 1. For easy visual distinction of the different areas within this matrix, we chose $\sigma_{\text{views}} = 0.1$ in this illustration. . . 54
4.2 Results of SNF (fused similarity), MHKSC and OSLOM on simulated multi-view data that exhibits a symmetrical multi-level hierarchical community structure with balanced cluster sizes. 56
4.3 Results of MHKSC and OSLOM on a simulated network exhibiting multi-level hierarchical community structure with unbalanced cluster sizes. 58

4.4 Comparison of the distributions of r_{MAD}-derived similarity values for each view and the simulated data from the first part of the simulation study. 61
4.5 An example visualization of the finest-level community partition of a part of our GBM network in Gephi, together with a further zoomed in illustration of a single community that makes it possible to identify nodes by their labels. 63

5.1 A visualization of the fused GBM gene-gene association network using the multi-level hierarchical community structure found by OSLOM. Nodes are colored according to their community membership on each level of the hierarchy. The bottom right panel presents a more detailed view of part the ground level, zoomed in on the area indicated by the gray rectangle in the upper left panel. 68
5.2 A part of the visualized fused GBM gene-gene association network with gene labels, colored according to the OSLOM community partition at the finest level of hierarchy. 69
5.3 A representation of the fused network in which each edge is colored according to the view of the data in which that particular edge is supported strongest, relative to the other edges that survived the thresholding process. 70
5.4 The weighted degree distribution of the fused GBM network, along with a power law fit to the data. 73
5.5 Visualization of a community that could be particularly relevant for a better understanding of GBM. 75

List of Tables

1.1 Various "omics" molecular profiling data levels and their approximate amount of variables [Gligorijević et al., 2016]. 7

4.1 An example 2×2 contingency table for the illustration of Fisher's exact test to determine p-values for the overrepresentation of a certain biological annotation $\mathcal{A}$ in a community of genes $\mathcal{C}$. 64

5.1 Some basic network statistics describing the multi-level hierarchical structure identified by OSLOM in our fused GBM network. 72

5.2 A (non-comprehensive) list of communities on the base hierarchy in which gene sets with certain biological annotations were overrepresented. 74

List of Acronyms

BCC Bayesian Consensus Clustering 27

CCA Canonical Correlation Analysis 24
CCLE Cancer Cell Line Encyclopedia 4
CDKN2A Cyclin-Dependent Kinase Inhibitor 2A 73, 75, 76
ChIP-seq Chromatin Immunoprecipitation Sequencing 18
CNA Copy Number Aberration 4, 18, 25–29, 59, 60, 71, 75, 76, 78
CONEXIC COpy Number and EXpression In Cancer 26
CPCA Consensus Principal Component Analysis 23

DBSCAN Density-Based Spatial Clustering of Applications with Noise ... 10
DMA Dirichlet-Multinomial Allocation 27
DNA Deoxyribonucleic acid 2, 6, 16–19, 59, 78, 79

ECOC Error COrrecting Codes 41
EGFR Epidermal Growth Factor Receptor 73
ELAVL2 Embryonic Lethal Abnormal Vision-Like protein 2 75
EM Expectation Maximization (algorithm) 10, 23

FDR False Discovery Rate 65, 66, 72, 73, 76
FURS Fast and Unique Representative Subset selection 34, 40

GBM Glioblastoma Multiforme . xi, xii, 3, 4, 8, 19, 20, 23, 24, 27, 29, 53, 59, 63, 67–69, 71–73, 75, 76, 79
GO Gene Ontology (database) 65
GSEA Gene Set Enrichment Analysis 26

HGCC Human Glioblastoma Cell Culture Resource 4, 19

ICGC International Cancer Genome Consortium 4
IFN Interferon 73
IFNA Interferon Alpha 75

JGL Joint Graphical Lasso 26
JIVE Joint and Individual Variance Explained 23, 24

KKT Karush-Kuhn-Tucker (optimality conditions) 42
KLHL9 Kelch-like protein 9 75
kNN k-Nearest Neighbors 28, 29, 37–39

kPCA kernel Principal Component Analysis 34, 39–41
KSC Kernel Spectral Clustering 34, 40, 42–44

LASSO Least Absolute Shrinkage and Selection Operator 16, 26
LPP Locality Preserving Projection 29
LS-SVM Least Squares Support Vector Machine 34, 39

MAD Median Absolute Deviation ..... xi, 60, 61, 78
MCIA Multiple or Co-Inertia Analysis 23
MDI Multiple Data set Integration 27
MHKSC Multi-Level Hierarchical Kernel Spectral Clustering . x, 30, 34, 35, 37, 39–41, 44, 45, 52, 53, 55–59
miRNA micro Ribonucleic Acid 4, 23–29
miRNA-seq micro Ribonucleic Acid Sequencing 18
MKL-DR Multiple Kernel Learning for Dimensionality Reduction 29
mRNA messenger Ribonucleic Acid 2, 16–18, 23, 26, 28
MTAP S-Methyl-5'-thioadenosin phosphorylase 75

NFIB Nuclear Factor 1 B-type 75
NMF Non-negative Matrix Factorization 23
NMI Normalized Mutual Information 55, 57

OPTICS Ordering Points To Identify the Clustering Structure 10
OSLOM Order Statistics Local Optimization Method . x–xii, 30, 35–37, 39, 45, 46, 48, 49, 52, 55–59, 61, 62, 64, 67–69, 71, 72, 76, 79

PANTHER Protein ANalysis THrough Evolutionary Relationships 65
PARADIGM PAthway Representation and Analysis by DIrect Reference on Graphical Models 27, 28
PCA Principal Component Analysis 23, 24
PLS Partial Least Squares 24

RBF Radial Basis Function 38, 54, 61
rMKL-LPP regularized Multiple Kernel Learning using Locality Preserving Projections 29
RNA Ribonucleic Acid 6, 7, 17
RNA-seq Ribonucleic Acid Sequencing 18

sCCA sparse Canonical Correlation Analysis 24
SEC61G Protein transport protein, subunit gamma 75
SH3GL2 Endophilin-A1 75
SICS Sparse Inverse Covariance Selection 25, 26
sMBPLS sparse Multi-Block Partial Least Squares 24
SNF Similarity Network Fusion . x, 28, 29, 37, 39, 52, 54, 56, 61, 62, 71, 78
SNMNMF Sparse Network-regularized Multiple Non-negative Matrix Factorization 23
SNP Single-Nucleotide Polymorphism 24
SVM Support Vector Machine 29, 34

TCGA The Cancer Genome Atlas 4, 19, 59
TF Transcription Factor 16
TYRP1 5,6-dihydroxyindole-2-carboxylic acid oxidase 75, 76

UPGMA Unweighted Pair Group Method with Arithmetic Mean 9

WGCNA Weighted Gene Co-expression Network Analysis 16

1

Introduction

"The computer is incredibly fast, accurate, and stupid. Man is unbelievably slow, inaccurate, and brilliant. The marriage of the two is a challenge and opportunity beyond imagination."

– Stuart G. Walesh, 1989

This chapter briefly introduces the reader to the topic of this thesis, introduces general concepts and ideas, and reviews recent advances and emerging challenges.

1.1 Background

Biological cells contain a great variety of molecular structures, forming systems that can be investigated as complex dynamic networks [Barabasi and Oltvai, 2004]. Novel innovations in biotechnology, along with continually improving cost-efficiency of high-throughput sequencing methods, have made available a plenitude of molecular data for researching such systems [Metzker, 2010, Pe'er and Hacohen, 2011]. This recent flood of data in biology, however, has led to stark disparities between our technological ability to generate vast amounts of biomedical data and our capacity to properly analyze and understand it [Sboner et al., 2011]. Thus, the age of "big data" in biology has come with an ever increasing demand for efficient statistical tools and computational methods to increase interpretability of data for clinical research [Marx, 2013].

This section aims to introduce readers that are unfamiliar with the fields of systems biology, cancer research and biostatistics to general ideas and concepts necessary to understand current challenges in biological big data statistics and their application to cancer data. Section 1.1.1 covers the foundations of systems biology. Section 1.1.2 discusses molecular profiling and data analysis in cancer research. Section 1.1.3 concisely explains the meaning of "Big Data" and its role in biological research. Section 1.1.4 provides the reader with a general introduction to clustering methods. Section 1.1.5 then extends the introduced concepts to network models. Section 1.1.6 presents the concept of multi-view data sets, and discusses how multiple views

can improve our understanding of underlying data-generating processes. Readers familiar with any of the above topics can likely skip the respective sections.

1.1.1 Systems Biology

Systems biology is a scientific disciple that aims to quantitatively model complex biological systems. The following aims to cover important paradigms and issues in the field.

1.1.1.1 Holism Versus Reductionism

Systems biology follows a *holistic* approach, based on the idea that the structure of all components in biological systems and the dynamics of all their interactions need to be investigated to be able to explain the system's emergent function or behavior [Loscalzo and Barabasi, 2011]. This holistic approach is usually contrasted with *traditional reductionism*, which deals with complex systems by dividing them into smaller parts that are each manageable to be analyzed on their own [Mazzocchi, 2012, Noble, 2008]. Hence, a reductionist biologist would aim to understand the human body in the way that most of us will be familiar with from high school; a collection of organs that all have a certain role, which are themselves made up of certain types of tissue, which contains certain types of cells, which contain certain organelles, which themselves are characterized by the molecular processes happening inside them. The systems biologist, however, would try to answer a different question: If we identify the structure of all molecules inside a human body and measure the rate of sufficiently many respective molecular processes, can we reconstruct how the whole human body functions?

1.1.1.2 A Renewed Interest

The goal of a fundamental systems-level understanding of biological systems has been a recurrent theme in the literature since the early 20th century [Bertalanffy, 1931, Wiener, 1949]. Until recently, reaching this ambitious objective for the human body was considered unlikely or merely hypothetical by many. However, since the launch of the Human Genome Project [Venter et al., 2001] and with the rise of high-throughput sequencing methods, the systems approach to biology experienced a dramatic increase in scientific attention. Sequencing a human's whole genome can give us useful information about the individual organism by providing us with the set of all genes encoded in the DNA, the *genotype*. This, however, is not enough to determine all the resulting observable characteristics of the individual, the so-called *phenotype*. In order to be able to truly understand the genotype-phenotype relationship in organisms, systems biologists yet have to reconstruct many complex interactions and processes including genes, transcribed mRNA, proteins, metabolites, and environmental conditions among many components [Kitano, 2002]. All of these are active areas of research [Legrain et al., 2011, Wang et al., 2009, Wishart et al., 2007, Turnbaugh et al., 2007].

1.1.1.3 Prospects for Precision Medicine

The idea that measurable properties of biological tissue or body fluids can be indicative of underlying mechanisms inside the body – including disease – can be traced back to at least ancient Greece. By the middle ages so-called "pee charts", which related the color, smell, and even taste of one's pee to certain medical conditions, were widespread [Nicholson and Lindon, 2008]. The resulting concept of personalizing treatment based on chemical measurements is the same as what precision medicine aims to do today [Collins and Varmus, 2015]. With the advent of highly efficient molecular profiling of patients' tissue or body fluids, precision medicine has recently become a much more exact science [Mirnezami et al., 2012]. Systems biology stands a great chance at significantly improving its capabilities by uncovering increasingly many biological mechanisms underlying certain changes in measurements that are related to a disease.

1.1.2 Molecular Cancer Research

Cancer is a widespread and often fatal disease, affecting many people around the globe. The availability of modern sequencing technologies and the resulting biomedical big data have transformed the field of cancer research in the recent past. This section is meant as an introduction to cancer research in general, and to how molecular sequencing data is utilized to advance our knowledge on how cancer works and on how it may be treated more effectively.

1.1.2.1 Cancer: A Deadly Disease

While most of this thesis is concerned with mathematics, statistics and programming, it is important to keep in mind that the underlying data comes from real people suffering from cancer, one of the deadliest diseases to human kind, affecting the lives of many people around the globe. With over 8 million cases annually – and trend increasing – cancer is the second leading cause of death worldwide, with nearly 1 in 6 deaths being due to some kind of cancer [Ferlay et al., 2015]. This implies a societal loss of about 196.3 million disability-adjusted life-years annually [Fitzmaurice et al., 2015]. A better understanding of how cancer is caused, how different types of cancer are related, and how cancer can be effectively treated in individuals has the potential to save and improve millions of lives around the globe in the future. This report is mainly concerned with the analysis of Glioblastoma Multiforme (GBM), which is the most common and most lethal type of brain cancer [Parsons et al., 2008].

1.1.2.2 Cancer Data Initiatives and Profiling Approaches

Cancer research is a large scientific field, and consequently there is an exponential growth of related data originating from journal publications, genome-wide association studies, protein-protein interaction surveys, epigenomics, immunomics, and

many more. Due to the large amount and heterogeneity of available sources, the storage, acquisition and analysis of relevant data poses a great challenge to biostatisticians [Pavlopoulou et al., 2015]. Out of the need for an orchestrated effort to making cancer data easily accessible, some large-scale collaborative projects now aim to provide well-annotated and structured databases. The Cancer Genome Atlas (TCGA, Weinstein et al. [2013]) and the International Cancer Genome Consortium (ICGC, Hudson et al. [2010]) are the most prominent examples of large databases containing molecular data on sequenced cancer tissue. The TCGA database contains reliable data on many GBM patients, and is therefore used in this study.

The fact that tissue samples from tumors are often not pure, but contain a significant amount of non-tumor cells makes the analysis of molecular data in cancer non-trivial. A common approach to obtain pure samples of cancer cells is to use cell lines. This means that a sample from a tumor is grown in vitro (i.e. petri dishes), with repeated subsampling by a small fraction of cells, so that the sample becomes increasingly enriched in proliferating tumor cells. A drawback of this method is that artificial in vitro growth of a tumor may change the biological function and molecular make-up within the cancer cells [Kaur and Dufour, 2012]. The Cancer Cell Line Encyclopedia (CCLE, Barretina et al. [2012]) is one of the most prominent databases providing human cell line data for a large variety of different cancer types. The Human Glioblastoma Cell Culture Resource (HGCC, Xie et al. [2015]) provides cell line data specifically for GBM patients. A relatively new approach that has the potential to overcome the sample purity problem is single-cell sequencing, which can examine the molecular make-up of individual tumor cells and thus lead to a better understanding of the function of single cells in their environment [Eberwine et al., 2014, Navin et al., 2011].

1.1.2.3 Finding Meaning in the Data

The overall objective of analyzing molecular cancer data can be roughly divided in two distinct goals. The first goal is to categorize patients by finding groups whose molecular profiles are clearly distinct, which is known as cancer subtype discovery [Dai et al., 2015]. The associated clinical data can then be compared across the discovered groups of patients in terms of age, gender, habits, survival rates and response to treatment with a range of drugs, just to mention a few. The classification of patients based on molecular data, together with the knowledge of differences in related patient data and clinical outcomes, can then be used to identify risk factors, to improve diagnoses, to personalize treatments, and to predict survival of new patients [Yang et al., 2007, Iqbal et al., 2010, Van't Veer and Bernards, 2008, Rosenwald et al., 2002]. The second goal of molecular data analysis in cancer research is to gain a better understanding of how genetic lesions drive the phenotype of tumor cells and contribute to disease progression [Kling et al., 2015]. This means that characteristics of the cancer at hand (e.g. a certain mutation), need to be identified and then related to changes of molecular profiles on different levels, such as gene expression, copy number aberrations (CNA), methylation, microRNA (miRNA) expression, protein expression or changes in metabolism. If closely connected variables

can be identified in the data analysis, they can be cross-referenced with manually curated databases about biological function. Finding such highly associated variables in molecular profiling data of a certain cancer can help to detect new biomarkers for disease diagnosis, to identify new drug targets, and to better understand the dynamics of a certain type of disease in general [Kussmann et al., 2006, Yang et al., 2012]. Integrative statistical analysis on a wide range of cancer patient data – as proposed in this project – stands a great chance at improving our understanding of most of the relationships mentioned above, and is thus an indispensable tool in the fight against cancer.

1.1.3 Big Data

In the recent past, the digital revolution has brought forward a wide range of groundbraking technologies, which make it easier than ever to produce and store vast amounts of information [Freeman and Louçã, 2001]. With this has come the age of "big data", which has great potential to transform society at large, and is arguably already doing so [Walker, 2014]. While it is clear from its name that big data refers to large volume of data being available about nearly every aspect of our lives, there is more to it – it is now widely accepted that big data is characterized by "three V's": volume, velocity and variety [Gartner, 2001].

1.1.3.1 Volume

"Volume" refers to the fact that today there is large amounts of data available about nearly every aspect of our lives, ready for analysis. For instance, the amount of total Facebook posts that could be searched through the company's Graph Search was 2.5 trillion as of 2016 [Constine, 2016]. A similar trend can be seen in the data generated by sequencing human genomes. Since the sequencing of the first entire human genome [Venter et al., 2001], it is estimated that a total of 250,000 had been sequenced by 2015. Considering the current growth rate (doubling every 7 months) every human on this planet would be sequenced by 2024 [Stephens et al., 2015]. Even considering Illumina's more conservative estimate (doubling every year) or Moore's Law (doubling every 18 months), the amount of sequenced human genomes will at least approach 100 million by 2025 [Regalado, 2014, Stephens et al., 2015]. Another aspect to the large volume of molecular sequencing data is its high dimensionality, with the human genome being comprised of approximately 35,000 genes, and the human proteome containing more than 100,000 different proteins [Horgan and Kenny, 2011]. This further increases the volume of biological big data and consequently renders storage, distribution and analysis even more challenging. An important difficulty arising from big volume data in systems biology is that it cannot be easily explored by simply "looking at the data". This problem is being tackled by trying to detect certain structures in the data that can be categorized and visualized. A basic explanation of such approaches is given in sections 1.1.4 and 1.1.5.

1.1.3.2 Velocity

"Velocity" refers to the fact that there is a continuous stream of large amounts of data being produced every day. For example, a total of about 4.75 billion Facebook posts were added to the social network each day in mid-2016 [Fu et al., 2017]. New data is being generated faster than ever in the biological world as well. In 2014, more than 200,000 new human genomes were sequenced, and the amount of molecular profiling data grows exponentially every year [Regalado, 2014]. For data analysis purposes, this means that data sets can no longer be considered static. For biologists, this increasingly fast stream of new incoming data means that computational analysis tools should be able to be scaled up easily and to be updated in real time as new information becomes available [Marx, 2013]. This need for continuous data processing is a great challenge that the age of big data has laid upon systems biologists and biostatisticians.

1.1.3.3 Variety

"Variety" refers to the fact that more and more distinct types of data are now becoming available on the same subject. Examples from Facebook are that each user can upload text statuses, images and videos, establish friendships with others, create events, be part of groups, and message other people on the social network. Each type of data can be utilized to learn more about the user, but novel strategies for effective integration of the different data types have to be employed in order for the analysis to be able to ultimately provide a better "big picture" of the user's characteristics. Again, systems biologists face a similar challenge [Hwang et al., 2005]. Molecular sequencing techniques provide biostatisticians with a wealth of information on different types of molecular data, including the structure of the DNA, transcriptional RNA expression, protein concentrations, and metabolite concentrations. For an overview of such different "omics" data types, see table 1.1. The above data types roughly describe different levels at which the genotype-phenotype relationship is expressed, but there are countless complex feedbacks between many of the components of these different levels, thus rendering a correct integration of various data types in biostatistics immensely difficult [Joyce and Palsson, 2006]. A more detailed explanation on integration of multi-view molecular data sets can be found in section 1.1.6. For medical applications, it can be helpful to also integrate clinical data such as age, gender or behavioral habits (e.g. smoking) of the patients included in an analysis. All the above examples show that systems biologists are confronted with high-variety big data that is especially hard to deal with in an organized manner.

1.1.4 Machine Learning and Cluster Analysis

A long-held belief about the relationship between humans and computers was expressed by Stuart G. Walesh in his memorable quote presented at the beginning of this chapter. Computers are incredibly fast, accurate and stupid, whereas humans are slow, inaccurate and brilliant. Consequently, the combination of human intel-

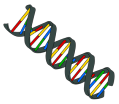
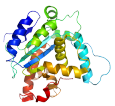
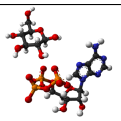
genomics & epigenomics		$\sim 25,000$ genes
transcriptomics		$\sim 10^5$ RNA transcripts
proteomics		$\sim 10^6$ proteins
metabolomics		$\sim 10^4$ metabolites
phenomics & exposomics		$\sim 10^8$ compounds
metagenomics		$\sim 10^{14}$ microorganisms

Table 1.1: Various "omics" molecular profiling data levels and their approximate amount of variables [Gligorijević et al., 2016].

ligence and computers' computational power is needed to solve difficult statistical problems [Walesh, 1989]. In the age of big data, the question arises about what to do if the data to be analyzed becomes too large and complex for even the most brilliant human to be able to tackle the problem by telling his stupid computer what to do. Machine learning tries to answer exactly this question: How can we make computers do what needs to be done without telling them precisely how to do it? [Samuel, 1959, Koza et al., 1996] Nowadays, machine learning is an immensely popular discipline in computer science, which uses methods from statistics to produce algorithms that are able to "learn" about the structure of some input data, and use the results for tasks such as prediction or categorization.

1.1.4.1 Supervised Versus Unsupervised Learning

The two main sub-fields of machine learning are supervised and unsupervised learning. The former tries to make inferences about some data with labeled responses, whereas the latter does the same if labels are not available [Friedman et al., 2001]. The difference between the two can easily be understood using a real-world analogy. For instance, when toddlers learn how to distinguish cats and dogs, they usually do so under the supervision of their parents or other people. If they see or hear either of the two animals, they will often be told which kind it actually is. What the toddler sees or hears can be seen as the data input, and the parent can be seen as the "teacher" who provides them with the correct label. With time, the toddler

then learns how to distinguish and recognize cats and dogs in a supervised manner. It is reasonable to believe, however, that with time a toddler would also be able to distinguish between cats and dogs if nobody were around to supervise them by giving them the correct labels. With time, the toddler would recognize certain features such as the shape of the snout or ears of the animals, or the sound of barking or meowing. At some point, the toddler would probably come to the conclusion that cats and dogs are two different types of animals. This process would then be called unsupervised learning, or "learning without a teacher".

In molecular cancer data analysis, an example of a supervised learning task would be to train a model to relate gene expression variables in a data set to the survival of the patients by using clinical outcome (dead vs. alive, or days survived) as label. This model could then be used to predict the survival of new cancer patients based on their gene expression data from a tumor sample. An example of an unsupervised learning task in molecular cancer research is trying to find previously unknown associations between different types of molecules or genes. During disease progression, a group of biological processes involving a certain set of genes could be deregulated in a certain way. Finding this group of genes using unsupervised learning may then – together with what is already known about those genes – shed some light on the underlying biological processes and potentially on the cause of the deregulation. The main goal of this thesis is to propose an approach to discover such groups of genes, and to apply it to GBM data. Hence, the following discussion is concerned with unsupervised learning methods suited for the task at hand.

1.1.4.2 Clustering Methods

One of the main tasks in unsupervised machine learning is finding clusters in the data at hand. A cluster can be broadly described as a group of certain objects that are more similar to each other than they are to objects belonging to a different cluster. If these objects are described by a list of measurements, cluster analysis can be defined as organizing the objects into groups based on some similarity measure on the multidimensional space spanned by the available measurements [Jain et al., 1999]. Figure 1.1 shows an illustrative example of one of the most intuitive tasks – clustering points in two-dimensional space based on their standard euclidean distance, where a small distance between two points indicates a high similarity between them. In the left panel the clustering task seems well-defined with tightly packed and well-separated clusters. In the right panel, however, objects within each cluster tend to be less similar (close) to the other objects in the same cluster, and the different clusters do not seem to be clearly separated. While there still seem to be clusters existent in the data, it is likely that some of the data points have been grouped into the "wrong" cluster with respect to the true underlying data generation process. Since there are many ways to define what it means for objects to be "similar" and since the research goal in clustering usually depends on the underlying data, a plethora of different algorithms have been proposed [Estivill-Castro, 2002]. The following covers the most prominent approaches and some popular algorithms.

Centroid-based clustering algorithms usually take as fixed input the number of de-

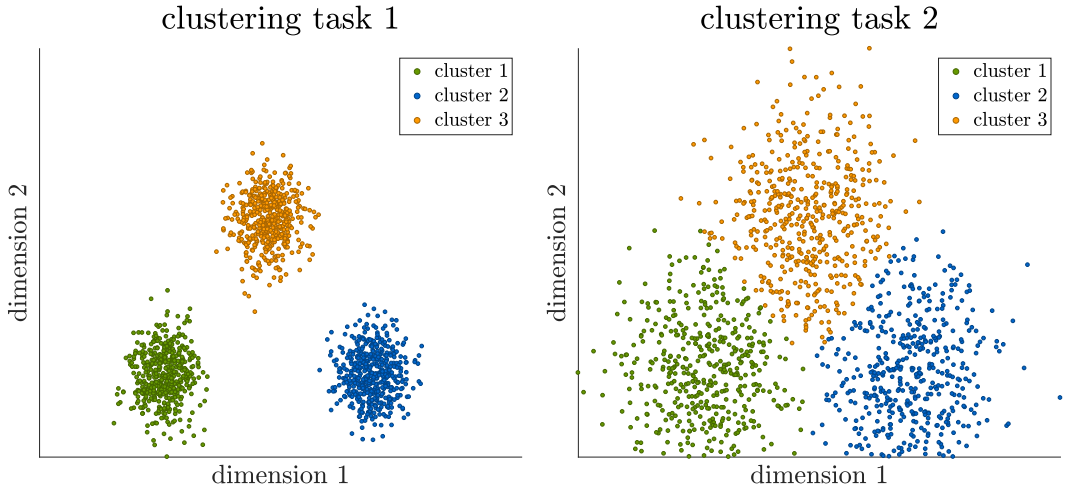


Figure 1.1: Two examples of clustering based on distance in two-dimensional space. The left panel shows a relatively "easy" clustering task with clearly separated clusters, while the panel on the left illustrates a harder example where clusters seem to overlap.

sired clusters k. They then find the k points in the underlying space that form the cluster centers (centroids) such that the sum of squared distances of all objects to their closest centroid are minimized. Popular centroid-based algorithm are the k-means and the k-mediods algorithms [Lloyd, 1982, Kaufman and Rousseeuw, 1987].

Hierarchical clustering algorithms rely on the idea that any object should rather be connected to a nearby object than to one at greater distance [Ward Jr, 1963]. Bottom-up hierarchical clustering algorithms start with a fully unconnected set of objects and then iteratively connect the objects that are close to each other, until a certain amount of connected components is found. In contrast, top-down hierarchical clustering algorithms start with removing connections between the objects that are furthest apart and then continue in the same fashion until a certain amount of connected clusters is identified. Hence, in both bottom-up and top-down hierarchical clustering algorithms the number of clusters in the data is generally determined by a distance threshold for adding or removing connections. By letting the threshold vary over the whole range of object distances, this produces a hierarchy of different cluster sets identified at different values for the threshold. Some common hierarchical clustering algorithms are based on the concepts of single-linkage, average-linkage (UPGMA) and complete linkage [Murtagh, 1983].

Distribution-based clustering algorithms are based on the idea that objects in a cluster should belong to the same probability distribution. Such algorithms make an assumption on the type and number of the underlying distributions and then maximize the likelihood of the data being generated by a mixture of them. While assuming normal distributions is often a quite strong assumption on the data, a popular choice are gaussian mixture models that are solved by the expectation-

maximization (EM) algorithm [Reynolds, 2015, Dempster et al., 1977, Rasmussen, 2000].

Density-based clustering algorithms rely on the assumption that a cluster is an area in which there is a higher density of objects in the underlying space than in the cluster's neighborhood [Kriegel et al., 2011]. This often implies that objects from low-density regions are not assigned to a cluster and considered noise. Popular density-based algorithms are DBSCAN, OPTICS and Mean-Shift [Ankerst et al., 1999, Comaniciu and Meer, 2002]

In the last decades, a vast amount of clustering algorithms have been proposed for different types of data and to accommodate for different shapes, densities, varying sizes or overlapping clusters [Ertöz et al., 2003]. A few of those have been described above, but a full review of the literature is not within the scope of this thesis. We therefore refer the reader to Xu and Tian [2015] for a comprehensive, accessible review of clustering techniques.

1.1.4.3 A Difficult, Subjective Task

The previous discussion has already been a prelude to the fact that clustering is usually a quite difficult problem. Since the main goal of clustering is the formulation of a hypothesis on the structure of the data at hand, algorithms require the user to make assumptions about the hypothesis to be learned [Kotsiantis and Pintelas, 2004]. This generally amounts to deciding what constitutes a group of items that are similar to all the other items within the group, and less similar to the rest. Hence, the same set of items often needs to be grouped differently, depending on the eventual goal of the user [Jain et al., 1999]. Another difficulty is that the interpretation of clusters may be very hard, so the clustering algorithm should also be designed to facilitate the analysis of the results that are obtained. Furthermore, finding exact solutions to clustering problems is often computationally very intensive since all or many pairwise similarities between the data points have to be processed [Koziel et al., 2014]. If data sets become too large to be handled computationally by exact algorithms, greedy heuristics or other approximate algorithms have to be employed [Swamy and Shmoys, 2004]. For all the above reasons, clustering should not be considered an application-independent mathematical problem, but rather a somewhat subjective approach that needs to be examined in the context of its ultimate use [Guyon et al., 2009].

Figure 1.2 shows an example of 1800 data points in two-dimensional euclidean space, whose clustering is quite subjective, and where several different results are equally viable without relying on any additional assumptions. The data set comprises three groups of points $\boldsymbol{X}^{(\text{upper left})}$, $\boldsymbol{X}^{(\text{upper right})}$ and $\boldsymbol{X}^{(\text{bottom})}$ drawn from different distri-

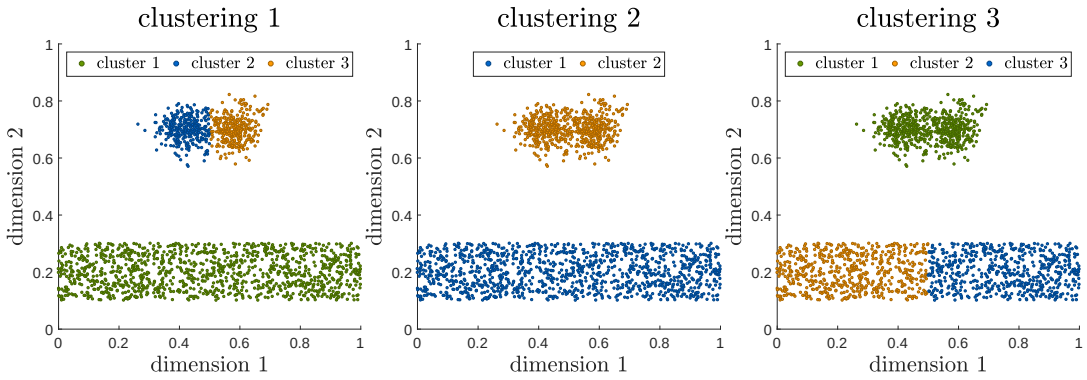


Figure 1.2: An illustrative example which shows that clustering is not a well-defined task, but rather a subjective classification of data that depends on the research goal.

butions [1]. Hence, with respect to the data generation process, the best clustering would retrieve three clusters of which each is mainly containing points of one of these groups. An according solution is shown on the left panel of figure 1.2. One may, however, argue that the two upper groups of points are quite similar, and that their members should be considered to belong to the same cluster. This results in the clustering shown in the central panel in figure 1.2. Another possible clustering is shown in the right panel. While the lower group seems to be arbitrarily divided into two parts here, this solution provides roughly equally sized clusters where points within each one are closest to all other points within it. This third option is actually the solution to the k-means algorithm for $k = 3$.

In this 2-dimensional example, the euclidean distance between two points gives rise to a very intuitive similarity measure between points, and results can easily be visually assessed in a two-dimensional plot. When it comes to clustering patients or their molecular profiling data in cancer research, however, it is often not clear what actually makes two sets of measurements similar. With data on thousands of genes or molecules for many patients, biostatisticians usually face the challenge to cluster data in an underlying space of thousands of dimensions. This renders an intuitive visualization or assessment of the results virtually impossible. Hence, similarity measures on molecular data and the clustering algorithms to be utilized have to be carefully chosen in the context of the origin of the data and the research goal. Parameters and settings in those algorithms are then often chosen somewhat subjectively or even arbitrarily so that the resulting clusters are considered meaningful by the researcher.

[1] Here, 300 of the data points are independently and identically distributed (i.i.d.) realizations of a bivariate normal random variable $\boldsymbol{X}^{(\text{upper left})} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu_1} = \begin{bmatrix} 0.42 \\ 0.7 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 0.002 & 0 \\ 0 & 0.002 \end{bmatrix}$. Another 300 data points are distributed as $\boldsymbol{X}^{(\text{upper right})} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu_2} = \begin{bmatrix} 0.58 \\ 0.7 \end{bmatrix}$. The remaining 1200 data points are uniformly distributed on the intervals $[0, 1]$ and $[0.1, 0.3]$ in the first and second dimensions, respectively.

