ABSTRACT

Title of Dissertation: NEW STATISTICAL METHODS FOR HIGH-DIMENSIONAL DATA WITH COMPLEX STRUCTURES

Qiong Wu
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Shuo Chen and Benjamin Kedem
Department of Mathematics

The overwhelming advances in biomedical technology facilitate the availability of high-dimensional biomedical data with complex and organized structures. However, due to the obscured true signals by substantial false-positive noises and the high dimensionality, the statistical inference is challenging with the critical issue of research reproducibility and replicability. Hence, motivated by these urgent needs, this dissertation is devoted to statistical approaches in understanding the latent structures among biomedical objects, as well as improving statistical power and reducing false-positive errors in statistical inference.

The first objective of this dissertation is motivated by the group-level brain connectome analysis in neuropsychiatric research with the goal of exhibiting the connectivity abnormality between clinical groups. In Chapter 2, we develop a likelihood-based adaptive dense subgraph discovery (ADSD) procedure to identify connectomic subnetworks (subgraphs) that are systematically associated with brain disorders. We propose the statistical inference procedure leveraging graph properties and combinatorics. We validate the proposed

method by a brain fMRI study for schizophrenia research and synthetic data under various settings.

In Chapter 3, we are interested in assessing the genetic effects on brain structural imaging with spatial specificity. In contrast to the inference on individual SNP-voxel pairs, we focus on the systematic associations between genetic and imaging measurements, which assists the understanding of a polygenic and pleiotropic association structure. Based on voxel-wise genome-wide association analysis (vGWAS), we characterize the polygenic and pleiotropic SNP-voxel association structure using imaging-genetics *dense* bi-cliques (IGDBs). We develop the estimation procedure and statistical inference framework on the IGDBs with computationally efficient algorithms. We demonstrate the performance of the proposed approach using imaging-genetics data from the human connectome project (HCP).

Chapter 4 carries the analysis of gene co-expression network (GCN) in examining the gene-gene interactions and learning the underlying complex yet highly organized gene regulatory mechanisms. We propose the interconnected community network (ICN) structure that allows the interactions between genes from different communities, which relaxes the constraint of most existing GCN analysis approaches. We develop a computational package to detect the ICN structure based on graph norm shrinkage. The application of ICN detection is illustrated using an RNA-seq data from The Cancer Genome Atlas (TCGA) Acute Myeloid Leukemia (AML) study.

NEW STATISTICAL METHODS FOR HIGH-DIMENSIONAL DATA WITH COMPLEX STRUCTURES

by

Qiong Wu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Shuo Chen, Chair/Advisor
Professor Benjamin Kedem, Co-Advisor
Professor Yan Li
Professor Vince Lyzinski
Professor Tianzhou Ma

Acknowledgments

I would like to acknowledge all the people who contributed and made this thesis possible.

First and foremost, I would like to extend my deepest gratitude to my advisor, Dr. Shuo Chen, for his invaluable supervision, continuous support, and unceasing encouragement throughout years of studying and researching. I acknowledge his guidance, which allows me to work on several interesting problems and build me in this field of expertise. I really appreciate his kindness in believing in my potentials and giving me the advice to be prepared for my future career. Dr. Chen has been always providing professional advice as a patient friend. I am feeling so lucky to have him as my advisor.

Further, yet importantly, I would like to thank my co-advisor, Dr. Benjamin Kedem for his continued guidance and advisory. I was exposed to interesting problems and fresh ideas through valuable discussions with him. I am deeply grateful to Dr. Tianzhou (Charles) Ma for being always supportive. Without his persistent help, the goal of this thesis would be hard to reach. I would also like to thank Dr. Yuan Zhang who worked closely with us in chapters of this thesis. I greatly benefit from his extensive knowledge of network theory and his passion for the field. In addition, I wish to thank Drs. Yan Li and Vince Lyzinski for their willingness to serve on my thesis committee and their invaluable time and effort in the process. Their insightful questions, constructive comments, and

valuable suggestions remarkably enriched this thesis.

I am also grateful for having the chance to study in the Statistics program. During the past years, the faculty I met stimulate me to not only become a knowledgeable statistician, but also an enthusiastic researcher. I also want to thank M. Cristina Garcia and Thomas J. Haines for being always helpful throughout my stay in the department.

Last but not least, I must give my most sincere thanks to my family and friends for their tremendous support, encouragement, and companionship. Without them, none of my work would have been possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACC	Anterior Cingulatr Cortices
ADSD	Adaptive Dense Subgraph Discovery
AML	Acute Myeloid Leukemia
BNMTF	Bounded Non-Negative Matrix Tri-Factorization
CNV	Copy Number Variants
DTI	Diffusion Tensor Imaging
FA	Fractional Anisotropy
FDR	False Discovery Rate
fdr	Local False Discovery Rates
fMRI	Functional Magnetic Resonance Imaging
FWER	Family-wise Error Rate
GCN	Gene Co-expression Networks
HC	Healthy Controls
HCP	Human Connectome Project
HK-relax	Heat Kernel-Based Community Detection
ICA	Independent Component Analysis
ICN	Interconnected Community Network
IGDB	Imaging-Genetics Dense Bi-clique
INS	Insular Gyri
IPL	Inferior Parietal Lobe
KL	Kullback-Leibler
LD	Linkage Disequilibrium
MLE	Maximum Likelihood Estimator
MMSB	Mixed Membership Stochastic Blockmode
MRI	Magnetic Resonance Imaging

NBS	Network Based Statistics
NMF	Non-negative Matrix Factorization
NP	Nondeterministic Polynomial-time
OCN	Overlapping Community Network
OrG	Orbitofrontal Cortex
PCA	Principal Component Analysis
PCL	Precentral Gryi
PPI	Protein-Protein Interaction
rfRMI	Resting state fMRI
RFT	Random Field Theroy
ROI	Reion of Interest
SBM	Stochastic Block Model
SFG	Superior Frontal Gyri
SN	Salience Network
SNP	Single Nucleotide polymorphism
SNR	Signal-to-Noise Ratio
STG	Superior Temporal Gyri
SZ	Schizophrenia
TCGA	The Cancer Genome Atlas
TNR	True Negative Rate
TPR	True Positive Rate
TPM	Transcripts Per Million
vGWAS	Voxel-wise Genome-Wide Association Analysis
WGCNA	Weighted Gene Co-expression Network Analysis

Chapter 1: Introduction

With the advent of advanced biomedical technology, high-dimensional data analysis has attracted widespread interest due to its applications in modern biomedical studies. For example, brain connectome analysis in neuropsychiatric research targets to identify functional connectivities related to brain disease (e.g., Schizophrenia) using functional magnetic resonance imaging (fMRI) data, which yields millions of brain voxels. The imaging-genetic studies with the goal of modeling the predictive mechanism of genetic markers on quantitative imaging measures rely on the analysis of billions of SNP-voxel pairs.

To date, the major works in high dimensional statistics have focused on various research topics. For instance, the dimensional reduction that transforms the data to a lower-dimensional representation (PCA, factor analysis, ICA, NMF, etc. [4, 5]), variable screening [6, 7], variable selection to determine set of non-zero coefficients (penalized regression [8, 9, 10], Bayesian models [11, 12]), multiple testing [13, 14, 15, 16], covariance matrix estimation [17, 18, 19, 20] among many others.

1.1 Multiple Testing Corrections

The modern applications of high-dimensional statistical inference motivate the need to consider a sequence of simultaneous hypothesis tests. In large-scale brain imaging data, the multiple testing problem arises when we are interested in testing the imaging observation at each voxel, specifically,

$$H_0^v : \theta_v \in \Theta_0^v \text{ v.s. } H_1^v : \theta_v \notin \Theta_0^v$$

with $v \in V$ corresponding to a set of brain voxels.

Bonferroni correction has been widely accepted to control the family-wise error rate (FWER), while a step-down procedure based on permutation test is developed by Westfall and Young in [21] for less conservative controls. Besides, the concept of False Discovery Rate (FDR) is first proposed in [13], and is employed through algorithms including: Benjamini–Hochberg procedure (BH-FDR) [13, 14], positive FDR (pFDR, [22]) and two-stage Benjamini, Krieger, & Yekutieli FDR procedure (BKY-FDR, [23]). Later, another alternate local false discovery rate (fdr) is proposed by Efron in [15] which relies on the empirical Bayes estimates of the mixture densities. These classical multiple testing corrections apply a universal threshold on test statistics/p-values of each voxel, such that the significant voxels may widespread among the whole brain which results in less biological interpretability. Meanwhile, since the true signals are often obscured by substantial false-positive noises in brain imaging data, a direct application of the classical multiple testing corrections may result in high false-positive findings.

Consequently, cluster-wise inference becomes increasingly popular in the field of neuroimaging. A cluster is defined as a set of suprathreshold voxels/connections that form a connected component spatially. Since a cluster of contiguous voxels surpassing a threshold has less probability to exist comparing with isolated suprathreshold voxels [24], the cluster-wise inference can be more powerful than voxel-wise inferences [25]. Subsequently, cluster size, voxels intensity and the combination [26] are developed as cluster-wise statistics. Parametric (Random Field Theory (RFT), [27, 28]) and nonparametric methods (permutation tests, [29, 30, 31]) are successfully applied to approximate the null distributions of test statistics and control the FWER.

1.2 Network Analysis in High-dimensional Data

Networks play an increasing role in characterizing complex interactive structures among high-dimensional objects. A network consists of a discrete set of study units and their pairwise relationships. For instance, the interactive relationships between pairs of genes are gathered in gene co-expression networks (GCN). The brain can be modeled as a complex network with brain regions and their functional connectivities. Hence, the advances of network analysis techniques proceed with the understanding of the complex structures among high-dimensional objects.

Community detection is of high significance in network analysis to obtain insight into valuable topological structures. A community is considered to be a group of units (nodes) that have a closer relationship (edges) with each other compared with others. Community detection has been widely studied during the past decades, and many community detection algorithms have been proposed including: clustering-based methods [32, 33], modularity-based methods [34, 35], spectral methods [1], etc.

Analyzing the gene co-expression patterns is one possible application of community detection [36, 37]. Genes are characterized as nodes in a network, while the interactive relationships are represented by edges. The strength of relationships is calculated by the similarity of co-expression patterns across subjects and reflected by edge weights. The community detection methods divide the set of genes into homogeneous groups, such that genes in the same community have similar expression patterns.

1.3 Overview

With the emphasis on brain connectome analysis and imaging-genetics studies, the application of classic multiple testing corrections may lead to no positive findings, since no single test statistics can pass the stringent cut-off due to the ultra-high dimensionality. Moreover, although the cluster-wise inference controls the FWER with maintaining high sensitivity, the available methods identify clusters as connected components which is spatially connected but not spatially constrained (or constrained in its conceptual network) in brain connectome data.

Hence, in Chapter 2, we focus on the group-level whole-brain connectome data, and target in extracting disease-related subnetworks with statistical inference. We propose a likelihood-based adaptive dense subgraph discovery (ADSD) model. Our method is robust to both false positive and false negative errors of edge-wise inference and thus can lead to a more accurate discovery of latent disease-related connectomic subnetworks.

We present the ADSD objective function, a generalization based on graph ℓ_0-norm, the statistical inference framework, and the associated algorithms in Section 2.2. Section 2.3 constructs theoretical results to guarantee the convergence properties of both objective functions. We apply the proposed approach to a brain fMRI study for schizophrenia research in Section 2.4, which identifies well-organized and biologically meaningful subnetworks that exhibit schizophrenia-related salience network centered connectivity abnormality. Analysis of synthetic data displayed in Section 2.5 also demonstrates the superior performance of the ADSD method for latent subnetwork detection in comparison with existing methods in various settings.

The purpose of Chapter 3 is the systematic investigation of genetic effects on brain structures and functions with spatial specificity using imaging-genetics data. We attempt to identify underlying organized association patterns of SNP-voxel pairs and understand the polygenic and pleiotropic networks on brain imaging traits. We develop computational strategies to detect latent SNP-voxel *bi-cliques* and inference model for statistical testing in Section 3.2 and 3.3. We further provide theoretical results to guarantee the performance of our computational algorithms and statistical inference. We validate our method by extensive simulation studies in Section 3.4, and then apply it to a voxel-wise genome-wide association analysis based on genetic data and white matter integrity data of 1042 participants from S1200 data release of the human connectome project (HCP) in Section 3.5.

In addition, the clustering of genes with similar expression patterns into groups implies a block-diagonal structure for the gene co-expression network. However, the real gene co-expression data may yield a more complicated network structure with interconnected communities. For example, genes from different communities might be enriched in signaling pathways, such that genes involve synergistic interactions with each other in a biological process. In Chapter 4, we develop a new computational package to extract interconnected communities from GCNs. We consider a pair of communities be interconnected if a subset of genes from one community is correlated with a subset of genes from another community. The interconnected community structure is more flexible and provides a better fit to the empirical co-expression matrix. To overcome the computational challenges, we develop efficient algorithms by leveraging advanced graph norm shrinkage approach in Section 4.2. We apply our interconnected community detection method to an RNA-seq data from The Cancer Genome Atlas (TCGA) Acute Myeloid Leukemia (AML) study and identify essential interacting biological pathways related to the immune evasion mechanism of tumor cells in Section 4.3. We validate and show the advantage of our method by extensive simulation studies in Section 4.4. Chapter 4 is a recall of recent work in [38].

Chapter 2: Statistical Inference for Group-level Brain Connectome Data

2.1 Introduction

Brain connectome analysis has become a powerful tool to understand the neurophysiology and neuropathology of brain diseases at a circuit level. These analyses focused on investigating patterns of functional and/or structural inter-connections between neural populations in the central nervous system associated with symptomatic phenotypes. Mounting evidence has shown that major neuropsychiatric disorders, including schizophrenia, Alzheimer's disease, and autism among others, are associated with disrupted structural and functional connectivity patterns [39].

Recent advances in neuroimaging statistics have facilitated group-level statistical analysis of structural and functional brain connectome data and the identification of disease-related brain connectome patterns [40, 41, 42]. In these analyses, the brain is often depicted as a graph [43], where each node corresponds to a brain region of interest (ROI) and an edge represents the connectivity linking any two nodes. An edge can represent functional connectivity based on functional magnetic resonance imaging (fMRI) data at rest or task, structural connectivity measuring white matter track connections, and weighted connection metric integrating multimodal brain connectivity [44]. These multivariate edges are the variables of interest in brain connectome analysis, which are constrained by the nodes in a weighted adjacency matrix and thus exhibit network topological properties [45]. Statistical inference for multivariate edge variables in an adjacency matrix remains challenging because of the need to account for multiple testing corrections and network topological structures simultaneously. Many statistical graph models have been developed and successfully applied to brain connectome data analysis yielding important findings (46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56 among many).

The current study focuses on extracting informative/signal subgraphs that are likely related to brain diseases from the whole brain connectome [46]. Our overarching goal is to accurately capture underlying signal subnetwork, such that the extracted subnetwork i) cover a high proportion of true positive edges (i.e., high sensitivity); (ii) include a few false positive edges (i.e., low false discovery rate (FDR)); and iii) are composed of highly organized network topological structures (i.e., biologically interpretable). In practice, however, this task is challenging because it is difficult to simultaneously balance the sensitivity and false positive findings while constraining all positive edges in organized subgraphs. The 'dense' subnetwork detection then becomes attractive because a subgraph of a small number of nodes in an organized network topological structure covering most signal edges can also lead to low FDR and high sensitivity. Although less discussed in the statistical literature, dense subgraph discovery in the field of computer science research has been carefully worked out (57), and thus may suit our needs for statistical analysis of brain connectome data.

Dense subgraph discovery methods are designed to identify a subgraph with a maximal density among all possible subgraphs, in short, the densest subgraph, in a binary graph. These methods rely on the assumption that the overall graph is non-random and there exists some subgraph where the edge ratios are much higher than the rest of the graph. [58] reduces the problem to a sequence of max-flow/min-cut computations, which requires a logarithmic number of min-cut calls. [59] propose a simple and fast greedy algorithm that is showed to have a 2-approximation guarantee by [60]. Nevertheless, existing dense subgraph discovery algorithms may not be directly applicable to our analysis due to the substantial noise in the group level brain connectome data. As demonstrated in Figure 2.1a, there may exist an enormous amount of false positive and false negative errors in edge-wise inference results that give rise to the difficulty of detecting dense subgraph using existing methods. Specifically, due to the noise, existing dense subgraph discovery algorithms tend to either identify over-sized subgraphs that may include a large proportion of false positive edges with low importance levels (high FDR) or detect over-conservative small-sized subgraphs that may not sufficiently cover signal edges (low sensitivity, 61). Moreover, the computational cost of many these dense subgraph density discovery algorithms is expensive, which may lead to intractable computational time for commonly used statistical inference methods for brain connectome analysis (e.g., permutation tests). Hence, we are motivated to integrate modern statistical techniques into dense graph discovery and mitigate these challenges for brain connectivity network analysis.

We propose a likelihood-based adaptive dense subgraph discovery (ADSD) model to extract informative connectomic subnetworks accurately. The new objective function is robust to edge-wise false positive and false negative noise by introducing a tuning parameter to balance the area density and degree density [61]. We optimize the tuning parameter objectively by maximizing the widely used likelihood function in statistical network/graph model (e.g. stochastic block model, 62, 63, 64). We develop efficient algorithms to implement the joint objective function of the ADSD. We further derive theoretical results which guarantee the approximation properties for any fixed graph and consistency for large graphs based on the proposed algorithm. In addition, we extend the adaptive density metric to a general multiple-subgraph setting via ℓ_0 graph norm shrinkage, and optimize the objective function through an efficient algorithm. We develop theoretical results to show the accuracy of the model estimation by the ℓ_0-norm based objective function. We then construct theoretical foundation for statistical inference on disease-related subgraph and implement permutation tests to approximate p-values with multiple testing correction [65, 66]. Our method is then applied to a resting state fRMI (rfRMI) brain connectomic study for schizophrenia research. The results of our real data analysis reveal for the first time systematic aberrant salience network centered connectivity patterns in schizophrenia patients using whole brain connectome network analysis. Although some of our findings coincide with previous studies using seed voxel-based method, our analysis is more comprehensive and less biased, because it does not require pre-selected seed sets or focuses on exclusively known networks. We perform extensive simulation studies to validate the proposed model and theoretical conclusions. The results demonstrate improved accuracy of informative subgraph detection in various settings.

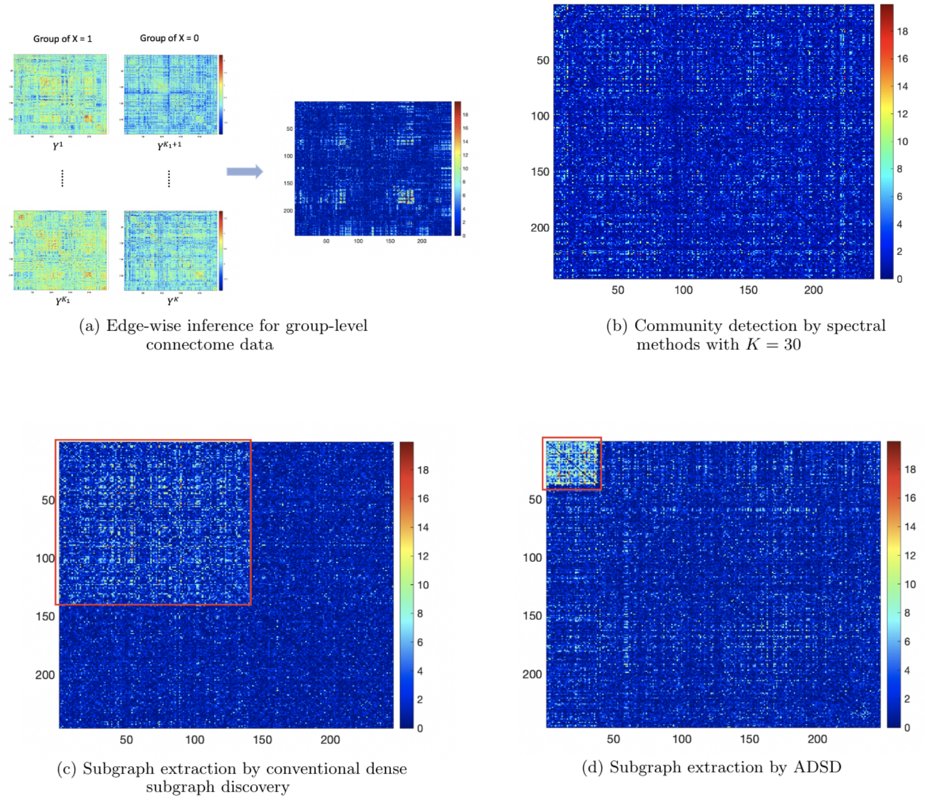
2.2 Methods

2.2.1 Background: group-level inference for multivariate edges in a graph space

Let $G = (V, E)$ be an undirected graph, where $V = \{v_i\}_{i=1}^n$ is a set of nodes representing brain areas and ROIs, and $E = \{e_{ij}\}_{i<j}^n$ denotes the set of edges between pairs of nodes (i.e, connections between brain areas). $G' = (V', E')$ is a subgraph of G if $V' \subset V$ and $E' \subset E$. Then, $G(S) = (S, E(S))$ is a 'nodes-induced' subgraph if $S \subset V$ and $E(S) = \{(u, v) \in E | u, v \in S\}$ being edges in E with endpoints in S.

We use $(\mathbf{Y}^k, X^k)_{k=1}^K$ to represent the group-level multivariate edge data in a graph space $G = (V, E)$, where $k = 1, \cdots, K$ is the subject index. $\mathbf{Y}^k_{n\times n}$ represents the brain connectome data in a binary/weighted adjacency matrix for subject k, and X^k is the corresponding vector of covariates (clinical and demographic variables). We assume that the location of nodes and edges are identical across subjects after spatial normalization. Thus, our goal is to perform statistical analysis and identify phenotype-related subnetworks with high sensitivity and well-controlled FDR [46, 47, 51, 52, 56]. Figure 2.1a demonstrates the procedure of group-level inference for brain connectome data.

Let $\boldsymbol{W} = \{w_{ij}\}_{i,j=1}^n$ denote the edge-wise inference matrix based on graph G, where each off diagonal entry w_{ij} represents the edge-wise statistical inference results on edge e_{ij} (e.g., test statistics t_{ij} and p values $-\log(p_{ij})$). For each edge e_{ij}, we denote a corresponding latent indicator variable δ_{ij} such that $\delta_{ij} = 1$ if edge e_{ij} is associated with the phenotype of interest and $\delta_{ij} = 0$ otherwise. We consider the edge-wise inference



(a) Edge-wise inference for group-level connectome data

(b) Community detection by spectral methods with $K = 30$

(c) Subgraph extraction by conventional dense subgraph discovery

(d) Subgraph extraction by ADSD

Figure 2.1: Motivation for informative subgraph extraction: (a) demonstrates the process of obtaining edge-wise inference matrix from the population level connectome data; (b) illustrates the commonly used community detection results (e.g. using stochastic block model) cannot detect any informative subgraph; (c) shows the results of existing dense subgraph discovery results; (d) describes a desirable informative subgraph detection procedure which can identify an organized and biologically interpretable topological structure consisting of informative edges. The results in (d) are based on the ADSD method (see details in the Results section).

results as our input data [67].

The goal of group-level brain network analysis is to identify a set of subnetworks $\{\hat{G}_c\}$ ($\hat{G}_c = (\hat{V}_c, \hat{E}_c)$) that are associated with a phenotype of interest, such that

1. $\Pr(\delta_{ij} = 1 | e_{ij} \in \hat{G}_c) > \Pr(\delta_{ij} = 1 | e_{ij} \notin \hat{G}_c)$ (dense subgraph);

2. The false discovery rate (FDR), $\frac{\sum_{i<j} I(\delta_{ij}=0|e_{ij} \in \hat{G}_c))}{\sum_{i<j} I(e_{ij} \in \hat{G}_c))}$, is low;

3. The sensitivity, $\frac{\sum_{i<j} I(\delta_{ij}=1|e_{ij} \in \hat{G}_c)}{\sum_{i<j} I(\delta_{ij}=1)}$, is high;

4. $\hat{G}_c$ is a well defined community (e.g., a node-induced subgraph that $\hat{G}_c = G(\hat{V}_c)$).

In practice, the task above is challenging. For example, the mass univariate methods including both FDR and family-wise error rate (FWER) controlling models apply an universal threshold on all edges, and yield a set of unrelated 'significant' edges. Thus, they can neither address the trade-off between sensitivity and false positive findings by leveraging the information of network or yield findings with an organized and biologically interpretale network topological structure. The network based statistics (NBS) method allows edges borrow strengths from each other, yet it yields an unorganized subgraph [39, 66]. Moreover, the signal subnetwork detected by NBS includes all nodes in G *almost surely*, i.e., $G(V_c) = G$, when n is larger than a handful of nodes [68], and thus less interpretable.

We also notice that the proportion of true positive edges $\frac{\sum I(\delta_{ij}=1)}{|E|}$ in G is often small in our motivated brain connectome data (e.g., around 5%), which may lead to the difficulty of applying the commonly used network models [69]. Figure 2.1b shows the results of the application of spectral methods which miss the network topological structure. Therefore, it is highly desirable to extract a ‘dense’ subgraph which is a node-induced subgraph $G(S_0)$, such that the edge density ρ_s is much higher than the overall density ρ:

$$\rho_s > k\rho,\, k > 1, \text{ with } \rho_s \triangleq \frac{\sum_{i<j, i,j \in S_0} I(\delta_{ij} = 1)}{|E(S_0)|} \text{ and } \rho \triangleq \frac{\sum_{i<j} I(\delta_{ij} = 1)}{|E|}, \tag{2.1}$$

and $G(S_0)$ includes most edges. The detected informative subgraph can either directly become the subnetwork of interest or intermediate results for further refined network analysis (e.g., using SBM). Since $\{\delta_{ij}\}$ is unknown, we adopt the weighted edge set $\{w_{ij}\}$ by assuming that $E(w_{ij}|\delta_{ij} = 1) > E(w_{ij}|\delta_{ij} = 0)$ for dense subgraph discovery.

2.2.1.1 Dense Subgraph Discovery

The conventional dense subgraph aims to detect a node-induced subgraph with maximized density. Two popular definitions of density function are also referred to as average degree and edge ratio [57, 59, 60]:

$$f_1 = \frac{|W(S)|}{|S|} \text{ and } f_2 = \frac{|W(S)|}{|E(S)|},$$

where $|W(S)| = \sum_{i,j \in S} w_{ij}$ and $|E(S)| = \binom{|S|}{2}$ for weighted graphs. The edge ratio agrees with our goal for informative subgraph detection. However, the implementation of dense subgraph discovery is not trivial. The direct optimization of edge ratio f_2 tends to detect a high-density subgraph with a tiny size. Meanwhile, it has been known the optimization of f_1 can lead to the detection of an over-sized subgraph [61], which may cause a high false positive rate for statistical inference. Figure 2.1c shows the results of conventional dense graph discovery by optimizing f_1. To address these challenges, we propose a likelihood based method for dense subgraph discovery.

2.2.2 Adaptive dense subgraph discovery (ADSD)

We consider $G = (V, E, \boldsymbol{W})$ as our input data that stores edge-wise inference results in a weighted adjacency matrix $\boldsymbol{W}$. Our goal is to extract a phenotype-related informative subgraph $G(S)$ induced by nodes set S in the sense that $E(w_{ij}|e_{ij} \in E(S)) \gg E(w_{ij}|e_{ij} \notin E(S))$ while maximally reducing false negative findings and improving the sensitivity.

To address the challenges in conventional dense subgraph discovery and improve the balance of the trade-off, we propose an adaptive density function:

$$f(S; \lambda) := \frac{|\boldsymbol{W}(S)|}{|S|^{\lambda}} \tag{2.2}$$

for $S \subset V$, where $\lambda \in [1, 2]$ is a tuning parameter, such that when $\lambda = 1$ and 2, the maximization of $f(S; \lambda)$ density function reduces to f_1 and f_2, respectively. For f_2, $|E(S)| = \binom{|S|}{2} \approx |S|^2/2$.

To better illustrate the impact of the tuning parameter λ on the FDR and sensitivity, we transform the optimization of objective function (2.2) to:

$$\arg\max_{S \subset V} \left\{ \log f_2(S) + \lambda' \log \frac{|\boldsymbol{W}(S)|}{|\boldsymbol{W}(G)|} \right\} \tag{2.3}$$

with $\lambda' \in [0, 1]$. The optimal solution is approximated by $f(S; \lambda)$ for large graphs with $\lambda' = \frac{2}{\lambda} - 1$. The first term in (2.3) is the true discovery rate $(1 - \text{FDR})$, while the second term is the sensitivity (power). In that, λ functions similarly to the tuning parameter in the shrinkage methods (e.g., LASSO) since f_1 is related to the loss function and f_2 implements the rule of parsimony. Increasing λ leads to a low FDR, while decreasing λ can improve the sensitivity. Therefore, our objective function is tailored for the four items of our overarching goal.

In practice, both $G(S)$ and λ need to be estimated, and λ is critical to balance the trade-off between FDR and sensitivity. We propose an iterative procedure to optimize the objective function (2.2) in subsection 2.2.2.1 and estimate λ in subsection 2.2.2.2. We name this new procedure adaptive dense subgraph discovery (ADSD).

2.2.2.1 Optimization with a known λ

We implement the objective function (2.2) using a greedy algorithm. The greedy algorithm has been the most commonly used technique to implement objective functions for dense subgraph discovery [59, 60]. Generally, a greedy algorithm removes a node with the minimum-degree at each iteration, and then selects the optimal dense subgraph from the process of node removal. The detailed procedure is described by Algorithm 1.

We denote the optimal dense subgraph based on our objective function (2.2) with a given λ by

$$S^*_\lambda = \arg\max_{S \subset V} f(S; \lambda),$$

Algorithm 1 Optimizing objective function (2.3) with a given λ

```
procedure ALGORITHM
    $S_1 \leftarrow V$
    for k=1 to $n-1$ do
        let v be the node in $G(S_k)$ with smallest degree: $v = \arg\min_{i \in S_k} \deg_{G(S_k)}(i)$;
        $S_{k+1} \leftarrow S_k / \{v\}$;
    end for
    Output the subgraph with largest objective function among $G(S_1), ..., G(S_{n-1})$;
end procedure
```

and the output of greedy algorithm as:

$$\tilde{S}_\lambda = \underset{S_1,\ldots,S_{n-1}}{\arg\max} f(S;\lambda).$$

A major advantage of the greedy algorithm is the low computational complexity, which is critical for our application. Although our greedy algorithm may not provide the exact solution, [60] proved the greedy algorithm has a 2-approximation as to $f_1(\cdot)$, that is $f_1(\tilde{S}_1) \geq 2f_1(S_1^*)$, where $\tilde{S}_1$ is the densest subgraph by greedy algorithm and S_1^* is the true maximizer for $f_1(\cdot)$. In section 3 of this chapter, we prove the theoretical approximation properties of $\tilde{S}_\lambda$ with regard to the maximization $f(S^*;\lambda)$ for various values of λ.

2.2.2.2 Likelihood-based method for λ estimation

Clearly, the performance of our greedy algorithm 1 relies on the unknown parameter λ (e.g. $\lambda = 1$ and 2 lead to the optimization $f_1(\cdot)$ and $f_2(\cdot)$ alone respectively). We propose a data-driven approach to automatically determine λ by maximum likelihood estimation. In statistical literature, the likelihood function of network/graph data has been well studied [64]. For example, a binary graph with K block can be defined by:

$$A_{ij}|\theta_i = a, \theta_j = b \sim \text{Bernoulli}(\pi_{ab})$$

where $\boldsymbol{A}^{n \times n}$ is a binary adjacency matrix, $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ is a latent vector of node labels, and $\boldsymbol{\pi} = (\pi_{ab})_{a,b=1}^{K}$ is a $K \times K$ symmetric probability generating matrix for generative for edges within and between blocks/communities.

We adopt the likelihood function of SBM because the dense subgraph structure in our ADSD model can be considered as a special case of the block diagonal structure in SBM. Specifically, in our model the graph $G = (V, E)$ includes an underlying true informative subgraph $G(S_0)$ and all other nodes are singletons. The number of communities of SBM is $K = n - n_s + 1$, where $n = |V|$ and $n_s = |S_0|$. We further assume the planted partition model that the parameters of Bernoulli distributions for edges between blocks are identical in SBM.

To construct the likelihood function for ADSD, we first binarize the input data matrix $\boldsymbol{W}$ using a threshold r and let $A_{ij} = \{\boldsymbol{W}(r)\}_{ij} = I(w_{ij} > r)$. We denote $\boldsymbol{\theta}(S)$ as a vector of node labels concerning the node set S for a dense subgraph $G(S)$, where an element $\theta_i(S) = 1$ if $i \in S$ and $\theta_i(S) = 0$ for $i \in V/S$. Then, the membership of edges regarding the nodes-induced subgraph $G(S)$ can be defined consequently as $\theta_{ij}(S) = \theta_i(S)\theta_j(S)$.

We let all edges in A follow a Bernoulli distribution with parameters π_{ij} that

$$\pi_{ij} = \begin{cases} \pi_s & \text{both } i, j \in S_0 \\ \pi_0 & o.w. \end{cases} \tag{2.4}$$

Using the mixture model representation, $\pi_{ij} = \theta_{ij}(S_0)\pi_s + (1 - \theta_{ij}(S_0))\pi_0$.

When the membership of informative subgraph is given, the MLE of the edge probabilities can be obtained by:

$$\hat{\pi}_s^{MLE} = \frac{|\boldsymbol{A}(S_0)|}{|E(S_0)|} \text{ and } \hat{\pi}_0^{MLE} = \frac{|\boldsymbol{A}| - |\boldsymbol{A}(S_0)|}{|E| - |E(S_0)|}.$$

In practice, S_0 is unknown and can be estimated by $\tilde{S}_\lambda$ from the Algorithm 1 with a given λ. The likelihood function based on $\tilde{S}_\lambda$ is in the form:

$$\begin{aligned} L_\lambda(\hat{\pi}_s^{MLE}, \hat{\pi}_0^{MLE}; \boldsymbol{\theta}(\tilde{S}_\lambda), \boldsymbol{A}) = & \prod_{i<j, i,j \in \tilde{S}(\lambda)} (\hat{\pi}_s^{MLE})^{a_{ij}} (1 - \hat{\pi}_s^{MLE})^{1-a_{ij}} \\ & \times \prod_{i<j, i \in V/\tilde{S}(\lambda) \text{ or } j \in V/\tilde{S}(\lambda)} (\hat{\pi}_0^{MLE})^{a_{ij}} (1 - \hat{\pi}_0^{MLE})^{1-a_{ij}} \end{aligned}$$

where $\boldsymbol{\theta}(\tilde{S}_\lambda)$ is the node label vector associated with $\tilde{S}_\lambda$. Therefore, λ can be estimated by two steps. First, for any $\lambda \in [1, 2]$, we can extract a dense subgraph $\tilde{S}_\lambda$ by the greedy algorithm 1. Next, $\hat{\lambda}$ is determined by the combination of λ and $\tilde{S}_\lambda$ that maximizes the likelihood function:

$$\hat{\lambda} = \arg\max_\lambda L_\lambda(\hat{\pi}_s^{MLE}, \hat{\pi}_0^{MLE}; \boldsymbol{\theta}(\tilde{S}_\lambda), \boldsymbol{A}),$$

The final result of dense subgraph discovery based on the MLE determined $\hat{\lambda}$ is $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\tilde{S}_{\hat{\lambda}})$.

We further consider the threshold r in $A_{ij} = \{\boldsymbol{W}(r)\}_{ij} = I(w_{ij} > r)$ as a random variable following a distribution $g(r)$ rather than a fixed value in order to avoid an arbitrary selection. We integrate the likelihood with respect to r based on the prior distribution $g(r)$, and thus our optimization is invariant to the selection of r. $g(r)$ can be a discrete distribution with a support $\{r_1, ..., r_m\}$ and corresponding probability $\{g(r_1), ..., g(r_m)\}$. In practice, the performance of our algorithm is robust to the prior distribution, given the reasonable support of r is used. By integrating r out, the likelihood function becomes:

$$L_\lambda(\hat{\pi}_s^{MLE}, \hat{\pi}_0^{MLE}; \boldsymbol{\theta}(\tilde{S}_\lambda), \boldsymbol{W}) = \int L_\lambda\left(\hat{\pi}_s^{MLE}, \hat{\pi}_0^{MLE}; \boldsymbol{\theta}(\tilde{S}_\lambda), \boldsymbol{W}(r)\right) g(r)dr$$

The general algorithm for ADSD is described in the Algorithm 2. Since Algorithm 1 is nested within the overall Algorithm 2, the low computational cost of Algorithm 1 is critical for the overall computational efficiency of ADSD. The complexity of the ADSD algorithm is $O(Mn^2)$ where M is a sufficient searching range of λ. The resulting subgraph $G(\hat{S}_{\hat{\lambda}})$ from our ADSD model can be further investigated for more delicate latent topological structures and statistically tested by permutation tests with family-wise error rate control [66, 70].

2.2.3 Subgraph extraction via ℓ_0 graph norm penalty

In the section, we resort to an ℓ_0 graph norm penalty based objective function to extract multiple dense subgraphs from $\boldsymbol{W}$. In this case, we express the topological

Algorithm 2 The complete ADSD algorithm

```
procedure ALGORITHM $\lambda \leftarrow 1$
    while $\lambda \leq 2$ do
        return the densest subgraph $\tilde{S}(\lambda)$ of $\boldsymbol{W}$ from Algorithm 1
        for $r = r_1$ to $r_m$ do
            calculate the likelihood: $L_\lambda(\hat{\pi}_s^{MLE}, \hat{\pi}_0^{MLE}; \boldsymbol{\theta}(\tilde{S}_\lambda), \boldsymbol{W}(r))$
        end for
        integrate w.r.t. $\lambda$:
    $L_\lambda(\boldsymbol{W}) = \sum_{i=1}^{m} L_\lambda(\hat{\pi}_s^{MLE}, \hat{\pi}_0^{MLE}; \boldsymbol{\theta}(\tilde{S}_\lambda), \boldsymbol{W}(r_i))g(r_i)$
    end while
    Output $\hat{\lambda}$ and $\tilde{S}_{\hat{\lambda}}$ with maximized $L_\lambda(\boldsymbol{W})$
end procedure
```

structure of G as

$$G = \left(\oplus_{c=1}^{C} G_c\right) \cup G_0$$

where each $G_c = \{V_c, E_c\}$ is a phenotype-related subnetwork and $G_0 = \{V_0, E_0\}$ is the rest of G. In other words, G is structured as a union of C phenotype-related *subnetworks* $G_1, \ldots, G_C$ and *singleton* nodes that do not belong to any subnetwork. Then, we follow similar idea as the ADSD: for any detected subgraph $\hat{G}_c$, we reward edge weights within this subnetwork while penalizing on its size (i.e., increasing density and subnetwork size). This objective function can lead to the discovery of a set of subgraphs with the maximal size and density. Specifically, we define

$$\boldsymbol{U} = (u_{ij})_{i,j} = \boldsymbol{W} * G, \quad \text{that is,} \quad u_{ij} = w_{ij} \cdot \delta_{ij} \tag{2.5}$$

where "$*$" denotes Hadamard (element-wise) matrix multiplication. Clearly, $\boldsymbol{U}$ depends on the specified structure of the underlying graph $G = (\delta_{ij})_{i,j}$. Define $\|\boldsymbol{U}\|_1 = \sum_{i,j} |u_{ij}|$ and $\|\boldsymbol{U}\|_0 = \sum_{i,j} I(|u_{ij}| > 0)$, where $\| \quad \|_1$ and $\| \quad \|_0$ are matrix element-wise ℓ_1 and ℓ_0 norms. Our core proposal is the following ℓ_0 graph norm shrinkage criterion:

$$\arg\max_{G,\tilde{C}} \log ||\boldsymbol{U}||_1 - \lambda_0 \log ||\boldsymbol{U}||_0 \tag{2.6}$$

where λ_0 is a tuning parameter. The objective function (2.6) jointly estimates the number of subgraphs and the subgraph memberships of all nodes in $G = \oplus_{c=1}^{C} G_c \cup G_0$. The objective function (2.6) maximize the edge weights with minimally sized subgraphs, which is mathematically equivalent to extract maximally sized subgraphs while maximizing the density. Therefore, the optimization of (2.6) is governed by two conflicting goals: covering high-weight informative edges and using minimally sized subgraphs. Maximizing the first term $||\boldsymbol{U}||_1$ can increase sensitivity by allocating a maximal number of high-weight edges into subgraphs, which promotes large subgraphs. In the meanwhile, we penalize the ℓ_0 graph norm to maximize the density of subgraphs. The second term can also suppress false positive noise because false positive edges tend to be distributed in a random pattern in G rather than an organized subgraph [70].

The tuning parameter λ_0 balances the two conflicting terms. Specifically $\lambda_0 = 0$ would send all nodes to one subnetwork, while a large λ_0 prefers small communities and singletons (nodes not in any community, thus contributing zero ℓ_0 graph norm) even to the true community structure. In our theoretical analysis, we find that for $\lambda_0 \in (0, 1)$, if μ_0/μ_1 is less than an upper bound dependent on λ_0, our criterion provides a consistent estimation of the community structure, thus well-controlling the rates of two types of errors in the multiple testing procedure. In practice, we can select λ_0 based on likelihood as ADSD.

2.2.3.1 Implementation

We optimize (2.6) and extract dense subgraphs using Algorithm 3. Specifically, we perform grid search for C. For each value of $C = C^\dagger$, let $\hat{G}(C^\dagger)$ be the estimated network structure by optimizing (2.6), and $\boldsymbol{U}_{\hat{G}(C^\dagger)}$ is the corresponding matrix from Hadamard matrix multiplication. $\boldsymbol{U}_c$ is the submatrix of $\boldsymbol{U}$ corresponding to G_c. Intuitively, both $\|\boldsymbol{U}_c\|_1$ and $\|\boldsymbol{U}_c\|_0$ decrease with an increasing value of C. The outcome provides a set of maximal subnetworks with high density. We provide the theoretical guarantee for the consistency and optimality of $\hat{C}$, $(\hat{G}_c)_{c=1,\ldots,\hat{C}}$ in section 2.3.

Algorithm 3 Implementation of ℓ_0-norm based subgraph extraction

```
procedure ALGORITHM $C^\dagger \leftarrow 2$
    while $C^\dagger =\leq n-1$ do
        Optimize $\arg\max_{G(C^\dagger)} \sum_{c=1}^{C^\dagger} \frac{\|\boldsymbol{U}_c\|_1}{\|\boldsymbol{U}_c\|_1^{\lambda_0}}$ through spectral methods
        Select $\hat{C}$ that $\arg\max_{C^\dagger=2,\cdots,n-1} \log\|\boldsymbol{U}_{\hat{G}(C^*)}\|_1 - \lambda_0 \log\|\boldsymbol{U}_{\hat{G}(C^*)}\|_0$
    end while
    Output $\hat{C}$, $(\hat{G}_c)_{c=1,\ldots,\hat{C}}$
end procedure
```

2.2.4 Statistical inference for phenotype-related subgraphs

We start to consider the primary problem of testing the existence of the subnetwork structure. Particularly, we are testing

$$\begin{aligned} &H_{G;0} : C = 0, \quad \text{that no phenotype-related subnetwork exists,} \\ &H_{G;a} : C > 0, \quad \text{that at least one phenotype-related subnetwork exists.} \end{aligned} \tag{2.7}$$

Recall the SBM likelihood we constructed for ADSD, $\boldsymbol{A}$ is the binarized adjacency matrix (corresponding to a binary network $G[r]$) of $\boldsymbol{W}$ using threshold r [16], such that $\boldsymbol{A}$ follows a mixture of Bernoulli distributions with probabilities π_s and π_0. Then, under $H_{G;0}$, we have that $G[r]$ is an Erdős-Renyi graph with parameter π_0. For any $\gamma \in (\pi_0, 1)$, we call a subgraph of this binary graph a "γ-quasi clique" if its observed edge density is at least γ. Define $G[r;\gamma]$ to be the largest-in-size γ-quasi clique in $G[r]$ and let $|G[r]|$ denote its size. Next, we show that we can reject the null based on these two properties of a subgraph (i.e., density γ and size $|G[r]|$), which can be conveniently extended to testing individual subgraphs.

Lemma 2.1. *Let $\boldsymbol{A}$ be a binary network with independent edges.*

- *Suppose $H_{G;0} : C = 0$ is true, that is $E[A_{ij}] = \pi_0$. Assume that for any $\gamma \in (\pi_0, 1)$, $v_0 = \omega(\sqrt{n})$ and n large enough such that $\left\{2/3 + 2(\gamma - \pi_0)^{-1}\right\}^{-1} v_0 \geq \log n$, we have*

$$(|G[r;\gamma]| \geq v_0 | H_{G;0}) \leq 2n \cdot \exp\left(-\left\{\frac{2}{(\gamma-\pi_0)^2} + \frac{2}{3(\gamma-\pi_0)}\right\}^{-1} \cdot v_0^2\right)$$

- *Suppose $H_{G;a} : C \geq 1$ is true. Assume that all subnetworks satisfy $G_c = \omega(\sqrt{n})$ and set $v_0 = c_0\sqrt{n}$ for a small enough constant $c_0 > 0$, such that $\min_{c=1,\ldots,C^*} |G_c| \geq v_0$, we have*

$$(|G[r;\gamma]| \geq v_0 | H_{G;a}) \geq 1 - \exp\left\{-\frac{(q-\gamma)^2 v_0(v_0-1)/4}{1+(q-\gamma)/3}\right\}$$

Lemma 2.1 states that i) the probability of a large and dense subnetwork G_c existing under H_0 is almost zero; whereas ii) the probability of a large and dense subnetwork

G_c existing under H_1 is approaching 1. Lemma 2.1 provides the theoretical foundation for our subnetwork-wise inference. Given an estimated $\hat{G}_c$ with high density and large size, we can conveniently reject the null hypothesis by applying the results of Lemma 2.1. Therefore, the statistical inference for a phenotype-related subnetwork $\hat{G}_c$ becomes testing on a statistic of the density and network size of $\hat{G}_c$. Built on this inference approach, the permutation tests [29, 66] can also effectively control the family-wise error rate to simultaneously testing multiple phenotype-related subgraphs $G_1, \ldots, G_C$ from $\boldsymbol{W}$ with multiple testing correction.

2.3 Theoretical Results

The theoretical work for conventional dense graph discovery has been well-established [57]. For example, [60] showed that the commonly used greedy algorithm proposed by [59] has a 2-approximation bound. In this chapter, we aim to extend the theoretical results for our new ADSD algorithms in 2.3 which generalizes the traditional objective function by introducing the parameter λ. Specifically, we discus the approximation bounds for ADSD with a full range of λ values in the following theorem 2.1.

Theorem 2.1 (Exact property of Algorithm 1)**.** *For a given graph $G = (V, E)$, with S^*_λ and $\tilde{S}_\lambda$ defined in section 2.2, the Algorithm 1 has a $\rho(\lambda, n)$-approximation, especially $f(S^*_\lambda; \lambda) \leq \rho(\lambda, n) f(\tilde{S}_\lambda; \lambda)$ with*

$$\rho(\lambda, n) = \begin{cases} c(\lambda) & \textit{if}\, \lambda \geq 2 \\ 2c'(\lambda)n^{(\lambda-1)(2-\lambda)} & \textit{if}\, 1 < \lambda < 2 \\ 2n^{1-\lambda} & \textit{if}\, 0.5 < \lambda < 1 \\ 2n^{\lambda}, & \textit{if}\, 0 < \lambda \leq 0.5 \end{cases}$$

where $c(\lambda) = 2^{\lambda-1}$ and $c'(\lambda) = 1 \vee 2^{1-\lambda}$.

The theorem 2.1 provides the performance of Algorithm 1 by guaranteeing the closeness of objective function in S^*_λ and $\tilde{S}_\lambda$. However, an optimal optimization may not result from a perfect recovery of informative subgraph for randomness (i.e. $S^*_\lambda \neq S_0$ for all λ). Hence, we further prove the asymptotic consistency of $\tilde{S}_\lambda$ from Algorithm 1 in following theorem 2.2. When the observed graph is generated from some underlying model with true informative subgraph S_0, there exist an λ such that $\tilde{S}_\lambda$ tends to S_0 with probability 1 asymptotically.

Theorem 2.2 (Asymptotic property of Algorithm 1)**.** *Assume the graph $G = (V, E)$ including an informative subgraph $G(S_0) = (S_0, E(S_0))$ is generated from the special SBM we defined in section 2.2.2.2, such that the edges are drawn from independent Bernoulli distributions with parameter $\pi_{ij} = \pi_{ij}(S_0) = \theta_{ij}(S_0)\pi_s + (1 - \theta_{ij}(S_0))\pi_0$, where $\theta_{ij}(S) = \theta_i(S)\theta_j(S)$, $\theta_i(S) = I(i \in S)$ and $\pi_s > \pi_0$. Let $|S_0| = O(|V|^{1/2+\epsilon})$ as $n \to \infty$ for any $\epsilon > 0$.*

Then, there exist some λ such that we will get exact recovery with probability 1 in

Algorithm 1, i.e. as $n \to \infty$,

$$\mathbb{P}(\forall i, \theta_i(\tilde{S}_\lambda) = \theta_i(S_0)) \to 1.$$

Theorem 2.2 provides the existence of parameter λ for a consistent estimator as the size of graph goes to infinity. We use the following Theorem 2.3 to demonstrate the performance of Algorithm 2 by illustrating the selected λ based on our likelihood-based criterion will lead to an estimator with negligible proportion of incorrect assignment for nodes.

Theorem 2.3. *Assume the graph $G = (V, E)$ includes an informative subgraph $G(S_0) = (S_0, E(S_0))$, such that the edges generate from independent Bernoulli distributions with parameter $\pi_{ij} = \pi_{ij}(S_0) = \theta_{ij}(S_0)\pi_s + (1 - \theta_{ij}(S_0))\pi_0$, where $\theta_{ij}(S) = \theta_i(S)\theta_j(S)$, $\theta_i(S) = I(i \in S)$ and $\pi_s > \pi_0$. Let $|S_0| = O(|V|^{1/2+\epsilon})$ as $n \to \infty$ for any $\epsilon > 0$.*

Then, as $n \to \infty$, the adaptive greedy algorithm with likelihood-based criterion results in an estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\tilde{S}_{\hat{\lambda}})$ with:

$$\hat{\lambda} = \arg\max_{\lambda} \sup_{\pi_s, \pi_0} L(\pi_s, \pi_0; \boldsymbol{\theta}(\tilde{S}_\lambda); \boldsymbol{A})$$

has incorrect assignment with probability converging to zero, i.e.

$$N_e(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} I(\hat{\theta}_i \neq \theta_i(S_0)) = o_p(n).$$

Furthermore, we present theoretical guarantees for the ℓ_0-norm based subgraph extraction. We first show the theoretical guarantee on the correctness of (2.6) optimization.

Theorem 2.4 (Consistency of subnetwork detection by (2.6))**.** *Let C^* be the true number of subnetworks and $\boldsymbol{\mu} = (\mu_0, \mu_1, ..., \mu_{C^*}, \mu_{C^*+1})$ denote their true proportions with $\|\boldsymbol{\mu}\|_1 = 1$. Suppose that the tuning parameter is set to be $\lambda_0 \in (0, 1)$, and assume*

$$\frac{\pi_0}{\pi_s} < \begin{cases} \frac{(C^*)^{\lambda_0} - 1}{C^* - 1}, & \textit{if } C^* \geq 2 \\ \lambda_0, & \textit{if } C^* = 1 \end{cases} \tag{2.8}$$

Then asymptotically, our criterion (2.6) is uniquely optimized by $C = C^*$ and $G_c = G_c^*$ for all $c = 1, \ldots, C^*$.

Theorem 2.4 ensures that by optimizing (2.6), we can learn the correct number of subnetworks. This optimization is combinatorial and difficult to carry out in practice. Yet, Theorem 2.4 also suggests that our criterion (2.6) can also serve model selection when combined with efficient subnetwork estimation procedures under each candidate C. In view of this, next, we present the theoretical guarantee for a computationally efficient estimation procedure for subnetwork detection under $C = C^*$. Let us define some notation. Recall the definitions of π_0, π_s, and define $\sigma_0^2 = (w_{ij}|\delta_{ij} = 0)$ and $\sigma_s^2 = (w_{ij}|\delta_{ij} = 1)$. Let $\boldsymbol{P} = [\boldsymbol{W}|G] = \boldsymbol{\Theta}\boldsymbol{\Omega}\boldsymbol{\Theta}^T$ denote the expectation matrix, where $\boldsymbol{\Theta} \in \{0, 1\}^{n \times (C+1)}$ is a membership matrix that contains exactly one "1" and all others "0" in each row. Here, $\boldsymbol{\Theta}_{i,(C+1)} = 1$ means that node i is a singleton node outside the subnetwork structure.

Theorem 2.5 (Consistency of spectral estimation under $C = C^*$)**.** *Assume that rank*$(\boldsymbol{P}) = C^*+1$*, and denote its smallest absolute nonzero eigenvalue by* ξ_n*. Assume* $(\pi_s \vee \sigma_s^2 \vee \sigma_0^2) \leq \alpha_n$ *for* $\alpha_n \geq c_0 \log n/n$ *and* $c_0 > 0$*. Then, if* $(2+\varepsilon)\frac{(C+1)n\alpha_n}{\xi_n^2} < \tau$ *for some* $\tau, \varepsilon > 0$*, the output* $\hat{\boldsymbol{\Theta}}_{C^*}$ *from the spectral estimation is consistent up to a permutation. Equivalently, if* $\hat{V}_c$ *is the estimated nodes set for subgraph* G_c*,* $c = 1, ..., C^*$*. Then* $\hat{V}_c \cap V_c$ *is the set in* V_c *that the assignment of nodes can be guaranteed, and with probability at least* $1 - n^{-1}$*, up to a permutation, we have*

$$\sum_{c=1}^{C} \left[1 - \frac{\left|\hat{V}_c \cap V_c\right|}{|V_c|} \right] \leq \tau^{-1}(2+\varepsilon)\frac{Cn\alpha_n}{\xi_n^2}.$$

The detailed derivations and proofs for the above theorems are provided in the Appendix.