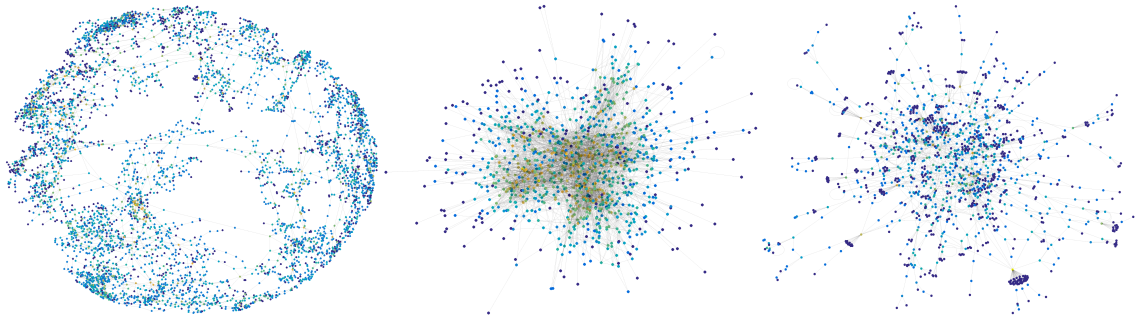


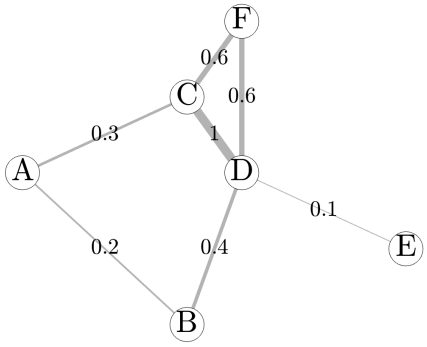
Figure 1.3: Visual 3-D representations of three examples of real world networks, plotted in Matlab. Left to right: a power grid, a network of email interactions, and a protein-protein interaction network.

1.1.5 Network Modeling

A network is a set of objects which are connected to some or all of the other objects in a certain way. They often arise naturally in the real world, and the scientific study of networks has recently become hugely popular [Barabasi and Oltvai, 2004]. To illustrate a few examples of real-world networks, figure 1.3 shows (from left to right) 3-D visualizations of the power grid of the western United States [Watts and Strogatz, 1998], a social network of email exchanges at a Spanish university [Guimera et al., 2003], and the largest cluster in a protein-protein interaction network in yeast [Jeong et al., 2001].

1.1.5.1 Terms and Definitions

A network or "a graph" $\boldsymbol{G}(\mathcal{V}, E)$ is a collection of nodes (or "vertices") $\mathcal{V}$ and a set of edges E, which each connect two elements of $\mathcal{V}$. In general, an edge can connect a node with a different node or with itself. Each edge between a node i to another node j can have a weight w_{ij}, which usually indicates the strength of the connection between the two nodes. A graph is called *unweighted* if all of its edge weights are the same (usually all equal to one). A graph with different edge weights is referred to as a *weighted* graph. Graphs can also be *directed*, which means that each edge represents an interaction of a specific direction from one node to another. The *degree* of a node in a graph is the number of edges that connect the node to others. In a weighted graph, the *weighted degree* is the sum of the weights of all such edges. In a directed graph, there is an out-degree and an in-degree, which only consider outgoing or incoming edges, respectively. The networks in figure 1.3 are unweighted and undirected, and nodes are colored by their degree, with blue indicating a low degree and yellow indicating a high degree. The left panel of figure 1.4 shows a simple example of a weighted, undirected network consisting of nodes A, B, C, D, E and F. Edges are labeled by their weight, and drawn with a thickness that is proportionate to the weight. A *connected component* of a network is the set of all points that can be reached from any node in the component by traveling along existing edges. Hence, two different components of an undirected network are sets



node 1	node 2	edge weight
A	B	0.2
A	C	0.3
B	D	0.4
C	D	1
C	F	0.6
D	E	0.1
D	F	0.6

	A	B	C	D	E	F
A	0	0.2	0.3	0	0	0
B	0.2	0	0	0.4	0	0
C	0.3	0	0	1	0	0.6
D	0	0.4	1	0	0.1	0.6
E	0	0	0	0.1	0	0
F	0	0	0.6	0.6	0	0

Figure 1.4: Three equivalent ways of describing a network. From left to right: A visualization of nodes and edges, an edge list with weights, and an adjacency matrix.

of nodes that are not connected by any edges. It is important to note that while the visualization in figure 1.4 is shown in two-dimensional space in a manner that makes it easy to look at the network's structure, the nodes have no actual fixed positions – they could be moved anywhere and the network would still be the same. In this thesis, only undirected networks are considered. Hence, throughout the remainder of the discussion, all networks will be assumed to be undirected.

For purposes of computational analysis, networks are usually represented as either an *edge list* or an *adjacency matrix*. In an edge list each row corresponds to an edge of the network. For unweighted networks, there are two columns. Each row lists the two nodes that are connected by the edge that this row corresponds to. In directed networks, the order of the node entries in the two columns indicates that the edge connects from the node in column one to the node in column two. If a network is weighted, its edge list has a third column indicating the weight of each edge. An adjacency matrix A of a network with n numbered nodes is an $n \times n$ matrix, where each entry A_{ij} corresponds to the edge between node i and node j. If no edge exists between two nodes, then the corresponding entry in the adjacency matrix is set to zero. In unweighted networks, an entry in the adjacency matrix is set to one if the corresponding edge exists. In weighted networks, the entries of the adjacency matrix are the weights of the edges. In directed networks, entry A_{ij} corresponds to the edge from node i to node j. In undirected networks, entries A_{ij} and A_{ji} are the same. Consequently the adjacency matrix of undirected networks is symmetric across its diagonal. The center and right panels in figure 1.4 show the edge list and adjacency matrix representations of the network on the left. Whether an edge list or an adjacency matrix is used for network analysis usually depends on the network at hand and the goal of the analysis. Adjacency matrices normally make it easy to process edge information in a structured way. For large networks where many of the entries of the adjacency matrix are zeros (a so-called sparse matrix), adjacency matrices often become prohibitively large. This problem can often be addressed by using sparse data structures. Yet, the edge list format allows for a shorter and often more intuitive description of the network in sparse settings, since only existing edges are represented as rows.

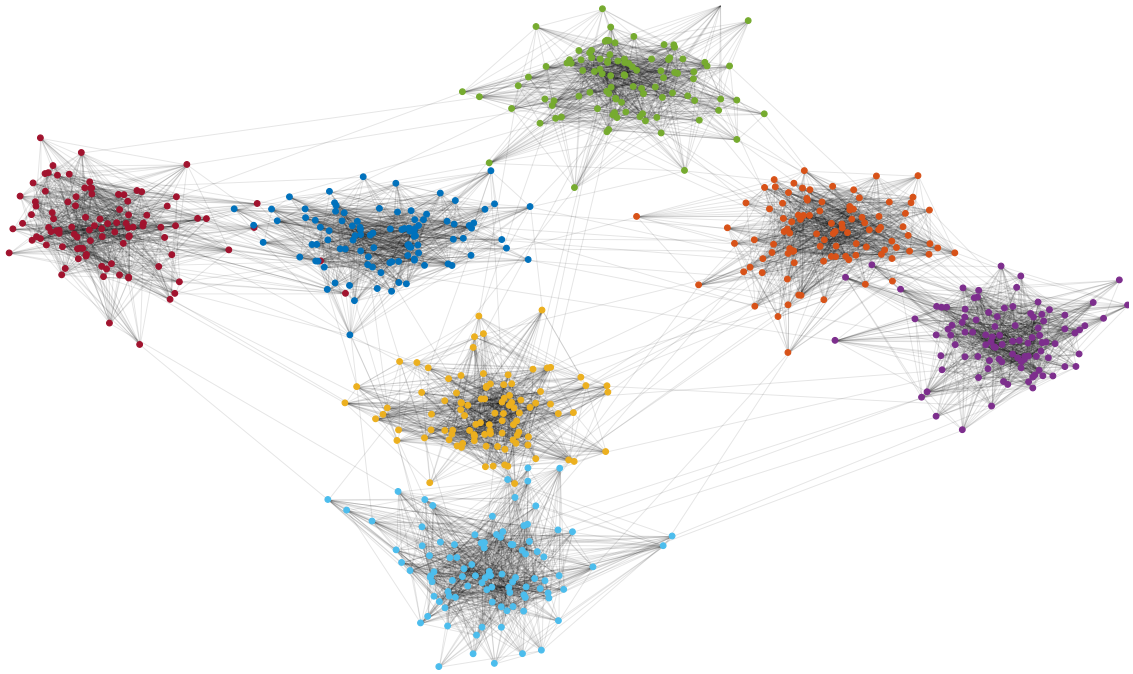


Figure 1.5: An example of communities detected in an unweighted network. Here, node color indicates community membership.

1.1.5.2 Community Detection

Many real-world networks exhibit community structures [Girvan and Newman, 2002]. For instance, the people in different university departments could be considered smaller communities within the larger university network. Detection of such communities is based on a similar idea as clustering. Intuitively, nodes within a community should be strongly interconnected, whereas there should be fewer edges connecting a node inside a community with nodes that belong to other communities. For unweighted networks, this loosely means that there should be more edges "inside" a community than edges connecting the community with the rest of the network [Fortunato, 2010]. For weighted networks the same applies to the sum of all the corresponding weights. Figure 1.5 shows some communities identified in an unweighted network.

Considering that the weight of an edge generally indicates the strength of the connection between the objects that two nodes represent, a network directly supplies us with the measure of similarity that is needed for clustering a set of objects. Hence, clustering and community detection are conceptually one and the same problem. It is therefore not surprising that community detection is often also referred to as "network clustering" or "graph clustering" [Schaeffer, 2007]. The main difference between the two tasks is that network nodes used for community detection do not have a fixed position in any underlying space, whereas data points used in traditional clustering do not have defined connections to each other that would explicitly embed them in a network [Zafarani et al., 2014]. Just as for clustering, a plethora of algorithms have been proposed to find community structures in networks. The most

common ones will be concisely presented in the following. For simplicity, we here generally consider unweighted networks, although extensions to weighted networks are often straightforward.

One of the oldest approaches to divide networks into communities – the minimum cut method – is based on finding a fixed number of groups of nodes such that the amount of edges between groups is minimized [Ford and Fulkerson, 1956]. A problem with this approach is that the solution often amounts to separating individual nodes from the rest of the network [Von Luxburg, 2007]. The ratio cut approach [Hagen and Kahng, 1992] circumvents this problem by normalizing, for each community, the number of "cut" edges between that community and the rest of the network by the number of nodes in the community. Another popular method is normalized cut [Shi and Malik, 2000], which instead normalizes each cut by the sum of all weights inside the community. Both ratio and normalized cut, however, are computationally very expensive.

A related, popular approach is spectral clustering, where standard clustering methods are applied to a relevant set of eigenvectors of the network's Laplacian matrix (the adjacency matrix, but with node degrees on its diagonal). It has been shown that spectral clustering is able to solve relaxed versions of the ratio and normalized cut conditions [Von Luxburg, 2007].

The Girvan-Newman algorithm [Girvan and Newman, 2002] relies on the concept of edge betweenness centrality, the number of shortest paths between pairs of nodes that run along a certain edge. If a network has a community structure, then many shortest paths between nodes should run along the few inter-community edges connecting them, and thus such edges will have a high edge betweenness centrality. The Girvan-Newman algorithm then iteratively removes the edge with the highest value and re-calculates the new edge betweenness centrality values. As edges are removed, this produces increasingly more connected components in a hierarchical fashion until no edges are left. These connected components can be identified as the communities of the network at any level of the resulting hierarchy.

Some other community detection methods are based on modularity maximization. Modularity is defined as the difference between the fraction of edges within a certain group and the expected fraction if edges were randomly distributed [Newman, 2006]. Communities with a particularly high value are then found by modularity maximization methods by employing approximate optimization algorithms. A popular method for community detection using modularity is the Louvain method [Blondel et al., 2008].

Other methods for community detection focus on the generative process underlying network formation, with the stochastic block model being a popular choice [Holland et al., 1983, Brownlees et al., 2017]. Finally, some community detection algorithms focus on cliques – sets of nodes that are fully connected – within a network [Bron and Kerbosch, 1973]. Since cliques can overlap, such algorithms generally find solutions where nodes can be part of multiple communities. As for clustering, community detection is a rather subjective task, and the choice of the right algorithm depends on the type of network at hand, as well as the ultimate goal of the research project.

1.1.5.3 Molecular Biological Networks

While the definition of many real-world networks such as social networks are often intuitive, things become a little more complicated when it comes to molecular biological networks. In online social networks such as facebook, users are the nodes and an edge may for example be drawn between them if they are friends. In a biological cell, there are many complex interactions of varying strength between genes, proteins, metabolites and other components.

For instance, a certain gene can be transcribed as mRNA, which then codes for the production of one or more proteins. Those proteins can be enzymes, which means that they catalyze certain reactions in the cell and thus further have an influence on metabolites. The rate of other metabolic reactions can then also be influenced by the relative concentration changes inside the cell. There are also isoenzymes, which are proteins of different structure which catalyze the same reaction, though often at different rates. Furthermore, some reactions require a multienzyme complex – consisting of multiple enzymes that are coded for by different genes – to be catalyzed. Proteins can also be transcription factors (TFs), which means that they can bind to a certain position on the DNA corresponding to some gene, and then either promote or inhibit the transcription of that gene's mRNA, which in turn results in the production of further proteins. There are more molecular mechanisms contributing to the function of a cell, but the clear take-away from the above examples is that molecular biological data gives rise to complex networks, which are difficult to describe with the established mathematical formalism in network science. While it is possible to consider each gene, protein or other compound a node, and each reaction an edge in a biological network, this does take into account the conditional statements introduced by isoenzymes and multienzyme complexes, nor does it take into account that reactions may involve multiple substrates or that enzymes do not take part in reactions but are still necessary for them to occur. Since a complete network on the level of a whole cell is virtually impossible to formalize, biologists have generally resorted to describing networks on a certain type of compound or mechanism, such as protein-protein interaction networks, gene regulatory networks, gene coexpression networks, or metabolic networks [Rual et al., 2005, Davidson and Levin, 2005, Stuart et al., 2003, Jeong et al., 2000].

When using molecular profiling data to reconstruct biological networks, a similarity is usually defined between different molecular variables across the data of multiple patients, using correlation or any other pairwise distance metric [Schadt, 2009]. Since most of the similarities between different variables are often nonzero but very low, it is usually desirable to introduce sparsity in the resulting similarity matrix in a way such that direct interactions are recovered from the data while distant interactions or noise are filtered out [August and Papachristodoulou, 2009]. Multiple approaches on how to tackle this challenge have been proposed, such as the graphical LASSO [Friedman et al., 2008] and WGCNA [Langfelder and Horvath, 2008].

1.1.6 Multi-View Data

Multiple views of data refer to the "variety" aspect of big data elaborated on in section 1.1.3.3. The general idea is that often data of different types can be collected about a certain object. Using a combination of those different types for a learning method in a smart way can then help us identify certain characteristics of that object with higher confidence.

1.1.6.1 The Promise and the Challenge

To illustrate the promises and challenges of using multi-view data for machine learning, let us shortly re-visit the Facebook example. Any user on the social network can upload text statuses, images and videos, and react to such uploads of others in certain ways. They can also establish friendships with other users, create events, be part of groups, and directly message other people. Users may also describe themselves in a biographical section, and include data such as their age, gender, interests, and places where they have lived, studied or worked. It is clear that each of those different types of user data can be utilized to learn about what "kind of person" the user actually is. The great promise of multi-view learning is that looking at all different types of data should make it much easier for us to understand by which characteristics we should describe the user, than by only looking at a single data type. The grand challenge in integrating multiple views is devising smart learning algorithms that make full use of the often complementary information in a structured, algorithmic way despite the heterogeneity of the data. What actually constitutes a "smart way" of integrating multiple views usually depends on the underlying data and on how the different views relate to each other. Due to this, multi-view learning has recently become a very active field of research, with a diverse range of new algorithms being proposed continuously [Bickel and Scheffer, 2004, Xu et al., 2013, Zhao et al., 2017].

1.1.6.2 Multiple Views of Molecular Profiling Data

As already mentioned in section 1.1.5.3, molecular profiling data consists of different types that can be described on different biological "omics" levels. Each biological level, however, may still give rise to multiple data types, which can be considered different views in data analysis. This is illustrated in figure 1.6. The *genomic level* concerns the structure of the DNA, which is a pair of macromolecules that are tightly held together and carry all the information needed to "instruct" any organism on how to grow, develop, function and reproduce [Watson et al., 1953]. The *transcriptomic level* describes the abundance of messenger RNA (mRNA) molecules in a cell, which are used to distribute the information stored in the DNA to the ribosomes, where they cause the production of certain proteins [Cooper and Hausman, 2004]. The *proteomic level* refers to the abundance of proteins, which can have multiple functions, such as being enzymes, being structural, or being regulatory [James, 1997]. The *metabolomic level* is associated with the abundance of metabolites, small molecules

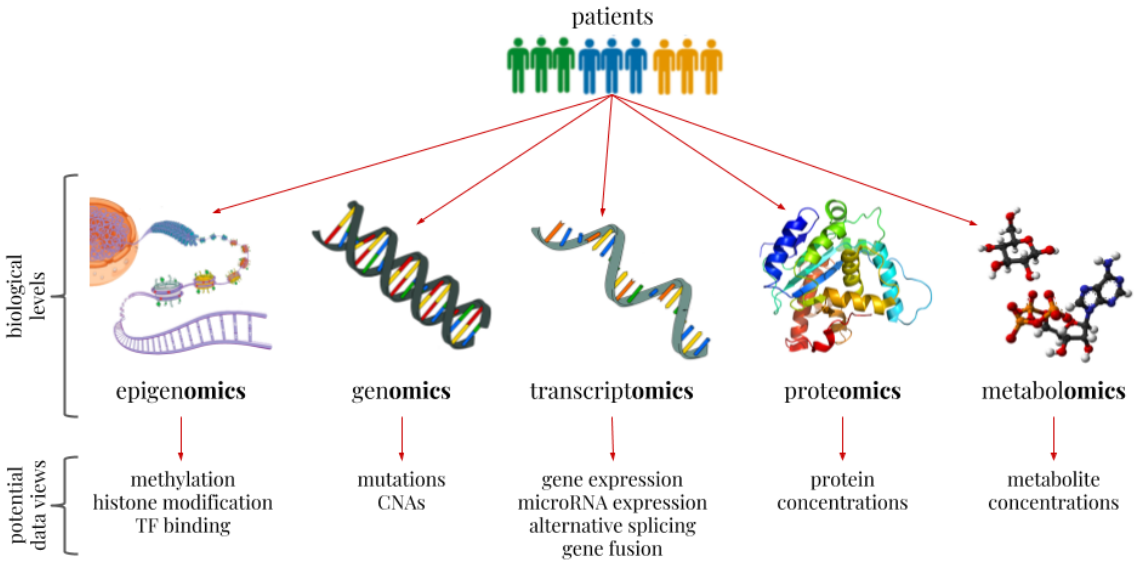


Figure 1.6: An overview of different "omics" levels in systems biology, and the sequencing data types (or views) that they give rise to.

that are used in reactions inside cells to various ends such as biomass production or energy conversion [German et al., 2005]. The *epigenomic level* deals with reversible structural changes of the DNA or histones, which influence gene expression without altering the underlying nucleotide sequence of DNA [Russell, 2010]. The integration of data on such different "omics" levels has become an important goal in systems biology [Gomez-Cabrero et al., 2014]. Therefore we here list some of the most important data types and how they are obtained.

Structural variations on the genomic level such as mutations and copy number aberrations (CNA) can be detected by whole genome sequencing [Wheeler et al., 2008]. Gene expression and other mechanisms on the transcriptomic level such as alternative splicing or gene fusion may be uncovered by RNA-seq [Nagalakshmi et al., 2008]. Another data type on the transcriptomic level is microRNA expression, which regulates the translation of mRNA and can be profiled by miRNA-seq [Creighton et al., 2009]. On the proteomic level, the concentrations of proteins can be revealed by protein arrays or mass spectrometry [Melton, 2004, Domon and Aebersold, 2006]. On the metabolomic level, the concentrations of metabolites can as well be determined by mass spectrometry [Dunn et al., 2013]. On the epigenomic level, histone modification can be identified by ChIP-seq, which makes it also possible to detect relationships between the proteome and the genome by locating transcription factor binding sites [Johnson et al., 2007]. Another data type on the epigenomic level is DNA methylation, which has an important role in gene regulation, and can be described by bisulfite sequencing [Lister et al., 2009].

Other views of data that could be integrated are given by molecular profiling data from different databases, which use distinct profiling techniques or distinct sampling methods such as direct tumor tissue sequencing versus cell line sequencing. It is also possible to integrate data from different types of cancer tumors or from different identified subtypes of the same cancer, to find shared mechanisms or specific

differences across disease types. Different approaches to multi-view integration in cancer research will be further discussed in the next section.

1.1.6.3 Different Types of View Integration in Cancer Data

The most common data integration task in cancer research is subtype identification based on multiple views of molecular data [Gligorijević et al., 2016]. In this task, the different patients are the "objects" that need to be clustered. Hence, any view for which data is available for the chosen set of patients can be used to construct a view-specific patient-to-patient similarity measure. The different view-specific similarity measures then need to be integrated before or during the clustering process. Since the number of patients in multi-view molecular cancer data sets is usually quite small, and since patients can be quite intuitively interpreted as objects that may be more or less similar, integrative subtype discovery can arguably be considered the most straightforward multi-view integration task in molecular cancer research.

Another integration task in cancer research is clustering all the available molecular variables based on their similarity across patients, especially if subsets of the patient samples come from distinct backgrounds and need to be assumed to exhibit different behavior. Then, these groups of patients become the multiple views of the data, and the molecular variables become the objects to be clustered based on similarity. In this case, the multiple views could be represented by two or more databases that use dissimilar sequencing technologies or different sample treatment, such as TCGA's GBM tissue data and HGCC's GBM cell line data [Weinstein et al., 2013, Xie et al., 2015]. The views could also be considered different diseases or disease subtypes, such as the four generally accepted subtypes of GBM [Verhaak et al., 2010]. Clustering such molecular variables can, for example, shed light on which genes or molecules interact in specific regulatory processes, and thus help characterize disease progression and identify appropriate therapeutical strategies. This integrative clustering of molecular variables is usually a more challenging task than subtype identification, since it is difficult to find a general, structured approach to simultaneously assess similarities arising from data sets that have different meaning and underlying generation processes. Furthermore, there are usually at least multiple thousand molecular variables to be clustered. This limits the amount of methods that are computationally feasible for clustering or makes it necessary to pre-select variables to be included in the analysis based on some more or less subjective approach, and further complicates an intuitive exploratory analysis of the results.

A more specific integrative task in cancer research is uncovering gene-gene associations from data types that can be related to a certain gene. Resulting gene-gene association networks can be seen as an integrative extension of widely studied gene co-expression networks, which relate genes to each other merely based on gene expression without attempting to integrate multiple views [Butte and Kohane, 1999, Margolin et al., 2006]. In addition to gene expression on the transcriptomic level, it is also possible to map copy number aberrations and mutations on the genomic level, as well as DNA methylation on the epigenomic level directly to the corresponding genes. This allows us to define genes as our objects to be clustered in patient space,

with the above molecular profiling data types representing our different views to be integrated. As opposed to clustering all molecular variables, the total number of genes is much smaller and consequently we have the choice between a larger number of clustering algorithms that are able to handle the task computationally. Furthermore, genes are quite well-annotated in terms of the molecular functions, biological processes or cellular components that they are involved in, which facilitates further analysis of any clusters identified in the data [Ashburner et al., 2000, Consortium et al., 2012, Maglott et al., 2005, Hubbard et al., 2002, Harrow et al., 2012]. The challenge in integrative gene-gene association clustering is that it is not generally clear how the different data types can be effectively merged to give a big picture of which genes are closely related to each other from any specific viewpoint, such as molecular processes or biological function. While large groups of closely interrelated genes should be reflected in all data types, chain-reaction like processes that transcend multiple omics layers will be virtually impossible to detect by clustering approaches, and generally cannot be directly inferred from the data in view of the large amount of noise in molecular profiling data [Arnold et al., 2013].

Due to the fact that multi-view exploratory analysis on molecular profiling data shows great promise but also faces many difficult challenges, the next section will establish a desirable framework for such analysis. Then, the following chapter will review the broad spectrum of literature which has attempted to tackle some of the arising challenges.

1.2 A Desirable Framework for Integrative Exploratory Molecular Data Analysis

This thesis aims to explore associations between different omics variables in tumor samples of GBM patients. The goal is to enable cancer researchers to identify biomarkers and to gain a better understanding of biological and molecular processes related to the cause and progression of the disease. The previous introductory sections have demonstrated that integrative unsupervised learning in a biological context and such high dimensional settings is a very difficult and often subjective task. Thus, this section will describe a hypothetical optimal framework for exploratory analysis of molecular variables in cancer research. In the following chapter, we then review recent advancements that have been made towards fulfilling one or more of the desired qualities listed below.

A highly desired property of approaches to identify associations between molecular variables is the capability to integrate multiple views of omics data, as the true underlying biological networks are formed by direct interactions that transcend virtually all levels (see sections 1.1.5.3 and 1.1.6.2). To make full use of the wealth of molecular data published by various profiling initiatives, it would also be expedient if the approach were able to integrate data originating from different databases. Furthermore, it would be advantageous if the approach were able to integratively make use of identified subtypes within the available set of GBM patients, to prop-

erly reflect differences in disease origin and progression across patient strata. While an algorithm capable of effectively integrating all of the above data types would be favorable, it has to be noted that the more omics data types are taken into account the more difficult it becomes to devise a structured way of integrating them in a meaningful manner. In addition, most data types are only available for a certain subset of patients or genes, and therefore including many data types for integration often implies reducing the amount of samples available for network estimation or clustering. When aiming to design a potent algorithm for exploratory integrative omics data analysis, we therefore face a trade-off between using many distinct data types to gain comprehensive coverage of molecular processes on different levels, the possibility to integrate them in a reasonable way, and the amount of samples available for robust estimation.

In addition to an effective integration of multi-view data, it would be desirable if our approach could estimate similarities between molecular variables in a sparse manner, since the true underlying networks are assumed to be formed by direct or close interactions of only a handful of entities at a time. A sparse estimation of the resulting statistical network would also facilitate exploratory analysis and biological interpretation, and makes it possible to employ efficient community detection algorithms for large-scale complex networks.

For easy interpretation of the results, it would be of benefit if communities were detected in a hierarchical fashion, which would make it possible to investigate only a handful of communities and their relationships to each other on any given level of resolution. Annotation of genes and other entities in terms of biological processes or molecular function also follows a roughly hierarchical structure from broad terms such as "response to stimulus" to highly specific terms such as "mitochondrial double-strand break repair via homologous recombination" [Consortium, 2014]. An approach that detects communities in a hierarchical fashion would therefore also make it possible to associate each resulting set of molecular entities with those biological processes or molecular functions that are over-represented in their corresponding annotations [Mi et al., 2016]. Such a multi-level hierarchical clustering method should be able to identify high-quality clusters that are biologically meaningful across all levels of resolution, from very coarse to much finer partitions of the data set at hand.

Optimally, such an unsupervised algorithm should also not require any fixed parameters that determine the number of clusters or the thresholds giving rise to different levels in the hierarchy. This is because such information is not available prior to an exploratory analysis of omics data, and there is no straightforward way to tune parameters since correct labels are unknown.

It would be extremely difficult to simultaneously honor all of the desired qualities that are listed above in a single algorithm. Therefore, in the next chapter, we focus on a review of recently suggested approaches that attempt to tackle some of these challenges separately.

2

Related Literature

"The answers you get from literature depend on the questions you pose."

– Margaret Eleanor Atwood

This chapter provides a broad overview of recent advances in unsupervised learning, that are relevant to exploratory molecular data analysis in cancer research. Many of the algorithms introduced here are not directly applicable to the data and aim of this study, but are presented to give the reader an idea of what is currently feasible in the domains of multi-view learning and multi-level hierarchical community detection. While it is difficult to sort all the approaches into fixed categories (a hard clustering problem!), the literature will be presented in the following order. Section 2.1 presents unsupervised learning methods that have mainly been introduced to deal with multi-omics integration, with a specific focus on cancer research. Section 2.2 introduces hierarchical community detection methods for large-scale complex networks, that show potential to be applied to networks estimated from molecular profiling data.

2.1 Unsupervised Multi-Omics Learning in Cancer Research

This section aims to serve as a concise survey of the most relevant and common approaches to integrative unsupervised learning based on multiple views of omics data. For more comprehensive reviews of the topic, we refer the reader to Kristensen et al. [2014], Bersanelli et al. [2016], Gligorijević et al. [2016] and Huang et al. [2017]. The methods presented in the following can often be assigned to multiple categories of conceptual approaches. Here, the classification into different section headlines is largely adopted from the latter reference, but it is important to note that there is often overlap between the different sections. Section 2.1.1 introduces matrix factorization approaches, section 2.1.2 presents methods based on analysis of correlation and covariance, section 2.1.3 covers bayesian methods, section 2.1.4 focuses on network-based approaches, and section 2.1.5 explains a few multiple kernel and multi-step procedures.

2.1.1 Matrix Factorization Approaches

Some of the most intuitive methods for unsupervised omics integration are based on Non-negative Matrix Factorization (NMF). NMF aims to represent a non-negative data matrix in a lower-dimensional space by decomposing it into non-negative loading and factor matrices, such that the reduced-dimension decomposion still captures as much of the variation in the data as possible [Lee and Seung, 2001]. Zhang et al. [2011] suggested Sparse Network-Regularized Multiple Non-negative Matrix Factorization (SNMNMF) for omics integration, where data matrices from multiple views are represented in a common lower-dimensional space with the requirement that all views to be integrated have to share the same non-negative factor. The authors used the method to identify perturbed pathways by integrating gene expression, methylation, and miRNA data from ovarian cancer patients and finding heterogeneous variables weighted highly in the same projected direction [Zhang et al., 2012]. The drawbacks of this method are the high computational complexity and memory requirements of NMF, and the fact that it requires carefully normalized, non-negative input data. An implementation of their NMF-based algorithm is available on the author's webpage[1].

Similarly to NMF, iCluster [Shen et al., 2009] aims to represent multiple views of the data in a joint latent variable space, but without the non-negativity constraints. The joint latent variables are estimated by gaussian likelihood-based inference using the EM algorithm. The authors successfully used iCluster for subtype discovery in GBM [Shen et al., 2012]. A software implementation is available as an R package with the same name [Shen, 2012]. A drawback of iCluster is that the algorithm is quite time-consuming, and consequently patient samples or genes need to be preselected to make an analysis computationally feasible. iCluster+ [Mo et al., 2013] is an extension of iCluster that allows for the integration of continuous, binary, categorical, and sequential data by assuming distinct underlying distributions for different views, including logistic, normal linear, multilogit, and Poisson distributions. The algorithm is implemented in the iClusterPlus R package [Mo and Shen, 2016]. A related approach is moCluster [Meng et al., 2015]. This method finds joint latent variables using either a sparsity-inducing version of Consensus Principal Component Analysis (CPCA) [Wold, 1987, Westerhuis et al., 1998] or multiple co-inertia analysis (MCIA) [Meng et al., 2014]. Then, any type of clustering may be employed in the common reduced-dimensional space. The authors were able to identify four subtypes of colorectal cancer by applying this method to methylation, mRNA and protein data. The moCluster algorithm is implemented as part of the R package mogsa [Meng, 2016].

Another NMF-based algorithm is Joint and Individual Variation Explained (JIVE) [Lock et al., 2013]. JIVE decomposes the data into different low-rank matrices, including one that approximately captures the joint variation across all views, one that approximately captures the individual structured variation for each of the views, and one that represents residual noise. The algorithm is an extension of Principal Component Analysis (PCA), and also includes an L1 penalty for dimension reduction.

[1] `http://zhoulab.usc.edu/\acrshort{SNM\acrshort{NMF}}/`

An application of JIVE on GBM tumor gene expression and miRNA data helped the authors to characterize gene-miRNA associations and resulted in clinically useful disease subtypes. Since JIVE is based on PCA, a drawback of the method is that it is not robust to outliers. The method is implemented in the R package r.jive [O'Connell and Lock, 2017]. Similarly to JIVE, a method called the Joint Bayes Factor [Ray et al., 2014] decomposes the input data into joint, individual and noise terms. In contrast to JIVE, the Joint Bayes Factor assumes shared loadings for both individual and joint factors, and uses the student-t sparseness-promoting prior [Tipping, 2001] to impose sparsity on those. Instead of using L1 penalties for regularization, the model assumes a beta-Bernoulli process [Thibaux and Jordan, 2007, Ghahramani and Griffiths, 2006] for joint and individual factors. This approach makes it possible to extract features that are common to all data types as well as view-specific features for subsequent analysis. A drawback of the Joint Bayes Factor is that it assumes linear relationships between the latent variable and the observational spaces, and that it requires very high concordance between the different views of the data [Huang et al., 2017]. A software implementation is available on the author's webpage[2].