2.4 Data Example

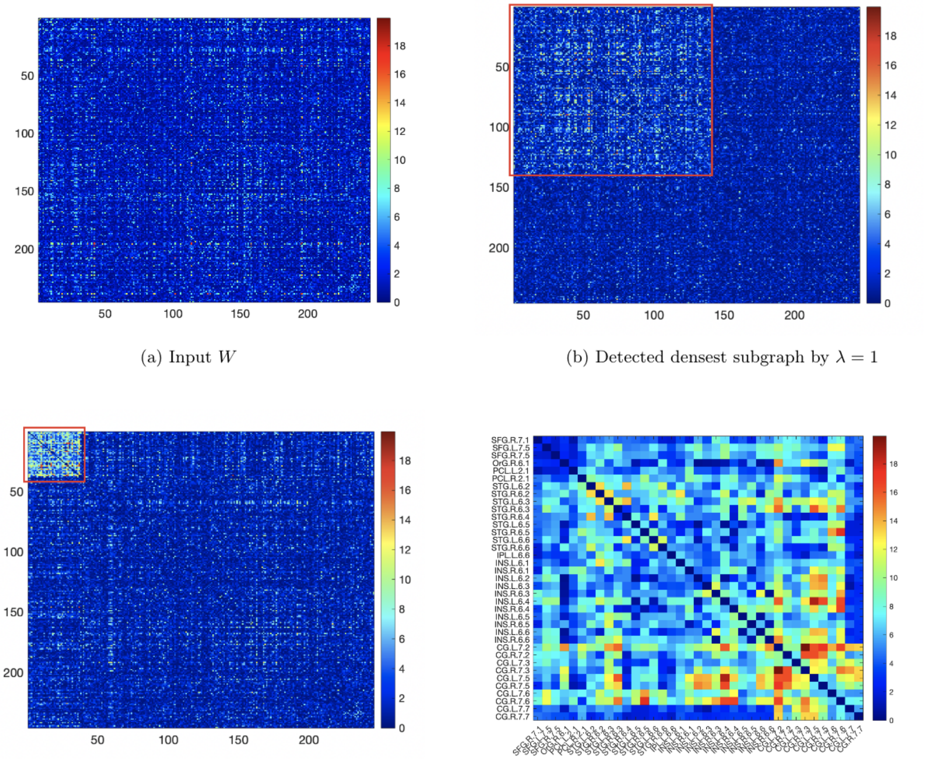
We apply the proposed ADSD method to the neuroimaging data collected from patients with schizophrenia and healthy controls. This data set includes 104 patients with schizophrenia (SZ) (age 36.88 ± 14.17, 62 males and 41 females, 1 other) and 124 healthy controls (HC) (age 33.75 ± 14.22, 61 males and 63 females). There are no systematic differences in age (test statistic 1.64, p value 0.10) or gender (test statistic 1.67, p value 0.10) between the two groups. The imaging acquisition and preprocessing details are described in Adhikari et al. [71]. A brain connectivity-based atlas is used to denote 246 regions of interest (ROIs) as nodes in a brain connectome graph [72]. The functional connection (edge) between a pair of nodes for each subject is calculated by the covariation between averaged time series from the two corresponding brain ROIs. The Fisher's Z transformed Pearson correlation coefficient then is applied for each edge. We perform non-parametric group level testing on each edge, although alternative inference methods can be used as well.

We focus on the input matrix $\boldsymbol{W}$ reflecting the importance levels ($-\log(p_{ij})$) on all edges, as demonstrated in Figure 2.2a. We first apply the greedy algorithm (e.g., Charikar's method that is equivalent to the proposed greedy algorithm with an ad-hoc $\lambda = 1$) for dense subgraph extraction. The results in Figure 2.2b seem to be an over-inflated subnetwork without clear biological interpretation and a large set of false positive edges. We also applied other popular subgraph detection methods, for example, breadth first search in network-based statistics, stochastic block model, and various community detection methods [66, 73, 74]). However, these algorithms either detect a subgraph including all brain regions or yield no findings. In contrast, by implementing our ADSD method (2), we obtain a subnetwork $\hat{S} = \tilde{S}_{\hat{\lambda}}$ with $\hat{\lambda} = 1.2$. We note that the detected subgraph is robust to the prior distribution of $G(r)$ as long as a reasonable support is used.

The computation is efficient, and it takes 2.21 seconds to implement the ADSD algorithm on a Mac with CPU Core i5 and memory 8GB. We further calculate the p-value of the network based on permutation test [45, 66, 70]. The p-value for the network is significant $p < 0.001$ with family wise error rate adjustment.

The results show a subnetwork with reduced functional connectivity in patients with schizophrenia compared to healthy controls (see in Figures 2.3), which is consistent with the current knowledge that schizophrenia is possibly a degenerative disorder and associated with hypoconnectivity [75]. This subnetwork is centered around the well-known salience network (SN) which is primarily composed of bilateral insular gyri (INS) and anterior cingulate cortices (ACC). The salience network contributes to complex and integrative brain functions including emotions, cognition, and self-awareness [76]. Numerous previous studies have reported that decreased functional connectivity in the salience network is related to several core symptoms of schizophrenia using seed voxel methods [77]. Our findings are well aligned with these established results. In addition to SN, our subnetwork extracted by ADSD involves several other brain regions including bilateral superior temporal gyri (STG), superior frontal gyri (SFG), precentral gryi (PCL), inferior parietal lobe left (IPL), and orbitofrontal cortex right (OrG). These regions have been identified to associate with auditory perceptual abnormalities (STG, IPL), voluntary movement (PCL), and sensory and cognition (SFG, OrG) [39]. Jointly, our detected subnetwork reveals a comprehensive and systematic brain connectivity aberrance in patients with schizophrenia, which is related to the impaired capability to integrate and comprehend information (e.g., multiple external stimuli) and to respond appropriately. The detected schizophrenia-related brain connectome subnetwork is biologically plausible. It provides evidence to combine prior isolated findings, and thus enhances our understanding of the complex brain connectomic patterns and clinical symptoms.

Thus, our novel analytic approach revealed a neural sub-network that has been previously shown to both differentiate healthy controls and patients with schizophrenia and has been critically linked to core symptoms of the disorder. Since our results do not depend on the arbitrary selection of seed voxels and pre-specified networks of interest, our results are subject to less selection bias and thus more reliable and comprehensive.



(a) Input W

(b) Detected densest subgraph by $\lambda = 1$

(c) Detected densest subgraph by ADSD

(d) Enlarged subgraph with labels

Figure 2.2: Results of data example: (a) is the input matrix $\boldsymbol{W}$; (b) shows the results of existing dense graph discovery; (c) demonstrates the results by applying ADSD; (d) illustrates refined topological structure based on results of ADSD.

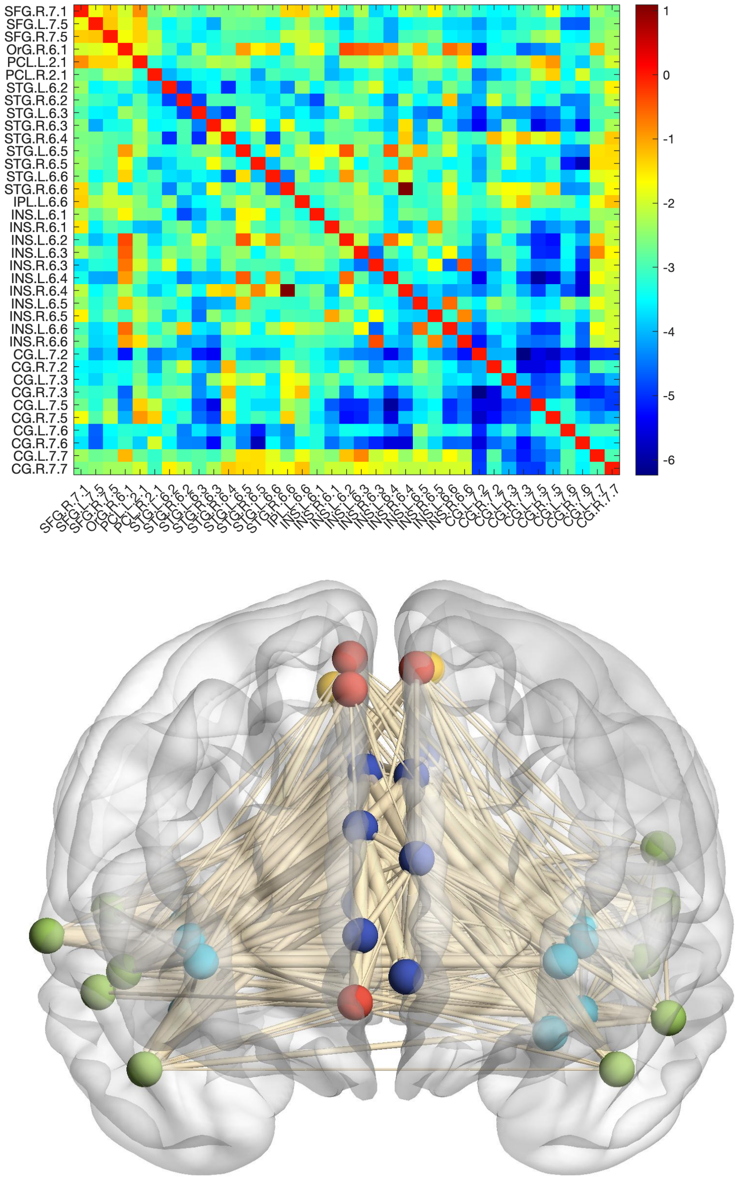


Figure 2.3: Results of data example: (a) illustrates the enlarged and labeled informative subgraph in t-statistics detected by ADSD which indicates decreased functional connectivity of SZ. (b) is a 3D demonstration of the subgraph: red nodes represent superior frontal gyrus (SFG) + orbitofrontal cortex right (OrG); yellow nodes are precentral gryi (PCL); green nodes are superior temporal gyrus (STG)+inferior parietal lobe left (IPL); blue nodes represent insular gyrus (INS); navy nodes represents cingulate cortex (CG).

2.5 Simulation Studies

In the simulation study, we generate multiple brain connectivity data sets under several settings. We consider a graph G with $|V| = 100$, and set an informative subgraph in a community structure with two possible sizes $|S_0| = 15$ and 30. We simulate connectivity matrices with different sample sizes (cases v.s. control): 30 v.s. 30 and 60 v.s. 60. We assume that most edges in the informative subgraph are differentially expressed between cases and healthy controls. We let the connectivity weights of edges inside the informative subgraph follow a normal distribution with mean μ_1 and variance σ^2, while all other edges have normal μ_0 and σ^2 for the case group. In the control group, we let all edges follow a normal distribution of μ_0 and σ^2. Specifically,

$$x_{ij(s)}^{\text{case}} | \{i < j, i, j \in S_0\} \sim N(\mu_1, \sigma^2), \quad x_{ij(s)}^{\text{case}} | \{i < j, i \text{ or } j \notin S_0\} \sim N(\mu_0, \sigma^2),$$

$$\text{and} \quad x_{ij(s')}^{\text{control}} | \{i < j, i, j \in V\} \sim N(\mu_0, \sigma^2),$$

where $x_{ij(s)}^{\text{case}}$ represents the edge linking node i and j for the sth subject in case group, and $x_{ij(s')}^{\text{control}}$ defines the edge weight for the s'th subject in control group.

We apply various standard effect sizes (i.e., signal-to-noise ratios - SNRs) by setting $\sigma = 1$, and $\mu_0 = 0$, $\mu_1 = 0.6$ and 0.8. We further consider a more realistic scenario by letting the proportion q_1 of edges inside informative-subgraph be non-differentially expressed (i.e. $N(\mu_0, \sigma^2)$ for both cases and controls). Similarly, we set a q_2 proportion of edges outside informative-subgraph are differentially expressed (i.e. $N(\mu_1, \sigma^2)$ for cases and $N(\mu_0, \sigma^2)$ for controls). (q_1, q_2) represent the practical non-perfect distribution of informative edges in the overall graph. In the simulation data, two sets of parameters $(q_1, q_2) = (0.8, 0.1)$ and $(0.9, 0.05)$ are used.

We compare the ADSD method with the two most popular dense subgraph discovery methods including Greedy algorithm with $\lambda = 1$ and Goldberg's algorithm. The results are evaluated by node-assignment accuracy in terms of true positive rate (TP) and true negative rate (TN) defined as follows:

$$TP = \frac{\sum_{i=1}^{n} I(\theta_i = \hat{\theta}_i = 1)}{\sum_{i=1}^{n} I(\theta_i = 1)}, \quad TN = \frac{\sum_{i=1}^{n} I(\theta_i = \hat{\theta}_i = 0)}{\sum_{i=1}^{n} I(\theta_i = 0)}$$

The mean and standard errors of TP and TN for three methods across 30 replicates for all settings are displayed in the following Tables 2.1 and 2.2. For ADSD, we report the estimated tuning parameter $\hat{\lambda}$ and the size of the selected subgraph $|\tilde{S}_{\hat{\lambda}}|$. $|\tilde{S}_1|$ denotes the size of subgraph detected by the Greedy algorithm and $|\hat{S}|$ by Goldberg's method.

Tables 2.1 and 2.2 demonstrate results of sample sizes 30 v.s. 30 and 60 vs. 60 respectively. In general, the performance of all algorithms is satisfactory when sample size, subgraph size, and effect size is large. When noise presents and either and subgraph size is small (i.e., the scenario for most brain connectome data analysis), ADSD outperforms the competing methods with much improved sensitivity.

We further performed permutation tests on the detected subgraph for network-level inference. We summarize the results in terms of False Negative error (n-FN) rate and False Positive error (n-FP) rate in Table 2.3. In general, the performance of ADSD inference is satisfactory except when sample size, subgraph size, and effect size are all small. The average computational time for each simulated data set is around one minute on a PC with

Table 2.1: The node-assignment accuracy of three methods under varied SNRs ($\mu_1 = 0.6, 0.8$) and subgraph sizes for 30 cases and 30 controls

			$(q_1, q_2) = (0.8, 0.1)$		$(0.9, 0.05)$	
$\lvert S_0 \rvert$	Methods		0.6	0.8	0.6	0.8
15	ADSD	TP	0.903 (0.107)	0.979 (0.044)	0.963 (0.055)	0.997 (0.020)
		TN	0.873 (0.093)	0.989 (0.024)	0.982 (0.050)	0.999 (0.004)
		$\hat{\lambda}$	1.206 (0.050)	1.150 (0.046)	1.032 (0.066)	1.001 (0.037)
		$\lvert\tilde{S}_{\hat{\lambda}}\rvert$	24.31 (17.48)	15.63 (2.20)	16.00 (4.48)	15.07 (0.47)
	Greedy	TP	1 (0)	1 (0)	0.985 (0.037)	0.997 (0.013)
		TN	0.081 (0.197)	0.061 (0.032)	0.746 (0.366)	0.999 (0.004)
		$\lvert\tilde{S}_1\rvert$	93.11 (3.33)	94.85 (2.68)	36.40 (31.26)	15.05 (0.38)
	Goldberg	TP	0.989 (0.025)	0.989 (0.025)	0.973 (0.046)	0.986 (0.027)
		TN	0.093 (0.039)	0.073 (0.033)	0.764 (0.351)	0.999 (0.004)
		$\lvert\hat{S}\rvert$	91.96 (3.27)	93.67 (2.79)	34.67 (29.98)	14.89 (0.55)
30	ADSD	TP	0.987 (0.024)	1 (0)	0.997 (0.010)	1 (0)
		TN	0.991 (0.015)	0.999 (0.004)	0.998 (0.005)	1 (0)
		$\hat{\lambda}$	1.035 (0.089)	0.998 (0.023)	0.985 (0.051)	1 (0)
		$\lvert\tilde{S}_{\hat{\lambda}}\rvert$	30.25 (1.37)	30.06 (0.28)	30.04 (0.47)	30 (0)
	Greedy	TP	0.996 (0.012)	1 (0)	0.999 (0.007)	1 (0)
		TN	0.988 (0.017)	0.999 (0.003)	1 (0)	1 (0)
		$\lvert\tilde{S}_{\hat{\lambda}}\rvert$	30.70 (1.29)	30.06 (0.24)	29.95 (0.22)	30 (0)
	Goldberg	TP	0.985 (0.020)	0.989 (0.016)	0.988 (0.019)	0.989 (0.016)
		TN	0.987 (0.026)	0.999 (0.003)	1 (0)	1 (0)
		$\lvert\hat{S}\rvert$	30.49 (1.88)	29.74 (0.52)	29.63 (0.56)	29.68 (0.47)

an i7 CPU 3.60 GHz and 64GB memory.

In summary, the simulation results clearly show that likelihood-based ADSD approach is more robust to both false positive and false negative noise and can better capture smaller subnetworks with a high sensitivity and a low false positive rate. These properties are critical for the brain connectome analysis in practice because the real data sets are often mixed with substantial noise and include a small proportion of signal edges.

Table 2.2: The node-assignment accuracy of three methods under varied SNRs ($\mu_1 = 0.6, 0.8$) and subgraph sizes for 60 cases and 60 controls

			$(q_1, q_2) = (0.8, 0.1)$		$(0.9, 0.05)$	
$\lvert S_0\rvert$	Methods		0.6	0.8	0.6	0.8
15	ADSD	TP	0.985 (0.035)	0.994 (0.023)	0.999 (0.009)	1 (0)
		TN	0.995 (0.014)	1.000 (0.003)	0.999 (0.003)	1 (0)
		$\hat{\lambda}$	1.139 (0.052)	1.105 (0.054)	0.998 (0.021)	1.002 (0.018)
		$\lvert\tilde{S}_{\hat{\lambda}}\rvert$	15.17 (1.25)	14.96 (0.42)	15.06 (0.28)	15 (0)
	Greedy	TP	1 (0)	1 (0)	0.999 (0.009)	1 (0)
		TN	0.063 (0.031)	0.067 (0.036)	1.000 (0.002)	1.000 (0.001)
		$\lvert\tilde{S}_1\rvert$	94.68 (2.64)	94.27 (3.07)	15.02 (0.24)	15.01 (0.10)
	Goldberg	TP	0.985 (0.028)	0.985 (0.028)	0.983 (0.029)	0.985 (0.028)
		TN	0.073 (0.032)	0.077 (0.036)	0.999 (0.003)	1.000 (0.002)
		$\lvert\hat{S}\rvert$	93.55 (2.60)	93.23 (2.98)	14.80 (0.45)	14.79 (0.41)
30	ADSD	TP	1.000 (0.003)	1 (0)	1 (0)	1 (0)
		TN	1.000 (0.002)	1 (0)	1 (0)	1 (0)
		$\hat{\lambda}$	1.000 (0.004)	1.001 (0.011)	1 (0)	1 (0)
		$\lvert\tilde{S}_{\hat{\lambda}}\rvert$	30.02 (0.20)	30 (0)	30 (0)	30 (0)
	Greedy	TP	1.000 (0.003)	1 (0)	1 (0)	1 (0)
		TN	1.000 (0.002)	1.000 (0.002)	1 (0)	1 (0)
		$\lvert\tilde{S}_{\hat{\lambda}}\rvert$	30.02 (0.20)	30.02 (0.14)	30 (0)	30 (0)
	Goldberg	TP	0.990 (0.016)	0.990 (0.015)	0.990 (0.015)	0.990 (0.015)
		TN	1.000 (0.002)	1.000 (0.002)	1 (0)	1 (0)
		$\lvert\hat{S}\rvert$	29.72 (0.53)	29.72 (0.49)	29.70 (0.46)	29.70 (0.46)

Table 2.3: The accuracy of permutation test under varied scenarios

(cases, controls)	(q_1, q_0)	$\lvert S_0 \rvert$		0.6	0.8
$(30, 30)$	$(0.8, 0.1)$	15	n-FP	0.340 (0.474)	0.010 (0.100)
			n-FN	0.010 (0.100)	0 (0)
		30	n-FP	0.020 (0.140)	0 (0)
			n-FN	0 (0)	0 (0)
	$(0.9, 0.05)$	15	n-FP	0.030 (0.171)	0 (0)
			n-FN	0 (0)	0 (0)
		30	n-FP	0 (0)	0 (0)
			n-FN	0 (0)	0 (0)
$(60, 60)$	$(0.8, 0.1)$	15	n-FP	0 (0)	0 (0)
			n-FN	0 (0)	0 (0)
		30	n-FP	0 (0)	0 (0)
			n-FN	0 (0)	0 (0)
	$(0.9, 0.05)$	15	n-FP	0 (0)	0 (0)
			n-FN	0 (0)	0 (0)
		30	n-FP	0 (0)	0 (0)
			n-FN	0 (0)	0 (0)

2.6 Discussion

In this chapter, we compare brain connectome matrices between diagnostic groups (e.g. schizophrenia and healthy subjects) to understand connectivity patterns altered by psychiatric illness. As in our motivation data example, however, phenotype-related subnetworks can be overwhelmed by substantial noise in the connectome data and thus difficult to extract. The noise heavily influences statistical inference by introducing enormous edge-wise false positive and negative errors that are constrained in a weighted adjacency matrix, and thus impose difficulty in understanding the network topology of phenotype-related brain circuits and in yielding valid statistical inference.

To overcome these challenges, we develop a novel ADSD method to reliably and robustly identify signal subgraphs (related to the phenotypes of interest) from the whole brain connectome network. The overall brain connectome inference network is often over-sized with a small proportion of signal edges which are not compatible with existing statistical network models. Therefore, it is desirable to detect a dense subnetwork maintaining most signal edges in a clique with a much smaller number of nodes (nodes induced subnetwork) and discarding a large proportion of false positive edges from the overall network. Dense graph discovery has been a popular research topic in network analysis for a couple of decades. Dense graph discovery methods are distinct from existing statistical methods for network analysis (e.g. various versions of community detection) because they focus on a network with a far fewer number of connections than a highly connected network consisting of communities. The dense graph discovery method is well suited for our application because the number of edges from the non-null distribution is relatively small [16]. A key limitation of the current dense graph discovery methods is sensitive to noise. Due to the substantial noise in brain connectome data, the existing dense graph discovery methods tend to extract over-sized dense subgraphs which can lead to a high FDR, potentially incorrect biological findings, and low replicability. The proposed ADSD method integrates the concept of shrinakge into dense graph discovery by introducing a balance parameter to include the most informative edges into the subgraph (high sensitivity) while maintaining a low FDR. The balance parameter can be estimated based on the likelihood function which is commonly used in network statistics. We develop efficient algorithms to implement the objective function that is compatible with computationally intensive inference methods (e.g., permutation tests and bootstraps) . In the current research, we apply permutation test based statistical inference on the dense subgraph.

Both the simulation and data example results show that the proposed method is robust to the false negative and positive edges and can accurately detect the target dense subgraph with high sensitivity and low false positive rates. Therefore, our goal of brain connectome analysis can be well met by applying ADSD.

Our work makes several contributions to the field: first, the ADSD and ℓ_0-norm based objective functions and algorithms provide new dense subgraph detection tools for noisy, weighted, large, and less dense graphs, which may have wide applications in data mining and knowledge discovery. Secondly, for ADSD algorithms, we derive theoretical results to provide the bounds for the approximation in a full range of the balance parameter. The asymptotic property of subgraph detection and balance parameter estimation are also developed. For ℓ_0-norm based algorithms, we provide the optimality of the objective function and error bounds. Last, the biological findings are novel, integrative, and clinically meaningful. Although part of these findings has been found in previous studies, only edge-wise results (i.e. links between regions to a fixed seed) are reported without fully investigating the interactive nature of network-level inference.

In this chapter, the hypo-connections in the salience network centered subnetwork groups in patients with schizophrenia are detected for the first time by whole brain connectome network analysis with explicit network topology. The reported network reveals the novel links between aberrant functional connectivity networks and impaired capability to integrate information from multiple sources (cognition deficits) in patients with schizophrenia, which may assist to further understand the underlying biological mechanism for multiple schizophrenic disorder symptoms.

In summary, we develop a likelihood-based adaptive dense graph detection method to extract the dense subgraph from a large and noisy network (weighted and/or binary). Our ADSD method outperforms existing dense subgraph discovery methods when the overall graph includes a small proportion of edges with high importance levels, and thus is well-suited for group-level brain connectome analysis. ADSD can also serve as a screening step for group level network analysis to effectively extract a dense subnetwork from a large overall network for further analysis. In addition, ADSD can be applied to other biological network data (e.g. interactive networks of genomics and proteomics data) and yield findings revealing latent and complex co-expression subnetworks. Therefore, ADSD can become a new useful tool for statistical analysis of large and less dense networks.

Chapter 3: A Multivariate-to-Multivariate Approach for Voxel-wise Genome-wide Association Analysis

3.1 Introduction

Imaging-genetics studies have garnered increased interest in the field of neuropsychiatric research. The joint application of whole genome sequencing and high-resolution imaging techniques is appealing because it can reveal the genetic effects on spatially specific brain functions and structures [78, 79, 80, 81]. The imaging-genetics analysis has becomes a new avenue to understand the genetic and neurological mechanisms for complex neuropsychiatric traits.

In imaging-genetics studies, both brain imaging data and genome sequence are measured for each participant. The genetic measurements can characterize genetic variations using single nucleotide polymorphism (SNP) and copy number variants (CNVs). The non-invasive brain imaging techniques assess the brain structures by magnetic resonance imaging (MRI), diffusion tensor imaging (DTI), and brain functions by functional magnetic resonance imaging (fMRI). The recent development of neuroimaging technology provides high-resolution imaging data with improved spatial specificity.

To date, the voxel-wise genome-wide association analysis (vGWAS) is a main approach in imaging-genetics studies to assess genetic architecture of structural brain imaging. However, the ultra-high dimensionality by combining imaging space (i.e., voxels) with whole genome (i.e., SNP) poses considerably computational challenges. Specifically, a typical imaging-genetics study collects roughly 10^7 SNPs and 10^5 voxels, which jointly contributes trillions (10^{12}) of SNP-voxel pairs [82, 83]. Thus, the statistical inference involves simultaneous massive-scale tests. The classic multiple testing methods [82, 84, 85] and voxel-wise inference methods [65, 86] have been first applied. However, the direct application of multiple testing correction, for example, false discovery rate (FDR), may lead to none positive findings, because no single SNP-voxel pairwise test p-value can pass the stringent cut-off due to the ultra-high dimensionality. In addition, various noise and heterogeneity in imaging-genetics data can further impede accurate inference. Other approaches such as advanced regression shrinkage models incorporating group sparse regularization [87, 88], and low rank regression models [80, 89] have been developed to fit multiple voxels in a joint model. Although enjoying numerous theoretical advantages, these methods are only applicable in summarized imaging features at ROIs due to computational burdens.