2.1.2 Correlation and Covariance-Based Approaches

Canonical-correlation analysis (CCA) [Hotelling, 1936] is a traditional method to infer the linear relationship between two multidimensional variables, by finding linear combinations of the multidimensional entries which have maximum correlation with each other [Härdle and Simar, 2007]. CCA and related approaches have been modified in many ways to accommodate for molecular data integration of two and sometimes more views. To this end, penalization and regularization terms can be included for both stable and sparse calculation of loading factors. Modified approaches for L1-penalized sparse CCA (sCCA), as well as elastic net CCA have been proposed to integrate two views of data and simultaneously increase biological interpretability due to the introduction of sparsity [Parkhomenko et al., 2009, Witten and Tibshirani, 2009]. Both are implemented in the R package PMA [Witten et al., 2013]. Additional CCA approaches for high-dimensional data have been suggested to consider the joint effects of groups of variables, which are selected based on prior biological information [Chen et al., 2012, Lin et al., 2013]. The algorithms have been used to study associations between nutrient intake and bacterial abundance in human gut microbiome data, and to uncover relationships between single nucleotide polymorphisms (SNPs) and gene expression in human gliomas.

Partial Least Squares (PLS) is a similar approach to CCA, but instead of focusing in correlations, it aims to find the loading factors that maximize the covariance. This makes PLS-based approaches less likely to suffer too much from outliers in the data [Huang et al., 2017]. The sparse PLS solutions analogous to sCCA and elastic net CCA have, however, been shown to perform similarly to their correlation-based counterparts [Lê Cao et al., 2009]. Sparse Multi-Block Partial Least Squares (sMBPLS) [Li et al., 2012] extends PLS to more than two types of data by using a

[2] `https://sites.google.com/site/jointgenomics/`

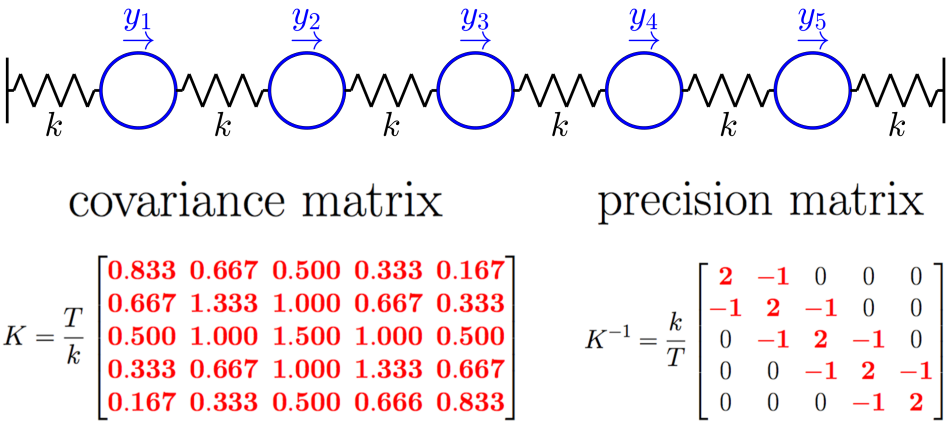


Figure 2.1: An illustration of how the precision matrix can often provide more useful information about the underlying system than the covariance matrix. Non-zero entries in the precision matrix of a chain harmonic oscillator represent direct interactions via a spring connection.

weighted sum of latent variables of those views in their objective function. The approach was used to discover multi-dimensional regulatory modules in ovarian cancer across data on the CNA, methylation and miRNA levels. An implementation of the algorithm is available on the author's webpage[3].

Another important covariance-related task in molecular data analysis is the estimation of the inverse covariance matrix (or precision matrix). This is due to the fact that the precision matrix contains the partial correlations between the variables under consideration. Under the assumption that the underlying data is normally distributed, this implies that entries are zero if and only if the two corresponding variables are conditionally independent [Das et al., 2017]. Hence the non-zero entries of the precision matrix indicate direct interactions between two variables, which are often of high relevance to researchers. Figure 2.1 illustrates this by considering a one-dimensional chain harmonic oscillator, where five particles with unit mass are coupled to each other by springs with the same spring constant k between two fixed walls as shown in the upper panel. The lower two panels show the covariance and precision matrices for the forces y_i acting on the respective particles, where T is a physical constant [MacKay, 2006]. While the covariance matrix is difficult to interpret since covariances between all variables are positive, the off-diagonal non-zero values of the precision matrix clearly indicate direct interactions between the corresponding particles (i.e. a connection via a spring).

Since the normality constraint virtually never holds exactly for real-world data, no entries of the inverse of an empirical covariance matrix become zero in practice. Furthermore, since high-dimensional data sets such as those encountered in omics research generally exhibit nearly perfect multicollinearity, the precision matrix is near-singular and thus its calculation is often not possible due to numerical instabilities. Sparse Inverse Covariance Selection (SICS) copes with both problems by estimating the precision matrix under sparsity constraints such as an L1 penalty

[3] `http://zhoulab.usc.edu/\acrshort{sMBPLS}/`

[Dempster et al., 1977]. Given the data, the variables with the highest gaussian log likelihood of being conditionally independent could then be selected as the non-zero entries of the estimated matrix. Meinshausen and Bühlmann [2006] described an approximate solution to this problem by essentially running an L1-penalized LASSO regression [Tibshirani, 1996] on each row of the covariance matrix. The graphical lasso was proposed by Friedman et al. [2008] as an exact solution to the problem.

The Joint Graphical Lasso (JGL) [Danaher et al., 2014] is an extension of the graphical lasso that allows for the integration of multiple views of data by estimating multiple graphical models while encouraging similar topology and edge weights. The problem is solved by maximizing a penalized log-likelihood with generalized fused lasso or group lasso penalties, using an alternating directions method of multipliers algorithm. The authors were able to identify differential edges between a graphical model estimated from lung cancer gene expression data and from a control group. The JGL algorithm is implemented in an R package of the same name [Danaher, 2013]. Kling et al. [2015] recently proposed a multi-view generalization of the graphical lasso for modeling genome-wide networks of multiple cancers and data types. The method solves a log likelihood maximization problem with both a sparsity penalty and a network differential penalty for different cancer types, while concatenating data matrices of different omics levels, such as mRNA, miRNA, CNA, and methylation. The resulting networks have been shown to overlap well with known pathway interactions, and have been successfully used to detect novel targets against brain tumor stem cells [Kling et al., 2016]. While this "augmented SICS" method is very useful for integrative exploratory data analysis, the inevitable multicollinearity in the high-dimensional setting creates instability of estimation. Moreover, improved scalability will be needed to deal with future big data. The results of this method are publicly available online[4], and a Matlab implementation is part of the supplementary data[5].

2.1.3 Bayesian Approaches

Bayesian methods in multi-omics integration can make biologically informed assumptions on different views of the data with various underlying probability distributions, and also on the specific relationships between the distinct views.

COpy Number and EXpression In Cancer (CONEXIC) [Akavia et al., 2010] is a bayesian network-based method that integrates CNA and gene expression data to identify aberrations that promote disease progression in cancer. The algorithm produces a ranked list of candidate driver genes (modulators) by finding those which are correlated with respect to their differential expression and also present in significantly abberant regions. CNA data helps determine the direction of the the influence, which cannot be inferred from correlations in gene expression alone. The authors used a CONEXIC-derived list of modulators in Melanoma data together with gene set enrichment analysis (GSEA) [Subramanian et al., 2005] to identify

[4] `cancerlandscapes.org`

[5] `https://academic.oup.com/nar/article/43/15/e98/2414280#supplementary-data`

driver mutations of the disease and the processes that they influence.

Multiple Data set Integration (MDI) [Kirk et al., 2012] is a bayesian method that is capable of integrating data from a diverse range of data sets and data views simultaneously. The method uses a Dirichlet-Multinomial Allocation (DMA) mixture model [Green and Richardson, 2001], using parameters describing the concordance of the different data sets to capture relationships between these models. This implies that the clustering results in each view of the data have an influence on the clustering in all the other views. MDI is capable of discovering groups of genes that tend to cluster together in one, some or all of the views. This means that associations between different clusters of genes can be related to a specific subset of the underlying data views. The authors applied MDI to multiple Saccharomyces cerevisiae data sets and focused on finding groups of genes that are co-expressed while their protein products also appear in the same complex. A Matlab implementation of MDI is available online[6].

A related approach is Bayesian Consensus Clustering (BCC) [Lock and Dunson, 2013], which simultaneously models the dependence and the heterogeneity of the the different data views. Similarly to MDI, it is capable of producing separate clusterings which are encouraged to adhere to an overall consensus clustering. it employs finite dirichlet mixture models modified for multi-view data and a Gibbs sampling procedure for consensus clustering. The authors employ BCC for subtype discovery, using gene expression, methylation, miRNA and proteomic data from of breast cancer tumor samples. An R implementation of BCC is available on the author's webpage[7].

2.1.4 Network-Based Approaches

Unsupervised integrative network models focus on associations or interactions between different molecular variables, and can be used for tasks such as detecting important genes in pathways, discovering communities of highly connected variables, or describing disease-associated mechanisms (modules) [Huang et al., 2017].

PAthway Representation and Analysis by DIrect Reference on Graphical Models (PARADIGM) [Vaske et al., 2010] is a probabilistic graphical model for inferring genetic activities that are specific to patients, by making use of manually curated pathway interactions between genes. The pathway data is used to determine whether two entities such as protein-coding genes, small molecules, complexes, gene groups or abstract processes should normally be correlated positively or negatively. All entities are then used as nodes in a directed acyclic network, with pathway-derived edges representing either expected up- or down-regulation. The actual interaction between a pathway's entities is measured by comparing levels of genomic variables such as gene expression and CNAs. For each pathway, PARADIGM then calculates a patient-specific score that represents how much the actual interaction between its entities deviates from the expected value. The authors found that clustering GBM

[6] `http://www2.warwick.ac.uk/fac/sci/systemsbiology/research/software/`
[7] `http://people.duke.edu/%7Eel113/software.html`

patients based on their pathways that show statistically significant perturbations resulted in a set of clinically relevant subtypes with significantly different survival profiles. An implementation of PARADIGM is available online[8].

Similarity Network Fusion (SNF) [Wang et al., 2014] is a method that constructs patient-similarity matrices for each view of the data, and then fuses them into a patient-similarity network. SNF first constructs a k-Nearest-Neighbor (kNN) patient network from each different view's similarity matrix. Those "affinity matrices" are then used to fuse the similarity matrices using a nonlinear method based on message-passing theory [Pearl, 2014]. The patient-to-patient similarities are iteratively updated, until they converge to a common similarity network. In this process, strong similarities that are supported by all or most types of the data reinforce each other, while weak similarities that are considered noise fade away. The fused network can then be used to identify subtypes among the patients by employing any community detection algorithm, such as spectral clustering. SNF is also able to identify which omics views of the underlying data support the existence of any edge in the fused network, thus offering deeper insight into which biological mechanisms could be responsible for the existence of distinct subtypes. The authors applied SNF to methylation, mRNA and miRNA data of five different cancer data sets, and found that the method was able to discover subtypes with statistically significant survival profiles for all of them. The algorithm is available in the R package SNFtool [Wang et al., 2017] and a Matlab implementation is available on the author's webpage[9]. The mathematics behind Similarity Network Fusion are discussed in detail in the theory chapter (section 3.1) of this thesis.

Lemon-Tree [Bonnet et al., 2015] is an unsupervised algorithm aiming to reconstruct module networks. The method builds an ensemble of co-expressed gene clusters by repeatedly using a model-based Gibbs sampler [Joshi et al., 2007]. It then identifies consensus gene modules by clustering the pairwise frequencies of genes belonging to the same cluster with a spectral edge clustering algorithm [Michoel and Nachtergaele, 2012]. Then, an individual additional candidate regulator view such as miRNA, CNA and methylation can be added to the consensus module to estimate a regulatory score calculated by a decision tree approach [Joshi et al., 2009]. The authors claim that Lemon-Tree performs especially well when attempting to infer closely related short-path networks. Since regulatory scores are only able to be assigned individually to distinct omics views, Lemon-Tree is unable to consider causal relationships between different regulator types themselves. The software implementation of this method is available online under a public license[10].

2.1.5 Multiple Kernel and Multi-Step Approaches

Integrative methods that are carried out in multiple steps are often used to first infer relationships between different views of molecular data, and to then relate them to certain phenotypes [Ritchie et al., 2015]. Kernel methods make use of the so-called

[8] `http://sbenz.github.com/Paradigm`

[9] `http://compbio.cs.toronto.edu/\acrshort{SNF}/\acrshort{SNF}/Software.html`

[10] `http://lemon-tree.googlecode.com`

"kernel trick"[Theodoridis, 2008]: kernel functions allow these methods to operate in a high-dimensional, implicit feature space by only calculating the inner products between the corresponding images of all pairs of data [Hofmann et al., 2008]. Due to the fact that kernel-based data integration is often carried out step-wise, it makes sense to present the two together here [Huang et al., 2017]. We present two relevant examples in the following.

Multiple Kernel Learning for Dimensionality Reduction (MKL-DR) is a method that uses multiple kernels to learn features in a reduced-dimensional common subspace [Lin et al., 2011]. Speicher and Pfeifer [2015] proposed an extension to this method (called rMKL-LPP), which introduces a regularization term and uses Locality Preserving Projections (LPP) [He and Niyogi, 2004] to conserve the aggregate distance for each sample's kNN. The method is able to use a variety of kernels for each omics view to learn features in a shared, reduced dimension and then cluster multiple networks based on those features using a support vector machine (SVM) or other traditional clustering methods. Applying rMKL-LPP to GBM gene expression, methylation and miRNA data, the authors found six disease subtypes that captured similarities and differences in both established expression and methylation subtypes, and gave better p-values for survival analysis than both iCluster and SNF. A software implementation is available upon request from the authors[11].

CNAmet [Louhimo and Hautaniemi, 2011] is a multi-step integration method for CNA, methylation, and gene expression data, which aims to detect genes that are either amplified in terms of CNA and also upregulated by hypomethylation, or deleted and also downregulated by hypermethylation. In the first step CNAmet links expression values to CNA and methylation data using the signal-to-noise ratio statistic [Hautaniemi et al., 2004]. The second step consists of calculating a score that captures which genes' differential expressions are due to both changes in methylation and CNA. The last step then derives adjusted p-values of the scores using a permutation test. The authors used CNA, methylation and gene expression data from GBM and ovarian cancer tumor samples to demonstrate that this approach can help to characterize genes and to gain a better understanding of biological processes during disease progression. CNAmet is available as an R package of the same name under public license[12].

2.2 Multi-Level Hierarchical Community Detection in Large-Scale Complex Networks

Hierarchical structures are ubiquitous in human societies and they are often considered necessary for an efficient governance of large organizations [Bavelas, 1950, Weber, 1978, Frank, 1985, Van Vugt et al., 2008]. It has been pointed out in general that the presence of a hierarchical organization makes complex systems especially stable and robust, thus often resulting in a long-run evolutionary advantage [Simon,

[11] nora@mpi-inf.mpg.de or npfeifer@mpi-inf.mpg.de

[12] `http://csbi.ltdk.helsinki.fi/\acrshort{CNA}met`

1991]. It is therefore not surprising that hierarchy and self-similarity on multiple levels are often considered universal characteristics of complex biological networks [Girvan and Newman, 2002, Ravasz and Barabási, 2003, Song et al., 2005] and that hierarchy has been successfully explained as an emergent property of complex evolutionary processes [Clune et al., 2013, Alcocer-Cuarón et al., 2014, Mengistu et al., 2016]. The majority of community detection approaches, however, aim to find the single "best" partition of a network. For our purpose of finding associations between molecular variables in omics data sets, as well as for various other purposes, a method should instead be capable of recognizing hierarchical structures (if present), and detect the corresponding levels of hierarchy [Sales-Pardo et al., 2007, Clauset et al., 2007, 2008]. A convenient side-effect of the identification of such hierarchical structures is that they are particularly intuitive to scrutinize and interpret at different levels of hierarchy in the framework of exploratory big data analysis. Figure 2.2 shows an example of such an approach applied to a social network of public figures, where each node represents a person, and an edge is present between two nodes if there exists a link from one of the two people's Wikipedia page to the other one's [Biddulph, 2012]. Communities were detected using the Louvain algorithm (see section 2.2.1). We note, for example, a prominent (light blue) cluster of politicians, which is closely connected to the turquoise cluster of members of royal families. The upper left orange cluster is mainly composed of fictional superheros and comic figures. The clusters on the central to upper right are mainly formed by sportspeople, yet subdivided into different communities of mainly male athletes (blue), female athletes (purple) and specific disciplines such as football (yellow) or golf and baseball (orange). The green cluster on the top contains musicians such as Elvis Presley, Bob Dylan, Kanye West and Eminem. A further possible subdivision is also recognizable, with the former two and the latter two falling into different subcommunities. A cluster that is somewhat harder to interpret is the central yellow one, which groups William Shakespeare, Albert Einstein, Jesus and Adolf Hitler into the same community. Nevertheless it becomes clear from figure 2.2 that methods capable of finding multi-level hierarchical community structures in networks have great potential to help us gain a better understanding of the real-world processes underlying big data sets.

This section predominantly focuses on the most popular and relevant approaches that are able to identify multi-level hierarchical structures in large-scale complex networks. Section 2.2.1 introduces modularity-based approaches, with a focus on the Louvain method [Blondel et al., 2008]. Section 2.2.2 presents techniques based on information theory, with a focus on the Infomap algorithm [Rosvall and Bergstrom, 2008]. Section 2.2.3 elaborates on relevant spectral clustering approaches, with a focus on Multilevel Hierarchical Kernel Spectral Clustering (MHKSC) [Mall et al., 2014]. Section 2.2.4 covers methods that rely on the statistical significance of communities, with a focus on the Order Statistics Local Optimization Method (OSLOM) [Lancichinetti et al., 2010].

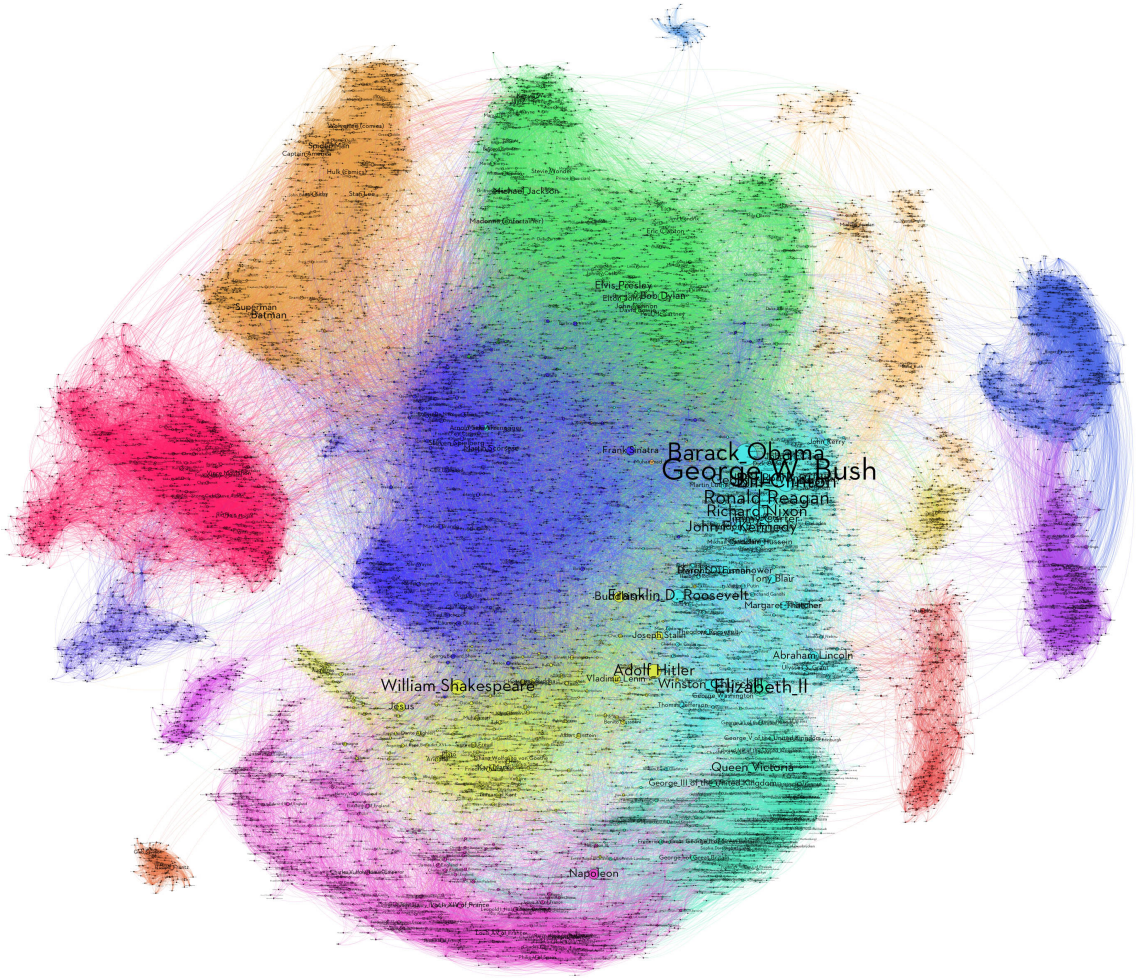


Figure 2.2: An example of communities detected in a large-scale social network using the Louvain algorithm, based on Wikipedia pages about famous public figures (nodes) and links between their pages (edges). The figure is reproduced from Biddulph [2012] under Creative Commons Attribution-ShareAlike 2.0 license (CC BY-SA 2.0).

2.2.1 Modularity Optimization Approaches

A popular choice for community detection is the maximization of the quality function *modularity* (Q) [Newman, 2004, Danon et al., 2005]. Modularity measures the difference between the density of edges within communities and the expected density for the same partition if edges were randomly distributed across the whole network. Since the exact optimization of modularity is a computationally hard problem, heuristics are required when dealing with large networks [Brandes et al., 2006]. Newman [2004] and Clauset et al. [2004] proposed greedy agglomerative hierarchical algorithms that start by assigning each node to different communities and then iteratively merge those communities that maximize the resulting difference in modularity ΔQ. Newman [2006] demonstrated that modularity can be expressed in terms of the eigenvectors of a network-characteristic matrix. This led to a faster and more effective spectral community detection method, which is, however, unable to find hierarchical structures in networks.

A heuristic modularity-based community detection technique for large scale networks, which is able to detect multiple layers of hierarchy is *Louvain*. It was proposed by Blondel et al. [2008] from the Université Catholique de Louvain, which has given the algorithm its name. Louvain consists of two steps. In the first step, all nodes are assigned a community using an efficient version of the agglomerative ΔQ-maximizing approach from [Clauset et al., 2004]. In the second step, a smaller new network is constructed, in which nodes are now the communities detected in the first step, and new edge weights are the sum of the weights of the edges between nodes belonging to the respective two communities in the original network [Arenas et al., 2007]. The two steps are then repeated until only a single community is found. The authors demonstrate that Louvain is able to find high-quality solutions in terms of modularity by applying the method to two large real-world networks – a Belgian mobile phone network comprised of 2 million users and different language communities, and a web graph of 118 million nodes. Louvain is publicly available online[13]. While Louvain performs well on coarse levels of hierarchy where communities are sufficiently large, it has been shown that the method suffers from a resolution limit, meaning that it is unable to find small communities below a certain size threshold even if they are unequivocally defined [Fortunato and Barthelemy, 2007, Good et al., 2010]. This is due to the fact that Louvain uses a global criterion to decide what the network structure would be like if edges were distributed at random. Networks with community structure, however, generally exhibit heterogeneity, with edge densities varying locally [Guimera et al., 2004, Reichardt and Bornholdt, 2006]. Approaches aiming to overcome the resolution problem by using modified definitions of modularity have been shown to still suffer from the resolution limit [Fortunato and Barthelemy, 2007].

[13] `https://sourceforge.net/projects/louvain/`

2.2.2 Methods Based on Information Theory

Information theory is mainly concerned with how information can be represented and quantified [Cover and Thomas, 2012]. An important topic in information theory is how data can be represented in a highly compressed way without loosing any important details. An example of this is a relatively long text message that says `"Hahahahahahaha!"`, but could be nearly perfectly compressed to `"ha(×7)!"`. In this example, the information to be compressed is given by a sequence of letters or other symbols. In network models, such sequences can be generated by a random walk on the network diagram by sequentially recording the nodes visited. Various community detection approaches rely on the idea that random walks on a network provide a proxy for information flow on its topology [Ziv et al., 2005, Pons and Latapy, 2005, Lai et al., 2010, Wang et al., 2013]. Such random walkers are statistically likely to spend long periods of time within certain highly interconnected communities that are less strongly connected to the rest of the network. This property can be used to find a (compressed) description of the random walk which is as short as possible. In this framework, it has been shown that community detection is equivalent to solving such compressed coding problems [Rissanen, 1978, Rosvall and Bergstrom, 2007].

Infomap [Rosvall and Bergstrom, 2008, Rosvall et al., 2009] is a community detection technique based on this information-theoretical framework. The method makes use of a modified version of Huffman coding [Huffman, 1952], which saves space by encoding common events using short codewords while using long codewords for rare events. To detect communities, the network is described on two levels: each community is assigned a unique identifier and all nodes are given a name that is unique within their community, but node names can be re-used across different communities. This is much like international phone numbers, where each country is assigned a distinct dialing code. While phone numbers without that dialing code are unique on each national level, they may not be on the international level if the country code is omitted. The Infomap algorithm then finds the community partition of the network that minimizes the expected description length of a random walk in this two-level coding framework.

In Rosvall and Bergstrom [2011], the authors extend the Infomap algorithm to allow for an arbitrary number of hierarchically nested index codebooks that specify movements between communities, sub-communities, sub-sub-communities, and so on. The resulting multi-level hierarchical community detection method – called the *hierarchical map equation* – minimizes the expected description length of a random walk across those multi-level coding frameworks. The solution to the problem then yields a multi-level hierarchical community structure of the network at hand. To solve the minimization problem at hand, the authors developed a fast stochastic and recursive search algorithm, which is implemented in C++ and available online[14]. An advantage of the approach based on random walks is that they can easily be described on both weighted and directed networks. The authors use their algorithm to discover hierarchical organizations in a journal citation network of science, the

[14] `http://www.mapequation.org/code.html`

global air traffic network, and the human disease network.

2.2.3 Kernel Spectral Clustering Methods

A popular choice in unsupervised learning are spectral clustering approaches [Chung, 1997, Ng et al., 2002, Zelnik-Manor and Perona, 2005], which obtain cluster assignments from an eigen-decomposition of the Laplacian matrix of a similarity measure between certain objects (see section 1.1.5.2). A difficulty with classical spectral clustering methods is the requirement to construct a full similarity matrix for all objects to be clustered, which limits such approaches to relatively small data sets. This challenge can be overcome by Kernel Spectral Clustering (KSC) [Alzate and Suykens, 2010], which relies on a formulation of kernel Principal Component Analysis (kPCA) [Schölkopf et al., 1998] in a dual-primal framework. Here, the "primal" and the "dual" refer to two different formulations of the same problem. This framework relies on the "kernel trick" introduced in section 2.1.5, where the primal optimization problem is formulated in a high-dimensional feature space, but the kernel method only implicitly operates in that space by solving an easier, equivalent problem in the dual. KSC formulates the primal weighted kPCA problem in the context of least squares support vector machines (LS-SVMs) [Suykens et al., 2002], which corresponds to an eigen-decomposition of a centered Laplacian matrix in the dual. This eigen-decomposition results in a clustering model in the dual. Due to its use of SVMs, the obtained model can provide cluster assignments for out-of-sample observations. This out-of-sample extension makes it possible to train a clustering model on a representative, smaller subset of the data at hand. KSC was first used to detect communities in networks by Langone et al. [2012], but the associated computationally expensive subset and model selection still rendered an application to large-scale networks prohibitive. The proposal of a fast and unique representative subset selection approach (FURS) [Mall et al., 2013] made it possible to apply KSC to complex big data networks [Mall et al., 2013,]. While a KSC-based approach for the detection of communities on levels of different resolution in large networks was proposed by Alzate and Suykens [2012], those levels are determined by user-defined values of a kernel parameter. Furthermore, communities identified on different levels do not generally form a natural hierarchy, with nodes that belong to the same cluster on one level being assigned to different clusters on a coarser level.

Multilevel Hierarchical Kernel Spectral Clustering (MHKSC) [Mall et al., 2014] is an agglomerative method that generates a natural multi-level hierarchical organization of large scale complex networks. The method creates two sub-networks that are representative of the entire large-scale network's hierarchical community structure by employing the FURS subsampling scheme. These two sub-networks are used as training and validation set, and are both about 15% of the size of the whole network in terms of their number of nodes. MHKSC first trains a predictive KSC model on the training set, which is able to project any possibly unseen node of the full network into a weighted kPCA eigenspace that is indicative of the network's community partition. The algorithm then uses the predictive KSC model on the nodes in the validation set, and creates an affinity matrix between those nodes'

projections in the eigenspace by calculating their pairwise cosine distances. On the ground level of the hierarchy, an initial distance threshold with respect to the entries of the affinity matrix is defined. The node that has most neighbors within that distance is identified, and that node together with all its neighbors form a community. The indices corresponding to that community are removed from the affinity matrix, and the procedure is repeated until the affinity matrix is empty. Just as for Louvain, communities on the lower level of hierarchy are then treated as nodes on the next-coarser level. For MHKSC, the affinity between two clusters is taken as the average of all pairwise distances between them. The new distance threshold is then chosen to be the mean of the minimum distances to each cluster's closest neighboring cluster. This agglomerative procedure is repeated until there is only one community on the coarsest level of hierarchy. The resulting monotonously increasing sequence of distance thresholds obtained from the validation set is then used on the cosine distance based affinity matrix of the entire network. This results in an identification of the multi-level hierarchical community structure of the whole network at hand. An implementation of MHKSC using both Matlab and Phyton is available online[15]. The mathematics behind MHKSC are discussed in detail in the theory chapter (section 3.2) of this thesis.

2.2.4 Techniques Based on Statistical Significance

Methods focused on the statistical significance of clusters are based on the fact that mere random fluctuations can account for larger-than-usual concentrations of edge weights within some groups of nodes, which then clearly do not represent any meaningful communities. Due to such fluctuations, many common community detection methods identify communities even in purely random graphs [Hu et al., 2010]. The issue of statistical significance of communities in networks is a research topic that has emerged only recently, and few approaches have been suggested to define such statistical significance or to use it in community detection algorithms [Spirin and Mirny, 2003, Reichardt and Leone, 2008, Bianconi et al., 2009, Lancichinetti et al., 2010].

The Order Statistics Local Optimization Method (OSLOM) [Lancichinetti et al., 2010] is a stochastic community detection method that aims to avoid the resolution limit problem of modularity-based approaches by relying on the statistical significance of communities. It is based on the proposal presented by Lancichinetti et al. [2010], which uses extreme and order statistics to define the significance of a node cluster as the probability that a general community detection algorithm finds such a cluster in a random network. OSLOM takes a user-defined significance threshold as input. The method starts with a single-community significance analysis. To that end, an initial single node in the network is picked at random, and then a certain number of neighbors that are considered most significant are added to the community. The algorithm then performs a two-step stochastic "clean-up procedure" on the community, which is roughly described below. In step 1, it considers whether it is possible to increase the significance of the community by adding external nodes. In

[15] `https://www.esat.kuleuven.be/stadius/ADB/mall/software\acrshort{MHKSC}.php`

step 2, non-significant nodes are pruned. The community is considered significant by the clean-up procedure if it results in a non-empty set of nodes, and insignificant otherwise. To obtain a robust estimate, the clean-up procedure is repeated multiple times, and the overall significance of the single community is determined by the majority of outcomes. To obtain multiple communities and to explore different regions of the network, the above-described single-community significance analysis is performed repeatedly for different randomly picked initial nodes, until similar communities are found over and over again. The result is a set of usually overlapping communities. Those communities are then reduced to significant minimal communities, which means that they exhibit no significant internal community structure themselves. To obtain a reasonable cover of the network, unions of minimal communities are checked for significant sub-communities and merged if none exist. In addition, out of highly similar communities the larger ones are picked. The whole procedure up to here is then again repeated multiple times to find a consensus cover of the network. This consensus cover represents the solution to the OSLOM algorithm.

The OSLOM method is able to handle directed, weighted, and time-dependent (dynamic) networks, and it can identify overlapping communities. It is also capable of detecting multi-level hierarchical structures by employing the same strategy as the Louvain method, where community detection is repeatedly applied to networks on coarser levels of hierarchy, in which nodes correspond to the communities on the closest finer level. Again, this multi-level hierarchy detection procedure is repeated multiple times in OSLOM to provide a stable consensus solution. A C++ implementation of the OSLOM method is available online[16].

[16] `http://www.oslom.org`

3

Theory

"Mathematics is the art of giving the same name to different things."

– Jules Henri Poincaré

The exploratory data analysis approach proposed in this thesis makes use of the Similarity Network Fusion (SNF) method that was briefly introduced in section 2.1.4, as well as the Order Statistic Local Optimization Method (OSLOM) whose foundations were presented in section 2.2.4. Multi-Level Hierarchical Kernel Spectral Clustering (MHKSC), which was introduced in section 2.2.3, was also considered for use on cancer data, and compared to OSLOM on simulated data. Therefore, this chapter describes the underlying theoretical details of all three approaches. Section 3.1 discusses the mathematics of SNF. Section 3.2 provides a rigorous treatment of MHKSC. Section 3.3 gives the theoretical details behind the assessment of significance of communities and its use in OSLOM.