Most current statistical inference approaches treat each imaging-genetic interaction as an individual unit and disregard the systematic nature of genetic influence on human brains. Comparing with massive SNP-voxel pairs, the polygenic and pleiotropic pattern formed by genetic variants from different chromosomes and multiple distant brain areas is a more realistic characteristic of imaging-genetics associations. The detection and statistical test of systematic association patterns help with the interpretability and replicability of biological findings.

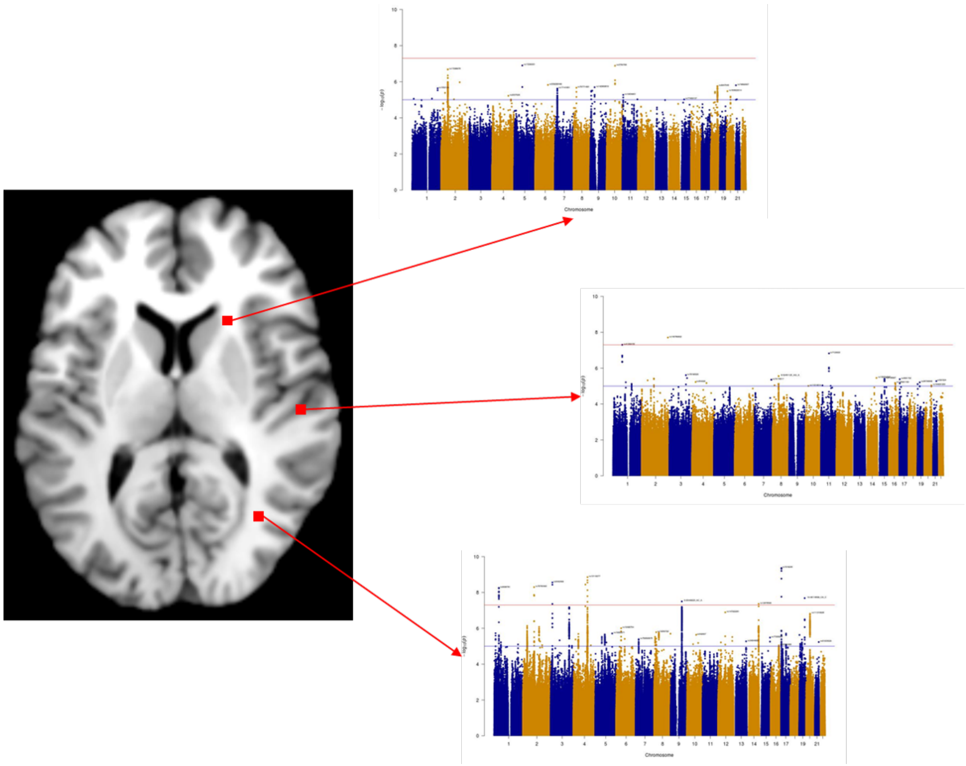


Figure 3.1: Data structure for vGWAS

In this chapter, we propose a new multivariate-to-multivariate method to detect and test polygenic and pleiotropic patterns based on vGWAS. Specifically, we consider associations between all SNP-voxel pairs as edges in a bipartite graph, and genetic variants and imaging voxels as two disjoint sets of nodes, correspondingly. We model the polygenic and pleiotropic SNP-voxel association structure as an imaging-genetics *dense* bi-clique (IGDB). IGDB is a node-induced subgraph consisting of a subset of SNPs and a subset of voxels, where the possibility of a SNP associated with a voxel is much elevated than the rest of the bipartite graph. Within an IGDB, each voxel can be considered as a polygenic imaging trait, and a SNP as a pleiotropic genetic variant. The existence of the polygenic and pleiotropic SNP-voxel association structure can be evaluated against a random bipartite graph. We then develop computationally efficient algorithms to extract the IGDB structure from the bipartite graph mixture model and thus provide sound estimates of parameters in the mixture model. Our inference on IGDB is constructed via likelihood ratio test based on the bipartite graph mixture model.

3.2 Methods

3.2.1 Background and notation

We consider an imaging-genetics data set collected from L independent subjects. We denote V as the set of brain imaging voxels with $|V| = n$. The imaging trait of a voxel $v \in V$ is $y_{v,l}$, and accordingly the vector of multivariate imaging traits is $\boldsymbol{y}_l = (y_{1,l}, ..., y_{n,l})^T$, for participant $l \in \{1, ..., L\}$. We let U be the set of genetic variants with $|U| = m$. Then, $\boldsymbol{x}_l = (x_{1,l}, ..., x_{m,l})^T$, $l = 1, ..., L$ represents the genetic variants for the participant l. Without loss of generality, we estimate the associations between multivariate imaging traits and multivariate genetic variants using a generalized linear regression model:

$$\mathbb{E}(\boldsymbol{y}_l|\boldsymbol{x}_l) = g^{-1}(\boldsymbol{B}^T\boldsymbol{x}_l + \boldsymbol{\alpha}^T\boldsymbol{z}_l),$$

where $g(\cdot)$ is a known link function with inverse function $g^{-1}(\cdot)$, $\boldsymbol{B} = \{\beta_{uv}\}_{u\in U, v\in V}$ is the $m \times n$ SNP-voxel association matrix, and β_{uv} represents the effect size of association between SNP u and voxel v with covariates $\boldsymbol{z}_l$ accounted. The goal of statistical inference is to accurately identify the subset of significant associations $\{\beta_{uv}\}$ via a sequence of hypotheses (15, 90, etc.):

$$H_0^{(u,v)} : \beta_{uv} = 0, \text{ versus } H_1^{(u,v)} : \beta_{uv} \neq 0,$$

for all $u \in U$ and $v \in V$.

A key distinction between univariate to multivariate inference (e.g., a single trait) and multivariate-to-multivariate inference is that the associations can be more constrained in multivariate-to-multivariate analysis when systematic association patterns present. For example, a cluster of genes can be associated with a cluster of voxels. The identification of the cluster-to-cluster association requires the joint modeling the global pattern and local/individual β_{uv} associations. Conventional inference methods (e.g., multiple testing correction or regression shrinkage) may only focus on a set of individual association pairs β_{uv} without recognizing the systematic patterns. To address this challenge, we propose a new multivariate-to-multivariate inference framework based on imaging-genetics dense bi-clique (IGDB).

3.2.2 Multivariate-to-multivariate inference from a graph perspective

We characterize the vGWAS association as a bipartite graph $G = (U, V, E)$, in which U and V are the two disjoint node sets representing sets of SNPs and voxels, respectively. Through the association matrix $\boldsymbol{B} = \{\beta_{uv}\}_{u \in U, v \in V}$, E denotes the edge set and $|E| \leq |U||V|$ such that $e_{uv} \in E$ if and only if $\beta_{uv} \neq 0$. In our application, we consider the global sparsity of connections between U and V and thus $|E| \ll |U||V|$.

In contrast to the inference on individual edges e_{uv}, we focus our study on the systematic association patterns with emphasis on pleiotropic and polygenic relationships. Particularly, we attempt to draw inference on the connections between neighborhoods of SNPs and voxels, $\mathcal{N}(e_{uv}) = \{e_{u'v'}, u' \in \mathcal{N}(u), v' \in \mathcal{N}(v)\}$. We restrict this structured associations in an IGDB subgraph in G. In general, a subgraph is defined as $G[S, T] = (S, T, E[S, T])$ with $S \subset U$, $T \subset V$, $E[S, T] = \{e_{uv} \in E | i \in S, j \in T\}$. We denote IGDB $G[S_0, T_0]$ as a special subgraph reduced by disjoint neighboring vertex sets:

$$S_0 = \mathcal{N}_{U,1} \cup ... \cup \mathcal{N}_{U,k_U} \text{ and } T_0 = \mathcal{N}_{V,1} \cup ... \cup \mathcal{N}_{V,k_V}$$

with concentrated imaging-genetic associations that:

$$\Pr(\beta_{uv} \neq 0 | \delta_{uv} = 1) > \Pr(\beta_{uv} \neq 0 | \delta_{uv} = 0), \tag{3.1}$$

where δ_{uv} is a binary variable indicating the IGDB-based network structure, i.e.,

$$\delta_{uv} \equiv \delta_{uv}(S_0, T_0) = I(e_{uv} \in G[S_0, T_0]),$$

or equivalently,

$$\mu_1 > \mu_0 \text{ for } \mu_1 = \frac{|E[S_0, T_0]|}{|S_0||T_0|} \text{ and } \mu_0 = \frac{|E| - |E[S_0, T_0]|}{|U||V| - |S_0||T_0|}.$$

Within an IGDB, genetic variants may come from different chromosomes, while imaging voxels may consist of multiple distant brain areas. This reflects that these imaging features (T_0) are polygenic traits and the genetic variants S_0 are pleiotropic alleles. The genetically correlated imaging features and functionally related SNPs jointly compose a functional biclique $G[S_0, T_0]$. In practice, a functional biclique $G[S_0, T_0]$ may further be decomposed into multiple sets of SNPs based on their linkage disequilibrium (LD) patterns and a few brain areas by the spatial contiguity constraint. Nevertheless, extracting a functional biclique $G[S_0, T_0]$ is critical because it provides a comprehensive association pattern between multivariate imaging and genetic features and the basis for the following steps (further decomposition). In the next subsection, we articulate that the IGDB enjoys several statistical advantages based on graph and combinatorics theories.

3.2.3 Graph properties of IGDB

Without loss of generality, we consider the probabilistic and graph properties of IGDB under the scenario of null hypothesis that G is a random bipartite graph (similar to

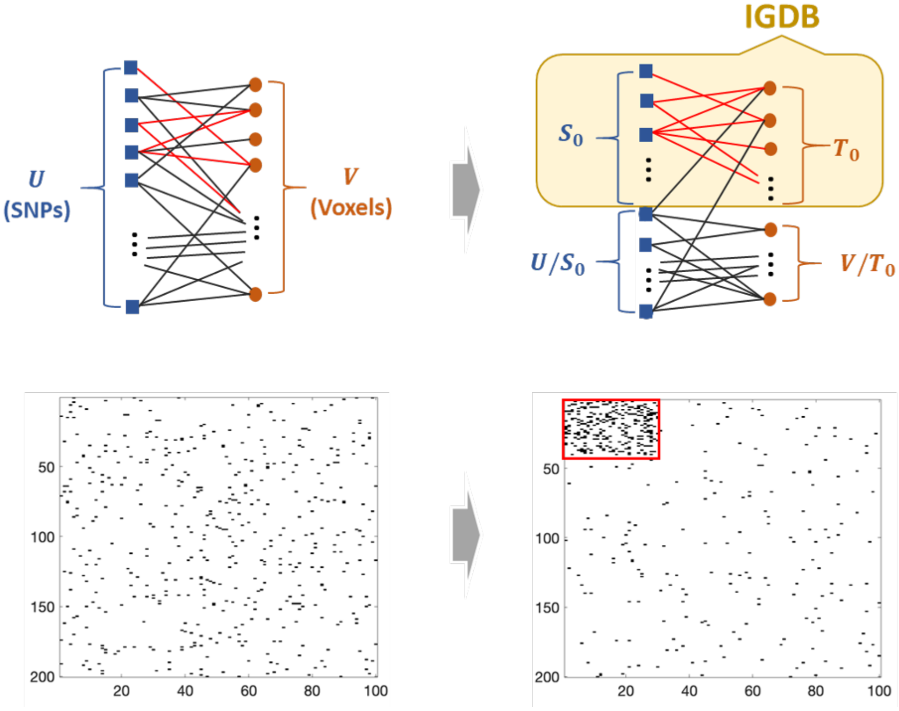


Figure 3.2: A demonstration of the bipartite graph with IGDB structure $G[S_0, T_0]$. The right subfigure indicates $G[S_0, T_0]$ in G with nodes reordered.

the derivation of exact tests). Specifically, the null and alternative hypotheses as follows:

$$\begin{aligned}
H_0: &\quad G \text{ is observed from a random bipartite graph } G(m, n, \mu_0),\\
H_1: &\quad \text{There exists an IGDB } G[S_0, T_0] \text{ such that}\\
&\quad e_{uv} \sim \begin{cases} \text{Bernoulli}(\mu_1), & \text{if } u \in S_0 \;\&\; v \in T_0 \\ \text{Bernoulli}(\mu_0), & \text{otherwise} \end{cases} \quad \text{with } \mu_1 > \mu_0.
\end{aligned}$$

We define a γ-quasi biclique as a subgraph with edge density at least γ, and denote as $G[S_\gamma, T_\gamma]$. Empirically, the probability to observe a γ-quasi biclique decrease with the subgraph size. In the following lemma, we specify the upper bound of the probability to observe a γ-quasi biclique via the subgraph sizes and edge densities under a random bipartite graph.

Lemma 3.1 (Under IGDB-wise null hypothesis)**.** *Suppose G is observed from a random bipartite graph $G(m, n, \mu_0)$. $G[S_\gamma, T_\gamma]$ is any subgraph with edge density $\frac{|E[S_\gamma, T_\gamma]|}{|S_\gamma||T_\gamma|} \geq \gamma \in (\mu_0, 1)$ (i.e., γ-quasi biclique). Let $m_0, n_0 = \Omega(\max\{m^\epsilon, n^\epsilon\})$ for some $0 < \epsilon < 1$. Then for sufficiently large m, n with $c(\gamma, \mu_0)m_0 \geq 8 \log n$ and $c(\gamma, \mu_0)n_0 \geq 8 \log m$, we have*

$$\mathbb{P}\left(|S_\gamma| \geq m_0, |T_\gamma| \geq n_0\right) \leq 2mn \cdot \exp\left(-\frac{1}{4}c(\gamma, \mu_0)m_0 n_0\right),$$

where $c(a, b) = \left\{\frac{1}{(a-b)^2} + \frac{1}{3(a-b)}\right\}^{-1}$.

Lemma 3.1 states that the probability to observe an IGDB decays on the edge density and exponentially on the size of the IGDB under the null. Thus, identifying an IGDB with a non-trivial size and high edge density from G suggests the rejection of the null hypothesis, which carries the theoretical background on the IGDB-wise inference. Our method is inspired by this finding and aims at the estimation and statistical inference of IGDB in imaging-genetics data.

3.3 Estimation and Inference

Let $\boldsymbol{W}_{m\times n}$ denote the inference result matrix (e.g., test statistics $w_{uv} = t_{uv}$ or $-\log(p_{uv})$) for the regression coefficients $\widehat{\boldsymbol{B}}_{m\times n}$. Then, our goal becomes to extract and test the IGDB structure from a weighted bipartite graph $G = (U, V, \boldsymbol{W})$.

3.3.1 IGDB estimation

We propose a new objective function for IGDB estimation which is inspired Lemma 3.1. We search for the maximal subgraph in G with a density constraint. Hence, we estimate the IGDB $G[S_0, T_0]$ based on edge weights matrix $\boldsymbol{W}$ by optimizing:

$$\max_{S\subseteq U, T\subseteq V} |S||T| \qquad \text{subject to } \frac{|\boldsymbol{W}[S,T]|}{|S||T|} \geq \gamma' \tag{3.2}$$

or the Lagrangian form after taking logarithm on both terms:

$$\max_{S\subseteq U, T\subseteq V} \log(|S||T|) + \lambda \log\left(\frac{|\boldsymbol{W}[S,T]|}{|S||T|}\right), \tag{3.3}$$

where γ' is the density constraint and the tuning parameter $\lambda \in (1, \infty)$.

The direct optimization of the objective function (3.3) is challenging because it is a nondeterministic polynomial (NP) problem [60, 91]. We propose computationally efficient (greedy) algorithms to implement the optimization while taking into account the spatial continuity. Our algorithm can be summarized as two steps. In step 1, we detect an initial IGDB without considering spatial constraint, which is developed based on the greedy algorithms in dense subgraph discovery [91]. In step 2, we then determine the IGDB based on the initial detection and its spatial connectivity by merging neighboring voxel clusters while preserving the maximization of the objective function. We describe the step 1 algorithm in the following Algorithm 4, and the details of step 2 are included in Appendix. In practice, the tuning parameter can be objectively selected by a likelihood method (see the Appendix for details). Multiple IGDBs can be extracted by performing algorithms repeatly with the detected IGDBs masked [92]. The computational complexity of Algorithm 4 is $O(C_1 mn)$, where C_1 is determined by the grid search of c.

To establish the approximation accuracy of the Algorithm 4 and its estimation of IGDB, let S^*_λ and T^*_λ be the true optimal maximizing the objective function (3.3):

$$(S^*_\lambda, T^*_\lambda) = \underset{S \subset U, T \subset V}{\arg\max}\, d_\lambda(S, T),$$

and $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ is the solution of Algorithm 4 with

$$(\tilde{S}_\lambda, \tilde{T}_\lambda) = \underset{c}{\arg\max} \underset{(S_1,T_1),\ldots,(S_{m+n-1},T_{m+n-1})}{\arg\max} d_\lambda(S, T),$$

where $d_\lambda(S,T) := \log(|S||T|) + \lambda \log\left(\frac{|\boldsymbol{W}[S,T]|}{|S||T|}\right)$.

The greedy algorithm defined in directed graph with average-degree based density (or equivalently $\lambda = 2$) is stated to guarantee a 2-approximation for the true optimal [60].

Algorithm 4 Optimizing objective function (3.3) with a given λ

```
**Input:** $G = (U, V, E, \boldsymbol{W})$, $\lambda$
**Output:** $G[\tilde{S}_\lambda, \tilde{T}_\lambda]$
procedure ALGORITHM
    for $c \in \{c_1, c_2, ..., c_L\}$ do
        $S_1 \leftarrow U, T_1 \leftarrow V$
        for k=1 to $n + m - 1$ do
            let $i \in S_k$ be the node with smallest degree: $i = \arg\min_{i' \in S_k} \deg_X(i'; S_k, T_k)$;
            let $j \in T_k$ be the node with smallest degree: $j = \arg\min_{j' \in T_k} \deg_Y(j'; S_k, T_k)$;
            if $\sqrt{c} \deg_X(i; S_k, T_k) \leq \frac{1}{\sqrt{c}} \deg_Y(j; S_k, T_k)$ then
                $S_{k+1} \leftarrow S_k / \{i\}$ and $T_{k+1} \leftarrow T_k$;
            else
                $S_{k+1} \leftarrow S_k$ and $T_{k+1} \leftarrow T_k / \{j\}$;
            end if
        end for
        Output $G[S^c, T^c]$ with maximized objective function among $G[S_1, T_1], ..., G[S_{n+m-1}, T_{n+m_1}]$;
    end for
    Output $G[\tilde{S}_\lambda, \tilde{T}_\lambda]$ with largest objective function among $G[S^{c_1}, T^{c_1}], ..., G[S^{c_L}, T^{c_L}]$;
end procedure
```

In short, $2d_2(\tilde{S}_2, \tilde{T}_2) > d_2(S_2^*, T_2^*)$. In this chapter, we investigate the approximation bounds in the bipartite graph setting. We present the approximation bounds for the proposed objective function (3.3) in terms of a parameter λ as the following theorem 3.1.

Theorem 3.1. *For a given bipartite graph $G = (U, V, E)$, with $(S_\lambda^*, T_\lambda^*)$ and $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ defined in Section 3.1.1, the greedy algorithm 4 has a $\rho(\lambda, m, n)$-approximation, i.e., $d_\lambda(S_\lambda^*, T_\lambda^*) \leq \rho(\lambda, m, n) d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)$ with*

$$\rho(\lambda, m, n) = \begin{cases} 2(mn)^{\frac{1}{\lambda}\left(1-\frac{2}{\lambda}\right)} & \textit{if } \lambda \geq 2 \\ 2(mn)^{\left(\frac{1}{\lambda}-\frac{1}{2}\right)} & \textit{if } \frac{4}{3} < \lambda < 2 \\ (mn)^{\left(1-\frac{1}{\lambda}\right)}. & \textit{if } 1 < \lambda \leq \frac{4}{3} \end{cases}$$

We then state that under the IGDB-based network structure, the optimization of the proposed objective function (3.3) leads to almost full recovery asymptotically.

Theorem 3.2. *Assume the graph $G = (U, V, E)$ with an IGDB $G[S_0, T_0] = (S_0, T_0, E[S_0, T_0])$ is generated from mixture of Bernoulli distributions: $e_{uv} \sim \delta_{uv}$Bernoulli$(\pi_1) + (1 - \delta_{uv})$Bernoulli$(\pi_0)$, $\delta_{uv} = I(e_{uv} \in G[S_0, T_0])$ and $\pi_1 > \pi_0$. For simplicity, we let $m = \Theta(n)$. Assume $|S_0| = O(|m|^{1/2+\epsilon})$ and $|T_0| = O(|n|^{1/2+\epsilon})$ as $n \to \infty$ for some $\epsilon > 0$. Denote*

$$e_S = \left(1 - \frac{\tilde{S}_\lambda \cap S_0}{S_0}\right) + \left(1 - \frac{\tilde{S}_\lambda^c \cap S_0^c}{S_0^c}\right)$$

and

$$e_T = \left(1 - \frac{\tilde{T}_\lambda \cap T_0}{T_0}\right) + \left(1 - \frac{\tilde{T}_\lambda^c \cap T_0^c}{T_0^c}\right)$$

be the error rates of node memberships based on $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ *from Algorithm 4.*

Then, there exists some λ *such that we will get almost full recovery in Algorithm 4, i.e. for all* $a \in (0, 1)$ *as* $n \to \infty$*,*

$$\mathbb{P}(e_S + e_T \geq a) \to 1.$$

3.3.2 Statistical inference of the IGDB

Recall the purpose of this chapter is to perform statistical inference on the pleiotropic and polygenic association pattern or the IGDB. We investigate the significant existence of an IGDB against a random bipartite graph as illustrated in section 3.2.3. In developing methodologies for the statistical test, without loss of generality, we assume that edge weights in $\boldsymbol{W}$ following a mixture marginal distribution [15]:

$$w_{uv} \sim \begin{cases} f_1(\cdot; \boldsymbol{\theta}_1), & \text{if } \beta_{uv} \neq 0 \\ f_0(\cdot; \boldsymbol{\theta}_0), & \text{if } \beta_{uv} = 0. \end{cases} \tag{3.4}$$

Subsequently, under the IGDB-based network model in section 3.2.2, $w_{uv}|\delta_{uv} = 1 \sim \mu_1 f_1 + (1 - \mu_1) f_0$, while $w_{uv}|\delta_{uv} = 0 \sim \mu_0 f_1 + (1 - \mu_0) f_0$. Empirically, we have the central tendency of $f_1(\cdot; \boldsymbol{\theta}_1)$ being greater than $f_0(\cdot; \boldsymbol{\theta}_0)$, in the sense that $\mathbb{E}_{\boldsymbol{\theta}_1}[w_{uv}|\beta_{uv} \neq 0] > \mathbb{E}_{\boldsymbol{\theta}_0}[w_{uv}|\beta_{uv} = 0]$.

Let r be a sound cutoff that dichotomize the weighted graph G into a binary graph $G^r = (U, V, \boldsymbol{A})$ using $a_{uv} = I(|w_{uv}| > r)$. Then, under IGDB structure indexed by node sets (S_0, T_0), the edges in G^r follow a mixture of two Bernoulli distributions:

$$a_{uv} | (S_0, T_0) \sim \text{Bernoulli}(\pi_{uv}) \tag{3.5}$$

where $\pi_{uv} = \delta_{uv}\pi_1 + (1 - \delta_{uv})\pi_0$ with the two parameters:

$$\pi_1 = \mu_1 \int_r^\infty f_1(w, \boldsymbol{\theta}_1) dw + (1 - \mu_1) \int_r^\infty f_1(w, \boldsymbol{\theta}_1) dw,$$

$$\pi_0 = \mu_0 \int_r^\infty f_1(w, \boldsymbol{\theta}_1) dw + (1 - \mu_0) \int_r^\infty f_1(w, \boldsymbol{\theta}_1) dw,$$

and $\pi_1 > \pi_0$. Then, the IGDB testing hypotheses boil down to

$$H_0 : \pi_1 = \pi_0 = \pi \quad \text{versus} \quad H_1 : \pi_1 > \pi_0,$$

based on our mixture distribution model (3.5).

We propose a likelihood-based approach for the IGDB-wise hypothesis testing. For a binarized graph G^r, we let

$$t_G = \log \frac{\sup_{H_0 \cup H_1} \mathcal{L}(\boldsymbol{\pi}; S, T, \boldsymbol{A})}{\sup_{H_0} \mathcal{L}(\pi; \boldsymbol{A})},$$

with likelihood given by Bernoulli distributions in (3.5) and a rejection region:

$$\Pr(t_G > \eta) \leq \alpha.$$

In determining the significance of IGDBs, the simultaneous testing needs to be accounted for all potential IGDBs. Besides, a rejection region (η) should be determined based on the distribution of t_G under null model. Hence, we employ the commonly used permutation test procedure in the field of neuroimaging [31, 66] to empirically approximate the distribution of the likelihood ratio statistic t_G under IGDB-wise null hypothesis and control the family-wise error rates (FWER). We describe the detailed testing procedure in Appendix.

3.4 Simulation Studies

3.4.1 Simulation settings

In this section, we evaluate the finite-sample performance of our proposed method, and compare with competing methods. We generate the synthetic data set with $m = 200$ genetic markers (i.e., $\boldsymbol{X} = (X_1, ..., X_{200})$), and $n = 100$ imaging features (i.e., $\boldsymbol{Y} = (Y_1, ..., Y_{100})$) on $L = 60$ subjects. We establish the associations between genetic markers and imaging features through a bipartite graph $G = (U, V, E)$. According to the IGDB structure in section 3.2.2, we observe the matrix of indicator variable $\boldsymbol{\Delta} = \{\delta_{uv}\}$ from the IGDB $G[S_0, T_0] = (S_0, T_0, E[S_0, T_0])$ with two possible sizes: $(|S_0|, |T_0|) = (50, 40)$ and $(30, 20)$. We let the proportion of significant associations within the IGDB μ_1 inside the IGDB (i.e. $1 - \mu_1$ as false positive findings), and μ_0 false positive edges at the background graph. Different parameters $(\mu_1, \mu_0) = (0.8, 0.2)$ and $(0.9, 0.1)$ are consider in this simulation study.