3.1 Similarity Network Fusion

SNF [Wang et al., 2014] is an integrative technique that is able to combine any set of similarity matrices that are based on different views of a given data set by taking advantage of both common and complementary information provided by distinct views of the same data set. The method makes use of parallel cross-diffusion processes, across kNN similarity graphs of the different views of the data, which is inspired by a multi-view learning framework originally developed for computer vision and image processing. The exact procedure for fusing the different similarity matrices is rigorously described below, mostly following the notation in Wang et al. [2014].

3.1.1 Construction of Similarity Matrices

Consider a data set that consists of m different views $\phi = (1, 2, \ldots, m)$. Each view is represented by an $n \times d^{(\phi)}$ data matrix, where n is the total number of patients $\{1, 2, \ldots, n\}$, and $d^{(\phi)}$ is the the dimension of view ϕ, i.e. the view-specific number

of variables that have been measured for all patients. The data of patient i in view ϕ is denoted by $\boldsymbol{x}_i^{(\phi)} \in \mathbb{R}^{d^{(\phi)}}$, here assuming that all data is continuous. A patient similarity network based on view ϕ of the data is then represented by a graph $G^{(\phi)}(\mathcal{V}, E^{(\phi)})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ are the nodes representing the set of patients, and $E^{(\phi)} \subseteq \mathcal{V} \times \mathcal{V}$ are the view-specific edges, which are weighted by how similar the patients are according to view ϕ. Hence, edge weights for each view can be represented by $n \times n$ similarity matrices $W^{(\phi)}$, where each entry $W_{ij}^{(\phi)}$ represents the similarity between patients i and j in view ϕ. To calculate $W^{(\phi)}$, the scaled exponential Radial Basis Function (RBF) similarity kernel

$$W_{ij}^{(\phi)} = \exp\left(-\frac{\boldsymbol{x}_i^{(\phi)^2}\boldsymbol{x}_j^{(\phi)^2}}{\mu\epsilon_{ij}^{(\phi)}}\right), \tag{3.1}$$

is used, where μ is a parameter that can be set empirically, and $\epsilon_{ij}^{(\phi)}$ is a scaling parameter defined as

$$\epsilon_{ij}^{\phi} = \frac{\frac{1}{n-1}\left(\sum\limits_{\substack{l=1\\l\neq i}}^{n}\sqrt{\boldsymbol{x}_i^{(\phi)^2}\boldsymbol{x}_l^{(\phi)^2}} + \sum\limits_{\substack{l=1\\l\neq j}}^{n}\sqrt{\boldsymbol{x}_j^{(\phi)^2}\boldsymbol{x}_l^{(\phi)^2}}\right) + \sqrt{\boldsymbol{x}_i^{(\phi)^2}\boldsymbol{x}_j^{(\phi)^2}}}{3}. \tag{3.2}$$

In other words, ϵ_{ij}^{ϕ} is given by the average of the mean euclidean distance between patient i and all its neighbors, the mean euclidean distance between patient j and all its neighbors, and the euclidean distance between patients i and j. The authors recommend to choose $\mu \in [0.3, 0.8]$. While using the euclidean distance often makes sense for continuous variables, the authors suggest using the chi-squared distance for discrete variables and an agreement-based measure for binary ones.

3.1.2 Fusion of Similarities

For the computation of a fused similarity, first a full and sparse kernel on the nodes $\mathcal{V}$ is defined for each view by

$$P_{ij}^{(\phi)} = \begin{cases} \dfrac{W_{ij}^{(\phi)}}{2\sum_{\substack{k=1\\k\neq i}}^{n} W_{ik}^{(\phi)}} & j \neq i \\ \dfrac{1}{2} & j = i. \end{cases} \tag{3.3}$$

The above normalization is chosen since it is free of the scale of self-similarity on the diagonal, and still satisfies $\sum_{j=1}^{n} P_{ij}^{(\phi)} = 1$. Denote by $N_{i,k}^{(\phi)}$ the set of v_i's k nearest neighbors in $G^{(\phi)}$, including itself. Then, for each view ϕ, local affinity is measured by the kNN affinity matrix

$$S_{ij}^{(\phi)} = \begin{cases} \dfrac{W_{ij}^{(\phi)}}{2\sum_{l\in N_{i,k}^{(\phi)}} W_{il}^{(\phi)}} & j \in N_{i,k}^{(\phi)} \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Here, the similarity of non-neighboring points is set to zero. The motivation behind this is that strong, local similarities between nodes are more reliable in noisy data than remote ones. Similarities to non-neighbors are then inferred by network diffusion. We now consider an iterative diffusion process, starting with $P^{(\phi)}(t_0) = P^{(\phi)}\ \forall \phi \in \{1, 2, \ldots, m\}$ at time step t_0, as defined in equation 3.3. We refer to $P^{(\phi)}(t)$ as the status matrix of the view ϕ at iteration t. SNF now repeatedly updates the status matrices $P^{(\phi)}$ for all views ϕ according to the recurrence relation

$$P^{(\phi)}(t+1) = S^{(\phi)} \times \left(\frac{\sum_{\substack{\xi=1 \\ \xi \neq \phi}}^{m} P^{(\xi)}(t)}{m-1} \right) \times S^{(\phi)^T} \quad \forall \phi \in \{1, 2, \ldots, m\}. \tag{3.5}$$

This approach updates all the status matrices at each time step, thus generating m parallel diffusion processes, which makes the status matrices of the different views increasingly similar with more iterations. Since the $S^{(\phi)}$ are kNN graphs of the $P^{(\phi)}(t_0)$, similarity information of two nodes propagates only through the common neighborhood, which makes SNF robust to noise and captures local structures of the similarity networks. After each iteration, the $P^{(\phi)}(t+1)$ are normalized using the same procedure as in equation 3.3. This ensures that throughout the SNF diffusion process each patient is always most similar to himself, and that the final fused network is full rank. Furthermore, the authors found that such normalization of the status matrices leads to faster convergence of SNF. After a user-defined number of iterations $t_{\max}$, the final, fused status matrix is calculated as

$$P = \frac{\sum_{\xi=1}^{m} P^{(\xi)}(t_{\max})}{m}. \tag{3.6}$$

The authors showed that the status matrices corresponding to the different views of the data usually converge within a few iterations of SNF. They recommend running SNF for about 10-20 iterations on real-world data. The final, fused status matrix P can then be used for further unsupervised machine learning tasks. This cross-diffusion process, which is here used for integrative omics analysis, is inspired by the theoretical multi-view learning framework that was originally developed for applications in computer vision and image processing [Wang et al., 2012]. In this thesis, we use a step-wise procedure that uses a sparse version of the output from SNF as the input for multi-level hierarchical community detection using OSLOM.

3.2 Multi-Level Hierarchical Kernel Spectral Clustering

MHKSC [Mall et al., 2014] is a kernel method that finds multi-level hierarchical structures in large, complex networks. It relies on the "kernel trick" by formulating a primal weighted kPCA problem in the context of LS-SVMs, and solving it as an eigen-decomposition of a centered Laplacian matrix in the dual, resulting in a community detection model with a powerful out-of-sample extension. The model is

trained on a representative subset of the whole network, and subsequently validated on another representative subset. A partition of the entire network into communities is then inferred using the out-of-sample extension based on the validation results. The hierarchical structure is inferred in an agglomerative fashion, thereby providing increasing distance thresholds that define the different levels of hierarchy. The exact algorithm is rigorously described below, mostly following the notation in Mall et al. [2014].

3.2.1 Predictive Kernel Spectral Clustering

This section describes the predictive Kernel Spectral Clustering (KSC) model, which is used as part of MHKSC, but is itself not designed to find hierarchical structures.

3.2.1.1 Representative Subset Selection

Consider a network $G(\mathcal{V}, E)$, where the set of nodes is denoted by $\mathcal{V}$ and the set of edges is denoted by $E \subseteq \mathcal{V} \times \mathcal{V}$. Let $|\mathcal{V}|$ denote the cardinality of $\mathcal{V}$, i.e. the number of nodes in the network. The first step of MHKSC consists of dividing the network G into a training set $\mathcal{V}_{\text{train}}$, a validation set $\mathcal{V}_{\text{valid}}$, and a test set $\mathcal{V}_{\text{test}}$. The sizes of the training and validation set are fixed at $|\mathcal{V}_{\text{train}}| = |\mathcal{V}_{\text{valid}}| = \lceil 0.15\,|\mathcal{V}| \rceil$, based on experimental results [Leskovec and Faloutsos, 2006]. The two sets are selected using Fast and Unique Representative Subset selection [Mall et al., 2013], which greedily selects nodes with high degree centrality. Such nodes are usually located in central regions of communities rather than in the periphery, and are therefore representative of the network's inherent community structure [Kang and Faloutsos, 2011]. MHKSC uses FURS to first select $\mathcal{V}_{\text{train}}$ from G, and then employs FURS again to select $\mathcal{V}_{\text{valid}}$ from nodes $\mathcal{V} \setminus \mathcal{V}_{\text{train}}$, yet based on the topology of the entire network G. Hence, both the training and the validation set are chosen to be representative of the community structure of the large-scale network. $\mathcal{V}_{\text{test}}$ is taken to be the entire network. Let $\boldsymbol{x}_i \in \mathbb{R}^N$ denote the adjacency list of node v_i, which contains the weights of edges connecting v_i to nodes $v_j \in \mathcal{V} = \left\{v_1, v_2, \ldots, v_{|\mathcal{V}|}\right\}$, and zero if no edge exists. Then the data underlying the training set $\mathcal{D}_{\text{train}} = \{\boldsymbol{x}_i\}_{i=1}^{|\mathcal{V}_{\text{train}}|}$ can then be efficiently used in memory since $|\mathcal{V}_{\text{train}}| \ll |\mathcal{V}|$ and since real-world networks are usually sparse.

3.2.1.2 Primal Formulation

We now present the primal formulation of the weighted kPCA using the "kernel trick". In the following, bold variables represent column vectors. Let $K(\cdot, \cdot)$ be a kernel function with an associated feature map $\varphi \colon \mathbb{R}^{|\mathcal{V}|} \to \mathbb{R}^{d_h}$. Here, $\mathbb{R}^{d_h}$ is a kernel-induced high-dimensional inner product space that the method implicitly operates in. We can define a kernel matrix Ω with respect to $\mathcal{D}_{\text{train}}$ by letting $\Omega_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x}_j)$ since that makes $K(\cdot, \cdot)$ a proper inner product on $\mathbb{R}^{d_h}$ and therefore ensures positive definiteness. Let $D_\Omega \in \mathbb{R}^{|\mathcal{V}_{\text{train}}| \times |\mathcal{V}_{\text{train}}|}$ be the diagonal and positive degree matrix associated to Ω, which is used as weighing

matrix for kPCA. Weighted kPCA then finds directions $\boldsymbol{w}$ in which the accordingly weighted variance of the projected variables $\boldsymbol{w}^T\varphi(\boldsymbol{x_i})$ is maximized. The choice of D_Ω for weighting is based on the idea that high-degree nodes are generally more representative of the community structure in a network [Alzate and Suykens, 2012]. Let $\Phi = \begin{bmatrix}\varphi(\boldsymbol{x}_1) & \varphi(\boldsymbol{x}_2) & \dots & \varphi(\boldsymbol{x}_{|\mathcal{V}_{\text{train}}|})\end{bmatrix}^T$ be the $|\mathcal{V}_{\text{train}}| \times d_h$ feature matrix with respect to $\mathcal{D}_{\text{train}}$. Then, considering a maximum amount $k^{(\max)}$ of eigenvectors that we want to include, the primal formulation of the weighted kPCA problem becomes

$$\begin{aligned}\min_{\boldsymbol{w}^{(l)},\boldsymbol{e}^{(l)},b_l} \quad & \left(\frac{1}{2}\sum_{l=1}^{k^{(\max)}-1}\boldsymbol{w}^{(l)^T}\boldsymbol{w}^{(l)} - \frac{1}{2\,|\mathcal{V}_{\text{train}}|}\sum_{l=1}^{k^{(\max)}-1}\gamma_l\boldsymbol{e}^{(l)^T}D_\Omega^{-1}\boldsymbol{e}^{(l)}\right) \\ \text{such that} \quad & \boldsymbol{e}^{(l)} = \Phi\boldsymbol{w}^{(l)} + b_l\mathbf{1}_{|\mathcal{V}_{\text{train}}|}, \quad l = 1,2,\dots,k^{(\max)}-1\end{aligned} \tag{3.7}$$

where $\boldsymbol{e}^{(l)} = \begin{bmatrix}\boldsymbol{e}_1^{(l)} & \boldsymbol{e}_2^{(l)} & \dots & \boldsymbol{e}_{|\mathcal{V}_{\text{train}}|}^{(l)}\end{bmatrix}^T$ are the projections onto the eigenspace, b_l are bias terms, $\gamma_l \in \mathbb{R}_+$ are regularization parameters, and the indices $l = 1,\dots,k^{(\max)}-1$ represent the number of score variables required to encode the $k^{(\max)}$ clusters [Alzate and Suykens, 2010]. Here, $\mathbf{1}_n$ denotes a of length n whose entries all all equal to one. Now, the clustering model in the primal is given by

$$e_i^{(l)} = \boldsymbol{w}^{(l)^T}\varphi(\boldsymbol{x}_i) + b_l, \quad i = 1,2,\dots,|\mathcal{V}_{\text{train}}|\,. \tag{3.8}$$

For each node $v_i \in \mathcal{V}_{\text{train}}$, its community membership is then binarily encoded in the sequence $\left\{\text{sign}\left(e_i^{(l)}\right)\right\}_{l=1}^{k^{(\max)}-1}$. Indicators for any user-defined number $k \leq k^{(max)}$ of communities can, for example, be obtained from these sequences using the Error Correcting Codes (ECOC) method [Baylis, 1997]. A technique that tunes k in MHKSC is presented in section 3.2.1.4.

MHKSC uses the kernel function $K\colon \mathbb{R}^{|\mathcal{V}|} \times \mathbb{R}^{|\mathcal{V}|} \to \mathbb{R}$ defined by $K(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\ \|\boldsymbol{y}\|}$, which means that the pairwise cosine similarities between adjacency lists provide us with the entries $\Omega_{ij} = \frac{\boldsymbol{x}_i^T\boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\ \|\boldsymbol{y}_j\|}$ of the kernel matrix. Using this positive definite normalized linear kernel function K is convenient, as it means that it is not necessary to choose any user-defined kernel parameter. In this framework, we can set $d_h = |\mathcal{V}|$ and utilize the explicit expression of the underlying feature map φ for large networks [Mall et al., 2013].

3.2.1.3 Dual Formulation

We now relate the primal formulation above to the dual problem. Consider the Lagrangian

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{w}^{(l)},\boldsymbol{e}^{(l)},b_l,\boldsymbol{\alpha}^{(l)}\right) = {} & \frac{1}{2}\sum_{l=1}^{k^{(\max)}-1}\boldsymbol{w}^{(l)^T}\boldsymbol{w}^{(l)} - \frac{1}{2\,|\mathcal{V}_{\text{train}}|}\sum_{l=1}^{k^{(\max)}-1}\gamma_l\boldsymbol{e}^{(l)^T}D_\Omega^{-1}\boldsymbol{e}^{(l)} \\ & - \sum_{l=1}^{k^{(\max)}-1}\boldsymbol{\alpha}^{(l)^T}\left(\boldsymbol{e}^{(l)} - \Phi\boldsymbol{w}^{(l)} - b_l\mathbf{1}_{|\mathcal{V}_{\text{train}}|}\right),\end{aligned} \tag{3.9}$$

of the primal optimization problem 3.7, with Lagrange multipliers $\boldsymbol{\alpha}^{(l)}$. Given a positive definite kernel function, positive regularization Karush-Kuhn-Tucker (KKT) optimality conditions [Karush, 1939, Kuhn et al., 1951], we obtain

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}^{(l)}} = 0 &\implies \boldsymbol{w}^{(l)} = \Phi^T \boldsymbol{\alpha}^{(l)} \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{e}^{(l)}} = 0 &\implies \boldsymbol{\alpha}^{(l)} = \frac{\gamma_l}{|\mathcal{V}_{\text{train}}|} D_\Omega^{-1} \boldsymbol{e}^{(l)} \\
\frac{\partial \mathcal{L}}{\partial b_l} = 0 &\implies \mathbf{1}_{|\mathcal{V}_{\text{train}}|}^T \boldsymbol{\alpha}^{(l)} = 0 \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}^{(l)}} = 0 &\implies \boldsymbol{e}^{(l)} = \Phi \boldsymbol{w}^{(l)} + b_l \mathbf{1}_{|\mathcal{V}_{\text{train}}|}.
\end{aligned} \tag{3.10}$$

Solving for the bias terms, we obtain

$$b_l = -\frac{1}{\mathbf{1}_{|\mathcal{V}_{\text{train}}|}^T D_\Omega^{-1} \mathbf{1}_{|\mathcal{V}_{\text{train}}|}} \mathbf{1}_{|\mathcal{V}_{\text{train}}|}^T D_\Omega^{-1} \Omega \boldsymbol{\alpha}^{(l)}, \quad l = 1, 2, \ldots, k^{(\max)} - 1, \tag{3.11}$$

where we make use of the fact that $\Phi\Phi^T = \Omega$. To be able to account for the bias terms in condensed matrix notation, we define the centering matrix

$$M_D = I_{|\mathcal{V}_{\text{train}}|} - \frac{\mathbf{1}_{|\mathcal{V}_{\text{train}}|} \mathbf{1}_{|\mathcal{V}_{\text{train}}|}^T D_\Omega^{-1}}{\mathbf{1}_{|\mathcal{V}_{\text{train}}|}^T D_\Omega^{-1} \mathbf{1}_{|\mathcal{V}_{\text{train}}|}}. \tag{3.12}$$

Eliminating the primal variables $\boldsymbol{w}^{(l)}$, $\boldsymbol{e}^{(l)}$, and b_l from equations 3.10, we conclude that the KKT conditions of equation 3.9 are satisfied by the eigenvectors of

$$D_\Omega^{-1} M_D \Omega \boldsymbol{\alpha}^{(l)} = \lambda_l \boldsymbol{\alpha}^{(l)}. \tag{3.13}$$

Since the KKT conditions are necessary but not sufficient for optimization of a not necessarily convex objective function, the relevant components from the solution for the eigenvectors $\alpha^{(l)}$ need to be selected. Considering this, equation 3.13 is the dual eigenproblem corresponding to the primal formulation in equation 3.7, where the eigenvectors $\boldsymbol{\alpha}^{(l)}$ are the dual variables. The out-of-sample predictive model in the dual for any unseen node $v \in \mathcal{V} \setminus \mathcal{V}_{\text{train}}$ with adjacency list $\boldsymbol{x}$ is then given by

$$\hat{\boldsymbol{e}}^{(l)}(\boldsymbol{x}) = \sum_{i=1}^{(l)} \alpha_i^{(l)} K(\boldsymbol{x}, \boldsymbol{x}_i) + b_l. \tag{3.14}$$

The predictive model can be validated using the representative subset $\mathcal{V}_{\text{valid}} \subset \mathcal{V}$, or it can be utilized to detect communities in the entire network by applying it to $\mathcal{V}_{\text{test}} \subset \mathcal{V}$.

3.2.1.4 Optimal Number of Clusters

Note that while the above-described KSC method derives a clustering model that allows us to identify at most $k^{(\max)}$ communities in the network, it does not specify which number of clusters is optimal. To overcome this, Mall et al. [2013] presented

a self-tuned approach for the selection of the number of clusters k. The predictive KSC model in the dual (equation 3.14) can be used on $\mathcal{V}^{\text{valid}}$ to obtain the latent variable matrix $P_{\text{valid}} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_{|\mathcal{V}_{\text{valid}}|} \end{bmatrix}^T$. Using P_{valid}, the authors create the non-negative affinity matrix

$$(A_{\text{valid}})_{ij} = 1 - \frac{\boldsymbol{e}_i^T \boldsymbol{e}_j}{\|\boldsymbol{e}_i\| \ \|\boldsymbol{e}_j\|}, \tag{3.15}$$

which assigns a value to each pair of nodes in the validation set by calculating the cosine distance of their projections in eigenspace. This means that the entries of A_{valid} corresponding to pairs of nodes which belong to the same community are very small. Hence, the affinity matrix is approximately block-diagonal. To identify the block-diagonal structure, a distance threshold t is defined. First, we find the index

$$\underset{i}{\text{argmax}} \left(\left\{ \sum_{j=1}^{|\mathcal{V}_{\text{valid}}|} \mathbb{1}\left(t - (A_{\text{valid}})_{ij}\right) \right\}_{i \in \{1,2,\dots,|\mathcal{V}_{\text{valid}}|\}} \right), \tag{3.16}$$

of the node v_i, which has most nodes at a cosine distance $< t$ in the projected eigenspace. Here, $\mathbb{1}(\cdot)$ denotes the function that takes the value 1 for positive numbers and the value 0 otherwise. Then, all nodes for which the cosine distance to v_{i} is smaller than t are considered to be in the same community. The rows and columns corresponding to all the nodes in this community are removed from A_{valid}, and the whole procedure is repeated until the affinity matrix is empty. For threshold t, denote by $k^{(t)}$ the number of steps that this procedure took and denote by $\boldsymbol{s}^{(t)} = \left(s_1^{(t)}, s_2^{(t)}, \dots, s_k^{(t)}\right)$ the sizes of the communities that were removed from the matrix. The ideal number of communities is found by using the notions of entropy and balance with respect to the size of the communities. The Shannon entropy for a set of communities of sizes $\boldsymbol{s}$ is defined as

$$H(\boldsymbol{s}) = -\sum_{i=1}^{k} \frac{s_i}{|\mathcal{V}_{\text{valid}}|} \log\left(\frac{s_i}{|\mathcal{V}_{\text{valid}}|}\right). \tag{3.17}$$

The balance can be defined as

$$B(\boldsymbol{s}) = \sum_{i=1}^{k} \frac{s_i}{\max(\boldsymbol{s})}. \tag{3.18}$$

To measure the quality of a community partition, the harmonic mean of entropy and balance is used. This so-called F-measure is defined as

$$F(\boldsymbol{s}) = \frac{2H(\boldsymbol{s})B(\boldsymbol{s})}{H(\boldsymbol{s}) + B(\boldsymbol{s})}. \tag{3.19}$$

Then, a set of candidate distance thresholds $T = \{0.1, 0.2, \dots, 1\}$ is taken into consideration, and the optimal number of communities is calculated as

$$\underset{k}{\text{argmax}} \left(\left\{ F\left(\boldsymbol{s}^{(t)}\right) \right\}_{t \in T} \right). \tag{3.20}$$

This approach is capable of determining the optimal number of clusters for a flat partition of the network. An approach for extraction of multi-level hierarchical structure is presented in the next section.

3.2.2 Muli-Level Hierarchy Detection

MHKSC is a bottom-up, agglomerative scheme that uses the above-described KSC methodology and extracts a monotonously increasing sequence of distance thresholds $\mathcal{T}$, which determine the partition of the network into communities at different levels of hierarchies. We here refer to the affinity matrix of the eigenprojections of the nodes in $\mathcal{V}_{\text{valid}}$ on the ground level with most fine-grained community partition as $A^{(0)}_{\text{valid}}$.

3.2.2.1 Selection of Distance Thresholds

The distance thresholds are evaluated using P_{valid}, which was obtained from the predictive model trained on $\mathcal{V}_{\text{train}}$. The threshold at ground level $t^{(0)}$ is set to a fixed value that results in a desired number of clusters at the finest level of hierarchy. While $t^{(0)}$ is technically a subjective, user-defined value, the authors in Mall et al. [2014] empirically found that $t^{(0)} \in [0.1, 0.2]$ works well for large-scale networks. If $t^{(0)} \ll 0.1$, the resulting community partition will include many singletons at the ground level, whereas for $t^{(0)} \gg 0.2$ most nodes would fall into one community, resulting in one huge connected component. However, the exact value of $t^{(0)}$ should be chosen to obtain a desired granularity at the ground level hierarchy, based on the research goal at hand. Given a fixed value of $t^{(0)}$, first the node v_i with maximal $\sum_{j=1}^{|\mathcal{V}_{\text{valid}}|} \mathbb{1}\left(t^{(0)} - \left(A^{(0)}_{\text{valid}}\right)_{ij}\right)$ is identified. Then, the first cluster at the ground level is defined as

$$\mathcal{C}^{(0)}_{1,\text{valid}} = \left\{ v_j \;\middle|\; \left(A^{(0)}_{\text{valid}}\right)_{ij} < t^{(0)} \right\}. \tag{3.21}$$

The rows and columns corresponding to the nodes in $\mathcal{C}^{(0)}_{1,\text{valid}}$ are then removed from $A^{(0)}_{\text{valid}}$ while keeping the indices of the entire affinity matrix. This procedure is repeated until the affinity matrix is empty, thus resulting in a ground-level community partition $\mathfrak{C}^{(0)}_{\text{valid}} = \left\{\mathcal{C}^{(0)}_{1,\text{valid}}, \mathcal{C}^{(0)}_{2,\text{valid}}, \ldots, \mathcal{C}^{(0)}_{k^{(0)},\text{valid}}, \right\}$, where $k^{(0)}$ denotes the number of communities on this finest level of hierarchy. To identify communities on the next coarser level of hierarchy, a new network is constructed by treating each community on the lower level as a node, with distances between clusters being defined as the mean distance between their nodes. At each level h of the hierarchy, this results in an affinity matrix $A^{(h)}_{\text{valid}}$, which is defined by

$$\left(A^{(h)}_{\text{valid}}\right)_{ij} = \frac{\sum_{m \in \mathcal{C}^{(h-1)}_{i,\text{valid}}} \sum_{l \in \mathcal{C}^{(h-1)}_{j,\text{valid}}} \left(A^{(h-1)}_{\text{valid}}\right)_{ml}}{\left|\mathcal{C}^{(h-1)}_{i,\text{valid}}\right| \times \left|\mathcal{C}^{(h-1)}_{j,\text{valid}}\right|}. \tag{3.22}$$

The distance threshold $t^{(h)}$ on hierarchy h is then taken to be the mean of the minimum distances

$$t^{(h)} = \frac{1}{k^{(h)}} \sum_{i=1}^{k^{(h)}} \min_j \left(\left\{ \left(A^{(h)}_{\text{valid}}\right)_{ij} \;\middle|\; j \neq i \right\} \right). \tag{3.23}$$

This process is repeatedly applied until there is only one single cluster on the coarsest level of hierarchy $h^{(max)}$. This provides us with a monotonously increasing sequence

of distance thresholds $\mathcal{T} = \left\{t^{(0)}, t^{(1)}, \ldots, t^{(h^{(max)})}\right\}$ and with the corresponding hierarchy of community partitions $\mathfrak{C}_{\text{valid}} = \left\{\mathfrak{C}^{(0)}_{\text{valid}}, \mathfrak{C}^{(1)}_{\text{valid}}, \ldots, \mathfrak{C}^{(h^{(max)})}_{\text{valid}}\right\}$.

3.2.2.2 Identification of Communities for the Whole Network

As noted in section 3.2.1.1, the set $\mathcal{V}_{\text{valid}}$ is a representative subset of the entire network $\mathcal{V}$. Hence, it is possible to use the distance thresholds $\mathcal{T}$ to infer a multi-level hierarchical structure on the entire network. To render the method self-tuned, the thresholds $t^{(h)} \in \left\{t^{(1)}, t^{(2)}, \ldots, t^{(h^{(max)})}\right\} > t^{(0)}$ are used. The detection of each community partition is then essentially done in the same manner as on the validation set. However, since the entire affinity matrix $A^{(1)}_{\text{test}}$ on the ground level of the hierarchy of the test network is often too large to be stored in memory for large-scale networks, MHKSC employs a greedy search to find a node v_i with as large as possible $\sum_{j=1}^{|\mathcal{V}_{\text{test}}|} \mathbb{1}\left(t^{(1)} - \left(A^{(1)}_{\text{test}}\right)_{ij}\right)$. Then, the first community is again defined as the set of nodes that have a distance smaller than $t^{(1)}$ to v_i. The only exception to that is given if a resulting community is too large to store its entire affinity matrix in memory. In that case, the maximal community size on the ground level is limited to the maximum possible number $n^{(1)}_{\text{max}}$ of nodes for which an affinity matrix can be stored, and only the $n^{(1)}_{\text{max}}$ closest nodes to v_i are added to the community. Besides these constraints, the hierarchy of community partitions on the entire network $\mathfrak{C}_{\text{test}} = \left\{\mathfrak{C}^{(1)}_{\text{test}}, \mathfrak{C}^{(2)}_{\text{test}}, \ldots, \mathfrak{C}^{(h^{(max)})}_{\text{test}}\right\}$ is constructed in the same way as was done on the validation set.

3.3 Order Statistics Local Optimization Method

The Order Statistics Local Optimization Method (OSLOM) [Lancichinetti et al., 2010] is a community detection method aiming to avoid the resolution limit problem of modularity-based approaches by relying on the local statistical significance of communities within their neighborhoods. The OSLOM method is able to handle directed, weighted, and time-dependent (dynamic) networks, and it identifies overlapping communities. The multi-level hierarchical structure is inferred in an agglomerative manner, where community detection is repeatedly applied to networks in which nodes correspond to communities on the closest finer level of the hierarchy. The exact algorithm is rigorously described below, mostly following the notation in Lancichinetti et al. [2010].

3.3.1 Statistical Significance of Communities

In OSLOM, the statistical significance of a particular community is defined as the probability of finding that community in a null model that is given by a random network without any community structure. The null model chosen here is the *configuration model* [Molloy and Reed, 1995], which generates a network by randomly

joining nodes under the constraint that the network's resulting degree distribution needs to attain a given pre-assigned shape.

Consider a network $G(\mathcal{V}, E)$, where $\mathcal{V}$ denotes the set of $|\mathcal{V}| = N$ nodes, and E denotes the set of edges. Denote by $\mathcal{C}$ the community whose significance is to be assessed. Consider also a node $i \notin \mathcal{C}$, which is considered for inclusion in $\mathcal{C}$. Let $m_{\mathcal{C}}$ be the degree of $\mathcal{C}$, which here means the sum of the degrees of all nodes included in $\mathcal{C}$. Let k_i be the degree of node i, and let M be the degree of $\mathcal{V} \setminus (\mathcal{C} \cup i)$. Write $m_{\mathcal{C}} = m_{\mathcal{C}}^{\text{in}} + m_{\mathcal{C}}^{\text{out}}$, where $m_{\mathcal{C}}^{\text{in}}$ accounts for edges within $\mathcal{C}$, and $m_{\mathcal{C}}^{\text{out}}$ accounts for edges that connect $\mathcal{C}$ to the rest of the network. Similarly, write $k_i = k_i^{\text{in}} + k_i^{\text{out}}$, where k_i^{in} accounts for edges that connect i to $\mathcal{C}$, and k_i^{out} accounts for edges that connect i to $\mathcal{V} \setminus (\mathcal{C} \cup i)$. Denote by $M^\star$ the internal degree of $\mathcal{V} \setminus (\mathcal{C} \cup i)$. Then, we can write $M = m_{\mathcal{C}}^{\text{out}} - k_i^{\text{in}} + k_i^{\text{out}} + M^\star$.

For weighted networks, OSLOM assumes that the probability of the existence of an edge between two nodes i and j with a certain weight w_{ij} is separable in two distinct terms in the configuration model [Radicchi et al., 2010]. The term for the network topology and the ranking procedure for the significance of nodes in the absence of edge weights is described in section 3.3.1.1. The inclusion of the term for edge weights is described in section 3.3.1.2.

3.3.1.1 Topological Relations

The topological relation between a community and a node depends only on how many edges exist between them, but not on their weights. Hence, we here describe the definition of statistical significance for unweighted networks. The next two sections then describe how edge weights can be included in the significance assessment.

Assume that $\mathcal{C}$ is a community in a graph that was generated by the configuration model with the constraint that each node maintains its degree as given in G. Then, the probability of i having k_i^{in} neighbors that belong to $\mathcal{C}$ is given by [Radicchi et al., 2010]

$$p\left(k_i^{\text{in}} \middle| i, \mathcal{C}, G\right) = A \frac{2^{-k_i^{\text{in}}}}{k_i^{\text{out}}!\ k_i^{\text{in}}!\ (m_{\mathcal{C}}^{\text{out}} - k_i^{\text{in}})!\ (M^\star/2)!}, \tag{3.24}$$

where $M^\star = 2E - m_{\mathcal{C}} - m_{\mathcal{C}}^{\text{out}} - 2k_i + 2k_i^{\text{in}}$, and A is a normalization factor ensuring that

$$\sum_{\left\{k_i^{\text{in}} \in \mathbb{N}\ \mid\ M^\star \geq 0\right\}} p\left(k_i^{\text{in}} \middle| i, \mathcal{C}, G\right) = 1. \tag{3.25}$$

Let $r^{(t)}\left(k_i^{\text{in}}\right)$ be the cumulative probability of node i having $\geq k_i^{\text{in}}$ edges that connect it to the community $\mathcal{C}$, so

$$r^{(t)}\left(k_i^{\text{in}}\right) = \sum_{j=k_i^{\text{in}}}^{k_i} p\left(k_i^{\text{in}} \middle| i, \mathcal{C}, G\right). \tag{3.26}$$

If the probability in equation 3.26 is low for a node i, then the connection between that node and the community $\mathcal{C}$ is unexpectedly strong with respect to the null model, and the node should therefore be considered for inclusion in the community.