We then consider the genetic markers (X_u) from Bernoulli distributions, while imaging features (Y_v) following normal distributions. Subsequently, for significant associations (i.e., $\beta_{uv} \neq 1$), the distribution of Y_v should be significantly different in two groups of X_u, whereas for $\beta_{uv} = 0$, Y_v follow a common distribution for values of X_u. Particularly, for the pair with $\beta_{uv} \neq 0$, Y_v is assumed to follow $N(\theta_1, \sigma^2)$ and $N(\theta_0, \sigma^2)$ in two groups of X_u (i.e., $Y_{v,l} \sim N(\theta_1 X_{u,l} + \theta_0(1 - X_{u,l}), \sigma^2), l = 1, ..., L$). For $\beta_{u'v'} = 0, Y_{v'l} \sim N(\theta_0, \sigma^2)$ for all $l = 1, ..., L$. Hence, we generate the edge weights matrix $\boldsymbol{W}$ with

$$w_{uv}|\delta_{uv} = 1 \sim \mu_1 t_{L-2}(\nu) + (1 - \mu_1)t_{L-2}$$

$$w_{uv}|\delta_{uv} = 0 \sim \mu_0 t_{L-2}(\nu) + (1 - \mu_0)t_{L-2}$$

with noncentral parameter $\nu = \frac{\theta_1 - \theta_0}{\sigma\sqrt{\frac{4}{L}}}$. We apply different signal and noise power (i.e., Signal-to-Noise Ratio (SNR)) by letting $\sigma = 1$, $\theta_0 = 0$ and $\theta_1 = 0.8, 1.0$ and 1.2. We replicate all scenarios for 100 times.

3.4.2 Performance metrics

In the simulation study, we evaluate the performance of proposed inferential procedure by the IGDB-wise testing results, and the extraction of IGDB based on the edge-wise accuracy. For IGDB-wise inference, we consider a detected IGDB $G[\hat{S}, \hat{T}]$ is a recovery of the underlying IGDB $G[S_0, T_0]$ if it is rejected in the proposed likelihood ratio test and has high similarity with $G[S_0, T_0]$. Specifically, we consider $G[\hat{S}, \hat{T}]$ is a true positive detection of $G[S_0, T_0]$ if $J_{\boldsymbol{X}} \wedge J_{\boldsymbol{Y}} > 0.6$ with

$$J_{\boldsymbol{X}} = \frac{S_0 \cap \hat{S}}{S_0 \cup \hat{S}} \text{ and } J_{\boldsymbol{Y}} = \frac{T_0 \cap \hat{T}}{T_0 \cup \hat{T}},$$

and we succeed to reject the IGDB-wise null hypothesis in the permutation test. Therefore, the detected IGDB leads to a false negative finding if the p-value in the permutation test is not lower than the a significant level (i.e., 0.05). Besides, we observe a false positive error if $G[\hat{S}, \hat{T}]$ has low similarity to $G[S_0, T_0]$ even we rejected the IGDB-wise null hypothesis. We report the accuracy of inference by False Positive Rate (FPR) and False Negative Rate (FNR) among replications.

Furthermore, we investigate the edge-wise accuracy by comparing to the standard multiple testing methods: pFDR by [93] and Bonferroni correction. We compare the true $\boldsymbol{\Delta}$ with estimated $\hat{\boldsymbol{\Delta}}$ from varied methods. For the proposed method, we obtain the $\hat{\boldsymbol{\Delta}}$ based on the extracted IGDB $G[\hat{S}, \hat{T}]$ and the hypothesis testing. Particularly, if we reject the IGDB-wise null hypothesis with a detected bicluster $G[\hat{S}, \hat{T}]$, we let $\hat{\boldsymbol{\Delta}} = \{\hat{\delta}_{uv}\} = \{I(e_{uv} \in G[\hat{S}, \hat{T}])\}$. In the case that we fails to reject, we consider $\hat{S}, \hat{T}$ as empty sets such that $\hat{\boldsymbol{\Delta}} = \mathbf{0}_{m \times n}$. For both pFDR and Bonferroni, we observe $\hat{\delta}_{uv}$ by a cutoff 0.2 of q values and adjusted p values.

Subsequently, based on the $\hat{\delta}_{uv}$ observed from different methods, and true parameters δ_{uv}, we calculate true positive rate (TPR) and true negative rate (TNR) as:

$$\text{TPR} = \frac{\sum_{u,v} I(\delta_{uv} = \hat{\delta}_{uv} = 1)}{\sum_{u,v} I(\delta_{uv} = 1)}, \quad \text{TNR} = \frac{\sum_{u,v} I(\delta_{uv} = \hat{\delta}_{uv} = 0)}{\sum_{u,v} I(\delta_{uv} = 0)}.$$

The associated means and standard deviations are reported based on 100 replications for each simulation scenario.

3.4.3 Results

The results from the IGDB-wise inference are summarized in Table 3.1. The power of the IGDB-wise inference relies on the size as well as the intensities (by different SNRs and parameters of mixture distributions) of the underlying IGDB $G[S_0, T_0]$, which confirms our theoretical conclusions. We only fails to reject the IGDB-wise null hypothesis with size $(30, 20)$, low SNR 0.8, and higher rates of noisy edges $(0.8, 0.2)$.

The comparative edge-level results from the proposed method and competing methods are displayed in Table 3.2 and Table 3.3 for different sizes of the IGDB. All three methods have improved performance when we have higher SNRs and less noisy edges. The proposed method out performs pFDR and Bonferroni correction for both TPR and TNR under different scenarios. Both pFDR and Bonferroni methods have high TNR but low TPR indicating a stringent cutoff, while the proposed method achieves a higher TPR maintaining a similar or even higher TNR than the others. The Bonferroni method is even more stringent where the TPR is even smaller than 10% when we have low SNRs (e.g., 0.8) for all cases.

Table 3.1: IGDB-wise inference results under varied SNRs and noises

			0.8	1.0	1.2
$(50, 40)$	$(0.9, 0.1)$	FPR	0 (0)	0 (0)	0 (0)
		FNR	0 (0)	0 (0)	0 (0)
	$(0.8, 0.2)$	FPR	0 (0)	0 (0)	0 (0)
		FNR	0 (0)	0 (0)	0 (0)
$(30, 20)$	$(0.9, 0.1)$	FPR	0 (0)	0 (0)	0 (0)
		FNR	0 (0)	0 (0)	0 (0)
	$(0.8, 0.2)$	FPR	0 (0)	0 (0)	0 (0)
		FNR	0.0600 (0.2375)	0 (0)	0 (0)

Table 3.2: Edge-wise accuracy under varied SNRs and noises with $(|S_0|, |T_0|) = (50, 40)$

(q_1, q_2)	Methods		0.8	1.0	1.2
(0.9, 0.1)	IGDB	TPR	0.9879 (0.0184)	0.9942 (0.0124)	0.9968 (0.0097)
		TNR	1 (0)	1 (0)	1 (0)
	pFDR	TPR	0.7453 (0.0090)	0.8686 (0.0045)	0.8995 (0.0023)
		TNR	0.8858 (0.0020)	0.8667 (0.0018)	0.8619 (0.0018)
	Bonferroni	TPR	0.0520 (0.0048)	0.1739 (0.0092)	0.3941 (0.0096)
		TNR	0.9942 (0.0005)	0.9806 (0.0008)	0.9562 (0.0012)
(0.8, 0.2)	IGDB	TPR	0.9938 (0.0126)	0.9982 (0.0064)	0.9984 (0.0061)
		TNR	0.9998 (0.0006)	1.0000 (0.0003)	1.0000 (0.0004)
	pFDR	TPR	0.7032 (0.0067)	0.7903 (0.0039)	0.8095 (0.0027)
		TNR	0.7842 (0.0021)	0.7577 (0.0019)	0.7517 (0.0018)
	Bonferroni	TPR	0.0458 (0.0043)	0.1557 (0.0084)	0.3506 (0.0097)
		TNR	0.9884 (0.0007)	0.9612 (0.0014)	0.9125 (0.0020)

Table 3.3: Edge-wise accuracy under varied SNRs and noises with $(|S_0|, |T_0|) = (30, 20)$

(q_1, q_2)	Methods		0.8	1.0	1.2
(0.9, 0.1)	IGDB	TPR	0.9987 (0.0081)	0.9992 (0.0060)	1 (0)
		TNR	1.0000 (0.0001)	1 (0)	1(0)
	pFDR	TPR	0.7043 (0.0176)	0.8537 (0.0085)	0.8954 (0.0042)
		TNR	0.9017 (0.0019)	0.8799 (0.0015)	0.8741 (0.0014)
	Bonferroni	TPR	0.0517 (0.0082)	0.1741 (0.0163)	0.3946 (0.0175)
		TNR	0.9942 (0.0005)	0.9807 (0.0009)	0.9561 (0.0012)
(0.8, 0.2)	IGDB	TPR	0.8527 (0.2248)	0.9645 (0.0398)	0.9778 (0.0287)
		TNR	0.9996 (0.0009)	0.9995 (0.0009)	0.9997 (0.0005)
	pFDR	TPR	0.6891 (0.0114)	0.7857 (0.0075)	0.8069 (0.0045)
		TNR	0.7952 (0.0022)	0.7661 (0.0017)	0.7596 (0.0019)
	Bonferroni	TPR	0.0473 (0.0095)	0.1563 (0.0144)	0.3525 (0.0173)
		TNR	0.9884 (0.0008)	0.9610 (0.0013)	0.9123 (0.0017)

3.5 Data Example

The Human Connectome Project (HCP) sponsored by National Institutes of Health (NIH) aims to construct the underlying neuro pathways with healthy human brain functions. The HCP becomes an important public resource for structural and functional brain connectivity data together with demographic, behavior, genetic data, and etc. We demonstrate the effectiveness of the proposed method in identifying systematic associations between genetic markers and imaging features using the S1200 data release from 1206 young adults. The brain imaging and genetics data of HCP were acquired for 1142 participants. The diffusion tensor imaging fractional anisotropy (FA) measures at 29,627 voxels were used in this chapter to characterize the white matter integrity by following an ENIGMA FA imaging processing pipeline [94]. Regarding genetic variants, 1,580,643 imputed SNPs passed the quality control filters in the data set (MAF< 0.014; HQE$< 1e^{-6}$; r-squared$>$ 0.03; call rate> 0.95).

We focus our study on the associations between 1,580,643 SNPs and 29,627 voxels. We applied hard-thresholding on p-values with cutoff 0.001. 13,498 SNPs with at least 0.5$\%$ non-zero p-values are included in the study.

We illustrate our analysis in chromosome 1 based on the matrix of association strength $\boldsymbol{W}_{1178\times 29627}$. By implementing the proposed IGDB-wise statistical inference procedure on the weight matrix $\boldsymbol{W}$ (i.e., Figure 3.3 (a)), we detect an IGDB with 384 SNPs and 3803 voxels via optimizing the objective function (3.3) as Figure 3.3 (b). The computation is efficient, which takes 20 minutes on a PC with an i7 CPU 3.60 GHz and 64GB memory. We further calculate the p value for the IGDB-wise statistical inference

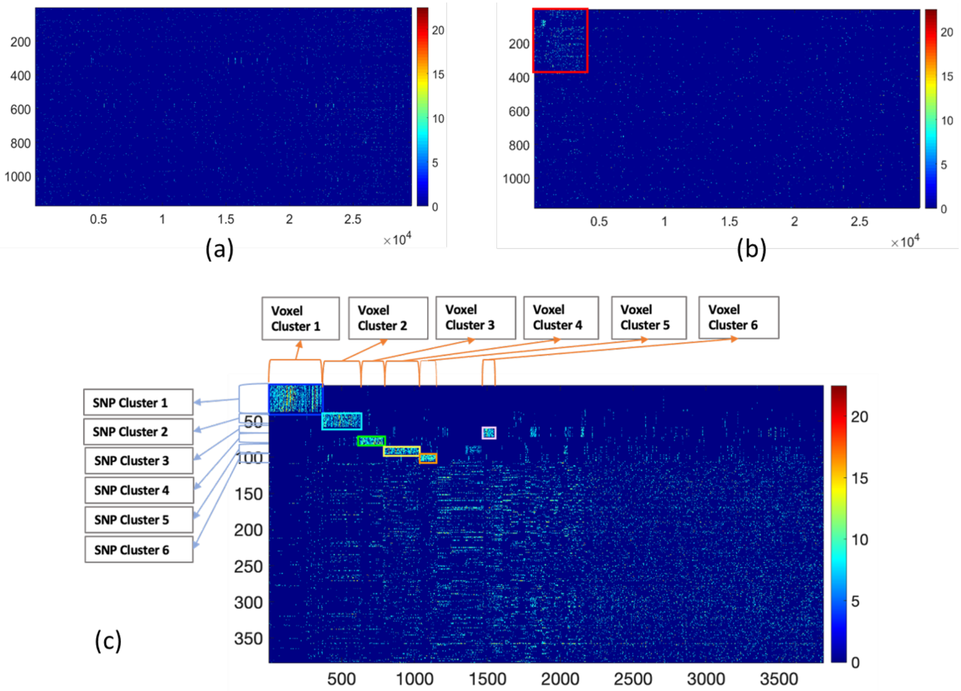


Figure 3.3: HCP data example: (a) is the input matrix $\boldsymbol{W}$; (b) demonstrates the detected IGDB; (c)displays the refined pattern of the IGDB

via the permutation test, which results in a significant existence of an IGDB with p value < 0.001.

Although the IGDB is an irreducible subgraph, it can be further refined based data-driven algorithms and spatial information of imaging data. We apply the existing community detection algorithms [69] on similarity matrices observed from the detected IGDB. The refined pattern in Figure 3.3 (c) displays 6 distinct SNP-voxel association clusters. We illustrate the cluster-wise association pattern in Figure 3.4, which shows that the SNP cluster is constructed from neighboring SNPs while the voxel clusters also preserve the spatial structure in imaging features. Note that the refined structure can not be identified without revealing the IGDB by the proposed algorithm.

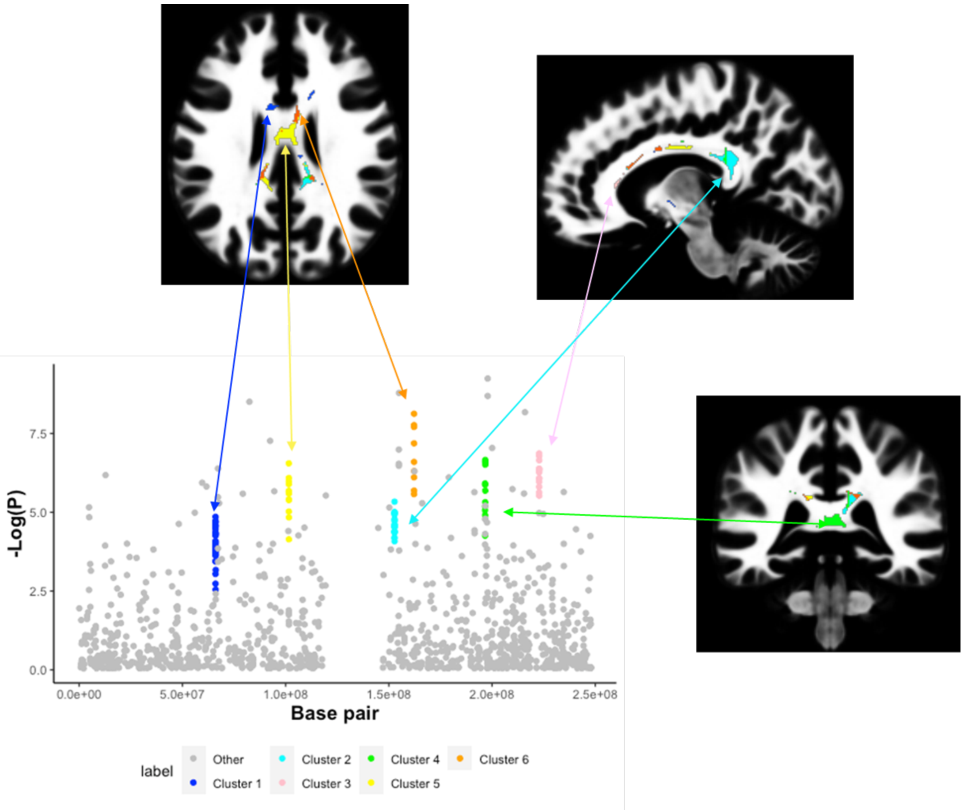


Figure 3.4: HCP data example: illustration of the association pattern

3.6 Discussion

The imaging-genetic studies aim in modelling the predictive mechanism of genetic markers on quantitative imaging measures. The multiple testing problems are challenging to identify the important effects in the dimension of imaging-genetic studies. Methods are proposed to reduce the number of multiple comparisons by identifying significant genetic variants with important contribution to all voxels or vice versus. In this chapter, we have developed a bipartite graph model to identify the IGDB for imaging-genetics association analysis. The IGDB is a bipartite subgraph consisting mainly of genetic variant-voxel pairs with strong associations. The subset of genetic variants or voxels can be further grouped to genes or subregions to provide more biological insights or more refined patterns. Thus, it can be considered as a functional subnetwork of genes and brain areas with pleiotropic and polygenic mechanisms. Such inference results provide a systematic and comprehensive view of imaging-genetics association.

We develop theoretical results to show that an IGDB structure with a non-trivial size is unlikely to be false positively detected. Built on these graph combinatorics properties, we propose likelihood ratio based inference for IGDB with a bipartite graph mixture model. A major contribution of this framework is that the statistical power of IGDB inference is invariant to the high-dimensionality of imaging-genetics features with controlled false positive error rate. Therefore, it alleviates the requirement of extremely large (even impractical) sample size for the imaging-genetics study. Our IGDB detection algorithm is computationally efficient, and thus is compatible with the permutation test. We provide theoretical bound to guarantee the approximation of our algorithm. Although our method is originally developed for imaging-genetics association analysis, it can be generalized to other modal multivariate-multivariate association analysis (e.g., eQTL).

Chapter 4: Extracting Interconnected Communities in Gene Co-expression Networks

4.1 Introduction

Gene co-expression network (GCN) analysis has been widely used to study the systematic interactions among high throughput genomic features [95]. GCN is often represented by an undirected graph, where each node denotes a gene and an edge between two nodes indicates the interactive relationship between a pair of genes. The edge weight (the strength of the interactive relationship) is calculated by metrics including various versions of correlation coefficients and mutual information measures across group samples [96]. Co-expressed genes in the network are often simultaneously active in the same biological processes. The community structures found in a large gene co-expression network can reveal the system-level property of genes and assist in understanding the underlying regulatory mechanism and the complex biology behind. For example, conserved functional modules have been identified from yeast co-expression network which contain a number of hub genes essential for yeast viability [97].

In the last two decades, clustering algorithms and network models have been developed and applied to learning the latent patterns of GCN data [36, 37, 98]. The results of commonly used GCN analysis tools (e.g., WGCNA) often yield a number "modules/communities" as the detected network structure. In the field of network analysis, the independent community structure has been extensively studied, yielding many efficient and theoretically justified estimation procedures, such as Newman-Girwan Modularity [74], also see [32, 99, 100, 101] for comprehensive reviews. The independent community structure implies a block-diagonal structure for the GCN adjacency matrix, as illustrated in Figure 4.1(d). For GCN analysis, the detected communities can be subsequently investigated by functional enrichment analysis to identify the functional categories overrepresented by the genes in the communities.

Independent communities vs. interconnected communities: Despite its usefulness, the independent community model may be over-simplified. In the current research, we propose a new interconnected community network (ICN) structure that relaxes the constraint of the independence between communities and allow connections between genes from different communities. Our model is more flexible and enjoys several advantages. First, genes from the different communities might be enriched in signalling pathways that have cross-talks between each other and play synergistic roles in a biological process. For example, signal propagation among several important kinase signalling pathways are well-documented in the literature [102, 103]. Second, by allowing between-community interactions, the ICN structure is compatible with the well-known properties of the real-world complex networks (e.g., protein-protein interaction (PPI) networks and metabolic networks) including hub nodes, small-worldness, and high efficiency among many others [104, 105, 106]. In contrast, the independent community structure shows no small-worldness and much lower global efficiency. Last, the flexibility of the ICN structure provides a better fit to our data, because it can more precisely characterize the latent topological structure (see Figure 4.1(a, c, e, f)). As a comparison, the independent community structure based algorithms may yield false positive interaction estimations by incorrectly merging two interconnected communities into one jumbo community (Figure 4.1(b)) or miss the interconnections between two interactive communities (Figure 4.1(d)).

Solving the exact ICN parameter estimation problem requires a combinatorial optimization and is intractable in practice. To overcome the computational challenge, we develop a set of new and efficient algorithms to extract the ICN structure and provide a user-friendly software package. Our algorithm operates as follows. We first detect a set of 'dense' communities where genes are highly correlated. Then, we perform statistical tests to identify interconnected community pairs. Finally, for each interconnected community pair, we identify the connecting edges by a shrinkage method. We implement the above steps using a unified objective function with ℓ_0 graph norm penalty to ensure a low false discovery rate. To authors' best knowledge, this is the first computationally efficient algorithm for detecting the ICN structure. We perform extensive simulation studies to validate the proposed method. We also apply our approach to a human RNA-seq data set for Acute Myeloid Leukemia research. The results specify important interconnections between communities and gene interactions among modules which is related to the immune evasion mechanism of tumor cells.

Overlapped communities vs. interconnected communities: The ICN model bears a resemblance to the overlapping community network (OCN) model [107, 108, 109], as in both structures genes can be shared in multiple communities. The main distinction lies in the connection patterns between communities. Under OCN models, a node from community 1 that shows nonzero overlapping membership towards community 2 typically has positive connection probabilities with most, if not all, members of community 2. In contrast, this node only needs to be connected with a proportion of members in community 2 under the ICN model. Moreover, our model allows two nodes from community 1 differ in their correspondence node sets from community 2, thus yields greater flexibility in modeling the between-community connections that exhibits both sparsity and non-uniformity while keeping the interpretation simple and clean. Thus, the OCN model can be viewed as a special case of the ICN model. As an empirical evidence of the ICN model's better fit to real-world data, Figure 4.1(f) shows a real-world GCN example: a gene in the conjunction may connect to all other genes in its own community, while only connects to a proportion of genes in the other community. There are few genes showing overlapping memberships. In simulation studies and data analysis, we demonstrate that the ICN model extracts the underlying network structure more accurately and provides a better fit.

4.2 Methods

We denote a preprocessed and normalized gene expression data set with n subjects and p genes by a matrix $\boldsymbol{X}_{n\times p}$. The weighted adjacency matrix $\boldsymbol{W}_{p\times p}$ for GCN analysis can be calculated across subjects based on $\boldsymbol{X}_{n\times p}$. For example, w_{ij}, the entry at the ith row and jth column of $\boldsymbol{W}_{p\times p}$, is a Fisher's Z transformed correlation coefficient between genes i and j with $1 \leq i \neq j \leq p$. We can also apply alternative pairwise association metrics to calculate w_{ij}, including Spearman correlation coefficient, Kendall's tau, and

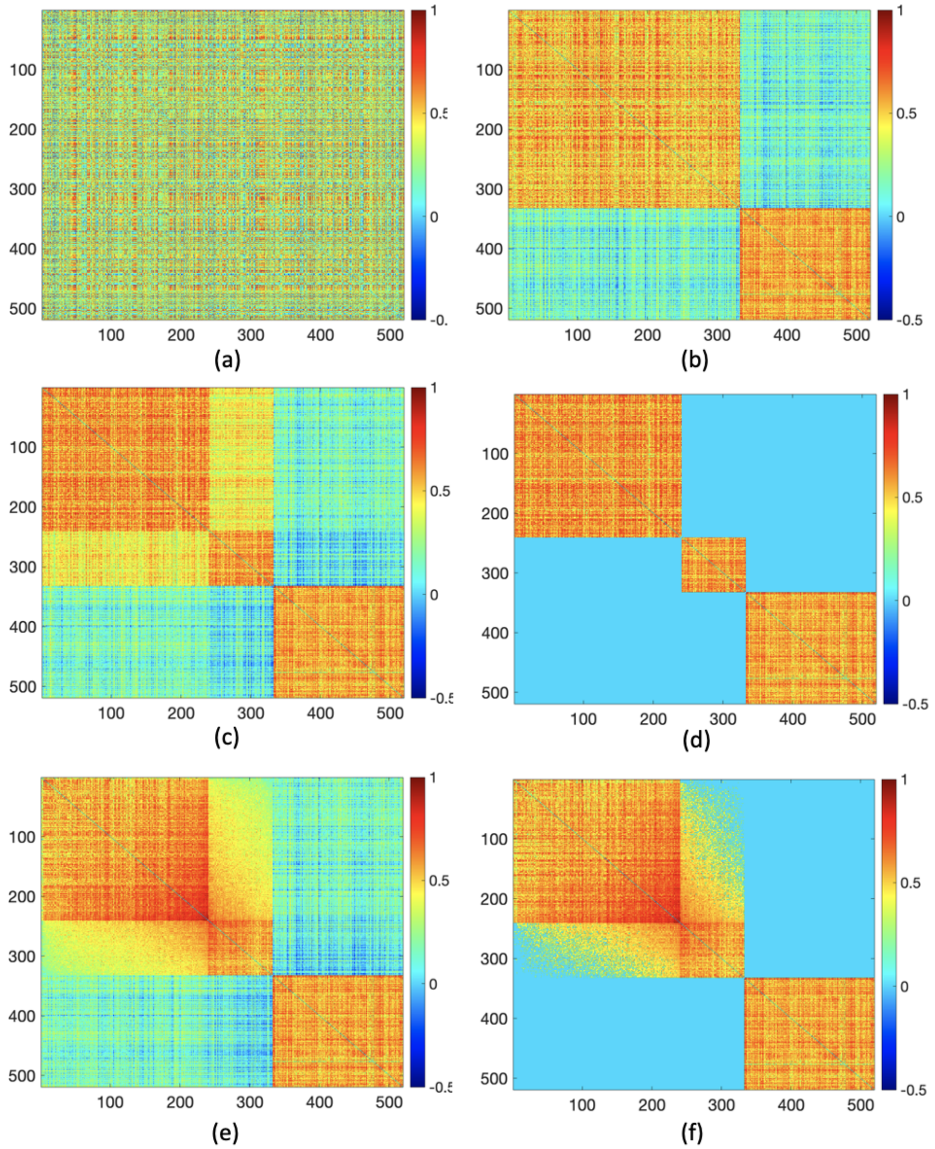


Figure 4.1: Demonstration of ICN using a subset of genes from data example, and results from comparing independent community network structure vs. ICN structure based on a gene expression data. (a) is the input gene co-expression matrix; (b) shows the results of two independent communities (based on a relaxed threshold); (c) illustrates that the first community in (b) can be further decomposed into two interactive communities; (d) however, under the independent community assumption, only three independent communities can be detected; (e) in contrast, the ICN method can recognize the interconnected communities; (f) the ICN method can further identify connecting edges between the two interconnected communities.

mutual information coefficient. Our goal is to extract the latent network topological structure of GCN based on the input data $\boldsymbol{W}_{p\times p}$, and enhance our understanding of the underlying complex biological processes [37, 110].

4.2.1 Gene co-expression networks with independent communities

A graph notation $G = (V, E)$ is often used to represent the structure of a co-expression network, where the node set V denotes p genes with $|V| = p$ and E represents the pair-wise interactive relationships among the p genes such that $|E| = p(p-1)/2$. We consider $\boldsymbol{W}_{p\times p}$ as the weighted adjacency matrix of G. In the last two decades, numerous algorithms have been developed to extract network structures from $\boldsymbol{W}_{p\times p}$ [37, 104, 110, 111, 112, 113]. In these models, the independent community structure of G is assumed. Specifically, the whole network can be partitioned into a set of disjoint communities $G = \{G_1, \ldots, G_C\}$ where each $G_c = \{V_c, E_c\}$ is a community, and edges only exist between nodes in the same community.

Here, we define the network topology by $\mathcal{T}(G)$, which characterizes the network layout of graph G. In other words, when G has an independent community structure, $\mathcal{T}(G)$ describes the assignment of all nodes in G into communities $\{G_1, \ldots, G_C\}$ and the corresponding allocation of edges in G.

We let edge weights $\{w_{ij}\}$ in $\boldsymbol{W}_{p\times p}$ follow a two-component mixture distribution [114]:

$$f(w_{ij}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_0, \delta_{ij}) = \delta_{ij} f_1(w_{ij}; \boldsymbol{\theta}_1) + (1-\delta_{ij}) f_0(w_{ij}; \boldsymbol{\theta}_0), \tag{4.1}$$

where f_1 is the distribution of the intra-community edge weights, and f_0 is the null distribution for background edge weights. θ_1 and θ_0 are distribution parameters for f_1 and f_0 respectively. We let the central tendency of f_1 be greater than f_0, $\mathbb{E}_{\boldsymbol{\theta}_1}[w_{ij}|\delta_{ij} = 1] > \mathbb{E}_{\boldsymbol{\theta}_0}[w_{ij}|\delta_{ij} = 0]$. Here, the binary indicator variable δ_{ij} is determined by the network topology $\mathcal{T}(G)$, and $\delta_{ij} = 1$ indicates the existence of correlation between i and j. Specifically, for an independent community structure, we have

$$\delta_{ij} = \begin{cases} 1, & \text{if } i, j \in G_c \text{, for some } c; \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

We denote the above function by $\{\delta_{ij}\} = h(\mathcal{T}(G))$. The function $h(\mathcal{T}(G))$ links the underlying network structure with the marginal distribution of edge weights in $\boldsymbol{W}_{p\times p}$, thus plays a central role in GCN analysis.

The goal of gene co-expression network analysis is to estimate the $\{\hat{G}_c\}_c$ (i.e., $\widehat{\mathcal{T}}(G)$) from $\boldsymbol{W}_{p\times p}$. We note that $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ can be easily estimated given $\widehat{\mathcal{T}}(G)$, and vice versa. Therefore, both non-parametric methods (e.g., clustering and community detection models) and parametric models (e.g., infinite mixture models using iterative algorithms) have been successfully applied for GCN data analysis and yield interesting biological findings [112, 115, 116]. However, the independent community assumption may over-simplify the true nature of complex biological networks. Network estimation tools for more general models are less developed and computationally expensive. To address this challenge, we develop novel and efficient numerical methods for flexible models that better capture the graph topological structures for GCN analysis.

4.2.2 Gene co-expression networks with interconnected communities

In this chapter, we consider a more general network structure: interconnected communities. We alleviate the constraint of independence between communities. Formally, we call a pair of communities "interconnected" if some of their nodes form between-community connections. We use $G_c \Leftrightarrow G_{c'}$ to denote that communities $G_c = \{V_c, E_c\}$ and $G_{c'} = \{V_{c'}, E_{c'}\}$ are interconnected. Then, $G_c \Leftrightarrow G_{c'}$ iff there exist at least two adjacent nodes, for example, i and j (i.e. $i \leftrightarrow j$) for $i \in G_c, j \in G_{c'}$. We use a subgraph $I_{c,c'}$ to denote the interconnection between G_c and $G_{c'}$ (see connected edges between the first two blocks in Figure 4.1(f)). Specifically, $I_{c,c'}$ is an edge-induced subgraph by the edge set $E^*_{c,c'} = \{e_{ij} | i \leftrightarrow j, i \in V_c, j \in V_{c'}\}$. We let $I_{c,c'} = \{V^*_c, V^*_{c'}, E^*_{c,c'}\}$, where $V^*_c \subset V_c$ and $V^*_{c'} \subset V_{c'}$ are two disjoint sets of nodes. By definition, $G_c \Leftrightarrow G_{c'}$ iff $0 < |E^*_{c,c'}| \leq |V_c| \times |V_{c'}|$.

The network topological structure of ICN, $\mathcal{T}_{\text{ICN}}(G)$, determines the indicator variable δ_{ij} as follows:

$$\delta_{ij} = \begin{cases} 1, & i, j \in G_c, \text{ for some } c; \\ k, & \text{if } e_{ij} \in E^*_{c,c'}, k = 2, ..., K; \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

where $k = 2, ..., K$ are indices for the interconnected community pairs, such that $K \leq C(C-1)/2 + 1$. The connection strengths of edges can vary in different interconnection subgraphs (e.g., the distributions of edges can be different in interconnection subgraph $I_{c,c'}$ and interconnection subgraph $I_{c^*,c^{**}}$). Thus, interaction edges can follow a distribution with $K-1$ mixture components instead of a single component.

In this case, the edge weight distribution of the observed co-expression weighted matrix $\boldsymbol{W}_{p\times p}$ can be modeled as a $K+1$-component mixture distribution with density function:

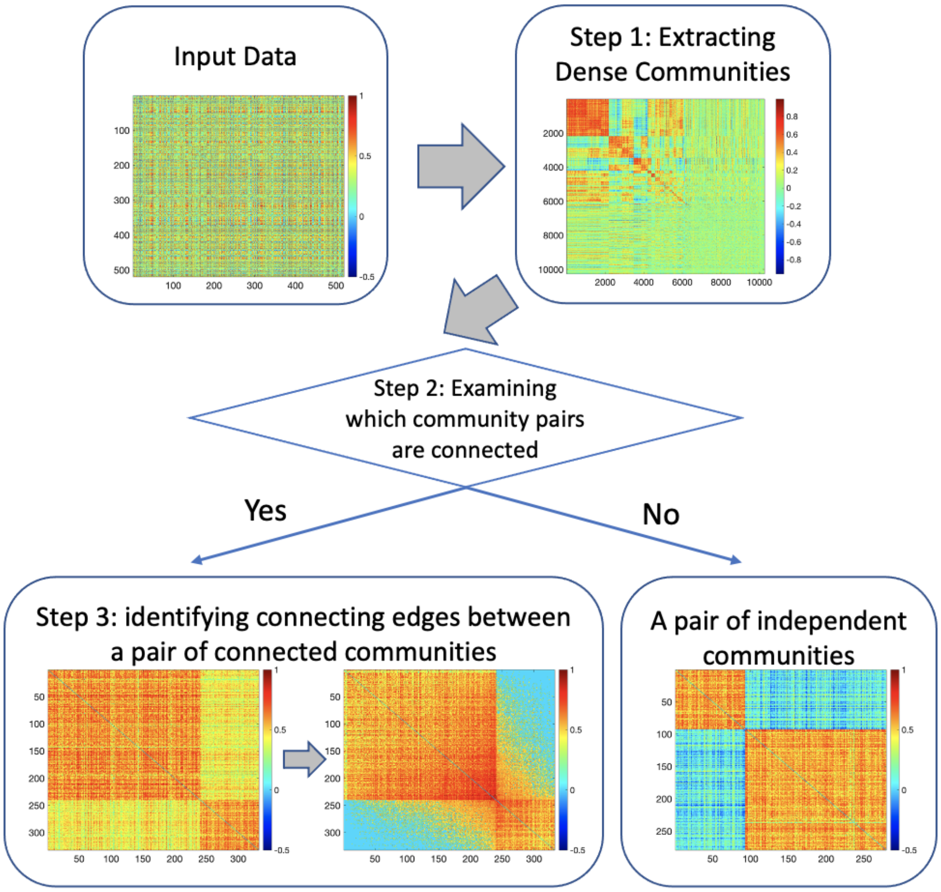
$$\begin{aligned} f(w_{ij}|\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K, \delta_{ij}) = f_1(w_{ij}; \boldsymbol{\theta}_1) I(\delta_{ij}=1) \\ + f_0(w_{ij}; \boldsymbol{\theta}_0) I(\delta_{ij}=0) + \sum_{k=2}^{K} f_k(w_{ij}; \boldsymbol{\theta}_k) I(\delta_{ij}=k), \quad (4.4) \end{aligned}$$

where f_1 is the density function for the intra-community edge weights, $f_k, k = 2, \cdots, K$ are the density functions for the edge weights in $I_{c,c'}$, and f_0 is the distribution for the background edge weights. To ensure the model identifiability, we let $\mathbb{E}_{\boldsymbol{\theta}_k}[w_{ij}|\delta_{ij}=1] > \mathbb{E}_{\boldsymbol{\theta}_k}[w_{ij}|\delta_{ij}=k] > \mathbb{E}_{\boldsymbol{\theta}_0}[w_{ij}|\delta_{ij}=0]$ for $k = 2, ..., K$. The parameter estimation of (4.4) is challenging because the indicator variable δ_{ij} is unknown and entangled with the network topological structure.

Our primary goal is to estimate the underlying network structure $\widehat{\mathcal{T}}_{\text{ICN}}(G)$ based on $\boldsymbol{W}_{p\times p}$. The estimation of $f_0, f_1, ..., f_k$ becomes straightforward with a known $\mathcal{T}_{\text{ICN}}(G)$. Recently, network/graph models have been developed to estimated the mixture model with independent community structure [117]. In this chapter, we focus on developing tools to estimate the latent ICN structure from a noisy weighted adjacency matrix $\boldsymbol{W}_{p\times p}$.

4.2.3 Detecting interconnected communities

Our goal is to extract both communities $\{\hat{G}_c\}$ and their interconnections $\{\hat{I}_{c,c'}\}$ from $\boldsymbol{W}_{p\times p}$. However, the true ICN structure detection tends to be disturbed by the noise



Input Data
100
200
300
400
500
100
200
300
400
500
1
0.5
0
-0.5
Step 1: Extracting Dense Communities
2000
4000
6000
8000
10000
2000
4000
6000
8000
10000
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Step 2: Examining which community pairs are connected
Yes
No
Step 3: identifying connecting edges between a pair of connected communities
50
100
150
200
250
300
1
0.5
0
-0.5
A pair of independent communities
50
100
150
200
250

Figure 4.2: The flowchart presents a brief overview of the three-step detection procedure for ICN structure. The input data is a weighted adjacency matrix of gene co-expression network. In step 1, the densely connected communities are detected. In step 2, we evaluate the connectivity between each pair of communities from step 1. If the pair of communities are significantly interconnected, we identify the connecting edges in step 3. Otherwise, we consider this pair of community to be independent.

in $\boldsymbol{W}_{p\times p}$ and the entangled network structure. For example, Figure 4.1(b) shows that merging two distinct but interconnected communities into one jumbo community can lead to false positive errors, while Figure 4.1(d) illustrates that the missing interconnections between communities can introduce false negative errors.

To mitigate this challenge, we develop a new objective function to cover the maximum edge weights in $\boldsymbol{W}_{p\times p}$ by an ICN structure with a minimal number of non-zero edges. By minimizing the size of ICN structure, we can effectively control the false positive errors (Figure 4.1(b)). Meanwhile, the maximized cover of edge weights would favor the interconnected communities (Figure 4.1(d)) and reduce the false negative errors.

To link the underlying ICN structure $\mathcal{T}_{\text{ICN}}(G)$ with the input data $\boldsymbol{W}_{p\times p}$, we introduce a matrix $\boldsymbol{U} = \{u_{ij}\}_{1\leq i,j\leq p}$ obtained by thresholding $\boldsymbol{W}_{p\times p}$:

$$u_{ij} = \begin{cases} w_{ij}, & \text{if } \delta_{ij} > 0 \text{, where } \delta_{ij} = h(\mathcal{T}_{\text{ICN}}(G)); \\ 0, & \text{otherwise.} \end{cases} \tag{4.5}$$

Then, our new objective function is:

$$\underset{C,\{G_c\},\{I_{c,c'}\}}{\arg\max} \ \log \|\boldsymbol{U}\|_1 - \lambda_0 \log \|\boldsymbol{U}\|_0, \tag{4.6}$$

where $\|\boldsymbol{U}\|_1 = \sum_{i,j} |u_{ij}|$ is the element-wise ℓ_1 matrix norm, $\|\boldsymbol{U}\|_0 = \sum_{i,j} I(|u_{ij}| > 0)$ is the element-wise ℓ_0 matrix norm, and λ_0 is a tuning parameter. We maximize $\|\boldsymbol{U}\|_1$ to include a maximal number of edges with high correlation values (letting $\delta_{ij} > 0$ for these edges), which can increase sensitivity and avoid false negative errors. In the meanwhile, we penalize the subgraph sizes (i.e., controlling the number of edges with $\delta_{ij} > 0$) through the ℓ_0 shrinkage term $\|\boldsymbol{U}\|_0$. Note that $\delta_{ij} > 0$ only if $e_{ij} \in G_c$ or $e_{ij} \in I_{c,c'}$ for the ICN structure. Therefore, we regulate the size of each G_c and the number of edges in $I_{c,c'}$ to control the false positive rate. In other words, we seek to include highly correlated edges using the minimally-sized cover of $\{G_c, I_{c,c'}\}_{c,c'=1}^{C}$ [118, 119, 120]. By implementing the ℓ_0 graph norm shrinkage, we can accurately recognize the underlying ICN structure without false positively combining interconnected communities into mega-communities.

Optimizing (4.6) may lead to estimating some nodes as singletons, or isolated nodes. The singleton genes have weak correlations between other genes in G, which are impossible to glue up small communities or be added to existing communities. This also reflects the shrinkage force of our method in controlling the false discovery rate. In practice, singleton genes are less interactive with other genes, and involved in more isolated biologic processes.

The shrinkage level of our estimation is regulated by λ_0. Generally, a smaller λ_0 leads to larger-sized networks covering more correlated edges (i.e., relatively higher sensitivity and false positive rates) while a greater λ_0 to yield more parsimonious networks covering less correlated edges (i.e., relatively lower sensitivity and false positive rates). λ_0 can be selected based on cross-validation [121] or by prior knowledge. Note that C in our objective function (4.6) is automatically determined by the optimization function rather than prespecified.

When the ICN structure presents in $\boldsymbol{W}_{p\times p}$, our objective function (4.6) favors the ICN structure over the independent community structure. The independent community structure either merges two interconnected communities into one large community with a larger ℓ_0 graph norm or treats them as independent communities with a smaller ℓ_1

matrix norm $\|\boldsymbol{U}\|_1$. Therefore, the objective function (4.6) is tailored for ICN structure extraction. We develop an iterative three-step procedure to optimize (4.6):

1. We first detect a set of communities with highly correlated intra-community edges $\{\hat{G}_c\}$;
2. We then examine which pairs of communities are interconnected $\hat{G}_c \Leftrightarrow \hat{G}_{c'}$;
3. For each pair of interconnected communities, we identify the edge set connecting pairs of communities: $\hat{E}^*_{c,c'}$ of $\hat{I}_{c,c'} = (\hat{V}^*_c, \hat{V}^*_{c'}, \hat{E}^*_{c,c'})$.

Step 1: detecting community structure

We first detect a set of densely connected, non-overlapping communities $\{G_c\}$ as the backbone of ICN structure. In step 1, we aim to only assign highly correlated genes into communities (see Figure 4.1(c)) and avoid the disruption from the between-community connections and false positive noise. We optimize the objective function:

$$\underset{C,\{G_c\}}{\arg\max} \log \|\boldsymbol{D}\|_1 - \tau_0 \log \|\boldsymbol{D}\|_0, \tag{4.7}$$

where τ_0 is a tuning parameter. The matrix $\boldsymbol{D} = \{d_{ij}\}$ differs from $\boldsymbol{U} = \{u_{ij}\}$ in (4.6) in that it only captures within-community connections, while ignoring the inter-community connections. In step 1, we assume that $\mathcal{T}(G)$ is an independent community structure. Thus, we have $d_{ij} = w_{ij}$ for the within-community edges $e_{ij} \in G_c, \forall c$, and $d_{ij} = 0$ otherwise. Similar to the objective function (4.6), the step 1 objective function can effectively suppress false positive noise and only detect densely connected communities by implementing the ℓ_0 graph norm shrinkage. In step 1, we focus on controlling false positive errors regardless of sensitivity. However, this has little impact on the overall sensitivity since we'll identify edges interconnecting communities in the following steps. Like λ_0, the tuning parameter τ_0 can also be selected by cross-validation. We would refer the audience to [69] for the details to implement (4.7) and corresponding theoretical results.

Step 2: testing whether a pair of communities are interconnected

Step 1 yields a set of estimated communities $\{\hat{G}_c\}$. We next examine which of these non-trivial communities are interconnected by statistical tests. For a pair of communities G_c and $G_{c'}$, the genes in one community G_c can be connected to some or all genes in another community $G_{c'}$. We denote $G(V_c, V_{c'}) = (V_c, V_{c'}, E(V_c, V_{c'}), \boldsymbol{W}_{c,c'})$ as the bipartite subgraph induced by all possible edges between communities $E(V_c, V_{c'})$. $\boldsymbol{W}_{c,c'} = \{w_{ij} | i \in V_c, j \in V_{c'}\}$ is the corresponding weighted matrix. We shall test $H_0 : G_c \not\Leftrightarrow G_{c'}$ vs. $H_a : G_c \Leftrightarrow G_{c'}$. Under H_0, the distribution of edge weights in $\boldsymbol{W}_{c,c'}$ is simply the distribution of $\{w_{ij} | \delta_{ij} = 0\}$. A viable strategy to approximate the distribution of $\{w_{ij} | \delta_{ij} = 0\}$ is to sample from edges connecting singletons based on step 1 community detection results.

We denote $G_R = (V_R, E_R, \boldsymbol{W}_R)$ as the singleton-induced subgraph. Then, the edge weights in $\boldsymbol{W}_R$ follow an f_0 distribution. The difference between the two distributions of edge weights in $G(V_c, V_{c'})$ and G_R can be measured by the Kullback–Leibler (KL) divergence. We reject H_0 for a large enough KL divergence and conclude that $G_c \Leftrightarrow G_{c'}$.

We perform this test non-parametrically by resampling. Specifically, we sample M sets of edge weights $B_1, ..., B_M$ (e.g., $M = 2,000$) from edges outside communities $\{w_{ij} | e_{ij} \notin \cup_{c=1}^{C} E_c\}$ with $|B_m| = |E(V_c, V_{c'})| = |V_c||V_{c'}|$. The p value is calculated as the percentile of KL divergence ($\boldsymbol{W}_{c,c'}$ vs. $\boldsymbol{W}_R$) among random samples d_m (B_m vs. $\boldsymbol{W}_R$). If p is less than a specific type 1 error rate α (e.g., $\alpha = 0.05$), then $G_c \Leftrightarrow G_{c'}$. We describe the details of Step 2 in the following Algorithm 5. The computational complexity for the pairwise tests of interconnectivity is $O(Mp^2)$, where M is the number of random samples and p the number of genes.

Algorithm 5 Step 2: Testing whether two communities (e.g. G_c and $G_{c'}$) are interconnected

```
procedure ALGORITHM
    Calculate the KL-divergence between $G(V_c, V_{c'})$ and a random graph $G_R = (V_R, E_R, \boldsymbol{W}_R)$ and denote it as $d_0$:
        $$d_0 = D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$
    where $P$ and $Q$ are estimated densities for edge weights from $G_R$ and $G(V_c, V_{c'})$ on support $\mathcal{X}$, respectively.
    for m = 1 to M do
        Get a random sample of edge weights $B_m$ of size $|V_c||V_{c'}|$ from edges outside communities $\{w_{ij} | e_{ij} \notin \cup_{c=1}^{C} E_c\}$;
        Calculate the KL-divergence between $B_m$ and $G_R$ and denote as $d_m$;
    end for
    $p_0$ is the percentile of $d_0$ in $\{d_m\}_{m=1}^{M}$;
    if $p_0$ is smaller than $\alpha$ then
        $G_c \Leftrightarrow G_{c'}$.
    end if
end procedure
```

Step 3: identifying connecting edges between each pair of interconnected communities

For a pair of connected communities $G_c \Leftrightarrow G_{c'}$, we further identify the connection edges $E^*_{c,c'}$ from all possible edges between these two communities, $E(V_c, V_{c'})$. Consistent with the overall objective function (4.6), our goal is to recover maximally correlated edges with parsimonious estimated community sizes. The ℓ_0 graph norm shrinkage is used to suppress false positive noise while maintain high sensitivity.

Let $G_{(c+c')}$ denote the subgraph induced by nodes in G_c and $G_{c'}$, and $\boldsymbol{W}_{(c+c')}$ be the associated edge weights. $\boldsymbol{W}_{(c+c')}$ differ from the edge weight matrix $\boldsymbol{W}_{c,c'}$ of $G(V_c, V_{c'})$ (defined in step 2) as $\boldsymbol{W}_{(c+c')}$ includes both intra- and inter- community edges whereas $\boldsymbol{W}_{c,c'}$ is only for inter-community edges.

We further introduce an indicator variable δ_{ij} for each edge in $\boldsymbol{W}_{(c+c')}$, such that $\delta_{ij} = 1$ if $e_{ij} \in G_c, G_{c'}$ or $I_{c,c'}$ and $\delta_{ij} = 0$ otherwise. Then, the matrix $\boldsymbol{S} = \{s_{ij}\}$ integrates the topological structure of $G_{(c+c')}$ into the observed data $\boldsymbol{W}_{(c+c')}$ by letting $s_{ij} = w_{ij}\delta_{ij}$. The step 3 objective function coincides with the primary objective function (4.6)

$$\underset{I_{c,c'}}{\arg\max} \left(\log \|\boldsymbol{S}\|_1 - \gamma_0 \log \|\boldsymbol{S}\|_0 \right), \tag{4.8}$$

where γ_0 is the tuning parameter. The optimization of (4.8) can be implemented by hard thresholding. The detailed procedure is described in Algorithm 6.

The computational complexity for the three-step procedure of ICN detection is

$O(p^4)$, where p is the number of genes.

Algorithm 6 Step 3: Identify connecting edges between interconnected communities (e.g. $G_c \Leftrightarrow G_{c'}$)

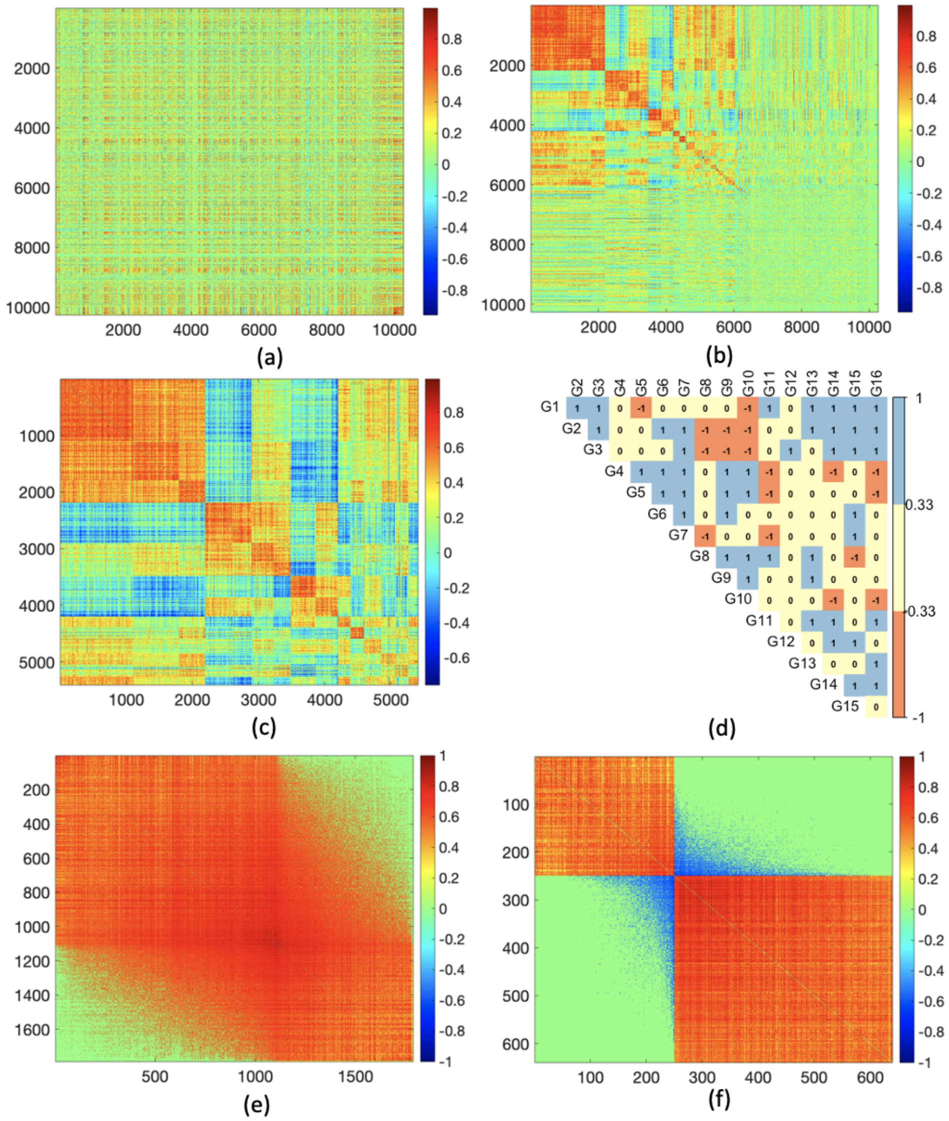
```
procedure ALGORITHM
    Given a sequence of cut-offs $r_1$, $r_2$,. . . ,$r_L$;
    Extract the weight matrix of the subnetwork $G_{(c+c')}$ for the pair of communities (e.g. $c$ and $c'$): $\boldsymbol{W}_{(c+c')} = \{w_{ij} | i, j \in V_c \cup V_c'\}$;
    for $l = 1$ to $L$ do
        Obtain the estimated set of connecting edges by thresholding: $\hat{E}^{(l)}_{c,c'} = \{e_{ij} | i \in V_c, j \in V_{c'}, w_{ij} > r_l\}$;
        Obtain the estimation of indicator variable based on estimated subnetwork: $\hat{\delta}_{ij} = 1$ if $e_{ij} \in \hat{E}_c, \hat{E}_{c'}$ or $\hat{E}^{(l)}_{c,c'}$;
        Observe $\hat{\boldsymbol{S}}$ as $\hat{s}_{ij} = (\boldsymbol{W}_{(c+c')})_{ij}\delta_{ij}$ and calculate the value of objective function as $h_l = \log \|\hat{\boldsymbol{S}}\|_1 - \gamma_0 \log \|\hat{\boldsymbol{S}}\|_0$;
    end for
    Select the $l^*$ for the largest value of objective function: $h_l$, $l = 1, ..., L$;
    The detected connecting edges between community $c$ and $c'$ are $\{e_{ij} | i \in V_c, j \in V_{c'}, w_{ij} > r_{l^*}\}$;
end procedure
```

4.3 Data Example

We demonstrate the effectiveness of our proposed method on an RNA-seq gene expression data set from The Cancer Genome Atlas (TCGA) Acute Myeloid Leukemia (AML) study [122]. The processed RNA-seq data of 173 leukemia patients measured in Transcripts Per Million (TPM) values were retrieved from the TCGA data repository on Broad GDAC Firehose (https://gdac.broadinstitute.org/). We performed standard preprocessing by filtering out genes with low means or low variance and took a $\log_2$ transformation. A total of 10259 genes survived our preliminary filtering for community detection. We then calculate the pairwise Pearson correlation matrix $\boldsymbol{W}_{10259 \times 10259}$ between the genes and input it into our algorithm. Depending on the raw data type, one may also use other similarity measures, such as Spearman correlation coefficient and maximum information coefficient, to replace the Pearson correlation.