The values of $r^{(t)}$ for different nodes can then be used to rank all nodes in $E \setminus \mathcal{C}$ according to their probability of being part of $\mathcal{C}$ in terms of its topological association to the group.

Since the node degree is a discrete value, however, the cumulative distribution has a step-wise shape. To compare nodes with different degrees in this setting, a bootstrap is implemented where in each run a value $r_i^{(t)}$ is assigned to each node i by randomly drawing a number from the interval $\left[r^{(t)}\left(k_i^{\text{in}}\right), r^{(t)}\left(k_i^{\text{in}}+1\right)\right]$. The variable $\boldsymbol{r}^{(t)} = \left\{r_i^{(t)}\right\}_{i\in\mathcal{V}\setminus\mathcal{C}}$ captures the likelihood of the topological relation of each external node with community $\mathcal{C}$. For the null model, it takes the form of a random variable that is uniformly distributed on the unit interval. Let $n_{\mathcal{C}}$ be the number of nodes in $\mathcal{C}$ and let $r_q^{(t)}$ be the value of $\boldsymbol{r}^{(t)}$ with rank q in increasing order. Then the cumulative distribution of the variable $\boldsymbol{r}^{(t)}$ is given by

$$\Omega_q^{(t)}\left(\boldsymbol{r}^{(t)}\right) = p\left(r_q^{(t)} < x\right) = \sum_{i=q}^{N-n_{\mathcal{C}}} \binom{N-n_{\mathcal{C}}}{i} x^i (1-x)^{N-n_{\mathcal{C}}-i}. \tag{3.27}$$

The values of the $\Omega_q^{(t)}$ then carry information about how much each node in $\mathcal{V}\setminus\mathcal{C}$ is compatible with the topological statistics expected in the null model. For evaluation of the entire group, define $c_m = \min_q\left(\Omega_q^{(t)}\left(\boldsymbol{r}^{(t)}\right)\right)$ among all neighbors of $\mathcal{C}$. The distribution of c_m is then tabulated numerically. It is denoted by $P\left(c_m < x\right) = \phi\left(x, N-n_{\mathcal{C}}\right)$ and called the "score" of the community $\mathcal{C}$. Since the score is defined as the minimum of the Ω-values, we also refer to it as the "best score" (bs).

3.3.1.2 Edge Weights

In the presence of edge weights, an additional variable $r_i^{(w)}$ is defined for each node i in the neighborhood of $\mathcal{C}$, analogous to the topological relation $\boldsymbol{r}^{(t)}$. Define the strength s_i of a node i as the sum of the weights of all edges that connect it to other nodes. The assumption for the distribution of weights in the null model is then that the weight of an edge is proportional to the average weight of the nodes that it is connected to, which is defined as $\langle w_i\rangle = s_i/k_i$. The cumulative probability of having an edge with a certain minimum weight between two nodes in the null model is then assumed to be

$$p\left(w_{ij} > x \mid k_i, k_j, s_i, s_j\right) = \exp\left(\frac{-x}{\langle\langle w_{ij}\rangle\rangle}\right), \tag{3.28}$$

where $\langle\langle w_{ij}\rangle\rangle = 2\langle w_i\rangle\langle w_j\rangle / \left(\langle w_i\rangle + \langle w_j\rangle\right)$ is the harmonic mean of the average weights of nodes i and j. Here, the harmonic mean was used since it is more sensitive to small values in $\langle w_{ij}\rangle$. Denote by $N(\mathcal{C})$ the neighborhood of community $\mathcal{C}$, meaning the set of all nodes in $\mathcal{V}\setminus\mathcal{C}$ that are connected to any node that is a member of $\mathcal{C}$. Consider now a node $i \in N(\mathcal{C})$ and denote by l the number of edges connecting it to $\mathcal{C}$. Write the normalized weight of all these l edges as $\omega_s = w_s/\langle w_s\rangle$, where w_s is the weight on the s-th edge for $s = 1, 2, \ldots, l$. For the given value l, define

$$\Psi_i = \sum_{s=1}^{l} \omega_s. \tag{3.29}$$

Ψ_i then follows the Erlang distribution since it is the sum of l exponentially distributed random variables [Evans et al., 2001]. The variable accounting for the relation between node i and the community $\mathcal{C}$ with respect to edge weights is then defined as the cumulative distribution

$$r_i^{(w)} = p(\Psi_i > x) = e^{-x} \sum_{z=1}^{l-1} \frac{x^z}{z!}. \tag{3.30}$$

Just as $\boldsymbol{r}^{(t)}$, the variable $\boldsymbol{r}^{(w)} = \left\{r_i^{(w)}\right\}_{i \in N(\mathcal{C})}$ is then random uniformly distributed on the unit interval for the null model.

3.3.1.3 The Combined Significance Score

If the network at hand is weighted, the topological variable $\boldsymbol{r}^{(t)}$ and the weight-related variable $\boldsymbol{r}^{(w)}$ are used to form a combined score $\boldsymbol{r}^{(t,w)}$ for each node i in the neighborhood of $\mathcal{C}$. A difficulty is that $\boldsymbol{r}^{(w)}$ is defined on the set of those N_n neighbors of $\mathcal{C}$, but that $\boldsymbol{r}^{(t)}$ is defined on all the $N^\star = N - n_\mathcal{C} \geq N_n$ nodes in $\mathcal{V} \setminus \mathcal{C}$. Hence, $\boldsymbol{r}^{(t)}$ is re-scaled to an equivalent random variable $\boldsymbol{r}'^{(t)}$, which is defined on the smaller sample of size N_n. Given an index $i \in 1, 2, \ldots, N^\star$, this amounts to mapping the variable to a new index $j \in 1, 2, \ldots, N_n$ such that the cumulative distribution $\Omega_q^{(t)}$ will coincide with the analogously defined cumulative distribution $\Omega_q^{(w)}$ on the subsample of the N_n nodes in the neighborhood of $\mathcal{C}$. This is approximated by re-scaling $\boldsymbol{r}'^{(t)}$ according to

$$\boldsymbol{r}'^{(t)} = \boldsymbol{r}^{(t)} \frac{N^\star + 1}{N_n + 1}. \tag{3.31}$$

For each node i in $N(\mathcal{C})$, the combined score for ranking variables according to how much they belong to the community is then computed as

$$r_i^{(t,w)} = p\left(r_i'^{(t)} r_i^{(w)} < x\right), \tag{3.32}$$

where $r_i^{(t,w)} = x(1 - \log x)$ in the null model, since it was assumed that the the two different $\boldsymbol{r}$-variables are independent and both random uniformly distributed on the unit interval. The set of variables $\boldsymbol{r}^{(t,w)} = \left\{r_i^{(t,w)}\right\}_{i \in N(\mathcal{C})}$ can than be used to rank nodes and to compute their respective cumulative probabilities $\Omega_q^{(t,w)}$ similarly to the unweighted case, with the difference that here the distribution of the topological term was re-scaled to fit the nodes in the neighborhood of $\mathcal{C}$.

3.3.2 Single Community Analysis

We here describe the approach that OSLOM takes to optimize the significance score of a single community $\mathcal{C}$. The next section then presents how the significance of all communities in the entire network is optimized.

OSLOM takes as input a significance level t_{OSLOM}, below which a community score is considered significant. The single community analysis consists of two steps. In the

first step, vertices external to $\mathcal{C}$ are considered for inclusion within the community. In the second step, non-significant nodes are removed.

In step one, for each node i in $N(\mathcal{C})$, the variable $r_i^{(t,w)}$ is calculated. Then, for the node m with $r_m^{(t,w)} = \min\left(\boldsymbol{r}^{(t,w)}\right)$, the score $\Omega_1^{(t,w)}$ is computed. In the case that $\phi\left(\Omega_1^{(t,w)}\left(r_m^{(t,w)}\right), N_n\right) < t_{\text{OSLOM}}$, the corresponding node is added to $\mathcal{C}$. In case $\phi\left(\Omega_1^{(t,w)}\left(r_m^{(t,w)}\right), N_n\right) > t_{\text{OSLOM}}$ the following next-best nodes are checked sequentially. If then there exists some q-best node j for which $\phi\left(\Omega_q^{(t,w)}\left(r_j^{(t,w)}\right), N_n\right) < t_{\text{OSLOM}}$, all of the $(1,\ldots,q)$-best nodes are added to $\mathcal{C}$. If $\nexists i \in \{1,\ldots,N_n\}$: $\phi\left(\Omega_{N_n}^{(t,w)}\left(r_i^{(t,w)}\right), N_n\right) < t_{\text{OSLOM}}$, then no additional nodes are added to the community $\mathcal{C}$. Denote the potentially larger community that is obtained after this first node-adding step by $\mathcal{C}'$. In step two, for each node $i \in \mathcal{C}'$, the variable $r_i^{(t,w)}$ is computed with respect to $\mathcal{C}' \setminus i$. Then, the node v with the highest value is picked, and its significance is checked by repeating step one for the sub-network $\mathcal{C}' \setminus i$. In case v is significant, it is kept inside community $\mathcal{C}'$, and there are no other nodes to be removed since v was picked to be the node with the "worst" r-value. In case v is insignificant, the node is removed from $\mathcal{C}'$ and the procedure is repeated by searching for the "worst" node within $\mathcal{C}' \setminus v$. This is repeated until at some point the worst node identified proves to be statistically significant. The new community without insignificant nodes is denoted $\mathcal{C}^\star$.

This two-step "clean-up" procedure of a community $\mathcal{C}$ is not deterministic due to the stochasticity in the computation of the cumulative probabilities $\boldsymbol{r}^{(t)}$. Hence, single community analysis needs to be repeated multiple times and a consensus result needs to be formed from all runs. To this end, a participation frequency f_i is calculated for each node i, which is defined as the ratio between the times that i was included in a non-empty community $\mathcal{C}^\star$ and the total number of runs leaving the resulting $\mathcal{C}^\star$ non-empty. The final "cleaned" community is then considered statistically significant if $\mathcal{C}$ resulted in a non-empty $\mathcal{C}^\star$ in more than half of the runs, and it contains all those nodes for which $f_i > 0.5$.

3.3.3 Network Analysis

For the identification of significant clusters on the entire network, OSLOM starts with some randomly picked node i. Then, it joins node i into a community $\mathcal{C}$ together with a certain number of q nodes among its neighbors that are considered most significant. While q could in principle be picked arbitrarily, the authors chose a power law with exponent -3. Then, single community analysis as defined in section 3.3.2 is performed on the community $\mathcal{C}$. This procedure is repeated many times for distinct initial nodes to explore all different regions of the entire network. This results in a large set of the communities that were found in all of the runs. Define two of those communities $\mathcal{C}_1$ and $\mathcal{C}_2$ to be similar if $|\mathcal{C}_1 \cup \mathcal{C}_2| \,/\, \min\left(|\mathcal{C}_1|, |\mathcal{C}_2|\right) > P_1$. Here, the authors chose $P_1 = 0.5$. The OSLOM algorithm then stops exploring the entire network when similar communities are found repeatedly.

The resulting collection of significant communities is a cover of the set $\mathcal{V}$, but many

communities overlap and are similar to each other. Hence, to obtain a reasonable solution for the community structure, some of those communities have to be kept and others have to be discarded. This is done by checking whether any community contains any significant sub-communities or whether any group of distinct communities forms a set of significant sub-communities within their union. Consider k communities $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$, as well as their union $\mathcal{C}_u = \cup_{i=1}^{k} \mathcal{C}_i$. If for the k communities that were cleaned up with respect to $\mathcal{C}_u$, it holds that $\left|\cup_{i=1}^{k} \mathcal{C}_i^{\star}\right| < P_2 \left|\mathcal{C}_u\right|$, then $\mathcal{C}_u$ is discarded and the set of smaller communities are kept. Otherwise, all smaller communities $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ are discarded, and their union $\mathcal{C}_u$ is kept. Here, the authors chose $P_2 = 0.7$.

Checking for significant sub-communities within each community in the above-described fashion results in a set of *minimal communities*, which means that none of them have a statistically significant internal community structure themselves. This set of minimal communities still contains many similar ones, so for all pairwise similar communities, their union is checked for significant sub-communities. If such an internal community structure is not found, they are merged. Otherwise, the larger of the two communities is kept, and the other one is discarded. In the case of equal sizes, the community with the lower score is kept. The output of this procedure is a sensible cover of the entire network. Due to the stochasticity of the method, multiple covers are created in the above-described manner and eventually joined by checking again for unions and similar communities, as described before. Nodes that are found not to be a part of any significant community in the final results are left as singletons and are called "homeless nodes".

3.3.4 Hierarchical Structure

Once the final covering of the network consisting of minimal significant communities has been found, the multi-level hierarchy is inferred in an agglomerative, bottom-up manner. Starting from the base (finest) level, a new "supernetwork" is constructed, where each of the communities on the finer level forms a "supernode" on the coarser one. Two supernodes are connected by a "superedge" if the two respective communities on the finer level are connected by some edge between any two nodes within them. The weights of the superedges are assigned based on the the sum of the weights of edges that connect the representative communities on the finer level. Edges incident on nodes that are members of multiple communities on the finer level contribute to the weights of the superedges of both communities on the coarser level, but contributions are downscaled by the number of memberships of the associated nodes. Consider an edge e_{ij} between nodes i and j, which are members of communities $\mathcal{C}_i$ and $\mathcal{C}_j$, respectively. If i and j belong to a total of v_i and v_j different communities, then then the contribution of e_{ij} to the weight of the superedge between $\mathcal{C}_i$ and $\mathcal{C}_j$ is given by $w_{ij}/\left(v_i v_j\right)$. Once the new supernetwork has been constructed on the coarser level of the hierarchy, the entire procedure for finding communities is applied in the same way as for the base level. This procedure is repeatedly used until the method finds no more significant communities in the network on the final, coarsest level of the hierarchy. The final output of the method

is then a set of potentially overlapping communities as well as homeless nodes on each level of the inferred hierarchy.

4

Methods

"Truth has nothing to do with the conclusion, and everything to do with the methodology."

– Stefan Basil Molyneux

This chapter describes the methods used in this project. Section 4.1 presents an evaluation of SNF, MHKSC and OSLOM on simulated data. Section 4.2 describes the retrieval and pre-processing of the cancer data. Section 4.3 covers how similarities between genes were estimated for each omics view of the data. Section 4.4 details how the similarities based on the distinct data views were fused to a single network. Section 4.5 explains how the community structure of the network was inferred on multiple resolutions. Section 4.6 describes the visualization method of the final network, using its multi-level hierarchical community structure. Section 4.7 presents how gene communities detected in the network are related to biological function.

4.1 Simulation Study

This section assesses the performance of SNF on multiple views of simulated data, and compares MHKSC and OSLOM for detecting the multi-level hierarchical community structure. Since there is no universal definition of how both mutual and complementary information in multiple views of data determine a multi-level hierarchical community structure, we divide the simulation study in two parts. In the first part, we construct nine different views of data in a symmetric fashion to obtain a two-level hierarchy of three equally sized communities on the coarse level, which each contain another three equally sized communities on the fine level. Here, each cluster on the coarse hierarchy is present in three out of the nine views and each cluster on the fine hierarchy is present in only a single view. This construction of different views neither allows for unbalanced community sizes nor for different numbers of finer-level communities within the different coarse-level communities. Hence, in the second part of the simulation study, we leave out the similarity fusion step and directly construct a similarity matrix with a well-defined unbalanced hierarchical community structure to further compare MHKSC and OSLOM with different

community sizes and non-symmetrical hierarchies. The simulation study was implemented in MATLAB [version R2016b, 2016], calling the Python [version 3.6.0, 2016] scripts included in the MHKSC-implementation published by Mall et al. [2014].

4.1.1 Fusion and Hierarchical Community Detection on Balanced Data with Symmetrical Hierarchical Structure

To obtain multiple views with a clear multi-level hierarchical community structure, we construct views in the form of nine distinct $N \times N$ similarity matrices $S^{(i)}$ $i \in \{1, 2, \ldots, 9\}$, where N is divisible by nine. In the combined data, we place three equally-sized coarse-level clusters, each of which contains another three equally-sized fine-level communities. Our approach is described below and exemplified in figures 4.1 and 4.2.

Each community on the coarse level is given size $N_{\text{coarse}} = N/3$, and is chosen to be present in the data of three out of the nine views. In each of these three views, this is achieved by drawing the entries corresponding to edges between the coarse-level community and the rest of the network from a normal random distribution $\mathcal{N}(0, \sigma_{\text{views}})$ with mean zero and variance σ_{views}, while drawing values within the coarse-level community from normal distributions with the same variance but higher means. The similarity values for all edges between nodes not included in the coarse-level community are drawn from $\mathcal{N}(0.2, \sigma_{\text{views}})$.

Each community on the fine level is given size $N_{\text{fine}} = N_{\text{coarse}}/3 = N/9$, and is chosen to be present in the data of only one out of the nine views. Similarly to the approach for coarse-level clusters, this is achieved by drawing entries corresponding to edges connecting the fine-level community and other nodes within the respective coarse-level community from $\mathcal{N}(0.2, \sigma_{\text{views}})$, while drawing values within the fine-level community from a normal distribution with higher mean $\mathcal{N}(0.9, \sigma_{\text{views}})$. All remaining values (i.e. those corresponding to edges between all nodes within the coarse-level community but not within the fine-level one) are drawn from $\mathcal{N}(0.4, \sigma_{\text{views}})$. We chose the above means of the different normal distributions in a way such that the distribution of the simulated similarity values on the unit interval resembles the distribution of the actual cancer data we used later in our application to GBM (see figure 4.4 for a comparison).

We make the similarity matrices of all views symmetric by updating them according to

$$S^{(i)} \leftarrow \frac{S^{(i)} + S^{(i)^T}}{2} \quad i \in \{1, 2, \ldots, 9\}. \tag{4.1}$$

Any values < 0 or > 1 are then wrapped back onto the unit interval and values on the diagonal were set to one, giving

$$S^{(i)}_{jk} \leftarrow \begin{cases} 1, & j = k \\ -S^{(i)}_{jk}, & S^{(i)}_{jk} < 0 \\ 2 - S^{(i)}_{jk}, & S^{(i)}_{jk} > 1 \\ S^{(i)}_{jk} & \text{otherwise} \end{cases} \quad i \in \{1, 2, \ldots, 9\},\ j, k \in \{1, 2, \ldots, N\}. \tag{4.2}$$

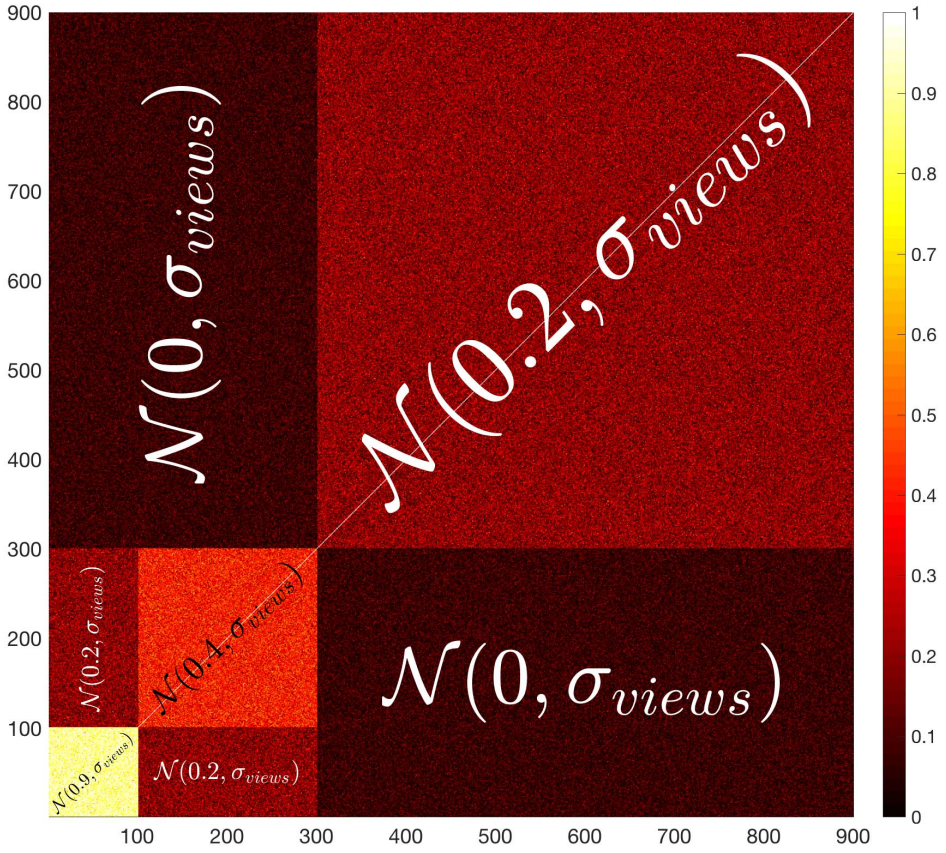


Figure 4.1: The distributions from which entries in the different views were drawn, here illustrated for view 1. For easy visual distinction of the different areas within this matrix, we chose $\sigma_{\text{views}} = 0.1$ in this illustration.

Any values that are still not within the unit interval after this update are assigned a random value between zero and one. The different areas of the similarity matrices corresponding to different random normal distributions are illustrated for view 1 in figure 4.1. All resulting views are shown in the top nine panels of figure 4.2.

We here choose $\sigma_{\text{views}} = 0.3$ and then use SNF to fuse the nine views of the data. For affinity matrix construction, we choose the number of neighbors to be $K = \lfloor N/10 \rfloor$, which was recommended in Wang et al. [2014]. The RBF kernel parameter is set to $\sigma_{\text{RBF}} = 0.5$, which lies within the range of values suggested by the authors, and proved to result in informative affinity matrices. SNF is run for $T = 15$ iterations, as suggested by the authors. The fused similarity matrix obtained by SNF is then thresholded to keep the $p = 5\%$ of edges with largest weights. The resulting fused and thresholded similarity matrix is shown in the lower left panel of figure 4.2. We note that SNF successfully recovers the multi-level hierarchical cluster structure from all nine views even though the views are chosen here to contain very little shared and mainly complementary information. Furthermore, the top p percent of edges effectively capture the community structure. For large-scale real-world networks it should be possible to choose a significantly smaller p, since they usually contain

more communities on more levels of hierarchy and approximately follow power-law degree distributions.

We apply both MHKSC and OSLOM to the fused similarity matrix. We tried different base-level distance thresholds $t_0 \in [0.05, 0.25]$ for MHKSC, but these choices did not significantly influence the quality of the results on the top two levels of hierarchy. We ran OSLOM on a community significance level of $\sigma_{\text{OSLOM}} = 0.1$, as suggested in Lancichinetti et al. [2011]. For obtaining OSLOM consensus partitions, we used ten runs on the base level and 50 on higher levels. To quantify the quality of the community detection results on each level of hierarchy, we used Normalized Mutual Information (NMI, see Strehl and Ghosh [2002], Vinh et al. [2010]) between the ground truth and the community results of the respective algorithm. The NMI between two clusterings is a value between zero and one, with higher values indicating a better agreement.

The center bottom panel of figure 4.2 shows the result of an MHKSC run with $t_0 = 0.15$. Here, the similariy matrix is sorted by the community assingments found by the algorithm. The red dashed lines indicate the community boundaries on the coarse level of hierarchy, and the yellow dotted lines represent the community boundaries on the fine level. We note that MHKSC finds the right communities on the coarse level of hierarchy (NMI = 1), but that it splits the communities on the fine level within each coarse cluster rather arbitrarily (NMI $\approx$ 0.70). In this example, MHKSC finds an additional finer-level partition containing over 30 communities, which is clearly not supported by the data (not shown in figure 4.2). The quality of the MHKSC results did not change much across multiple runs or when varying the parameter t_0. Sometimes, however, the algorithm even placed two clusters on the coarse level into the same community. The bottom right panel of figure 4.2 shows the result of an OSLOM run on the same simulated data. We note that OSLOM also identifies the correct communities on the coarse level of hierarchy (NMI = 1), and that its results on the fine level are also very close to the ground truth (NMI $\approx$ 0.99). Furthermore, OSLOM only identifies the two levels of hierarchy that are indeed supported by the data. The quality of the OSLOM results did not change much across multiple runs.

4.1.2 Hierarchical Community Detection on Unbalanced Data with Asymmetrical Hierarchical Structure

Since we have not yet considered the performance of the two algorithms on asymmetric hierarchical structures and on unbalanced community sizes, we now directly construct a similarity matrix on which to compare MHKSC and OSLOM. We again construct three communities on the coarse level and nine communities on the fine level. In the constructed data set, denote community i on hierarchy j by $\mathcal{C}_i^j$. The

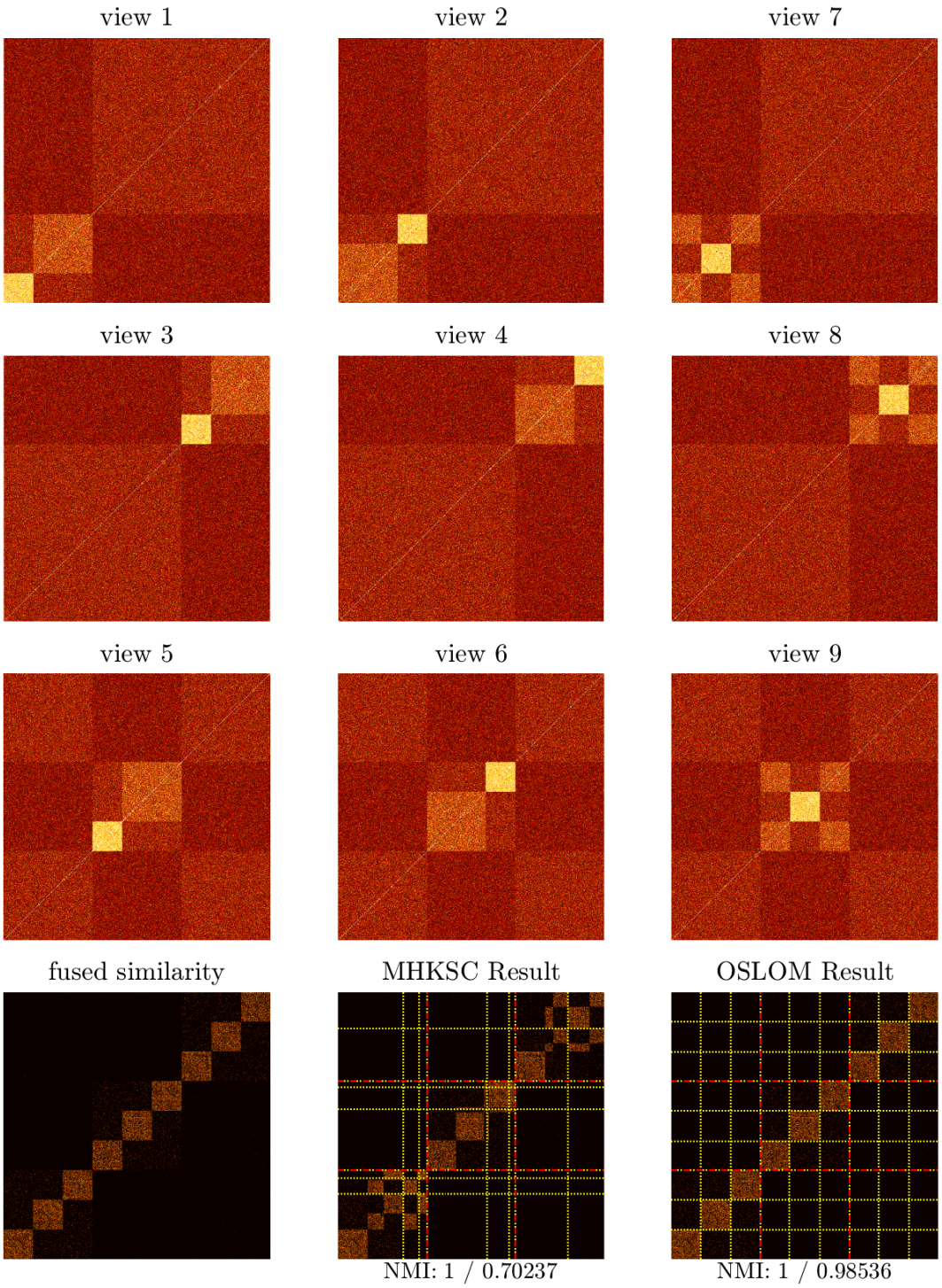


Figure 4.2: Results of SNF (fused similarity), MHKSC and OSLOM on simulated multi-view data that exhibits a symmetrical multi-level hierarchical community structure with balanced cluster sizes.

communities are distributed according to

$$\begin{aligned} \mathcal{C}_1^1 &= \left\{\mathcal{C}_1^0, \mathcal{C}_2^0, \mathcal{C}_3^0\right\} \\ \mathcal{C}_2^1 &= \left\{\mathcal{C}_4^0, \mathcal{C}_5^0\right\} \\ \mathcal{C}_3^1 &= \left\{\mathcal{C}_6^0, \mathcal{C}_7^0, \mathcal{C}_8^0, \mathcal{C}_9^0\right\} \end{aligned} \tag{4.3}$$

where cluster sizes are scaled by the maximum cluster size $s_{\max}$ (here chosen to be 150) according to

$$\begin{aligned} \left|\mathcal{C}_1^0\right| &= \lceil 2s_{\max}/3 \rceil = 100 \\ \left|\mathcal{C}_2^0\right| &= \left\lceil 2/3 \left|\mathcal{C}_1^0\right| \right\rceil = 67 \\ \left|\mathcal{C}_3^0\right| &= \left\lceil 2/3 \left|\mathcal{C}_2^0\right| \right\rceil = 45 \\ \left|\mathcal{C}_4^0\right| &= \lceil s_{\max}/2 \rceil = 75 \\ \left|\mathcal{C}_5^0\right| &= \left\lceil 2/3 \left|\mathcal{C}_4^0\right| \right\rceil = 50 \\ \left|\mathcal{C}_6^0\right| &= s_{\max} = 150 \\ \left|\mathcal{C}_7^0\right| &= \left\lceil 2/3 \left|\mathcal{C}_6^0\right| \right\rceil = 100 \\ \left|\mathcal{C}_8^0\right| &= \left\lceil 2/3 \left|\mathcal{C}_7^0\right| \right\rceil = 67 \\ \left|\mathcal{C}_9^0\right| &= \left\lceil 2/3 \left|\mathcal{C}_8^0\right| \right\rceil = 45. \end{aligned} \tag{4.4}$$

Entries of the similarity matrix corresponding to the communities on the fine level are drawn from a normal random distribution $\mathcal{N}(0.7, \sigma_{\text{sim}})$. All other entries that fall within a community on the coarse level are drawn from $\mathcal{N}(0.45, \sigma_{\text{sim}})$. The remaining entries are drawn from $\mathcal{N}(0, \sigma_{\text{sim}})$. We here set $\sigma_{\text{sim}} = 0.25$. Then, we make the similarity matrix diagonal and wrap entries onto the unit interval as was done in the first part of the simulation study. The matrix is again thresholded to only keep the top 5% of edges. The resulting similarity matrix is visualized in the top left panel of figure 4.3. We compare the performance of MHKSC and OSLOM in the same manner as in the first part of the simulation study.