We apply the proposed algorithm to the correlation matrix (Figure 4.3(a)). The community structure is detected by optimizing objective function (4.7) using algorithm of Step 1 (Figure 4.3(b)). The detected subnetwork includes 16 communities of interest displayed in Figure 4.3(c). We next perform Step 2 algorithm to find interconnected community pairs. As is shown in Figure 4.3(d), our method identified 16 blocks, where the blue and red squares indicate the significantly positive and negative interconnections between the communities of row and column indices. The yellow blocks represent the non-interconnection between some community pairs. Finally, we implement Step 3 algorithm to identify interconnected edges between each interconnected community pairs. To present more details of our estimation results, we also displayed the weights of estimated nonzero edges in the subnetwork between communities 1 and 2 in Figure 4.3(e), and that between communities 7 and 8 in Figure 4.3(f).

Communities can be either positively or negatively interconnected with each other. For example, a set of genes in the 1st community shows strong positive correlations with the 2nd community as Figure 4.3(e), while genes in the 7th and 8th community are negatively correlated as Figure 4.3(f). To gain biological insights of the interconnected communities by ICN, we followed a systematic approach to perform pathway analysis (a.k.a. gene set enrichment analysis) using curated KEGG [123] and Reactome pathway database [124]. For each pair of communities (e.g., community 1 and 2, community 7 and 8), we first performed pathway analysis to the union of genes from the two modules (e.g., $\hat{V}_1 \cup \hat{V}_2$) and identified enriched pathways with Fisher's exact test p-value < 0.05 and



(a)

(b)

(c)

	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16
G1	1	1	0	-1	0	0	0	0	-1	1	0	1	1	1	1
G2		1	0	0	1	1	-1	-1	-1	0	0	1	1	1	1
G3			0	0	0	1	-1	-1	-1	0	1	0	1	1	1
G4				1	1	1	0	1	1	-1	0	0	-1	0	-1
G5					1	1	0	1	1	-1	0	0	0	0	-1
G6						1	0	1	0	0	0	0	0	1	0
G7							-1	0	0	-1	0	0	0	1	0
G8								1	1	1	0	1	0	-1	0
G9									1	0	0	1	0	0	0
G10										0	0	0	-1	0	-1
G11											0	1	1	0	1
G12												0	1	1	0
G13													0	0	1
G14														1	1
G15															0

(d)

(e)

(f)

Figure 4.3: The procedure to detect interconnected communities: (a) demonstrates the input large correlation matrix $\boldsymbol{W}_{10259\times10259}$; (b) shows the step 1 analysis results by re-ordering the genes and revealing latent dense communities and singletons; (c) zooms in to show the 16 dense communities; (d) demonstrates the testing results for interconnection between 120 pairs of 16 communities (postively connected: blue; negatively connected: brown; not connected: yellow). We illustrate the results of step 3 using two examples, in (e) we identify positively connected edges between the 1st and 2nd community; and in (f) we find negatively connected edges between the 7th and 8th community.

at least 10 genes overlapping with the union set. We then conducted pathway analysis separately on the genes without interconnecting edges in each module (e.g., $\hat{V}_1 \setminus \hat{V}_1^*$ and $\hat{V}_2 \setminus \hat{V}_2^*$) as well as genes in the interconnection subgraph (e.g., $\hat{V}_1^* \cup \hat{V}_2^*$). The demonstrative results listing the Fisher's exact test in each part for the top 10 KEGG and Reactome pathways for community 1-2 and 7-8 pair are included in Appendix. The pathway enrichment patterns show both commonality and uniqueness of pathways enriched in the three parts, implying the specificity of each part identified by ICN. For example, Wnt signaling pathways and ERBB signaling pathway are enriched with genes from community 1-specific and interconnecting part while RNA degradation is only enriched in the interconnecting part.

Regarding the enriched pathways in the interconnection between community 1 and 2, we further investigated the genes contributed by each module in the pathway topology plots [125]. Among the top ranked pathways, for example, mitogen-activated protein kinase (MAPK) signaling pathway of KEGG plays a critical regulatory role in leukemia [126] and is enriched in the interconnection subgraph $\hat{V}_1^* \cup \hat{V}_2^*$. The topology plot of this pathway displayed in Figure 4.4 found the genes from community 1 and 2 highlighted in two different colors work synergistically to realize the shared function of the MAPK pathway.

For community 7 and 8, it turned out that the pathways enriched by the gene set in the 7th community include MAPK pathways and its downstream regulated processes such as RNA transcription by RNA Polymerase, while those enriched by the set in the 8th community are characterized by immune response-activating signal transduction. Their negative relationship implies the adverse impact of MAPK pathway dysregulation on host

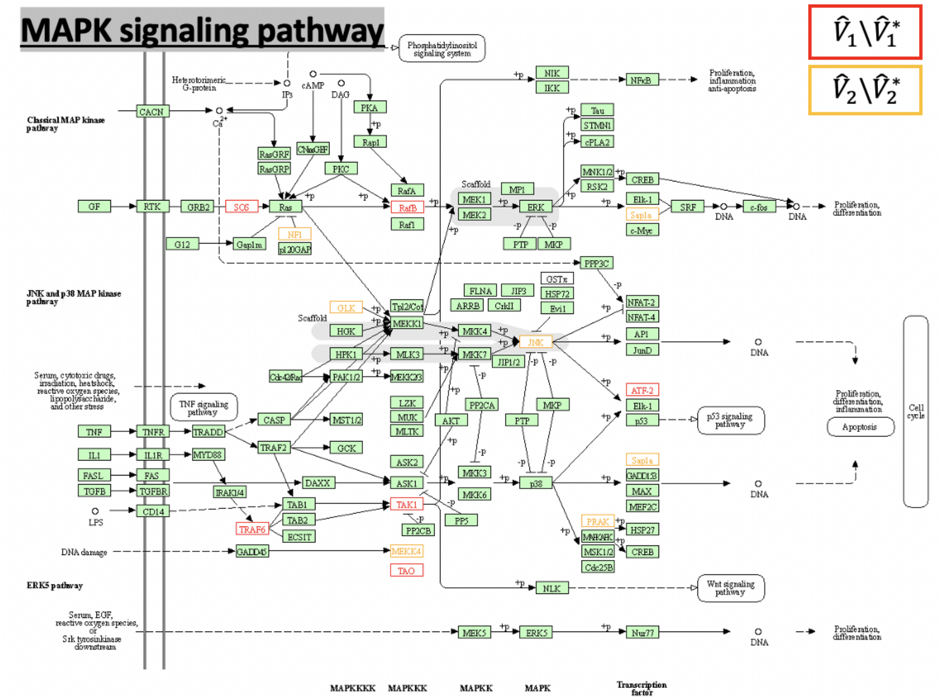


Figure 4.4: Topology plot of selected MAPK signaling pathway with the genes from community 1 and 2 highlighted by red and yellow, respectively.

immune response and represent a possible immune evasion mechanism by tumor cells in AML [127, 128].

4.4 Simulation Results

Table 4.1: Estimated edge-level FPR, FNR and standard errors for $C_0 = 3$ and $n = 200$ based on different detection methods. The FPR and FNR are calculated separately for between-community edges, within-community edges, and overall edges.

			ICN	MMSB	BNMTF	HK-relax
(60, 40, 40)	Between	FPR	0.008 (0.008)	0.221 (0.020)	0.027 (0.012)	0.031 (0.020)
		FNR	0.157 (0.129)	0.768 (0.041)	0.780 (0.112)	0.848 (0.109)
	Within	FPR	0.001 (0.003)	0.054 (0.009)	0.065 (0.015)	0.084 (0.028)
		FNR	0.025 (0.061)	0.944 (0.012)	0.120 (0.072)	0.204 (0.112)
	Overall	FPR	0.009 (0.009)	0.274 (0.019)	0.016 (0.008)	0.038 (0.029)
		FNR	0.039 (0.066)	0.716 (0.030)	0.061 (0.025)	0.126 (0.037)
(30, 20, 20)	Between	FPR	0.005 (0.003)	0.237 (0.023)	0.005 (0.002)	0.016 (0.011)
		FNR	0.189 (0.234)	0.739 (0.084)	0.781 (0.071)	0.815 (0.168)
	Within	FPR	0.001 (0.005)	0.059 (0.010)	0.015 (0.003)	0.035 (0.009)
		FNR	0.072 (0.154)	0.939 (0.013)	0.090 (0.038)	0.036 (0.023)
	Overall	FPR	0.005 (0.007)	0.295 (0.024)	0.005 (0.002)	0.034 (0.010)
		FNR	0.070 (0.161)	0.668 (0.039)	0.073 (0.029)	0.017 (0.017)

In this section, we focus on comparing the performance of our method to existing methods using synthetic data. We set up $p = 200$ genes with $C_0 = 3$ non-trivial communities. We set the three community sizes and the number of singleton nodes to be $(60, 40, 40; 60)$ and $(30, 20, 20; 130)$, respectively. Regarding the ICN structure, we assume that community 1 and 2 are positively interconnected with 40% connecting edges, and community 1 and 3 are negatively interconnected with 20% connecting edges. We first generated an adjacency matrix $\mathbf{\Delta}_{p\times p} = \{\delta_{ij}\}$ based on the ICN structure by (4.3). We then simulated edge weights in $\mathbf{\Omega}_{p\times p}$ from a 4-component mixture distribution by

(4.4). We let f_1, f_2, f_3 and f_0 follow normal distributions with means 0.8, 0.5, -0.6, 0, respectively, and standard deviation 0.1. To ensure the positive semi-definiteness of the covariance matrix, we then observed $\mathbf{\Sigma} = \mathbf{\Omega}^T\mathbf{\Omega}$ as our true covariance matrix. Finally, we generated the observable data matrix X from $N(0, \mathbf{\Sigma})$ with a sample size $n = 200$. We repeated each setting for 500 times.

We applied the ICN algorithm to the sample correlation matrix of $\boldsymbol{X}$. The results were stored as recovered network structure in an estimated matrix of $\hat{\mathbf{\Delta}} = \{\hat{\delta}_{ij}\}$. For each simulated data set, the computational time of the three-step procedure is around 65 seconds on a laptop with CPU Core i5 and memory 8GB.

We further compared our method with benchmarks of overlapping community detection methods, including Mixed Membership Stochastic Blockmodel (MMSB) [129], Bounded Non-Negative Matrix Tri-Factorization (BNMTF) [107] and Heat Kernel-Based Community Detection (HK-relax) [130]. We evaluated the performance of these methods by comparing $\hat{\Delta}$ with Δ in terms of false positive rates (FPR) and false negative rates (FNR). Specifically,

$$\text{FPR}_l = \frac{\sum_{i,j} I(\delta_{ij} \neq l, \hat{\delta}_{ij} = l)}{\sum_{i,j} I(\delta_{ij} \neq l)}; \text{FNR}_l = \frac{\sum_{i,j} I(\delta_{ij} = l, \hat{\delta}_{ij} \neq l)}{\sum_{i,j} I(\delta_{ij} = l)}$$

where $0 < l \leq K$ is a label for the subgraph type (e.g., community or interconnection subgraph). Then, FPR and FNR were evaluated separately for between- and within-community edges. Last, we measured the overall FPR and FNR by categorizing edges into two classes $\delta_{ij} > 0$ vs. $\delta_{ij} = 0$.

We summarize the performance of ICN and competing methods across 500 repeated experiments under each configuration by reporting means and standard deviations in Table

4.1. Additional simulation results under various settings are presented in Appendix. In Table 4.1, our method demonstrates clear systematical advantages, possibly due to its greater flexibility in modeling fully and partially connected between-community edges. Specifically, we find that the ICN method can better recover the Between-community edges and control the overall false positive error rate, and thus more accurately detect the latent network topology.

4.5 Discussion

Gene co-expression analysis has been at the heart of genomic data analysis to understand the complex biological processes and pathways. Currently available GCN tools typically focus on modeling block-diagonal correlation structures, while inter-block correlations widely present in many data sets. In this chapter, we presented a new model to capture inter-block correlations in a flexible way. Despite its simplicity, our objective function is well suited to revealing the organized GCN structure while preserving the high network modularity and efficiency. Controlling false positive findings is critical for GCN analysis because false positively connected edges severely interrupt the accurate extraction of interconnected network structure. To cope with such practical needs, we proposed an ℓ_0 graph norm penalty term to suppress false positive edges by covering connected edges using minimally-sized ICN structures. We also developed efficient algorithms to implement the optimization. The advantageous performance of our method was showcased on both synthetic and real-world data.

Besides the modeling and computational advantages, the ICN model may lead to new biological findings by providing a new angle to study the co-expression network in a hierarchical network structure from interactive genes (communities) to interactive communities. The proposed ICN method can also be applied to the other omics data with latent and complex interactive structures (e.g., protein-protein interactions). The accurately detected ICN covariance structure can also assist differential expression analysis by performing dependence adjusted multivariate inference [131].

Chapter 5: Conclusions

Nowadays, the analysis of high-dimensional biomedical data becomes crucial in various applications including genomics, genetics, neuroimaging, and so forth. However, the high dimensionality and the complex interactive relationships among biomedical objects bring new challenges to statistical inference.

In Chapter 2, since the edge-wise inference results from group-level brain connectome data can be overwhelmed by substantial false positive and negative errors, we develop the ADSD method to reliably and robustly extract the latent disease-related subnetworks. The proposed ADSD and ℓ_0-norm based objective functions are suitable for general applications in biomedical data. The two-step procedure can be conveniently adapted in cases with multiple clinical groups and regressions with covariates. We develop the asymptotic properties of the ADSD algorithm and ℓ_0-norm based algorithms regarding the recovery of true subgraphs and the optimality of the objective function. The application in the data example reveals the hypo-connections in the salience network-centered subnetwork for patients with schizophrenia, which assists the understanding of brain activities related to brain disorder.

In Chapter 3, with an emphasis on the pleiotropic and polygenic association pattern between genetic markers and imaging measures, we characterize the imaging-genetics association as a bipartite graph model with an IGDB structure. The extracted IGDB could exhibit the systematic genetic effects on quantitative imaging measurements, which produces more biologically interpretable findings. We construct theoretical results regarding the approximation bounds of the IGDB algorithm and the asymptotic power of the proposed likelihood ratio approach. However, the current study also points to several future directions. The screening methods accounting for the dependence structures among genetic markers and imaging features would be helpful for further statistical inferences or regression models. Statistical models to relate these genetic and brain activity associations with brain disorders still deserve further research.

In Chapter 4, we supplement the block-diagonal correlation structures in current GCN analysis with the inter-block correlations through ICN. The ICN method possesses favorable advantages in its flexibility of modeling gene interactions and computational efficiency. The proposed ICN approach is also robust to noise and effective to capture the delicate network structure, and thus can be applied to a wide range of bioinformatics research problems focusing on detections of interactive structures. These new pathway patterns can also serve as an input into further analysis. For example, the detected interconnected communities provide data-driven pathways which can be integrated into advanced machine learning algorithms.

Appendix A: Supplemental for Chapter 2

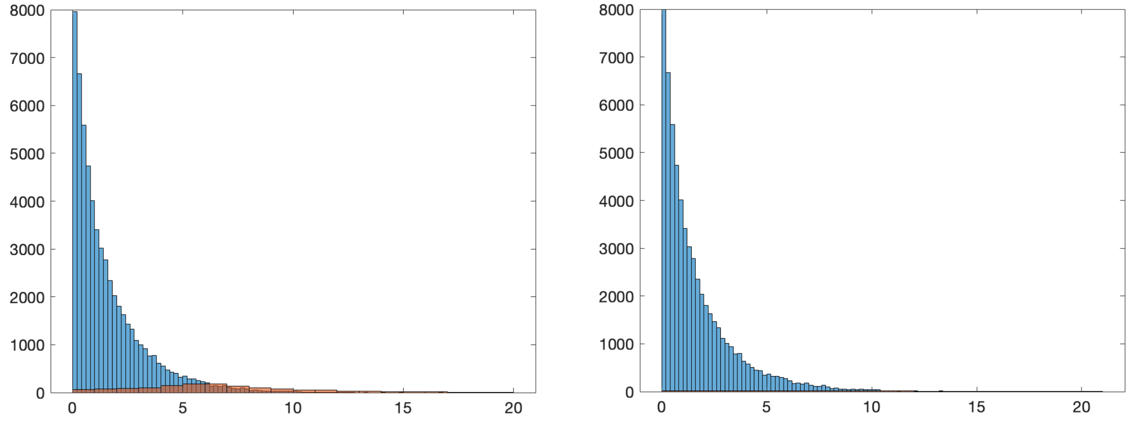
A.1 Additional Data Results

A.1.1 Comparison results with NBS method

We conducted an additional analysis using NBS with the option“intensity” and various threshold values on the real data example. We summarize the results in the following Table A.1. Indeed, these results are more comparable to ADSD than the option “extent”. The current results demonstrate that: i) when the threshold is more stringent, NBS tends to identify over-sparse (low density) subgraph with average node-degree close to 1; ii) when the threshold is more liberal, NBS selects a sparse subgraph with a large number of nodes. In general, edges detected by NBS are more sparse and randomly distributed (not restricted in a small node-induced subgraph) in the overall graph, which is limited to characterize the systematically altered connectivity patterns by the phenotype. We further compared edges selected by NBS and ADSD in Figure A.1. The sensitivity of ADSD is higher because it can select edges systematically rather than only relying on edge-wise inference as NBS.

Table A.1: Connected components detected by NBS function "intensity" option with varied thresholds.

Threshold	Number of nodes	Number of edges
2.5	196	624
2.8	138	305
3.1	93	151
3.3	71	92
3.5	36	48



(a) ADSD

(b) NBS "intensity" with threshold 3.3

Figure A.1: Histograms of $-\log(p)$ values of edges selected by ADSD and NBS with threshold 3.3

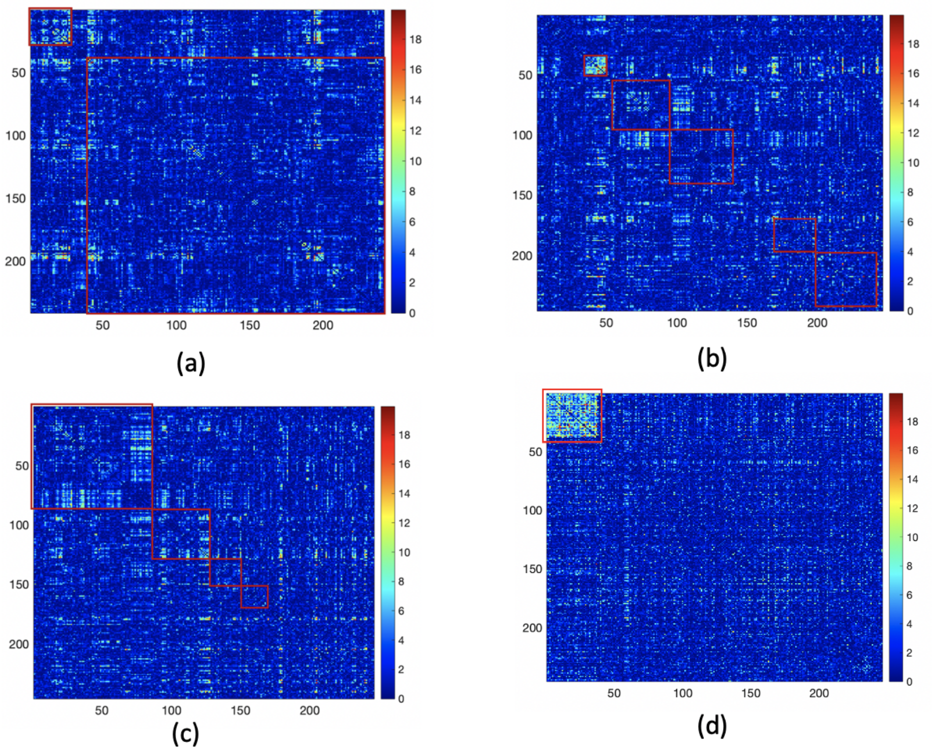


Figure A.2: Results from community detection methods: (a) spectral clustering ($K = 10$) in [1]; (b) Louvain method by [2]; (c) INFOMAP algorithm by [3]; (d) ADSD

A.1.2 Comparison results with community detection methods

We compared our approach to several popular community detection methods. We demonstrate the results in the following Web Figure 3. Our method outperforms the competing algorithms. The difference can be explained by the fact that dense subgraph discovery aims to extract a subgraph with a high density of phenotype-related edges, while community detection algorithms tend to assign similar nodes into balance-sized communities.

A.2 Detailed Proofs

Before we prove the Theorem 1, we first establish the following Lemma 3 by applying the idea of proof in [60] with density function $f_1(S)$ directly.

Lemma A.1. *The greedy algorithm will give a* $2n^{|\lambda-1|}$ *approximation, i.e.* $f(S^*) \leq 2n^{|\lambda-1|}v$.

Proof. For an undirected graph, we assign the edge e_{ij} to either node i or j. Denote the number of edges assigned to node i as $d(i)$. Then,

$$\sum_{i \in S^*} d(i) \geq |E(S^*)|,$$

since all edges in $E(S^*)$ will be assigned to node i or j which will be included in $\sum_{i \in S^*} d(i)$.

Consider a specific way to assign edges. Let the node gets assigned when it is removed in the greedy algorithm. In other words, $d(i)$ equals the degree of node i in the iteration that it is removed in greedy algorithm. Assume the subgraph at the iteration that i is removed is $G(S')$ such that at this iteration the nodes set changes from $S' \rightarrow S'/\{i\}$. Therefore, since the node i is deleted for the smallest degree, we will have the degree of i is smaller than the average degree in this iteration:

$$d(i) \leq 2\frac{|E(S')|}{|S'|}.$$

Case 1: For $\lambda < 1$,

$$d(i) \leq 2\frac{|E(S')|}{|S'|} \leq 2\frac{|E(S')|}{|S'|^\lambda} \leq 2\frac{|E(\tilde{S})|}{|\tilde{S}|^\lambda} = 2v,$$

$$\Rightarrow f(S^*) = \frac{|E(S^*)|}{|S^*|^\lambda} \leq \frac{\sum_{i \in S^*} d(i)}{|S^*|^\lambda} \leq 2v\frac{|S^*|}{|S^*|^\lambda} \leq 2vn^{1-\lambda}.$$

Case 2: For $\lambda > 1$,

$$d(i) \leq 2\frac{|E(S')|}{|S'|} \leq 2\frac{|E(S')|}{|S'|^\lambda} \times |S'|^{\lambda-1} \leq 2\frac{|E(\tilde{S})|}{|\tilde{S}|^\lambda} \times |S'|^{\lambda-1} = 2vn^{\lambda-1},$$

$$\Rightarrow f(S^*) = \frac{|E(S^*)|}{|S^*|^\lambda} \leq \frac{\sum_{i \in S^*} d(i)}{|S^*|^\lambda} \leq 2vn^{\lambda-1}\frac{|S^*|}{|S^*|^\lambda} \leq 2vn^{\lambda-1}.$$

□

Proof of Theorem 2.1. **Case 1:** $\lambda \geq 2$**.** The subgraph with only two nodes has $f(S;\lambda) = 1/2^\lambda$, otherwise all elements in this 2×2 matrix should be zero and inductively the raw graph have no edges (all elements in the corresponding matrix should be zero). Thus, $f(\tilde{S};\lambda) \geq 1/2^\lambda$.

On the other hand, when $\lambda \geq 2$,

$$f(S^*;\lambda) = \frac{|E(S^*)|}{|S^*|^\lambda} \leq \frac{|S^*|^2}{2|S^*|^\lambda} \leq \frac{1}{2}.$$

Hence, $f(\tilde{S};\lambda) \geq 2^{1-\lambda} f(S^*;\lambda)$ and $c(\lambda) = 2^{\lambda-1}$.

Case 2: $1 < \lambda < 2$**.** Assume there exists $\lambda > 0$ such that $f(S^*;\lambda) \leq 2n^a f(\tilde{S};\lambda)$ and we want to find such λ. It suffices to show

$$f(S^*;\lambda) = \frac{|E(S^*)|}{|S^*|^\lambda} \leq \frac{|S^*|^2}{2|S^*|^\lambda} \leq 2\frac{1}{2^\lambda}n^a$$

since from case 1, $f(\tilde{S};\lambda)$ has a lower bound $1/2^\lambda$. It's automatically true for $|S^*| \leq (2^{2-\lambda}n^a)^{1/(2-\lambda)} = 2n^{a/(2-\lambda)}$.

For $|S^*| > 2n^{a/(2-\lambda)}$, from the proof of case 2 in lemma 3, we have

$$\begin{aligned} f(S^*;\lambda) =& \frac{E(S^*)}{|S^*|^\lambda} \leq \frac{\sum_{i\in S^*} d(i)}{|S^*|^\lambda} \leq 2f(\tilde{S};\lambda)n^{\lambda-1}|S^*|^{1-\lambda} \\ &\leq 2f(\tilde{S};\lambda)n^{\lambda-1}(2n^{a/(2-\lambda)})^{1-\lambda} \end{aligned}$$

Let $2f(\tilde{S};\lambda)n^{\lambda-1}(2n^{a/(2-\lambda)})^{1-\lambda} = 2f(\tilde{S};\lambda)n^a$, we will get $a = (\lambda-1)(2-\lambda)$. Hence, $f(S^*;\lambda) \leq 2n^{(\lambda-1)(2-\lambda)}(1 \vee 2^{1-\lambda})f(\tilde{S};\lambda)$, then $c'(\lambda) = 1 \vee 2^{1-\lambda}$.

Case 3: For $0.5 < \lambda < 1$, the claim is automatically true from lemma 3.

Case 4: $0 < \lambda < 0.5$

$$\begin{aligned} f(S^*;\lambda) =& \frac{|E(S^*)|}{|S^*|^\lambda} \leq \frac{|E(V)|}{|S^*|^\lambda} = \frac{|E(V)|}{|S^*|^\lambda|V|^\lambda} \times |V|^\lambda \\ \leq& f(\tilde{S};\lambda)\frac{|V|^\lambda}{|S^*|^\lambda} \leq 2f(\tilde{S};\lambda)n^\lambda \end{aligned}$$

□

Proof of Theorem 2.2. Part 1. Let $\mathcal{C} = \{S_1, S_2, ..., S_{n-1}\}$ be the sequence of subgraphs generated by deleting the smallest-degree node in Algorithm 1. We first prove the true subgraph S_0 is included in $\mathcal{C}$ up to a permutation with high probability as $n \to \infty$. Denote

$n = |V|$ and $n_s = |S_0|$.