The lower left panel of figure 4.3 shows the results of a representative run of MHKSC. We note that the algorithm finds too many communities on the coarse level (NMI ≈ 0.85) and fails to identify most of the communities on the fine level (NMI ≈ 0.83). Again, MHKSC finds an additional finer-level partition containing over 20 communities that are not supported by the data (not shown in figure 4.2). Setting t_0 to higher values removes this third level of hierarchy, but then the algorithm splits the true communities in the data in a seemingly arbitrary manner. The lower right panel of figure 4.3 shows the results of a representative run of OSLOM. We note that OSLOM again identifies the correct communities on the coarse level of hierarchy (NMI = 1), and that its results on the fine level are still very close to the ground truth (NMI ≈ 0.98). Again, OSLOM identifies only the two levels of hierarchy which are actually supported by the data. A visualization of the community structure found by OSLOM is shown in the upper right panel of figure 4.3, where nodes are colored by their fine-level community assignments. Nodes that were

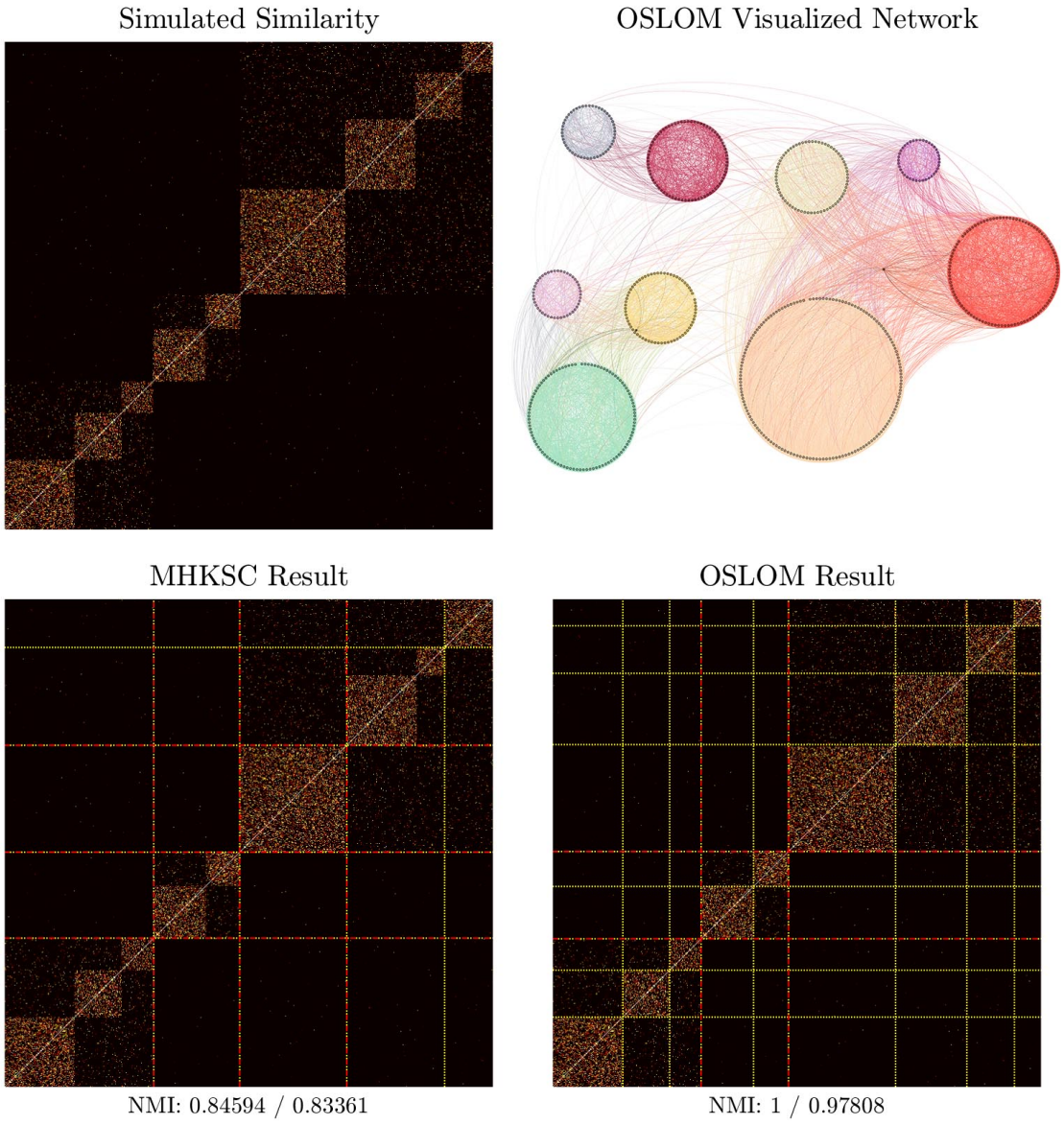


Figure 4.3: Results of MHKSC and OSLOM on a simulated network exhibiting multi-level hierarchical community structure with unbalanced cluster sizes.

assigned to more than one community are displayed in black. For more information on network visualization see section 4.6.

We assume that the relatively poor performance of MHKSC is due to the rather greedy sequential approach of selecting all neighbors of a node within the given distance threshold to be in the same community, thus disregarding the fact that any of those nodes may better fit into another community (see section 3.2.2). On the contrary, OSLOM is a more robust approach since it relies on the statistical significance of each community within the network and uses the consensus results of multiple runs. We conclude that OSLOM is most likely a better choice than MHKSC for our application to GBM sequencing data.

4.2 Data and Preprocessing

In this analysis we use gene expression, copy number aberration (CNA) and DNA methylation sequencing data from Glioblastoma Multiforme (GBM) tumor samples. All data was downloaded from The Cancer Genome Atlas database (TCGA, Weinstein et al. [2013]). We removed from the data set those tumor samples that had no data available for any of the three data types. While expression and CNA levels are given for each gene, methylation data is annotated by the Illumina human methylation sequencing platform (27k) probe identifier, thus making it necessary for further analysis to map methylation probes to the genes available in the rest of the data. Since such probes can be placed anywhere on those DNA's bases that can be methylated, and since genes may be overlapping, it is possible that multiple probes map to the same gene and also that a single probe maps to multiple genes. In case of a single probe mapping to more than one gene, we considered this probe's data a candidate for each of the genes. If multiple probes mapped to a single gene, out of the probes with no missing values, we selected the one that had the lowest correlation to that gene's expression data, as it was done in the TCGA GBM publication by Brennan et al. [2013]. After mapping methylation data to genes, we only kept those genes in the data set which had no missing values in any of the three data types. This resulted in expression, CNA and methylation data views for $n_{\text{genes}} = 7758$ genes, for each observation from at total of $n_{\text{patients}} = 270$ patients. Finally, we standardized the data of each view in a patient-wise fashion. This means that for each view and each patient, we subtracted the mean of the data of all genes from the given values and then divided them by their standard deviation. All data pre-processing was carried out in the R language for statistical computing [R version 3.5.0, 2018].

4.3 View-Specific Similarity Estimation

A straightforward way – and the most popular choice – to estimate the similarity between two genes is to use their standard "Pearson" correlation [Pearson, 1895, Qin et al., 2003, Fehrmann et al., 2015, Tzfadia et al., 2016]. High-dimensional molecular sequencing data, however, is generally rather noisy and often contains

outliers [Yang et al., 2002, Wang et al., 2005]. In addition, clustering results are highly dependent on the underlying similarity measure. Since Pearson's correlation is based on sample averages, it can be heavily affected by a few or even a single measurement [Maronna et al., 2006, Huber, 2011]. We would therefore prefer a pairwise similarity measure that is more stable to perturbations in the observed sample. Other popular choices that are thought to be more robust to outliers since they are based on ranks of the observations are Spearman's rank-order correlation [Spearman, 1904] and Kendall's tau [Kendall, 1938]. While these measures are more robust, they often exhibit low finite-sample efficiency and may not behave optimally in the case of high-dimensional sequencing data [Abdullah, 1990, D'haeseleer et al., 1998]. To solve the problem of finding a pairwise similarity estimator that is robust to outliers, various other approaches have been proposed [Shevlyakov and Smirnov, 2011].

We here choose the absolute median deviation-based MAD correlation coefficient [Pasman and Shevlyakov, 1987, Hampel, 1974], which has shown to yield good results in terms of robustness, bias in small samples, and scalability [Shevlyakov and Smirnov, 2011, Serra et al., 2018]. Given a sample of observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, the median absolute deviation is defined as

$$\mathrm{MAD}(\boldsymbol{x}) = \mathrm{med}\left(\left\{\, |x_i - \mathrm{med}(\boldsymbol{x})| \,\right\}_{i\in\{1,2,\ldots,n\}}\right), \tag{4.5}$$

where $|\cdot|$ denotes the absolute value and $\mathrm{med}(\cdot)$ is the median. The MAD correlation coefficient between two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is then given by

$$r_{\mathrm{MAD}}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{\mathrm{MAD}^2(\boldsymbol{u}) - \mathrm{MAD}^2(\boldsymbol{v})}{\mathrm{MAD}^2(\boldsymbol{u}) + \mathrm{MAD}^2(\boldsymbol{v})}, \tag{4.6}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the robust principal variables

$$\boldsymbol{u} = \frac{\boldsymbol{x} - \mathrm{med}(\boldsymbol{x})}{\sqrt{2}\mathrm{MAD}(\boldsymbol{x})} + \frac{\boldsymbol{y} - \mathrm{med}(\boldsymbol{y})}{\sqrt{2}\mathrm{MAD}(\boldsymbol{y})} \quad \text{and} \quad \boldsymbol{v} = \frac{\boldsymbol{x} - \mathrm{med}(\boldsymbol{x})}{\sqrt{2}\mathrm{MAD}(\boldsymbol{x})} - \frac{\boldsymbol{y} - \mathrm{med}(\boldsymbol{y})}{\sqrt{2}\mathrm{MAD}(\boldsymbol{y})}. \tag{4.7}$$

To obtain the similarity matrices $S^{(v)}$ for our three views $v \in \{$expression, CNA , methylation$\}$, we calculated all the $n_{\mathrm{genes}}(n_{\mathrm{genes}} - 1)/2$ pairwise MAD correlation coefficients between all genes, for each view. The MAD correlation matrix estimation was done in R, using parts of the code published by Serra et al. [2018].

4.4 Network Fusion

Given its good results in the simulation study, we employ Similarity Network Fusion [Wang et al., 2014] to fuse the view-specific similarity matrices into a single gene-gene similarity network. As input for affinity matrix construction, we use the three distance matrices $1 - \left|S^{(v)}\right|$, $v = \{$expression, CNA , methylation$\}$. We set the number of nearest neighbors for affinity matrix construction to $K = 500$, which is less than the approximately $n_{\mathrm{genes}}/10$ neighbors suggested by the authors, but this value seems reasonable given the large number of genes and the fact that genes

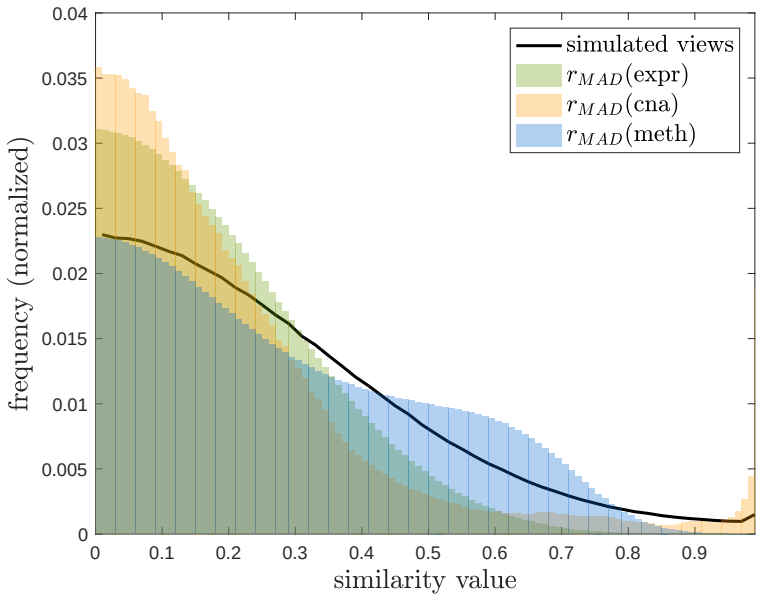


Figure 4.4: Comparison of the distributions of r_{MAD}-derived similarity values for each view and the simulated data from the first part of the simulation study.

usually do not interact with that many other genes. We also tried a few other values of K but found that the choice of this parameter did not significantly alter the community structure of the resulting fused network. We here use the same value of the RBF kernel parameter $\sigma_{\text{RBF}} = 0.5$ as in the simulation study due to the fact that the MAD correlations of our data are similarly distributed on the unit interval as the values we chose for the different views in the simulation study (see figure 4.4). We ran SNF for $T = 20$ iterations to ensure proper convergence.

The fused similarity matrix was thresholded to retain only the strongest $p = 1\%$ weights in the network, which resulted in a cutoff value of approximately 0.15. The motivation behind keeping a lower percentage of edges than in the simulation study was the fact the real-world biological networks tend to be very sparse [Boccaletti et al., 2006], and that values lower than the thresholds likely represent noise or artifacts from the network diffusion process in SNF. We tried a few different percentages, but we found that keeping $p \gg 1\%$ of edges did not result in detecting a very significant hierarchical community structure when using OSLOM, and it also made it virtually impossible to illustrate the resulting network with all its edges in a way that allows for intuitive visual exploration of the data. Furthermore, setting $p \ll 1\%$ resulted in too few edges to effectively detect the network's community structure using OSLOM. We use the edge list representation (see section 1.1.5.1) of the resulting thresholded (i.e. sparse) network to efficiently store all underlying information. Our network fusion step made use of the MATLAB implementation of SNF published by Wang et al. [2014].

4.5 Multi-Resolution Community Detection

Based on its remarkable results in our simulation study, we use (OSLOM, Lancichinetti et al. [2011]) for identifying the multi-level hierarchical community structure of our SNF-fused, thresholded gene-gene association network. We set the p-value for statistical significance of a cluster to $t_{\text{OSLOM}} = 0.1$. This choice is based on the fact $t_{\text{OSLOM}} \gg 0.1$ gave very few, large communities, many of which did not seem to be significantly related to any biological function. On the contrary, setting $t_{\text{OSLOM}} \ll 0.1$ resulted in too many very small communities, making it very difficult to visually explore the data. If the research goal were not the exploration of the large-scale structure emerging from all gene-gene associations in the network but rather the identification of the role of a few pre-selected genes, then smaller values of t_{OSLOM} would also be reasonable. To obtain robust estimates for the multi-resolution community structure in our network, we use the consensus community partition results of 100 runs on the finest level, and 200 runs on all coarser levels of the hierarchy. We make use of the implementation published by Lancichinetti et al. [2011], which was written in the C++ programming language [Stroustrup, 2000].

4.6 Visualization of Multi-Resolution Networks

For visualization purposes, we use additional C++ code published by Lancichinetti et al. [2011] to write files for each level of hierarchy in the *Pajek* (.net) format, which was developed for analysis and visualization of large networks [Batagelj and Mrvar, 2004]. To be able to handle custom edge colors in our visualization, we further convert Pajek .net-files to the GUESS graph exploration (.gdf) format [Adar, 2006]. Here, a two-dimensional position for the visual representation of each node is assigned the same value across the files for all hierarchies. Nodes that are uniquely assigned to the same community on the finest level of hierarchy are placed closest together. Nodes that get assigned to the same community on increasingly higher levels are visualized at growing distances from each other. Nodes that belong to multiple communities are assigned a position that is between the positions of the nodes assigned to these different communities. On levels higher than the finest level of hierarchy, entire communities are considered "super-nodes" with a joint central position for this purpose. In each Pajek/GDF file representing a certain level of hierarchy, the nodes are assigned different colors, according to their community membership on that particular level. Nodes (or super-nodes) that belong to multiple communities on the given level of hierarchy are assigned the color gray. Homeless nodes (i.e. those that are not assigned to any community) are assigned the color white. In addition, we here assign to each node a label, which makes it possible to identify the node by its respective gene identifier. Edge weights between two nodes are taken to be the pairwise similarity between the respective genes as defined by the corresponding values in the thresholded fused similarity matrix. For easy implementation of our entire approach presented up to here, we have compiled a main MATLAB script that calls all the appropriate R and Python files, as well as

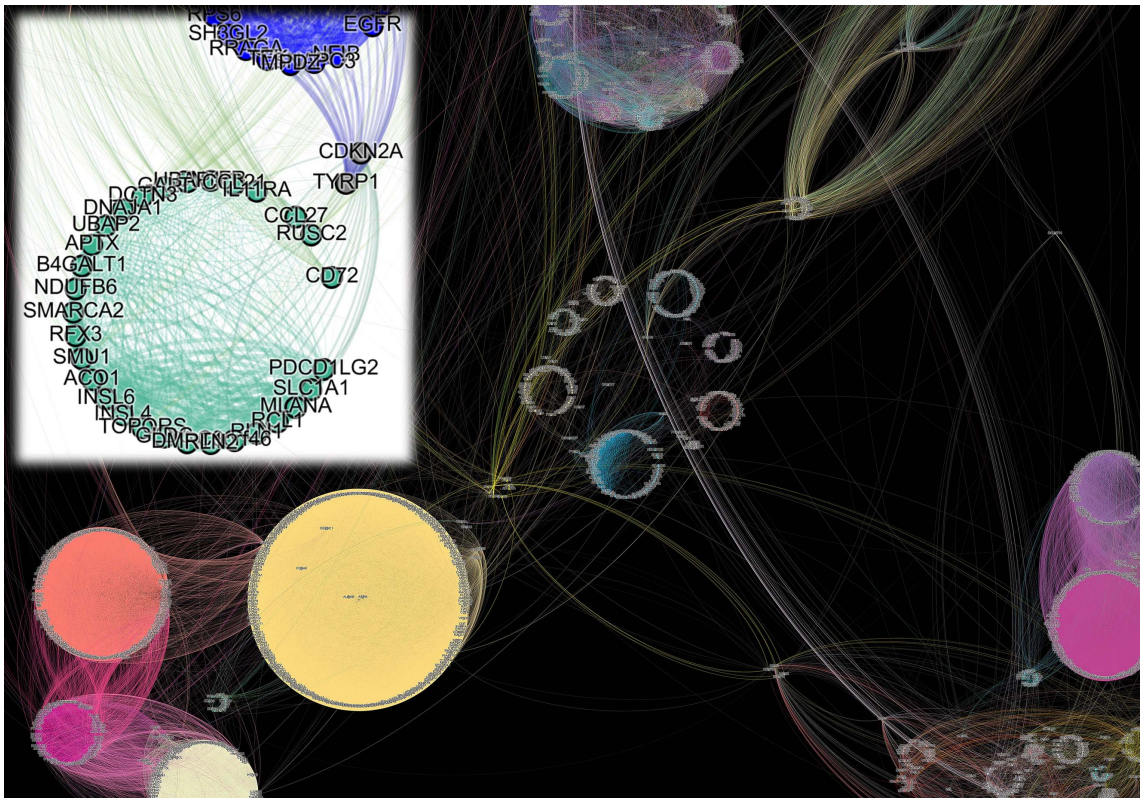


Figure 4.5: An example visualization of the finest-level community partition of a part of our GBM network in Gephi, together with a further zoomed in illustration of a single community that makes it possible to identify nodes by their labels.

the executables compiled from C++.

The Pajek and GDF network files obtained by this procedure can be loaded into graph visualization programs such as *Pajek*[1] [Batagelj and Mrvar, 1998, De Nooy et al., 2011] or *Gephi*[2] [Bastian et al., 2009]. We here chose Gephi for the simple reason that it is open-source software that runs on the Windows, Mac OS X and Linux operating systems, and since it supports both Pajek and GDF files. After loading our files into Gephi, we interactively changed the appearance of our graph to obtain a final illustration that is easily explorable and visually appealing. In particular, we modified node and label sizes, changed edge and label colors, and scaled the width of edges to be proportional to their respective weights. An example of such a visualization of a part of our fused gene-gene association network in GBM is shown in figure 4.5. Here, nodes are colored according to the community partition on the base level of the hierarchy, and edges gradually change their color between the two colors of the nodes that they are connecting. More visualizations are presented in chapter 5.

[1] `http://mrvar.fdv.uni-lj.si/pajek/`
[2] `https://gephi.org/`

	has $\mathcal{A}'$	$\overline{\text{has } \mathcal{A}'}$	sum
in $\mathcal{C}'$	a genes	b genes	$a + b = s_{\mathcal{C}'}$ genes
$\overline{\text{in } \mathcal{C}'}$	c genes	d genes	$c + d = s_{\text{ref}'} - s_{\mathcal{C}'}$ genes
sum	$a + c$ genes	$b + d$ genes	$a + b + c + d = s_{\text{ref}'}$ genes

Table 4.1: An example 2×2 contingency table for the illustration of Fisher's exact test to determine p-values for the overrepresentation of a certain biological annotation $\mathcal{A}$ in a community of genes $\mathcal{C}$.

4.7 Gene Set Overrepresentation Analysis

To check whether the community structure identified in our network is biologically meaningful, we carry out gene set overrepresentation analyses [Mi et al., 2013, Subramanian et al., 2005] on the different groups of genes that are defined by the OSLOM community detection results. For each community under consideration, this means that we analyze human (homo sapiens) biological annotations of the corresponding list of genes, and compare it to the annotations of a reference list containing all the genes that were used in our study.

The goal is to find out whether a certain biological annotation is overrepresented in the given community, which would mean that we find more genes with that particular annotation within the community than we would expect in a list of genes of the same size that was randomly subsampled from the reference list. The fold change of an overrepresented annotation indicates the magnitude of change between the number of respective genes that are expected for a random subsample of the reference list, and the actual number of such genes that are observed in the given community. For instance, consider a reference gene list of size $s_{\text{ref}} = 1000$, in which a total number of $s_{\mathcal{A}} = 100$ genes are biologically annotated with the class $\mathcal{A} =$ "immune response". Assume we are examining a community $\mathcal{C}$ of size $s_{\mathcal{C}} = 250$, and we find that 50 genes inside the community are annotated with "immune response". If we just randomly drew 250 genes from the reference list, however, we would expect to find about 25 such genes in the sample. This implies a 2-fold change because we found twice as many genes annotated with "immune response" as we would have expected.

Due to the stochastic nature of the process, it is always possible to obtain more than the expected number of genes when sampling from the reference list at random. Therefore, we also want to assess whether the overrepresentation of any biological annotation is statistically significant. To this end, we calculate p-values that give the probability of obtaining observed overrepresentations by pure chance. Hence, the lower the p-value, the more confident we can be that the overrepresentation of a certain biological annotation is due to our community detection method rather than random events. We use Fisher's exact test [Fisher, 1922, McDonald, 2009] to determine these p-values. Assume that we have a reference list of size $s_{\text{ref}'}$ and want to assess the statistical significance of some annotation $\mathcal{A}'$ in a community $\mathcal{C}'$ of size $s_{\mathcal{C}'}$. As illustrated in table 4.1, denote by a the number of genes inside $\mathcal{C}'$ with $\mathcal{A}'$,

by b the number of genes inside $\mathcal{C}'$ without $\mathcal{A}'$, by c the number of genes outside $\mathcal{C}'$ with $\mathcal{A}'$, and by d the number of those outside $\mathcal{C}'$ without $\mathcal{A}'$. Then, Fisher's exact p-value is given by hypergeometric distribution

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{s_{\text{ref'}}}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{s_{\text{ref'}}}{b+d}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ b!\ c!\ d!\ s_{\text{ref'}}!}. \tag{4.8}$$

In our above example for the annotation "immune response", we would obtain $a = 50$, $b = s_{\mathcal{C}} - a = 250 - 50 = 200$, $c = s_{\mathcal{A}} - a = 100 - 50 = 50$ and $d = s_{\text{ref}} - s_{\mathcal{C}} - c = 1000 - 250 - 50 = 700$. Plugging the values into equation 4.8, this gives $p \approx 7.40 \times 10^{-9}$. This means that the probability of finding the above-mentioned 2-fold change in "immune response" in our community by pure chance is approximately 7.40×10^{-9}. This is very unlikely, so we could be very confident that it is a result of our community detection method.

To decide which findings we consider significant, we have to define a certain statistical significance level α_p. A common approach is to consider results with $p < \alpha_p = 0.05$ to be statistically significant. Since we are, however, checking multiple biological annotations for possible overrepresentation, this will result in a number of false discoveries. For example, if we are checking 1000 different annotations for overrepresentation in a community based on a statistical significance level of $\alpha_p = 0.05$, we could still expect to make about 50 false discoveries. Hence, we use the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] to adjust the raw p-values obtained by Fisher's exact test for this False Discovery Rate (FDR). Given an FDR significance level α_{FDR} and a list of p-values $(p_1, p_2, \ldots, p_m)$ with $p_i \leq p_j\ \forall i, j \in \{1, 2, \ldots, m\}$ corresponding to all m hypotheses to be tested, the method finds the smallest number k such that $p_k \leq \frac{k}{m}\alpha_{\text{FDR}}$. Then, any raw p-value p_i with $i \leq k$ is considered significant on the level α_{FDR}. For our analysis, we choose $\alpha_{\text{FDR}} = 0.05$.

We used the Gene Ontology (GO) database [Consortium, 2016] to test for overrepresentation of annotations in molecular function, biological process and cellular component. The Reactome database [Croft et al., 2013] and the PANTHER database [Mi et al., 2016] were utilized to check for overrepresentation of pathways and protein class, respectively.

Since we are here checking several communities in the network for overrepresentations, a further adjustment for this additional layer of multiple testing would technically be necessary. This, however, is outside the scope of this thesis. The main goal here is to show that our approach does indeed produce communities that are biologically relevant. We therefore just check a handful of communities on the base level of the hierarchy and report FDR values that are not adjusted for for multiple community testing. If we would check all communities on the fine hierarchy for all annotations and all databases, a further upward adjustment of the lowest FDR values by a factor of roughly the total number of communities times the number of databases would be necessary. Our approach found a little over 200 communi-

ties on the base level of the hierarchy and we used five different databases, so this would imply an upward adjustment of the lowest FDR values of about three orders of magnitude. The lowest FDR values that we found by only checking a handful of communities, however, were many orders of magnitude lower. Hence, we can safely assume that our approach finds biologically relevant communities without having to check every single community in the network and further adjusting FDR values.

5

Results

'In some strange way, any new fact or insight that I may have found has not seemed to me as a "discovery" of mine, but rather something that had always been there and that I had chanced to pick up."

– Subrahmanyan Chandrasekhar

This chapter covers the main results of the application of our multi-view hierarchical community detection approach to the GBM data. Section 5.1 presents the visualization results of the network. Section 5.2 briefly reports a few network statistics. Section 5.3 presents the results of the overrepresentation analysis of biological annotations, for a few select communities on the base hierarchy. Section 5.4 discusses a particular community grouping together genes that seem to be of special importance in Glioblastoma Multiforme.

5.1 Visual Exploration of the Hierarchical Community Structure

We used Gephi to visualize the final network as described in section 4.6. Figure 5.1 shows illustrations of the entire network, each colored according to the different community positions on the indicated level of the hierarchy. We observe that our network does indeed exhibit a remarkable community structure, with many groups of nodes being very densely interconnected while at the same time being well-separated from the rest of the network. This striking community structure is predominantly captured by OSLOM on the finest level of the resulting hierarchy. We can, however, also visually verify that the algorithm joins those communities that share over-proportionally many edges earlier than those that are most separated as it moves up the hierarchy. We therefore conclude that multi-level hierarchical community structure of the fused GBM gene-gene association network can be effectively explored by our visualizations.

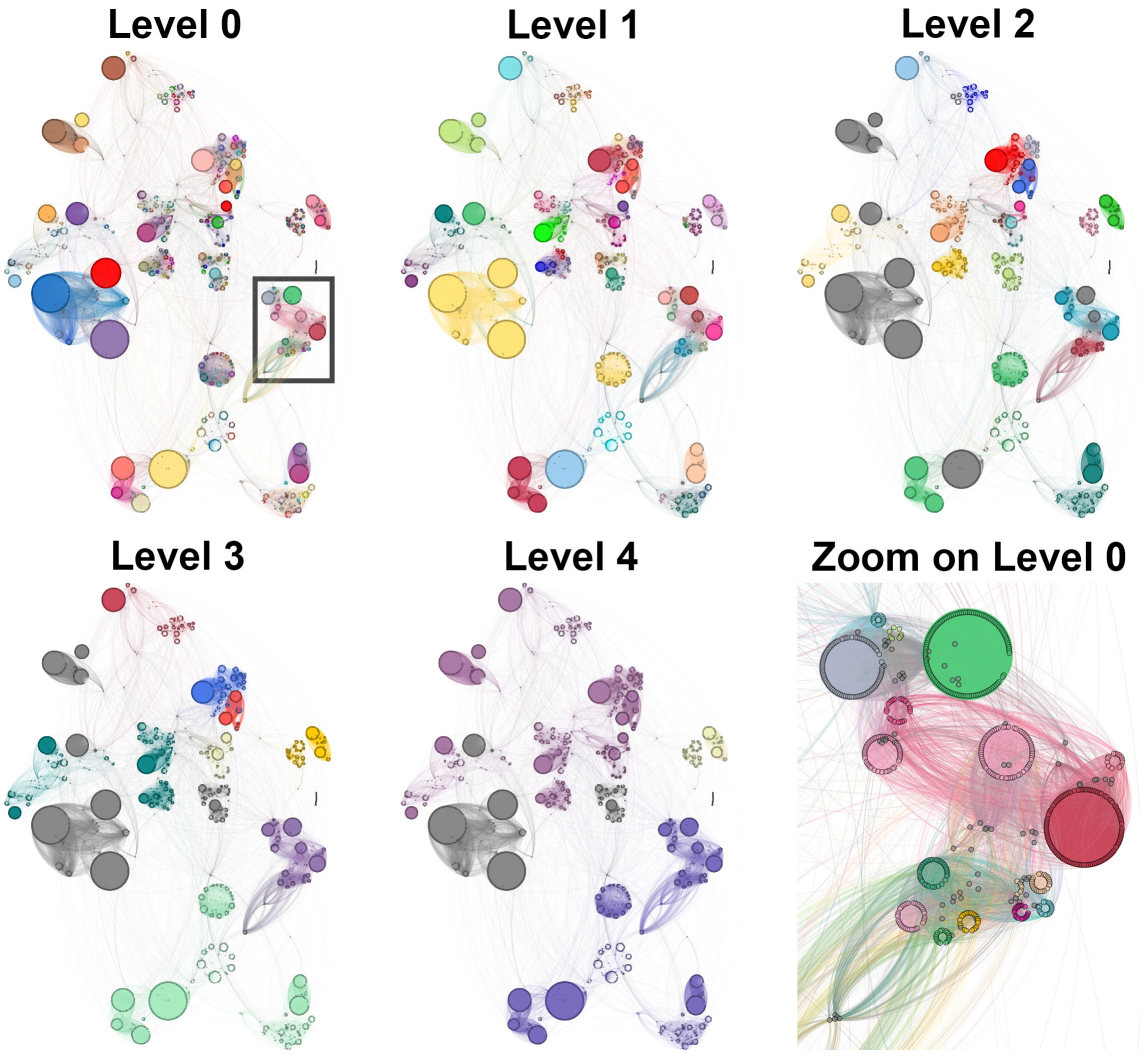


Figure 5.1: A visualization of the fused GBM gene-gene association network using the multi-level hierarchical community structure found by OSLOM. Nodes are colored according to their community membership on each level of the hierarchy. The bottom right panel presents a more detailed view of part the ground level, zoomed in on the area indicated by the gray rectangle in the upper left panel.

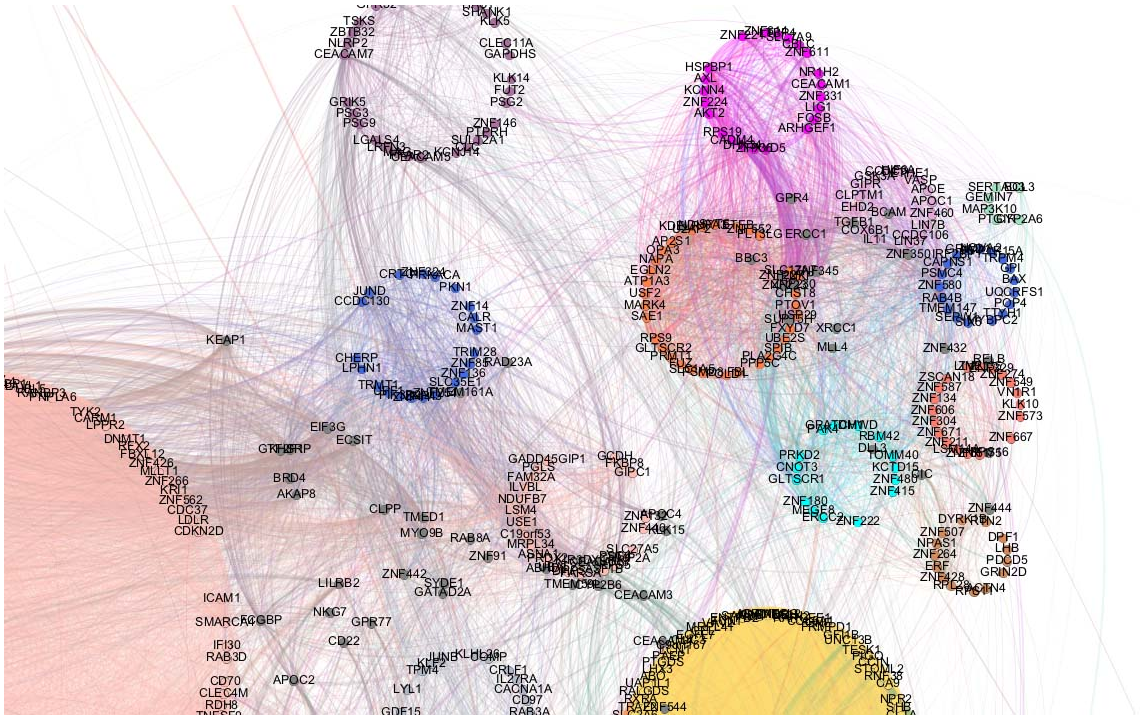


Figure 5.2: A part of the visualized fused GBM gene-gene association network with gene labels, colored according to the OSLOM community partition at the finest level of hierarchy.

Since Gephi makes it simple to zoom into the network visualization, and to manually adjust various aspects of its appearance, it is easy to magnify certain parts of the graph or single communities under interest. When eximining certain communities or groups of communities in detail, it is also possible to show individual gene identifiers directly on the graph in the form of text labels. Figure 5.2 shows and example of such a more detailed view on part of the graph. Particular gene identifiers may also be searched among the list of node labels and then selected and shown directly in a zoomed-in of the network visualization. It is then also possible to add all the neighbors of that particular gene to the selection, and examine the relationships between them. This means that the role of certain genes of assumed importance in GBM within the whole association network can be swiftly explored in an intuitive manner.

We also included the possibility to color edges according to the underlying view of the data in which that edge is supported strongest, relative to all other edges in that view which survived the thresholding process. Figure 5.3 shows the result of coloring edges in this fashion for the entire GBM network. We note that the majority of clusters seems to be supported by all underlying views of the data. There are, however, a few communities that seem to be supported by only a single view, or only by a combination of two of the views used here. Such information can provide further valuable biological insights about why certain groups of genes cluster together and what that implies for our understanding of how GBM "works".

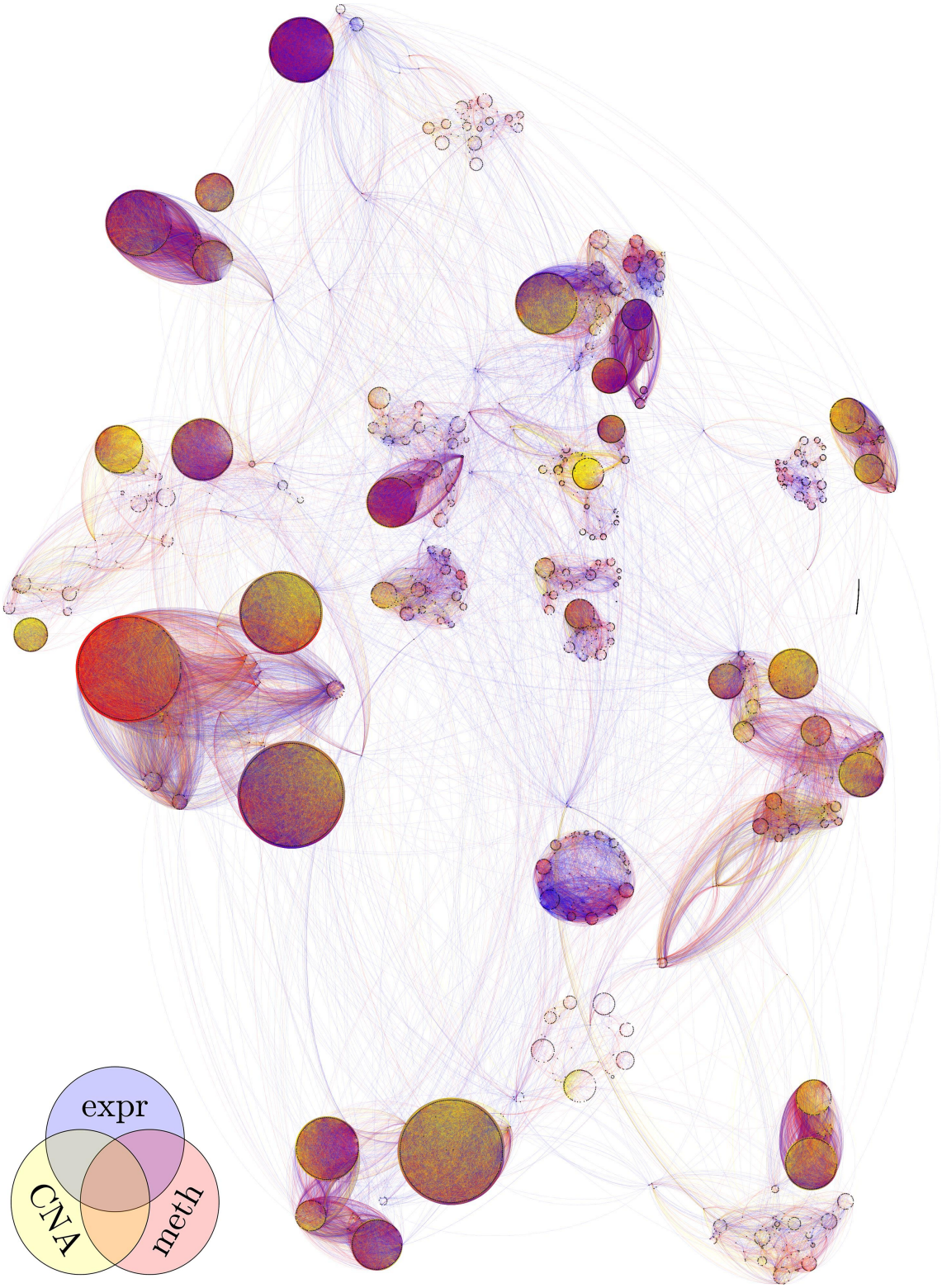


Figure 5.3: A representation of the fused network in which each edge is colored according to the view of the data in which that particular edge is supported strongest, relative to the other edges that survived the thresholding process.

While Wang et al. [2014] compared the raw values of the data underlying any particular edge in each of the views, we here decided to employ a relative, rank-based approach since the distributions of the similarity values in the different views have significantly different shapes (see figure 4.4). Due to the fact that the fusion and thresholding processes favor large values of similarity to be represented in the final network, the view-specific similarity values corresponding to the edges in the final network will mainly lie within the upper tails of the individual views' similarity value distributions. Since the CNA-based similarity matrix contains significantly more values in the interval $[0.9, 1]$ the approach of Wang et al. [2014] would indicate that the majority of the edges are supported by the CNA data. While it is reasonable to assume that the community structure of the final network is indeed significantly influenced by the heavy tail of the CNA similarity distribution, the non-linearity of SNF and the subsequent thresholding step make it difficult to assess the magnitude of the impact that each view has on the final results. For example, it is not only the within-cluster edges that determine the final community structure of the network, but also the absence of strong inter-community edges. While the presence of strong edges is favored by very high similarity values in the underlying data views, their absence is favored by rather low similarity values, which are more frequently encountered in the gene expression view. A way to determine the importance of the distinct views on the community structure of the SNF result would be to detect communities for each view of the data, and compare agreements of the results to communities identified in the fused data. Running OSLOM on the entire similarity matrix of each individual view and the fused view, however, proved to be too computationally expensive. Furthermore, applying OSLOM to the respective thresholded matrices would likely not result in any useful information since any of the non-zero similarities in a particular view can be quite significant in SNF, depending on their nearest neighbors and their corresponding similarities in the other views.

5.2 Network Statistics

We here briefly report a few basic statistics about our GBM network and the community structure identified by OSLOM. Table 5.1 summarizes some of the most important information. We observe that there are very few "homeless" nodes, which are not assigned to any community. Such nodes have a very low average degree, which means that the respective genes are not very similar to any of the the other genes included in this study. Furthermore, the vast majority of genes is uniquely assigned to a community on the fine levels of the hierarchy, which is important to allow for an intuitive and easy visual exploration of the data. Despite the network's large size and sparsity, the average path length between two nodes is very short, which implies that the network exhibits the small-world property, which is claimed to hold for most large-scale complex networks [Watts and Strogatz, 1998].

Figure 5.4 shows the weighted degree distribution of the final network. Since it is often claimed that real-world complex networks exhibit degree distributions that

level of hierarchy (if applicable)	**0**	**1**	**2**	**3**	**4**
number of nodes	7758				
number of edges	300933				
average weighted degree	15.23				
average path length	4.63				
number of communities	218	69	32	22	17
average community size	38.47	116.77	296.53	421.73	524.12
number of covered nodes	7575	7593	7593	7593	7593
fraction of homeless nodes	2.36 %	2.13%	2.13%	2.13%	2.13%
average memberships of covered nodes	1.11	1.06	1.25	1.22	1.17
average degree of homeless nodes	0.098	—	—	—	—

Table 5.1: Some basic network statistics describing the multi-level hierarchical structure identified by OSLOM in our fused GBM network.

follow a power law [Song et al., 2005], we fit a power-law distribution to the data. We note that the degree distribution in our network can be claimed to approximately follow a power law, but that there is a significant number of nodes that have higher weighted degrees on the interval $[20, 40]$. We attribute this to the fact that the network has many very distinct communities on the base level of the hierarchy that seem to be nearly fully connected by relatively strong weights and have an average community size of approximately 38 (see table 5.1).

5.3 Communities Related to Biological Function

The aim of this thesis is to propose a method for identifying the multi-resolution community structure in gene-gene association networks based on multi-view data, and not a complete biological analysis of the results obtained from its application. Therefore, it is here sufficient for us to show that the method indeed finds biologically relevant groups of interacting genes, by only checking a few communities for statistical overrepresentation of biological annotations, as described in section 4.7. Due to the fact that many communities on the finest level of the hierarchy seem to consist of genes that cluster together very well and are often also well-separated from other communities, we limited our analysis to the base hierarchy. Table 5.2 lists a few communities that we checked for biological overrepresentation and showed significant results based on an FDR-rate of $\alpha_{\text{FDR}} = 0.05$. P-values are FDR adjusted to correct for testing multiple annotations. The OSLOM score indicating the statistical significance of a community is denoted "bs" here. Any fold-overrepresentation that is denoted by 100 here means that the actual fold-change was ≥ 100. We point out that the lowest p-values that we obtain are on the order of 10^{-16}, and therefore highly statistically significant. Furthermore, it can be observed that some of the most significant results are related to immune responses, mitotic cell division and cell death – all highly relevant in cancer. We thus conclude that our method indeed provides a community structure in gene-gene association networks that is biologi-

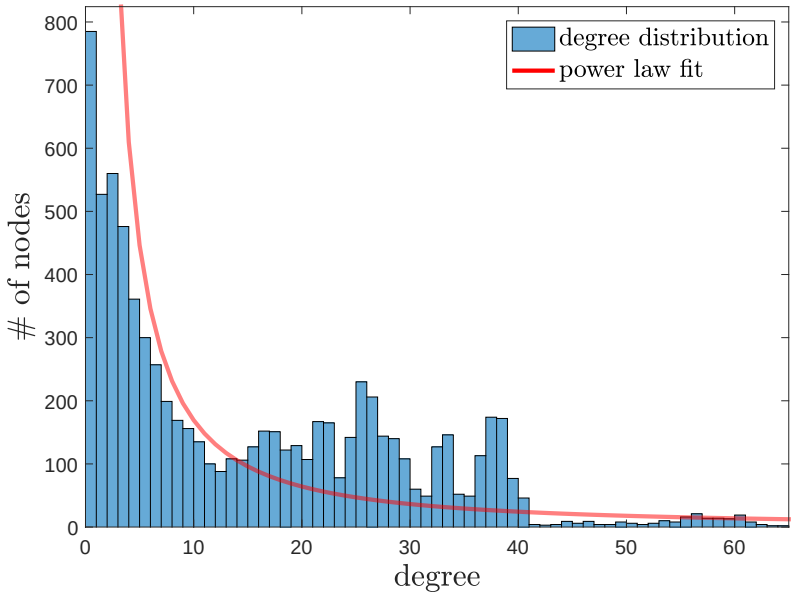


Figure 5.4: The weighted degree distribution of the fused GBM network, along with a power law fit to the data.

cally relevant and may be further explored to shed light on the inner workings of brain cancer.

5.4 A Potentially Important Community for Glioblastoma Multiforme

To exemplify how individual communities or genes could be further analyzed in the context of our fused GBM network, we now turn our focus on one particular community that shows promise to reveal further insight into the inner workings of GBM. The community we chose for this purpose is the blue one that is labeled "# 8" in table 5.2. We chose this particular community due to the fact that it demonstrates consistently high fold-changes for the biological annotations that were found to be overrepresented by the genes in the community, and due to the fact that the associated FDR-corrected p-values are all quite low. Furthermore, we found that the genes in this community were mainly related to immune system response and related processes such as interferon (IFN) signalling, which have been shown to play important roles in cancer, and brain cancer in particular [De Visser et al., 2006, Reiche et al., 2004, Friese et al., 2004].

In addition, the community includes the genes CDKN2A (Cyclin-dependent kinase inhibitor 2A) and EGFR (Epidermal growth factor receptor), which are among the genes that are most frequently altered in GBM [Ueki et al., 1996, Frederick et al., 2000, Wong et al., 1987, Parsons et al., 2008]. Figure 5.5 illustrates this community in two manners. The left panel shows nodes and edges in the colors that were assigned to the different communities on the finest level of the hierarchy (same

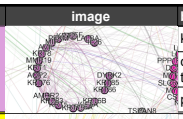
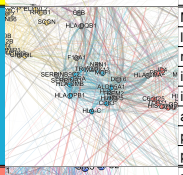
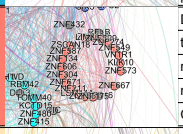
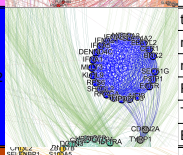
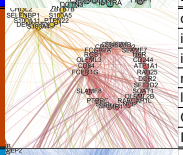
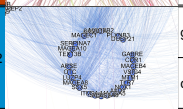
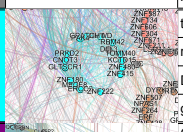
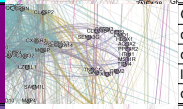
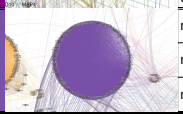
level	#	color	size	bs	image	biological annotation (GO/Panther/Reactome)	fold	p-value	example genes (max 20)
0	1	Plum	30	0.658836		keratine filament	100	1.57E-16	KRT1, KRT2, KRT5, KRT6A, KRT6B, KRT8, KRT75, KRT76, KRT83, KRT85, KRT86, ESPL1, DYRK2, LALBA, MMP19, MIP, AQP2, TSPAN8, IL26, INHBE
						cornification (type of programmed cell death that occurs in the epidermis)	39.51	8.78E-11	
						programmed cell death	5.34	1.75E-04	
0	2	Yellow	23	7.59E-01		cell cycle, mitotic	20.78	2.26E-15	CDCA8, KIF2C, KIF14, KIF23, CCNA2, CCNB2, PLK1, CENPA, BUB1B, NCAPH, AURKB, SPC25, BUB1, NEK2, ASPM, CENPE, ODC1, RRM2, MCM6, UCK2
						cell division	21.55	4.82E-15	
						sister chromatid segregation	46.06	6.40E-15	
						chromosome	28.88	2.82E-07	
						microtubule motor activity	50.15	3.22E-04	
0	3	Cerulean	23	9.16E-01		MHC protein complex	100	1.14E-12	HLA-DMA, HLA-DMB, HLA-C, HLA-DRA, HLA-E, HLA-DPB1, HLA-DQB1, HCP5, TRIM38, TREM2, C2, SERPINB9, SERPINB1, HLA-F, CFB, DEF6, CD83, AIF1, MAPK13, MDFI
						lumenal side of endoplasmatic reticulum membrane	100	6.47E-09	
						major histocompatibility complex antigen	100	7.05E-08	
						immune response	6.04	3.71E-07	
						inferon-gamma-mediated signalling pathway	53.06	6.19E-07	
						antigen processing and processing of peptide antigen	21.55	5.09E-06	
						positive regulation of T cell activation	16.53	1.70E-04	
						peptide antigen assembly with MHC class II protein complex	100	4.14E-04	
0	4	Bittersweet	23	1.13E-01		DNA binding transcription factor activity	7.92	6.16E-12	ZNF134, ZNF135, ZNF211, ZNF175, ZNF274, ZNF304, ZNF329, ZNF350, ZNF432, ZNF444, ZNF544, ZNF551, ZNF551, ZNF573, ZNF587, ZNF606, ZNF667, ZNF671, ZSCAN18, LSM14A
						transcription, DNA-templated	5.33	3.10E-08	
						regulation of transcription by RNA polymerase II	4.49	5.27E-07	
						metal ion binding	3.39	4.73E-06	
						C2H2 zinc finger transcription factor	23.06	3.12E-03	
0	5	Peach	31	7.22E-01		intermediate filament	40.82	6.60E-12	KRT12, KRT15, KRT19, KRT23, KRT24, KRT31, KRT32, KRT33B, KRT34, KRT36, KRT38, VTN, FOXN1, SOX15, OR1A1, OR3A1, OR3A3, CCL7, CCL13, DNAH17
						keratinocyte differentiation	30.59	2.75E-11	
						cornification	42.06	3.91E-11	
						skin development	17.92	4.23E-09	
0	6	Pine Green	33	7.79E-01		histone	89.07	1.31E-11	HIST1H4G, HIST1H2BJ, HIST1H4A, HIST1H4E, HIST1H2BK, HIST1H1E, HIST1H2BI, HIST1H2BE, HIST1H2BC, HIST1H1D, HIST1H2BD, TAF11, DAXX, PRL, HFE, RANBP9, SNRPC, JARID2, BTN3A1, C6orf62
						nucleosome	76.87	3.59E-10	
						nocleosome assembly	53.5	1.58E-09	
						chromatin assembly	46.76	1.66E-09	
						DNA damage / telomere stress induced senescence	87.67	5.01E-05	
						innate response in mucosa	100	4.70E-03	
						histone H3-K27 trimethylation	100	4.26E-02	
0	7	Fuchsia	27	1.78E-101		olfactory receptor activity	91.57	1.37E-08	OR10H1, OR7C1, OR7C2, OR7A5, OR10H2, OR10H3, OR7A17, PDE4C, INSL3, PTGER1, F2RL3, CYP4F11, CYP4F2, CYP4F2, CYP4F12, RNASEH2A, NCAN, DNAJB1, PSPN, NOTCH3
						detection of chemical stimulus involved in sensory perception of smell	91.57	4.95E-08	
						G-protein coupled serotonin signalling pathway	75.22	2.02E-02	
0	8	Blue	25	0.0000132		type I interferon receptor binding	100	2.08E-07	EGFR, CDKN2A, IFNA1, IFNA2, IFNA5, IFNA8, IFNA21, RPS6, CER1, TEK, PSIP1, MLLT3, BNC2, NFIB, MTAP, SNAPC3, RPS6, SH3GL2, RRAGA, TYRP1
						regulation of peptidyl-serine phosphorylation of STAT protein	100	2.96E-06	
						natural killer cell activation involved in immune response	100	2.97E-06	
						Regulation of IFNA signaling	100	4.58E-06	
						T-cell activation involved in immune response	85.59	5.87E-06	
						B-cell proliferation	86.92	1.96E-05	
0	9	Mahogany	30	7.78E-01		cell adhesion molecule	19.41	8.34E-07	CD1D, CD2, CD48, CD53, CD84, CD244, IL6R, RGS1, PTPRC, CTSS, FCER1G, SLAMF7, SELL, FCGR2A, FCGR3B, S100A8, MR1, CREG1, MNDA, NCF2
						immunoglobulin receptor superfamily	44.53	1.65E-06	
						immune response	4.95	1.28E-05	
						neutrophil mediated immunity	8.73	6.64E-05	
						leukocyte migration	10.41	6.56E-04	
						Cell surface interactions at the vascular wall	18.46	1.85E-03	
						IgG binding	100	1.79E-02	
0	8	Process Blue	34	9.54E-102		gamete generation	30.17	1.17E-06	MAGEC1, MAGEA5, MAGEB4, MAGEA10, MTM1, MAGEB2, MAGEA8, VSIG4, EDA2R, SERPINA7, PAGE1, F9, SLC6A8, SSX5, LUZP4, CDR1, AGTR2, GABRE, ITM2A
						cell adhesion molecule	13.68	1.12E-03	
0	11	Gray	21	1.29E-101	*[all nodes overlapping with other communities → no individual visualization of the community available]*	cell periphery	2.74	8.92E-06	CLEC4A, PIK3CG, VAMP8, ITGAM, SLC7A7, LAIR1, LRMP, BIN2, FCGR3B, CD37, CD86, HCLS1, PTPN6, NCKAP1L, NCF2, PLEK, AIF1, SELPLG, SLCO2B1, ADORA3,
						immune system process (17 out of 21 genes)	4.4	1.36E-05	
						myeloid leukocyte activation	10.15	9.04E-05	
						regulation of cell adhesion mediated by integrin	46.57	2.34E-02	
0	12	Cyan	26	1.07E-01		transcription, DNA-templated	4.31	2.08E-04	MEGF8, CARD8, ZNF83, ZNF180, ZNF222, ZNF223, ZNF227, ZNF230, ZNF345, ZNF415, ZNF480, ERCC2, SUPT5H, CIC, RBM42, MLL4, CNOT3, PRKD2, GPATCH1, XRCC1
						regulation of RNA biosynthetic process	3.75	2.70E-05	
						DNA binding transcription factor activity	5.02	1.20E-03	
						generic transcription pathway	9.92	9.14E-06	
0	13	Purple	27	4.61E-01		G-protein coupled chemoattractant receptor activity	100	9.22E-06	CCR3, CCR4, CCR9, CX3CR1, XCR1, SEMA3B, SEMA3G, GNAT1, PPP4R2, GRM2, MST1R, ACOX2, LIMD1, ITIH1, MOBP, CAV3, TGM4, TNNC1, CLEC3B, HESX1
						cell death	6.39	2.10E-02	
						cytokine receptor activity	46.57	2.32E-05	
						locomotion	11.55	2.58E-04	
						chemokine receptors bind chemokines	43.46	3.25E-04	
0	14	Royal Purple	131	6.14E-01		muscle structure development	6	3.10E-05	C1orf35, EGLN1, ACTA1, MYOG, AGT, ZNF238, BTG2, TGFB2, TNNI1, MYBPH, KCNH1, ACTN2, HNRNPU, OBSCN, CENPF, ATF3, RYR2, HLX, PROX1, CD55
						negative regulation of protein activation cascade	62.81	2.53E-03	
						negative regulation of complement activation	100	6.99E-03	

Table 5.2: A (non-comprehensive) list of communities on the base hierarchy in which gene sets with certain biological annotations were overrepresented.

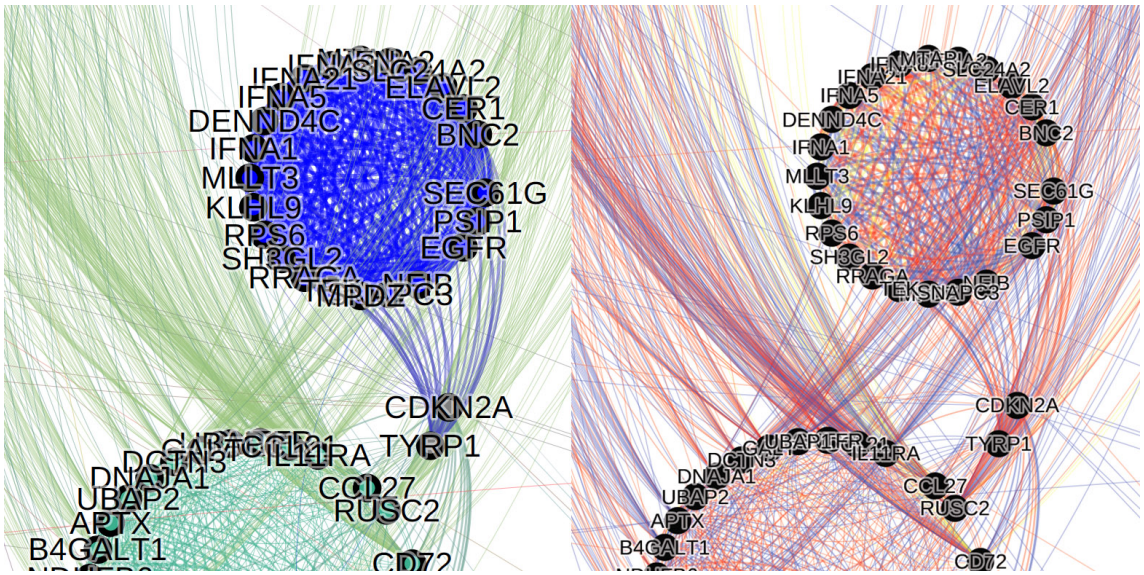


Figure 5.5: Visualization of a community that could be particularly relevant for a better understanding of GBM.

as in figures 5.2 and 5.1). It also features large node labels that make the gene identifier related to each node easily readable. The right panel shows nodes in black and edges in colors that represent the underlying view in which the edge is most strongly supported relative to all other similarity values in that view associated with an edge in the graph (same as in figure 5.3). Here, node labels were chosen to be smaller to make it easier to identify the color of the edges close to any particular node. The left panel of figure 5.5 makes it obvious that genes CDKN2A and TYRP1 (5,6-dihydroxyindole-2-carboxylic acid oxidase) belong to both the blue community we are examining here and the jungle green community shown right below it (not featured in table 5.2).

Since this community seems very important in GBM, it is worth exploring the function of other genes that are included. Many of these genes have been associated to GBM and cancer in general in the literature. For instance, interferon alpha (IFNA) has antiangiogenic activity, thus suppressing cell proliferation in cancer by inhibiting the formation of new blood cells that tumors need to grow [Von Marschall et al., 2003]. The gene KLHL9 (Kelch-like protein 9) has been identified as a causal genetic driver of GBM, with deletions being associated with poorest survival [Chen et al., 2014]. NFIB (nuclear factor 1 B-type) has been shown to induce differentiation and to inhibits the growth of Glioblastoma tumors [Stringer et al., 2013]. The community also includes the gene SH3GL2 (endophilin-A1), which is a candidate tumor suppressor gene which is particularly highly expressed in the central nervous system [Giachino et al., 1997, Ringstad et al., 1997]. The gene MTAP (S-methyl-5'-thioadenosine phosphorylase) has been found to be often deleted in GBM [Kryukov et al., 2016] Furthermore, an approach modeling network effects of CNA on transcription in glioblastoma has previously identified both MTAP and SEC61G (Protein transport protein, subunit gamma) as potentially tumorigenic [Jörnsten et al., 2011]. The community also includes the gene ELAVL2 (ELAV-like protein

2), which is related to neuronal proliferation and differentiation [Yano et al., 2005]. Given that all these genes are potentially important in GBM, further biological analysis of all genes in this community promises new insights into the underlying mechanisms of the disease. We can further analyze which views of the data support the edges within our community. The right panel of figure 5.5 follows the same color scheme as figure 5.3 (expression: blue, CNA: yellow, methylation: red). We note that edges across the whole community are supported by expression and methylation, whereas CNA only seems to connect a subset of the genes in the community. The edges connecting the blue community with the green community are almost exclusively supported by expression and methylation.

The cluster significance score that OSLOM assigns to this blue community is approximately bs $= 1.3 \times 10^{-5}$, which is well below our chosen significance threshold $t_{\text{OSLOM}} = 0.1$, yet much higher than for many other communities on the finest level of the hierarchy. In figure 5.5, we can clearly see that this is due to the fact that our blue community is still relatively strongly connected to the jungle green one below it through the shared genes CDKN2A and TYRP1. This means that it may be worthwhile to also explore the role of the green community, and how it biologically relates to the GBM-related biological processes overrepresented in the blue community. We find that the green community can be mainly related to hormone activity (27.09-fold, FDR= 3.61×10^{-3}) and cytokine activity (17.19-fold, FDR= 9.46×10^{-3}), which have both been related to biological processes associated with glioblastoma [Bonavia et al., 2010, Davis et al., 2006].

As expected if we move up one hierarchy, OSLOM joins the blue and the green community into one (see figure 5.1), and assigns it a very low cluster significance score (bs $\approx 3.6 \times 10^{-101}$). Since the two clusters are largely associated with distinct biological function, however, all statistically significant overrepresentations in this joint level-1 community are mainly inherited from the base communities and therefore result in lower fold-changes and less significant results. This implies that the "bridge" genes CDKN2A and TYRP1 likely have important roles in connecting the otherwise separated biological processes in the two communities.

Our example analysis of this blue community was able to find various genes that are known to play important roles in GBM, and makes it possible to identify further candidate genes and gene-gene interactions to be further investigated. Thus, we have shown that our proposed method effectively fuses different views of molecular profiling data and finds a multi-resolution community structure that is biologically relevant, and whose exploration promises to give better insights into the inner workings of GBM or other types of cancer.

6

Discussion

"Science never solves a problem without creating ten more."

– George Bernard Shaw

6.1 Future Work

We here proposed a two-step procedure that first fuses similarities from different views of the data, and then identifies the multi-resolution community structure of the resulting network. Recently, in the field of multi-view learning, a plentitude of promising methods have been proposed that simultaneously manage data fusion and community detection tasks [Kumar and Daumé, 2011, Christoudias et al., 2012, Cai et al., 2013, Liu et al., 2013, Wang et al., 2013, Xia et al., 2014, Cao et al., 2015, Li et al., 2015, Xu et al., 2015, Liu et al., 2016, Wang et al., 2016, Xu et al., 2016, Liao et al., 2017, Nie et al., 2017, Wang et al., 2017, Zong et al., 2017, Ni et al., 2018, Houthuys et al., 2018]. None of those methods, however, are able to identify hierarchies in the data, many rely on the number of clusters as a user-defined input, and the computational complexity of the majority of them renders applications to high-dimensional big data sets infeasible. A plausible option for an approach that could improve on the method suggested in this thesis could be such a simultaneous multi-view multi-level hierarchical community detection algorithm, which is self-tuned and scalable to big data networks. For instance, a possible extension of the Infomap algorithm [Rosvall and Bergstrom, 2011], allowing for random walks on multi-level networks would have the potential to solve the data integration and multi-level hierarchy detection problems simultaneously.

We proposed our method within the framework of gene-gene association networks, but in theory the approach is applicable to any data that has multiple views and on which a sensible similarity measure can be defined. For instance, with small modifications the method could be used to concatenate different omics data types and rather consider data from multiple databases or distinct patient strata as the views that need to be fused.

As mentioned in section 4.7, we manually checked only a handful of the identified communities for overrepresentation of biological annotations, and did not adjust

them for for the false discovery rate when testing multiple communities or multiple databases. A useful extension of our approach would thus be an automated testing for all biological annotations in multiple databases and for all communities, along with a sensible adjustment of the false discovery rate.

Finally, a rigorous method for determining view importance with regard to the resulting community structure could be of great help in further analysis of the results. We acknowledge that in our particular case it seems likely that the relatively large amount of high similarities in the CNA data could have a significant impact on the final community structure. These large correlations are mainly due to genes that are very close to each other on the DNA in the same chromosomal region and are therefore often copied together by chance. Hence these large similarities are not necessarily meaningful. Here we have not accounted for this, so a possible improvement on our method would include and adjustment of CNA similarity values based on the probability that two genes are copied together by chance.

6.2 Societal and Ethical Aspects

The reliance of big data cancer statistics on large amounts of detailed patient data brings with it both opportunities and risks. While the availability and analyzability of a continuous stream of abundant biological and health-related data for distinct individuals will bring about great advances in personalized treatment, a great emphasis has to be placed on data protection and ethical use of personal data. For instance, making patient data related to behavioral risk factors (such as smoking or an unhealthy diet) available to health insurers, has the potential to improve public health by encouraging healthy behavior through flexible premiums. If such data can be related to preexisting conditions, however, the availability of well-interpretable personalized data sets puts people with genetic predispositions for certain diseases at an unfair disadvantage. This means that healthcare professionals and researchers have a moral obligation to ensure proper data anonymization and protection at each step of cancer research. Results for certain patient groups should always be put into perspective with respect to their societal and personal impacts.

6.3 Conclusion

In this thesis, we have proposed a statistical method that is able to identify the multi-level hierarchical community structure in large-scale multi-view molecular sequencing data sets. To our knowledge, this is the first time that such a method has been proposed. Based on a comprehensive literature review and a simulation study, we suggested a step-wise procedure. First, the robust Median Absolute Deviation (MAD) correlation coefficient was used as a similarity measure on each view of the data. Then, Similarity Network Fusion (SNF) was employed to merge the joint and complementary information captured by the view-specific similarities. The multi-resolution community structure of the fused similarity was identified using the

Order Statistics Local Optimization Method (OSLOM). Finally, to allow for simple and intuitive exploratory analysis, community detection results were visualized and related to known biological functions.

The proposed method was successfully applied to gene expression, copy number aberration and DNA methylation data from Glioblastoma Multiforme (GBM) tumor samples. The analysis revealed a distinct community structure in the resulting gene-gene association network. The effectiveness of the method was demonstrated by identifying multiple communities related to biological functions that play key roles in cancer. While the effectiveness of the proposed method was verified, several improvements are thinkable, which were out of the scope of this thesis. We believe that further research focused on capitalizing on the wealth of information provided by multiple views of molecular data bears great promise for a better understanding of how cancer works and how it might be treated in the future.

Bibliography

[1] Mokhtar Bin Abdullah. On a robust correlation coefficient. *The Statistician*, pages 455–460, 1990.

[2] Eytan Adar. Guess: a language and interface for graph exploration. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 791–800. ACM, 2006.

[3] Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, Helen C Causton, Panisa Pochanard, Eyal Mozes, Levi A Garraway, and Dana Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017, 2010.

[4] Carlos Alcocer-Cuarón, Ana L Rivera, and Victor M Castaño. Hierarchical structure of biological systems: A bioengineering approach. *Bioengineered*, 5(2):73–79, 2014.

[5] Carlos Alzate and Johan AK Suykens. Multiway spectral clustering with out-of-sample extensions through weighted kernel pca. *IEEE transactions on pattern analysis and machine intelligence*, 32(2):335–347, 2010.

[6] Carlos Alzate and Johan AK Suykens. Hierarchical kernel spectral clustering. *Neural Networks*, 35:21–30, 2012.

[7] Carlos Alzate and Johan AK Suykens. A semi-supervised formulation to binary kernel spectral clustering. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

[8] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, volume 28, pages 49–60. ACM, 1999.

[9] Alex Arenas, Jordi Duch, Alberto Fernández, and Sergio Gómez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9(6):176, 2007.

[10] Hugh Arnold, Todd Smith, and N Eric Olson. What does take to identify the signal from the noise in molecular profiling of tumors? *Journal of Biomolecular Techniques: JBT*, 24 (Suppl):S33, 2013.

[11] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[12] Elias August and Antonis Papachristodoulou. Efficient, sparse biological network determination. *BMC Systems Biology*, 3(1):25, 2009.

[13] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101, 2004.

[14] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov,

Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

[15] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *Icwsm*, 8:361–362, 2009.

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

[17] Vladimir Batagelj and Andrej Mrvar. Pajek—analysis and visualization of large networks. In *Graph drawing software*, pages 77–103. Springer, 2004.

[18] Alex Bavelas. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

[19] D J Baylis. *Error Correcting Codes: A Mathematical Introduction*, volume 15. CRC Press, 1997.

[20] Yoav Benjamini and Yosef Hochberg. 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, 1995.

[21] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(2):S15, 2016.

[22] Ludwig Bertalanffy. Kritische theorie der formbildung. 1931.

[23] Ginestra Bianconi, Paolo Pin, and Matteo Marsili. Assessing the relevance of node features for network structure. *Proceedings of the National Academy of Sciences*, 106(28): 11433–11438, 2009.

[24] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *ICDM*, volume 4, pages 19–26, 2004.

[25] Matt Biddulph. Extracting a social graph from wikipedia people pages, 2012.

[26] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[27] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[28] Rudy Bonavia, Akitake Mukasa, Yoshitaka Narita, Dinah WY Sah, Scott Vandenberg, Cameron Brennan, Terrance G Johns, Robert Bachoo, Philipp Hadwiger, Pamela Tan, et al. Tumor heterogeneity is an active process maintained by a mutant egfr-induced cytokine circuit in glioblastoma. *Genes & development*, 24(16):1731–1745, 2010.

[29] Eric Bonnet, Laurence Calzone, and Tom Michoel. Integrative multi-omics module network inference with lemon-tree. *PLoS computational biology*, 11(2):e1003983, 2015.

[30] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. Maximizing modularity is hard. *arXiv preprint physics/0608255*, 2006.

[31] Cameron W Brennan, Roel GW Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2): 462–477, 2013.

[32] Coen Bron and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

[33] Christian T Brownlees, Gudmundur Gudmundsson, and Gábor Lugosi. Community detection in partial correlation network models. 2017.

[34] Atul J Butte and Isaac S Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. In *Proceedings of the AMIA Symposium*, page 711. American Medical Informatics Association, 1999.

[35] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI*, pages 2598–2604, 2013.

[36] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 586–594. IEEE, 2015.

[37] James C Chen, Mariano J Alvarez, Flaminia Talos, Harshil Dhruv, Gabrielle E Rieckhof, Archana Iyer, Kristin L Diefes, Kenneth Aldape, Michael Berens, Michael M Shen, et al. Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell*, 159(2):402–414, 2014.

[38] Jun Chen, Frederic D Bushman, James D Lewis, Gary D Wu, and Hongzhe Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2012.

[39] C Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. *arXiv preprint arXiv:1206.3242*, 2012.

[40] Fan RK Chung. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

[41] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[42] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Structural inference of hierarchies in networks. In *Statistical network analysis: models, issues, and new directions*, pages 1–13. Springer, 2007.

[43] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98, 2008.

[44] Jeff Clune, Jean-Baptiste Mouret, and Hod Lipson. The evolutionary origins of modularity. *Proc. R. Soc. B*, 280(1755):20122863, 2013.

[45] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

[46] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[47] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[48] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2014.

[49] Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic acids research*, 45(D1):D331–D338, 2016.

[50] Josh Constine. Facebook sees 2 billion searches per day, but it's attacking twitter not google, available at https://techcrunch.com/2016/07/27/facebook-will-make-you-talk/, 2016.

[51] Geoffrey M Cooper and Robert E Hausman. *The cell: Molecular approach*. Medicinska naklada, 2004.

[52] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[53] Chad J Creighton, Jeffrey G Reid, and Preethi H Gunaratne. Expression profiling of microrнas by deep sequencing. *Briefings in bioinformatics*, 10(5):490–497, 2009.

[54] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2013.

[55] Xiaofeng Dai, Ting Li, Zhonghu Bai, Yankun Yang, Xiuxia Liu, Jinling Zhan, and Bozhi Shi. Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, 5(10):2929, 2015.

[56] Patrick Danaher. *JGL: Performs the Joint Graphical Lasso for sparse inverse covariance estimation on multiple classes*, 2013. URL `https://CRAN.R-project.org/package=JGL`. R package version 2.3.

[57] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[58] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005 (09):P09008, 2005.

[59] Anup Das, Aaron L Sampson, Claudia Lainscsek, Lyle Muller, Wutu Lin, John C Doyle, Sydney S Cash, Eric Halgren, and Terrence J Sejnowski. Interpretation of the precision matrix and its application in estimating sparse brain connectivity during sleep spindles from human electrocorticography recordings. *Neural computation*, 29(3):603–642, 2017.

[60] Eric Davidson and Michael Levin. Gene regulatory networks, 2005.

[61] Faith B Davis, Heng-Yuan Tang, Ai Shih, Travis Keating, Lawrence Lansing, Aleck Hercbergs, Robert A Fenstermaker, Ahmed Mousa, Shaker A Mousa, Paul J Davis, et al. Acting via a cell surface receptor, thyroid hormone is a growth factor for glioma cells. *Cancer research*, 66(14):7270–7275, 2006.

[62] Wouter De Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory social network analysis with Pajek*, volume 27. Cambridge University Press, 2011.

[63] Karin E De Visser, Alexandra Eichten, and Lisa M Coussens. Paradoxical roles of the immune system during cancer development. *Nature reviews cancer*, 6(1):24, 2006.

[64] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[65] Bruno Domon and Ruedi Aebersold. Mass spectrometry and protein analysis. *science*, 312 (5771):212–217, 2006.

[66] Warwick B Dunn, Alexander Erban, Ralf JM Weber, Darren J Creek, Marie Brown, Rainer Breitling, Thomas Hankemeier, Royston Goodacre, Steffen Neumann, Joachim Kopka, et al. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9(1):44–66, 2013.

[67] Patrik D'haeseleer, Xiling Wen, Stefanie Fuhrman, and Roland Somogyi. Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. In *Information processing in cells and tissues*, pages 203–212. Springer, 1998.

[68] James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature methods*, 11(1):25, 2014.

[69] Levent Ertöz, Michael Steinbach, and Vipin Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 47–58. SIAM, 2003.

[70] Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75, 2002.

[71] Merran Evans, Nicholas Hastings, and Brian Peacock. Statistical distributions, 2001.

[72] Rudolf SN Fehrmann, Juha M Karjalainen, Małgorzata Krajewska, Harm-Jan Westra, David Maloney, Anton Simeonov, Tune H Pers, Joel N Hirschhorn, Ritsert C Jansen, Erik A Schultes, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature genetics*, 47(2):115, 2015.

[73] J Ferlay, I Soerjomataram, M Ervik, R Dikshit, S Eser, C Mathers, M Rebelo, DM Parkin, D Forman, and F Bray. Cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet]. lyon, france: International agency for research on cancer. globocan. 2013; 2012 v1. 0. *Available from: http://globocan. iarc. fr*, 2015.

[74] Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[75] Christina Fitzmaurice, Daniel Dicker, Amanda Pain, Hannah Hamavid, Maziar Moradi-Lakeh, Michael F MacIntyre, Christine Allen, Gillian Hansen, Rachel Woodbrook, Charles Wolfe, et al. The global burden of cancer 2013. *JAMA oncology*, 1(4):505–527, 2015.

[76] Lester R Ford and Delbert R Fulkerson. Maximal flow through a network. *Canadian journal of Mathematics*, 8(3):399–404, 1956.

[77] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[78] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[79] Robert H Frank. *Choosing the right pond: Human behavior and the quest for status.* Oxford University Press, 1985.

[80] Lori Frederick, Xiao-Yang Wang, Greg Eley, and C David James. Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas. *Cancer research*, 60 (5):1383–1387, 2000.

[81] Christopher Freeman and Francisco Louçã. *As time goes by: from the industrial revolutions to the information revolution*. Oxford University Press, 2001.

[82] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[83] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[84] Manuel A Friese, Jörg Wischhusen, Wolfgang Wick, Markus Weiler, Günter Eisele, Alexander Steinle, and Michael Weller. Rna interference targeting transforming growth factor-β enhances nkg2d-mediated antiglioma immune response, inhibits glioma cell migration and invasiveness, and abrogates tumorigenicity in vivo. *Cancer research*, 64(20): 7596–7603, 2004.

[85] Pei-Wen Fu, Chi-Cheng Wu, and Yung-Jan Cho. What makes users share content on facebook? compatibility among psychological incentive, social capital focus, and content type. *Computers in Human Behavior*, 67:23–32, 2017.

[86] Gartner. Big data. *https://www.gartner.com/it-glossary/big-data/*, 2001.

[87] J Bruce German, Bruce D Hammock, and Steven M Watkins. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, 1(1):3–9, 2005.

[88] Zoubin Ghahramani and Thomas L Griffiths. Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482, 2006.

[89] Claudia Giachino, Erica Lantelme, Letizia Lanzetti, Salvatore Saccone, Giuliano Della Valle, and Nicola Migone. A novel sh3-containing human gene family preferentially expressed in the central nervous system. *Genomics*, 41(3):427–434, 1997.

[90] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[91] Vladimir Gligorijević, Noël Malod-Dognin, and Nataša Pržulj. Integrative methods for analyzing big data in precision medicine. *Proteomics*, 16(5):741–758, 2016.

[92] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges, 2014.

[93] Benjamin H Good, Yves-Alexandre de Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[94] Peter J Green and Sylvia Richardson. Modelling heterogeneity with and without the dirichlet process. *Scandinavian journal of statistics*, 28(2):355–375, 2001.

[95] Roger Guimera, Leon Danon, Albert Diaz-Guilera, Francesc Giralt, and Alex Arenas. Self-similar community structure in a network of human interactions. *Physical review E*, 68 (6):065103, 2003.

[96] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.

[97] Isabelle Guyon, Ulrike Von Luxburg, and Robert C Williamson. Clustering: Science or art. In *NIPS 2009 workshop on clustering theory*, pages 1–11, 2009.

[98] Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

[99] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[100] Wolfgang Härdle and Léopold Simar. *Applied multivariate statistical analysis*, volume 22007. Springer, 2007.

[101] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[102] Sampsa Hautaniemi, Markus Ringnér, Päivikki Kauraniemi, Reija Autio, Henrik Edgren, Olli Yli-Harja, Jaakko Astola, Anne Kallioniemi, and Olli-Pekka Kallioniemi. A strategy for identifying putative causes of gene expression variation in human cancers. *Journal of the Franklin Institute*, 341(1-2):77–88, 2004.

[103] Xiaofei He and Partha Niyogi. Locality preserving projections. In *Advances in neural information processing systems*, pages 153–160, 2004.

[104] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[105] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[106] Richard P Horgan and Louise C Kenny. 'omic'technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3):189–195, 2011.

[107] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[108] Lynn Houthuys, Rocco Langone, and Johan AK Suykens. Multi-view kernel spectral clustering. *Information Fusion*, 44:46–56, 2018.

[109] Yanqing Hu, Yuchao Nie, Hua Yang, Jie Cheng, Ying Fan, and Zengru Di. Measuring the significance of community structure in complex networks. *Physical Review E*, 82(6):066106, 2010.

[110] Sijia Huang, Kumardeep Chaudhary, and Lana X Garmire. More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84, 2017.

[111] Tim Hubbard, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. The ensembl genome database project. *Nucleic acids research*, 30(1):38–41, 2002.

[112] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[113] Thomas J Hudson, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[114] David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[115] Daehee Hwang, Alistair G Rust, Stephen Ramsey, Jennifer J Smith, Deena M Leslie, Andrea D Weston, Pedro De Atauri, John D Aitchison, Leroy Hood, Andrew F Siegel, et al. A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17296–17301, 2005.

[116] Javeed Iqbal, Dennis D Weisenburger, Timothy C Greiner, Julie M Vose, Timothy McKeithan, Can Kucuk, Huimin Geng, Karen Deffenbacher, Lynette Smith, Karen Dybkaer, et al. Molecular signatures to improve diagnosis in peripheral t-cell lymphoma and prognostication in angioimmunoblastic t-cell lymphoma. *Blood*, 115(5):1026–1036, 2010.

[117] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[118] Peter James. Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly reviews of biophysics*, 30(4):279–331, 1997.

[119] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651, 2000.

[120] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[121] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[122] Rebecka Jörnsten, Tobias Abenius, Teresia Kling, Linnéa Schmidt, Erik Johansson, Torbjörn EM Nordling, Bodil Nordlander, Chris Sander, Peter Gennemark, Keiko Funa,

et al. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular systems biology*, 7(1):486, 2011.

[123] Anagha Joshi, Yves Van de Peer, and Tom Michoel. Analysis of a gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics*, 24(2):176–183, 2007.

[124] Anagha Joshi, Riet De Smet, Kathleen Marchal, Yves Van de Peer, and Tom Michoel. Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, 25(4):490–496, 2009.

[125] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating'omics' data sets. *Nature reviews Molecular cell biology*, 7(3):198, 2006.

[126] U Kang and Christos Faloutsos. Beyond'caveman communities': Hubs and spokes for graph compression and mining. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 300–309. IEEE, 2011.

[127] William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

[128] Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

[129] Gurvinder Kaur and Jannette M Dufour. Cell lines: Valuable tools or useless artifacts, 2012.

[130] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[131] Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24): 3290–3297, 2012.

[132] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[133] Teresia Kling, Patrik Johansson, José Sanchez, Voichita D Marinescu, Rebecka Jörnsten, and Sven Nelander. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic acids research*, 43(15):e98–e98, 2015.

[134] Teresia Kling, Roberto Ferrarese, Patrik Johansson, Dieter Henrik Heiland, Fangping Dai, Ioannis Vasilikos, Astrid Weyerbrock, Rebecka Jörnsten, Maria Stella Carro, Sven Nelander, et al. Integrative modeling reveals annexin a2-mediated epigenetic control of mesenchymal glioblastoma. *EBioMedicine*, 12:72–85, 2016.

[135] Sotiris Kotsiantis and Panayiotis Pintelas. Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications*, 1(1):73–81, 2004.

[136] John R Koza, Forrest H Bennett, David Andre, and Martin A Keane. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In *Artificial Intelligence in Design'96*, pages 151–170. Springer, 1996.

[137] Slawomir Koziel, Leifur Leifsson, and Xin-She Yang. *Solving computationally expensive engineering problems: methods and applications*, volume 97. Springer, 2014.

[138] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3): 231–240, 2011.

[139] Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313, 2014.

[140] Gregory V Kryukov, Frederick H Wilson, Jason R Ruth, Joshiawa Paulk, Aviad Tsherniak, Sara E Marlow, Francisca Vazquez, Barbara A Weir, Mark E Fitzgerald, Minoru Tanaka,

et al. Mtap deletion confers enhanced dependency on the prmt5 arginine methyltransferase in cancer cells. *Science*, 351(6278):1214–1218, 2016.

[141] HW Kuhn, AW Tucker, et al. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

[142] Abhishek Kumar and Hal Daumé. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 393–400, 2011.

[143] Martin Kussmann, Frédéric Raymond, and Michael Affolter. Omics-driven biomarker discovery in nutrition and health. *Journal of biotechnology*, 124(4):758–787, 2006.

[144] Darong Lai, Hongtao Lu, and Christine Nardini. Enhanced modularity-based community detection by random walk network preprocessing. *Physical Review E*, 81(6):066118, 2010.

[145] Andrea Lancichinetti, Filippo Radicchi, and José J Ramasco. Statistical significance of communities in networks. *Physical Review E*, 81(4):046110, 2010.

[146] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato. Finding statistically significant communities in networks. *PloS one*, 6(4):e18961, 2011.

[147] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[148] Rocco Langone, Carlos Alzate, and Johan AK Suykens. Kernel spectral clustering for community detection in complex networks. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

[149] Kim-Anh Lê Cao, Pascal GP Martin, Christèle Robert-Granié, and Philippe Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, 10(1):34, 2009.

[150] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[151] Pierre Legrain, Ruedi Aebersold, Alexander Archakov, Amos Bairoch, Kumar Bala, Laura Beretta, John Bergeron, Christoph H Borchers, Garry L Corthals, Catherine E Costello, et al. The human proteome project: current state and future direction. *Molecular & cellular proteomics*, 10(7):M111–009993, 2011.

[152] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2006.

[153] Wenyuan Li, Shihua Zhang, Chun-Chi Liu, and Xianghong Jasmine Zhou. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19):2458–2466, 2012.

[154] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *AAAI*, pages 2750–2756, 2015.

[155] Longlong Liao, Kenli Li, Keqin Li, Qi Tian, and Canqun Yang. Automatic density clustering with multiple kernels for high-dimension bioinformatics data. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 2105–2112. IEEE, 2017.

[156] Dongdong Lin, Jigang Zhang, Jingyao Li, Vince D Calhoun, Hong-Wen Deng, and Yu-Ping Wang. Group sparse canonical correlation analysis for genomic data integration. *BMC bioinformatics*, 14(1):245, 2013.

[157] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh. Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–1160, 2011.

[158] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271): 315, 2009.

[159] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 252–260. SIAM, 2013.

[160] Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. Multiple kernel k-means clustering with matrix-induced regularization. In *AAAI*, pages 1888–1894, 2016.

[161] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[162] Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20): 2610–2616, 2013.

[163] Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013.

[164] Joseph Loscalzo and Albert-Laszlo Barabasi. Systems biology and the future of medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(6):619–627, 2011.

[165] Riku Louhimo and Sampsa Hautaniemi. Cnamet: an r package for integrating copy number, methylation and expression data. *Bioinformatics*, 27(6):887–888, 2011.

[166] David MacKay. Gaussian process basics, 2006.

[167] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

[168] Raghvendra Mall, Rocco Langone, and Johan AK Suykens. Furs: Fast and unique representative subset selection retaining large-scale community structure. *Social Network Analysis and Mining*, 3(4):1075–1095, 2013.

[169] Raghvendra Mall, Rocco Langone, and Johan AK Suykens. Kernel spectral clustering for big data networks. *Entropy*, 15(5):1567–1586, 2013.

[170] Raghvendra Mall, Rocco Langone, and Johan AK Suykens. Self-tuned kernel spectral clustering for large scale networks. In *Big Data, 2013 IEEE International Conference on*, pages 385–393. IEEE, 2013.

[171] Raghvendra Mall, Rocco Langone, and Johan AK Suykens. Multilevel hierarchical kernel spectral clustering for real-life large scale complex networks. *PloS one*, 9(6):e99966, 2014.

[172] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, page S7. BioMed Central, 2006.

[173] RARD Maronna, R Douglas Martin, and Victor Yohai. *Robust statistics*, volume 1. John Wiley & Sons, Chichester. ISBN, 2006.

[174] Vivien Marx. Biology: The big challenges of big data, 2013.

[175] MATLAB. *MATLAB version 9.1.0 (R2016b)*. Natick, Massachusetts, version R2016b, 2016. URL `https://www.mathworks.com/products/matlab.html`.

[176] Fulvio Mazzocchi. Complexity and the reductionism–holism debate in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(5):413–427, 2012.

[177] John H McDonald. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

[178] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[179] Lisa Melton. Protein arrays: proteomics in multiplex. *Nature*, 429(6987):101, 2004.

[180] Chen Meng. *mogsa: Multiple omics data integrative clustering and gene set analysis*, 2016. R package version 1.8.0.

[181] Chen Meng, Bernhard Kuster, Aedín C Culhane, and Amin Moghaddas Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics*, 15 (1):162, 2014.

[182] Chen Meng, Dominic Helm, Martin Frejno, and Bernhard Kuster. mocluster: identifying joint patterns across multiple omics data sets. *Journal of proteome research*, 15(3):755–765, 2015.

[183] Henok Mengistu, Joost Huizinga, Jean-Baptiste Mouret, and Jeff Clune. The evolutionary origins of hierarchy. *PLoS computational biology*, 12(6):e1004829, 2016.

[184] Michael L Metzker. Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31, 2010.

[185] Huaiyu Mi, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. Large-scale gene function analysis with the panther classification system. *Nature protocols*, 8(8):1551, 2013.

[186] Huaiyu Mi, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D Thomas. Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, 45(D1):D183–D189, 2016.

[187] Tom Michoel and Bruno Nachtergaele. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E*, 86(5):056111, 2012.

[188] Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. Preparing for precision medicine. *New England Journal of Medicine*, 366(6):489–491, 2012.

[189] Qianxing Mo and Ronglai Shen. *iClusterPlus: Integrative clustering of multi-type genomic data*, 2016. R package version 1.10.0.

[190] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[191] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[192] Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[193] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, 2008.

[194] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90, 2011.

[195] Mark EJ Newman. Analysis of weighted networks. *Physical review E*, 70(5):056131, 2004.

[196] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[197] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[198] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[199] Jingchao Ni, Wei Cheng, Wei Fan, and Xiang Zhang. Comclus: A self-grouping framework for multi-network clustering. *IEEE Transactions on Knowledge and Data Engineering*, 30 (3):435–448, 2018.

[200] Jeremy K Nicholson and John C Lindon. Systems biology: metabonomics. *Nature*, 455 (7216):1054, 2008.

[201] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, pages 2408–2414, 2017.

[202] Denis Noble. *The music of life: biology beyond genes*. Oxford University Press, 2008.

[203] Michael J. O'Connell and Eric F. Lock. *r.jive: Perform JIVE Decomposition for Multi-Source Data*, 2017. URL `https://CRAN.R-project.org/package=r.jive`. R package version 2.1.

[204] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8(1):1–34, 2009.

[205] D Williams Parsons, Siân Jones, Xiaosong Zhang, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, I-Mei Siu, Gary L Gallia, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897): 1807–1812, 2008.

[206] VR Pasman and GL Shevlyakov. Robust methods of estimating the correlation coefficient. *Avtomatika i Telemekhanika*, (3):70–80, 1987.

[207] Athanasia Pavlopoulou, Demetrios A Spandidos, and Ioannis Michalopoulos. Human cancer databases. *Oncology reports*, 33(1):3–18, 2015.

[208] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[209] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[210] Dana Pe'er and Nir Hacohen. Principles and strategies for developing network models in cancer. *Cell*, 144(6):864–873, 2011.

[211] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

[212] Python. *Python Language Reference*. Python Software Foundation, Wilmington, Delaware, version 3.6.0, 2016. URL `http://www.python.org`.

[213] Jie Qin, Darrin P Lewis, and William Stafford Noble. Kernel hierarchical gene clustering from microarray expression data. *Bioinformatics*, 19(16):2097–2104, 2003.

[214] R version 3.5.0. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

[215] Filippo Radicchi, Andrea Lancichinetti, and José J Ramasco. Combinatorial approach to modularity. *Physical Review E*, 82(2):026102, 2010.

[216] Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[217] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, 2003.

[218] Priyadip Ray, Lingling Zheng, Joseph Lucas, and Lawrence Carin. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30(10):1370–1376, 2014.

[219] Antonio Regalado. Emtech: Illumina says 228,000 human genomes will be sequenced this year, available at http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/,
2014.

[220] Joerg Reichardt and Michele Leone. (un) detectable cluster structure in sparse networks. *Physical review letters*, 101(7):078701, 2008.

[221] Jörg Reichardt and Stefan Bornholdt. When are networks truly modular? *Physica D: Nonlinear Phenomena*, 224(1-2):20–26, 2006.

[222] Edna Maria Vissoci Reiche, Sandra Odebrecht Vargas Nunes, and Helena Kaminami Morimoto. Stress, depression, the immune system, and cancer. *The lancet oncology*, 5(10): 617–625, 2004.

[223] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, pages 827–832, 2015.

[224] Niels Ringstad, Yasuo Nemoto, and Pietro De Camilli. The sh3p4/sh3p8/sh3p13 protein family: binding partners for synaptojanin and dynamin via a grb2-like src homology 3 domain. *Proceedings of the National Academy of Sciences*, 94(16):8569–8574, 1997.

[225] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[226] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85, 2015.

[227] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346 (25):1937–1947, 2002.

[228] Martin Rosvall and Carl T Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007.

[229] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[230] Martin Rosvall and Carl T Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, 2011.

[231] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[232] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173, 2005.

[233] Peter J. Russell. *IGenetics: A Molecular Approach*. Benjamin Cummings, 2010.

[234] Marta Sales-Pardo, Roger Guimera, André A Moreira, and Luís A Nunes Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224–15229, 2007.

[235] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[236] Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach, and Mark B Gerstein. The real cost of sequencing: higher than you think! *Genome biology*, 12 (8):125, 2011.

[237] Eric E Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218, 2009.

[238] Satu Elisa Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007.

[239] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[240] Angela Serra, Pietro Coretto, Michele Fratello, and Roberto Tagliaferri. Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data. *Bioinformatics*, 2018.

[241] Ronglai Shen. *iCluster: Integrative clustering of multiple genomic data types*, 2012. URL `https://CRAN.R-project.org/package=iCluster`. R package version 2.1.0.

[242] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

[243] Ronglai Shen, Qianxing Mo, Nikolaus Schultz, Venkatraman E Seshan, Adam B Olshen, Jason Huse, Marc Ladanyi, and Chris Sander. Integrative subtype discovery in glioblastoma using icluster. *PloS one*, 7(4):e35236, 2012.

[244] Georgy Shevlyakov and Pavel Smirnov. Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics*, 40(1&2):147–156, 2011.

[245] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[246] Herbert A Simon. The architecture of complexity. In *Facets of systems science*, pages 457–476. Springer, 1991.

[247] Chaoming Song, Shlomo Havlin, and Hernan A Makse. Self-similarity of complex networks. *Nature*, 433(7024):392, 2005.

[248] Charles Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[249] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, 2015.

[250] Victor Spirin and Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.

[251] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[252] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[253] Brett Stringer, Bryan Day, Guy Barry, Michael Piper, Paul Jamieson, Kathleen Ensbey, Zara Bruce, Linda Richards, and Andrew Boyd. The glial differentiation factor nuclear factor one b (nfib) induces differentiation and inhibits growth of glioblastoma. In *Neuro-Oncology*, volume 15, pages 27–27. Oxford University Press, 2013.

[254] Bjarne Stroustrup. *The C++ programming language*. Pearson Education India, 2000.

[255] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249–255, 2003.

[256] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

[257] Johan AK Suykens, Tony Van Gestel, and Jos De Brabanter. *Least squares support vector machines*. World Scientific, 2002.

[258] Chaitanya Swamy and David Shmoys. *Approximation algorithms for clustering problems*. Citeseer, 2004.

[259] Sergios Theodoridis. *Pattern Recognition*. Elsevier, 2008.

[260] Romain Thibaux and Michael I Jordan. Hierarchical beta processes and the indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571, 2007.

[261] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[262] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[263] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804, 2007.

[264] Oren Tzfadia, Tim Diels, Sam De Meyer, Klaas Vandepoele, Asaph Aharoni, and Yves Van de Peer. Coexpnetviz: comparative co-expression networks construction and visualization tool. *Frontiers in plant science*, 6:1194, 2016.

[265] Keisuke Ueki, Yasuhiro Ono, John W Henson, Jimmy T Efird, Andreas von Deimling, and David N Louis. Cdkn2/p16 or rb alterations occur in the majority of glioblastomas and are inversely correlated. *Cancer research*, 56(1):150–153, 1996.

[266] Mark Van Vugt, Robert Hogan, and Robert B Kaiser. Leadership, followership, and evolution: Some lessons from the past. *American Psychologist*, 63(3):182, 2008.

[267] Laura J Van't Veer and René Bernards. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452(7187):564, 2008.

[268] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway

activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26 (12):i237–i245, 2010.

[269] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[270] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

[271] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[272] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

[273] Zofia Von Marschall, Arne Scholz, Thorsten Cramer, Georgia Schäfer, Michael Schirner, Kjell Öbeerg, Bertram Wiedenmann, Michael Höcker, and Stefan Rosewicz. Effects of interferon alpha on vascular endothelial growth factor gene transcription and tumor angiogenesis. *Journal of the National Cancer Institute*, 95(6):437–448, 2003.

[274] Stuart G Walesh. *Urban surface water management*. John Wiley & Sons, 1989.

[275] Saint John Walker. Big data: A revolution that will transform how we live, work, and think, 2014.

[276] Bo Wang, Jiayan Jiang, Wei Wang, Zhi-Hua Zhou, and Zhuowen Tu. Unsupervised metric fusion by cross diffusion. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2997–3004. IEEE, 2012.

[277] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

[278] Bo Wang, Aziz Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. *SNFtool: Similarity Network Fusion*, 2017. URL `https://CRAN.R-project.org/package=SNFtool`. R package version 2.2.1.

[279] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *International conference on machine learning*, pages 352–360, 2013.

[280] Huixia Wang, Xuming He, Mark Band, Carole Wilson, and Lei Liu. A study of inter-lab and inter-platform agreement of dna microarray data. *BMC genomics*, 6(1):71, 2005.

[281] Wenjun Wang, Dong Liu, Xiao Liu, and Lin Pan. Fuzzy overlapping community detection based on local random walk and multidimensional scaling. *Physica A: Statistical Mechanics and its Applications*, 392(24):6578–6586, 2013.

[282] Yang Wang, Wenjie Zhang, Lin Wu, Xuemin Lin, Meng Fang, and Shirui Pan. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. *arXiv preprint arXiv:1608.05560*, 2016.

[283] Yueqing Wang, Xinwang Liu, Yong Dou, and Rongchun Li. Multiple kernel clustering framework with improved kernels. *Discover*, 1(2):3–4, 2017.

[284] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57, 2009.

[285] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[286] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171 (4356):737–738, 1953.

[287] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[288] Max Weber. *Economy and society: An outline of interpretive sociology*, volume 1. Univ of California Press, 1978.

[289] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[290] Johan A Westerhuis, Theodora Kourti, and John F MacGregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of chemometrics*, 12(5):301–321, 1998.

[291] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189):872, 2008.

[292] Norbert Wiener. Cybernetics. or control and communication in the animal and the machine. 1949.

[293] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, et al. Hmdb: the human metabolome database. *Nucleic acids research*, 35(suppl_1):D521–D526, 2007.

[294] Daniela Witten, Rob Tibshirani, Sam Gross, and Balasubramanian Narasimhan. *PMA: Penalized Multivariate Analysis*, 2013. URL `https://CRAN.R-project.org/package=PMA`. R package version 1.0.9.

[295] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

[296] Svante Wold. *PLS modeling with latent variables in two or more dimensions*. 1987.

[297] Albert J Wong, Sandra H Bigner, Darell D Bigner, Kenneth W Kinzler, Stanley R Hamilton, and Bert Vogelstein. Increased expression of the epidermal growth factor receptor gene in malignant gliomas is invariably associated with gene amplification. *Proceedings of the National Academy of Sciences*, 84(19):6899–6903, 1987.

[298] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, pages 2149–2155, 2014.

[299] Yuan Xie, Tobias Bergström, Yiwen Jiang, Patrik Johansson, Voichita Dana Marinescu, Nanna Lindberg, Anna Segerman, Grzegorz Wicher, Mia Niklasson, Sathishkumar Baskaran, et al. The human glioblastoma cell culture resource: validated cell models representing all molecular subtypes. *EBioMedicine*, 2(10):1351–1363, 2015.

[300] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[301] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view intact space learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2531–2544, 2015.

[302] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[303] Yu-Meng Xu, Chang-Dong Wang, and Jian-Huang Lai. Weighted multi-view clustering with feature selection. *Pattern Recognition*, 53:25–35, 2016.

[304] Xiaohong R Yang, Mark E Sherman, David L Rimm, Jolanta Lissowska, Louise A Brinton, Beata Peplonska, Stephen M Hewitt, William F Anderson, Neonila Szeszenia-Dąbrowska, Alicja Bardin-Mikolajczak, et al. Differences in risk factors for breast cancer molecular subtypes in a population-based study. *Cancer Epidemiology and Prevention Biomarkers*, 16 (3):439–443, 2007.

[305] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4): e15–e15, 2002.

[306] Yongliang Yang, S James Adelstein, and Amin I Kassis. Target discovery from data mining approaches. *Drug discovery today*, 17:S16–S23, 2012.

[307] Masato Yano, Hirotaka J Okano, and Hideyuki Okano. Involvement of hu and heterogeneous nuclear ribonucleoprotein k in neuronal differentiation through p21 mrna post-transcriptional regulation. *Journal of Biological Chemistry*, 280(13):12690–12699, 2005.

[308] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[309] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2005.

[310] Shihua Zhang, Qingjiao Li, Juan Liu, and Xianghong Jasmine Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27(13):i401–i409, 2011.

[311] Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research*, 40(19):9379–9391, 2012.

[312] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[313] Etay Ziv, Manuel Middendorf, and Chris H Wiggins. Information-theoretic approach to network modularity. *Physical Review E*, 71(4):046117, 2005.

[314] Linlin Zong, Xianchao Zhang, Long Zhao, Hong Yu, and Qianli Zhao. Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *Neural Networks*, 88:74–89, 2017.