At stage $k(< n - n_s)$ of Algorithm 1, for $i \in S_0$, and $i' \in V/S_0$, we have

$$d_i = \sum_{j=1}^{n_s} X_{ij} + \sum_{j=n_s+1}^{n-k+1} Y_{ij} \quad \text{and } d_{i'} = \sum_{j=1}^{n-k+1} Y_{i'j}$$

where $X_{ij} \sim$ Bernouli(π_s), $Y_{ij} \sim$ Bernouli(π_0), and $Y_{i'j} \sim$ Bernouli(π_0)

From chernoff bound, for $i \in S_0$, $i' \in V/S_0$ and $\delta \in (0,1)$

$$\mathbb{P}(d_i \leq (1-\delta)(n_s\pi_s + (n-k-n_s)\pi_0)) \leq \exp\left[-\frac{\delta^2}{2}(n_s\pi_s + (n-k-n_s)\pi_0)\right],$$

and

$$\mathbb{P}(d_{i'} \geq (1+\delta)(n-k)\pi_0) \leq \exp\left[-\frac{\delta^2}{2+\delta^2}(n-k)\pi_0\right].$$

Hence, there exist $\epsilon_0(\delta) > 0$, such that

$$\mathbb{P}(d_i - d_{i'} \geq \epsilon_0(\delta)) \geq 1 - \exp\left[-\frac{\delta^2}{2}(n_s\pi_s + (n-k-n_s)\pi_0)\right] - \exp\left[-\frac{\delta^2}{2+\delta^2}(n-k)\pi_0\right].$$

At stage $k(< n - n_s)$ of Algorithm 1, the event of deleting a node outside S_0 is:

$$\cap_{i \in S_0} \cap_{i' \in S_k/S_0} \{d_i \geq d_{i'}\} = \cap_{i \in S_0} [\cup_{i' \in S_k/S_0} \{d_i < d_{i'}\}]^c$$

Then,

$$\mathbb{P}\left(\cap_{i\in S_0}\cap_{i'\in S_k/S_0}\{d_i \geq d_{i'}\}\right) \geq 1-\sum_{i\in S_0}\mathbb{P}\left(\cup_{i'\in S_k/S_0}\{d_i < d_{i'}\}\right)$$

$$\geq 1-\sum_{i\in S_0}\sum_{i'\in S_k/S_0}\mathbb{P}(d_i < d_i')$$

$$\geq 1-n_s(n-k-n_s)\left\{\exp\left[-\frac{\delta^2}{2}(n_s\pi_s+(n-k-n_s)\pi_0)\right]+\exp\left[-\frac{\delta^2}{2+\delta^2}(n-k)\pi_0\right]\right\}$$

Using similar argument, the class $\mathcal{C}$ includes the true subgraph up to a permutation Q:

$$\mathbb{P}(Q(S_0)\in\mathcal{C}) = \mathbb{P}\left(\cap_{k=1}^{n-n_s-1}\cap_{i\in S_0}\cap_{i'\in S_k/S_0}\{d_i\geq d_{i'}\}\right)$$

$$\geq 1-(n-n_s-1)n_s(n-k-n_s)\times$$

$$\left\{\exp\left[-\frac{\delta^2}{2}(n_s\pi_s+(n-k-n_s)\pi_0)\right]+\exp\left[-\frac{\delta^2}{2+\delta^2}(n-k)\pi_0\right]\right\}$$

$$\rightarrow 1 \text{ as } n\rightarrow\infty \text{ and } n_s = O(n^{1/2+\epsilon}).$$

Part 2. We then establish the true subgraph can be selected from $\mathcal{C}$ by density function $f(S;\lambda)$ for some λ. For $k < n-n_s$, from the proof of part 1, $S_0\subset S_k$ with high probability. From Chernouff bounds,

$$\mathbb{P}\left(2|A(S_k)|-2|A(S_0)|\leq(1+\delta)[(n-k-1)^2-(n-n_s-k-1)^2]\pi_0\right)$$

$$\geq 1-\exp\left[-\frac{\delta^2}{2+\delta^2}[(n-k-1)^2-(n-n_s-k-1)^2]\pi_0\right]$$

and

$$\mathbb{P}\left(2|A(S_0)|>(1-\delta)n_s^2\pi_s\right)\geq 1-\exp\left[-\frac{\delta^2}{2}n_s^2\pi_s\right].$$

Thus, with high probability,

$$\frac{|A(S_k)| - |A(S_0)|}{|A(S_0)|} < \frac{1+\delta}{1-\delta} * \frac{[(n-k-1)^2 - (n-n_s-k-1)^2]\pi_0}{n_s^2 \pi_s}$$

On the other hand,

$$\frac{|A(S_k)|}{(n-k-1)^\lambda} \leq \frac{|A(S_0)|}{(n_s)^\lambda} \Leftrightarrow \frac{|A(S_k)| - |A(S_0)|}{|A(S_0)|} \leq \frac{(n-k-1)^\lambda}{(n_s)^\lambda} - 1,$$

hence, it suffices to have

$$\frac{1+\delta}{1-\delta} * \frac{[(n-k-1)^2 - (n-n_s-k-1)^2]\pi_0}{n_s^2 \pi_s} \leq \frac{(n-k-1)^\lambda}{(n_s)^\lambda} - 1.$$

For $\pi_s > \pi_0$, there exist corresponding δ and λ to make the inequality holds.

For $k > n - n_s$, we could use similar argument and the claim is true. □

Proof of Theorem 2.3. We implement the proof of SBM under maximum likelihood fitting by [132] that constrain stochastic block model with $K \lesssim n^{1/2}$ communities and the average degree $M \gtrsim (\log(n))^{3+\delta}$. The growth restriction on K is utilized to bound the number of possible choices of assignment. Although the restriction is not satisfied in our model, the assumption such that all edges outside the dense subgraph are considered as singletons makes the number of possible assignments being the same as $K = 2$. The restriction on average degree is also automatically satisfied for fixed π_s and π_0.

Next, we prove the conclusion is valid when the assignment is maximized over a smaller class of possible solutions, which is generated by different values of λ.

The Theorems 1 and 2 in [132] also hold in our model because the maximization over a subset of parameter space is smaller than over the whole space. For their theorem 3, from the proof of our theorem 2, the true assignment is in the subset that we maximized with probability converging to 1, i.e. $\boldsymbol{\theta}(S_0) \in \{\boldsymbol{\theta}(\tilde{S}_\lambda), \lambda \in (0,2)\}$. Then theorem 3 holds with probability converging to 1, i.e. $\bar{L}_P(\boldsymbol{\theta}(S_0))$- $\bar{L}_P(\hat{\boldsymbol{\theta}}) = o_p(M)$. Hence, our claim is true. □

Proof of Lemma 2.1. We discuss the two cases.

- We first present the proof for the $H_{G;0}$ case. By Bernstein's inequality, for any $v \in \{v_0, \dots, n\}$ and any $V \subseteq [n]$ satisfying $|V| = v$, we have

$$\begin{aligned}\left(\left|v^{-2}\sum_{i,j\in V}(A_{ij}-\pi_0)\right| > \gamma-\pi_0\right) &\le 2\exp\left(-\frac{v^4(\gamma-\pi_0)^2}{2[v^2+\frac{1}{3}v^2(\gamma-\pi_0)]}\right)\\ &= 2\exp\left(-\left\{\frac{2}{(\gamma-\pi_0)^2}+\frac{2}{3(\gamma-\pi_0)}\right\}^{-1}\cdot v^2\right)\end{aligned}$$

  Therefore by a union bound, we have

$$\begin{aligned}&\left(\cup_{V\subseteq[n]:v=|V|\ge v_0}\left\{\left|v^{-2}\sum_{i,j\in V}(A_{ij}-\pi_0)\right| > \gamma-\pi_0\right\}\right)\\ \le&\sum_{v=v_0}^{n}\binom{n}{v}\cdot 2\exp\left(-\left\{\frac{2}{(\gamma-\pi_0)^2}+\frac{2}{3(\gamma-\pi_0)}\right\}^{-1}\cdot v^2\right)\\ \le&\sum_{v=v_0}^{n}2\exp\left(-\left\{\frac{2}{(\gamma-\pi_0)^2}+\frac{2}{3(\gamma-\pi_0)}\right\}^{-1}\cdot v^2+v\log n\right)\\ \le&\sum_{v=v_0}^{n}2\exp\left(-\left\{\frac{4}{(\gamma-\pi_0)^2}+\frac{4}{3(\gamma-\pi_0)}\right\}^{-1}\cdot v^2\right)\\ \le&2n\cdot\exp\left(-\left\{\frac{4}{(\gamma-\pi_0)^2}+\frac{4}{3(\gamma-\pi_0)}\right\}^{-1}\cdot v^2\right)\end{aligned}$$

- Now we prove for the $H_{G;a}$ case. The strategy is to simply consider G_c and show that with high probability, G_c would form a γ-quasi clique in $G[r]$. We have

$$
\begin{aligned}
&\left\{\binom{|G_c|}{2}^{-1}\sum_{i,j:(i,j)\in E(G_c)}(G[r])_{ij}\geq\gamma|H_{G;a}\right\}\\
=&\left\{\binom{|G_c|}{2}^{-1}\sum_{i,j:(i,j)\in E(G_c)}\{(G[r])_{ij}-\pi_s\}\geq\gamma-\pi_s|H_{G;a}\right\}\\
\geq&1-\left\{\left|\binom{|G_c|}{2}^{-1}\sum_{i,j:(i,j)\in E(G_c)}\{(G[r])_{ij}-\pi_s\}\right|\geq\pi_s-\gamma|H_{G;a}\right\}\\
\geq&1-\exp\left\{-\frac{\frac{1}{2}(\pi_s-\gamma)^2\binom{|G_c|}{2}^2}{\binom{|G_c|}{2}+\frac{1}{3}(\pi_s-\gamma)\binom{|G_c|}{2}}\right\}\\
=&1-\exp\left\{-\frac{\frac{1}{2}(\pi_s-\gamma)^2\binom{|G_c|}{2}}{1+(\pi_s-\gamma)/3}\right\}
\end{aligned}
$$

□

Proof of Theorem 2.4. We prove the population version, such that $\boldsymbol{W}$ satisfying $w_{ij}=\mu_1$ for $\delta_{ij}=1$ and $w_{ij}=\mu_0$ for $\delta_{ij}=0$. Denote $\boldsymbol{U}_C^*$ as the matrix under true network structure G_C^*, i.e., $\boldsymbol{U}_C^*=\boldsymbol{W}*G_C^*$, and $\hat{\boldsymbol{U}}_C$ related with the optimized network structure under C, i.e., $\hat{\boldsymbol{U}}_C=\boldsymbol{W}*\hat{G}_C$.

For each $C\neq C^*$, let x be the number of corresponding edges with $\{\hat{\boldsymbol{U}}_C\}_{ij}=\mu_1$ and y to be $\{\hat{\boldsymbol{U}}_C\}_{ij}=\mu_1$. In other words, $x=\|\hat{\boldsymbol{U}}_C*G_C^*\|_0$ and $y=\|\hat{\boldsymbol{U}}_C\|_0-\|\hat{\boldsymbol{U}}_C*G_C^*\|_0$. Then, the objective function (10) takes value:

$$
\begin{aligned}
J_{\hat{\boldsymbol{U}}_C}&=\log\|\hat{\boldsymbol{U}}_C\|_1-\lambda_0\log\|\hat{\boldsymbol{U}}_C\|_0\\
&=\log\frac{x\mu_1+y\mu_0}{(x+y)^{\lambda_0}}.
\end{aligned}
$$

On the other hand, under true network structure G_C^*, the objective function (10):

$$\begin{aligned} J_{\boldsymbol{U}_C^*} &= \log \|\boldsymbol{U}_C^*\|_1 - \lambda_0 \log \|\boldsymbol{U}_C^*\|_0 \\ &= \log \frac{\|\boldsymbol{U}_C^*\|_0 \mu_1}{\|\boldsymbol{U}_C^*\|_0^{\lambda_0}} \geq \log \frac{x\mu_1}{x^{\lambda_0}} = J_{\hat{\boldsymbol{U}}_{C^*G_C^*}}, \end{aligned}$$

since the right-hand side is increasing in x for $\lambda_0 \in (0,1)$, and $x \leq \boldsymbol{U}_C^*\|_0$ by definition.

Hence, to show our criterion (10) is optimized by $C = C^*$ and $G_c = G_c^*$ for all $c = 1, \ldots, C^*$, it suffices to have

$$\begin{aligned} J_{\hat{\boldsymbol{U}}_C} < J_{\hat{\boldsymbol{U}}_{C^*G_C^*}} &\iff \frac{x\mu_1 + y\mu_0}{(x+y)^{\lambda_0}} \leq \frac{x\mu_1}{x^{\lambda_0}} \\ &\iff \frac{\mu_0}{\mu_1} < \left[\left(1 + \frac{y}{x}\right)^{\lambda_0} - 1\right] \frac{x}{y}, \end{aligned} \tag{A.1}$$

for each $C \neq C^*$ and $\hat{\boldsymbol{U}}_C$. Let $h(t) = \left[(1+t)^{\lambda_0} - 1\right] \frac{1}{t}$, then, $h'(t) = \frac{1}{t^2}\left[1 - \frac{(1-\lambda_0)t+1}{(1+t)^{1-\lambda_0}}\right]$. Since $(1+t)^a < 1 + at$ for $a \in (0,1)$ and $t > 0$, $h'(t)$ is negative and $h(t)$ is decreasing for all $t > 0$.

Therefore, it suffices to have

$$\frac{\mu_0}{\mu_1} < \left[\left(1 + \sup_{x,y} \frac{y}{x}\right)^{\lambda_0} - 1\right] \frac{1}{\sup_{x,y} y/x}.$$

For each block $\hat{G}_c$, $c \in \{1, ..., C\}$, the nodes $\hat{V}_c$ of $\hat{G}_c$ are possible to have true memberships of at most C^* communities. Then, the number of edges in $\hat{G}_c$ with edge weight μ_1 would satisfy

$$\binom{g_1}{2} + \binom{g_2}{2} + ... + \binom{g_{C^*}}{2} \text{ with } g_1 + g_2 + ... + g_{C^*} = |\hat{V}_c|$$

where g_c is the number of nodes from the true community G_c^*. Consider a sufficiently large graph and the numbers of nodes change continuously, we have

$$\frac{\binom{g_1}{2} + \binom{g_2}{2} + ... + \binom{g_{C^*}}{2}}{\binom{|\hat{V}_c|}{2}} \geq C^*.$$

Therefore, for $C* \geq 2$, $y/x \leq C^* - 1$ and for $C^* = 1$, $y/x \leq 1$. Hence, the claim is true. □

Proof of Theorem 2.5. It suffices to show the consistency results are guaranteed for spectral clustering in our setting of a continuous stochastic block model. The proof of theorem 3.1 in [133] can be easily extended to a weighted case using continuous versions of Bernstein inequality and Chernoff bounds.

To bound light pairs, $u_{ij} = x_i y_j \mathbf{1}(|x_i y_j| \leq \sqrt{d}/n) + x_j y_i \mathbf{1}(|x_j y_i| \leq \sqrt{d}/n)$, then $|u_{ij}| \leq 2\sqrt{d}/n$, and $x^T W' y$ can be written as

$$\sum_{1 \leq i < j \leq n} w'_{ij} u_{ij}.$$

Then, for zero-mean independent random variables, apply Bernstein inequality,

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{i<j} w'_{ij} u_{ij}\right| \geq c_0\sqrt{d}\right] &\leq 2\exp\left(-\frac{\frac{1}{2}c_0^2 d}{\sum_{i<j}\sigma_{ij}^2 u_{ij}^2 + \frac{1}{3}\frac{2\sqrt{d}}{n}c_0\sqrt{d})}\right) \\
&\leq 2\exp\left(-\frac{\frac{1}{2}c_0^2 d}{\sigma_{\max}^2 \sum_{i<j} u_{ij}^2 + \frac{2c_0}{3}\frac{d}{n}}\right) \\
&\leq 2\exp\left(-\frac{c_0^2}{4 + \frac{4c_0}{3}} n\right).
\end{aligned}$$

In bounding heavy pairs, let $e(I, J)$ be the summation of edge weights in node sets I and

J: $e(I, J) = \sum_{(i,j) \in s(I,J)} w_{ij}$. Define $\mu(I, J) = \mathbb{E}e(I, J), \overline{\mu}(I, J) = p_{\max}|I||J|$. We could obtain continuous versions of Lemma 4.1 and 4.2 in supplementary material of [133].

Using Bernstein inequality:

$$\mathbb{P}\left(\sum_{j=1}^{n} w_{ij} \geq c_1 d\right) \leq \mathbb{P}\left(\sum_{j=1}^{n} w'_{ij} \geq (c_1 - 1)d\right) \leq \exp\left[-\frac{\frac{1}{2}(c_1-1)^2 d^2}{\sum_{j=1}^{n} \sigma_{ij}^2 + \frac{1}{3}(c_1 - 1)d}\right]$$

$$\leq \exp\left[-\frac{\frac{1}{2}(c_1-1)^2 d^2}{n\sigma_{\max}^2 + \frac{1}{3}(c_1 - 1)d}\right] \leq \exp\left[-\frac{\frac{1}{2}(c_1-1)^2 d}{1 + \frac{1}{3}(c_1 - 1)}\right] \leq n^{-\frac{3c_0(c_1-1)^2}{2c_1+4}}$$

We have for $c_0 > 0$, there exists constant $c_1 = c_1(c_0)$ such that with probability at least $1 - n^{-c_0}$, $\sum_{j=1}^{n} w_{ij} \leq c_1 d$.

From Chernoff Bound:

$$\begin{aligned}\mathbb{P}[e(I, J) \geq k\overline{\mu}(I, J)] &= \mathbb{P}\left[\sum_{(i,j)\in s(I,J)} w_{ij} \geq k\bar{\mu}(I, J)\right] \\ &\leq \exp(-\bar{\mu}(I, J)(k \ln k - (k - 1))) \\ &\leq \exp\left[-\frac{1}{2}(k \ln k)\bar{\mu}\right]\end{aligned}$$

the lemma 4.2 is true from exacly the same calculations.

Hence, our claim is true with stated assumptions from Theorem 3.1 of [133]. □

Appendix B: Supplemental for Chapter 3

B.1 Algorithms

We present the detailed algorithm considering the spatial constraint in IGDB extraction as follows. With the input of a set of voxels $\tilde{T}_\lambda$ from Algorithm 4, we refine the voxel sets as pieces of spatially connected voxel clusters which also preserve the high value of objective function. Let G_V be the graph corresponding to the spatial connectivity of voxels, then our goal is to identify connected components $\tilde{T}_{\lambda,1}, ..., \tilde{T}_{\lambda,k_{\tilde{V}}}$ in G_V while they are spatially distinct from each other. We propose the following procedure. Let $\tilde{T}^0_{\lambda,1}, ..., \tilde{T}^0_{\lambda,k_{\tilde{V}}}$ be the connected components at time 0, which is the decomposition of $\tilde{T}_\lambda$ in G_V. We merge the connect components by adding nodes from its shortest path (e.g., shortest path between $\tilde{T}^t_{\lambda,\hat{k}}$ and $\tilde{T}^t_{\lambda,\hat{k}'}$) iteratively until the distance of any two connected components exceed a give threshold. In order to preserve the high density of the refined voxel sets, the nodes should be added in the order of nodes degree (e.g., degrees in $G[\tilde{S}_\lambda, V]$). Hence, we adapt the Dijkstra's algorithm in searching for shortest path in a node weighted graph where the weights are defined as the reciprocal of the node degrees [134]. The computational complexity of the procedure is $O(n^2 + n\log(n))$.

Algorithm 7 IGDB extraction with spatial constraint

```
    Input: $\tilde{T}_\lambda$, $G_V$, $\boldsymbol{\rho} = (\rho_1, ..., \rho_n)$
    Output: $G[\tilde{S}_\lambda, \tilde{T}_\lambda]$
procedure ALGORITHM
    At $t = 0$: $T^0 \leftarrow \tilde{T}_\lambda$
    while $\min_{k_1,k_2}$ distance$(T^i_{k_1}, T^i_{k_2})) \leq$ threshold do
        Decompose $T^t$ to connected components $T^t_1, ..., T^t_{k^t}$ in $G_V$;
        Compute shortest path for each pair of connected components $L_{k,k'}(T^t_k, T^t_{k'})$, $k_1, k_2 = 1, ..., k^t$;
        Merge the two connected components with least weights in shortest paths: $T^{t+1} = T^t \cup L_{\hat{k},\hat{k}'}$ if $\hat{k}, \hat{k}' = \arg\min_{k,k'} L_{k,k'}$;
    end while
    Output $\hat{T}_\lambda \leftarrow T^{t_{\max}}$ ;
end procedure
```

Algorithm 8 Determine tuning parameter λ by likelihood function

```
procedure ALGORITHM
    Given a grid of tuning parameters: $\lambda_1, \lambda_2, ..., \lambda_J$, and a sequence of cutoffs $r_1, r_2, ..., r_R$
    while $\lambda \in \{\lambda_1, ..., \lambda_J\}$ do
        Return the IGDB $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ of $\boldsymbol{W}$ from Algorithm 4
        for $r = r_1$ to $r_R$ do
            calculate the likelihood: $\mathcal{L}_\lambda(\hat{\boldsymbol{\pi}}; \tilde{S}_\lambda, \tilde{T}_\lambda, \boldsymbol{W}(r))$
        end for
        integrate w.r.t. $r$:
      $\mathcal{L}_\lambda(\boldsymbol{W}) = \sum_{i=1}^{R} \mathcal{L}_\lambda(\hat{\boldsymbol{\pi}}; \tilde{S}_\lambda, \tilde{T}_\lambda, \boldsymbol{W}(r_i))g(r_i)$
    end while
    Output $\hat{\lambda}$ and $(\tilde{S}_{\hat{\lambda}}, \tilde{T}_{\hat{\lambda}})$ with maximized $\mathcal{L}_\lambda(\boldsymbol{W})$
end procedure
```

B.2 Permutation test

The Section 3.3.2 provides a way to detect an IGDB and maximize the likelihood function with the given the IGDB. Thus, given a bipartite graph with adjacency matrix $\boldsymbol{A}$, we observe the test statistic as:

$$t_G = \log \frac{m_1(\hat{\boldsymbol{\pi}}; \tilde{S}_{\hat{\lambda}}, \tilde{T}_{\hat{\lambda}}, \boldsymbol{A})}{m_0(\hat{\pi}; \boldsymbol{A})}.$$

Let $\phi(\cdot)$ be the vectorization of a matrix, such that $\phi(\boldsymbol{A})$ is an mn vector. Denote τ as a permutation of mn elements, and P_τ is the corresponding permutation matrix. Let $G_\tau = (U, V, E_\tau)$ is an edge-permuted graph from G. Then, under the IGDB-wise null hypothesis, in which G is observed from a uniform random bipartite graph model, the edge-permuted graph G_τ would be a realization from the same null model. We let $\tau(1), ..., \tau(B)$ be B random permutations. Then, the corresponding edge-permuted adjacency matrices are given by $A_{\tau(1)}, ..., A_{\tau(B)}$ with $\boldsymbol{A}_\tau = \phi^{-1}(P_\tau \phi(\boldsymbol{A}))$. The test statistics associated with edge-permuted adjacency matrices $A_{\tau(1)}, ..., A_{\tau(B)}$ forms a random sample of t_G under null hypothesis, which can be utilized to obtain the empirical distribution of t_G under null hypothesis. We illustrate whole procedure of the permutation test as follows:

Algorithm 9 Implementation of likelihood ratio statistic via permutation tests

```
Input: $G = (U, V, E, \boldsymbol{A})$
Output: $p$-value of the IGDB-wise hypothesis test
procedure ALGORITHM
    calulate the test statistic on $G$ and denote as: $t_0$
    for $b = 1$ to $B$ do
        generate permutation matrix $P_b$ on $mn$ elements
        observe adjacency matrix of edge-permuted graph $G_b$: $\boldsymbol{A}_b = \phi^{-1}(P_b \phi(A))$
        calculate the test statistic on $G_b$ as: $t_b$
    end for
    $p_0$ is the percentile of $t_0$ in $\{t_b\}_{b=1}^{B}$;
    if $p_0$ is smaller than $\tau$ then
        $G[S, T]$ is a significant subgraph of $G$.
    end if
end procedure
```

B.3 Detailed Proofs

Proof of Lemma 3.1. From Bernstein's inequality, for any $S \subseteq U$, $T \subseteq V$, we have

$$\begin{aligned}
\mathbb{P}\left(\left||S||T|^{-1}\sum_{i,j\in G[S,T]}(A_{ij}-\pi)\right| > \gamma-\pi\right) &\leq 2\exp\left(-\frac{(|S||T|)^2(\gamma-\pi)^2}{2[|S||T|+\frac{1}{3}|S||T|(\gamma-\pi)]}\right)\\
&= 2\exp\left(-\left\{\frac{2}{(\gamma-\pi)^2}+\frac{2}{3(\gamma-\pi)}\right\}^{-1}\cdot|S||T|\right)\\
&:= 2\exp\left(-\frac{1}{2}c(\gamma,\pi)\cdot|S||T|\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(|S_\gamma|\geq m_0, |T_\gamma|\geq n_0\right)\\
\leq&\mathbb{P}\left(\cup_{(S,T):|S|\geq m_0,|T|\geq n_0}\left\{\left||S||T|^{-1}\sum_{i,j\in G^r[S,T]}(A_{ij}-\pi)\right|>\gamma-\pi\right\}\right)\\
\leq&\sum_{|S|=m_0}^{m}\sum_{|T|=n_0}^{n}\binom{m}{|S|}\binom{n}{|T|}2\exp\left(-\frac{1}{2}c(\gamma,\pi)\cdot|S||T|\right)\\
\leq&\sum_{|S|=m_0}^{m}\sum_{|T|=n_0}^{n}2\exp\left(-\frac{1}{2}c(\gamma,\pi)\cdot|S||T|+|S|\log m+|T|\log n\right)\\
\leq&\sum_{|S|=m_0}^{m}\sum_{|T|=n_0}^{n}2\exp\left(-\frac{1}{4}c(\gamma,\pi)\cdot|S||T|\right)\\
\leq&2mn\exp\left(-\frac{1}{4}c(\gamma,\pi)\cdot|S||T|\right)
\end{aligned}$$

□

Lemma B.1. *The greedy algorithm will give a* $2(mn)^{|\frac{1}{2}-\lambda|}$*-approximation, i.e.*

$$d_\lambda(S^*_\lambda, T^*_\lambda) \leq 2d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)(mn)^{|\frac{1}{2}-\lambda|}.$$

Proof of Lemma B.1. For a bipartite graph, we assign the edge e_{ij} to either node i or j where $i \in U$ and $j \in V$. Denote the number of edges assigned to node i as $d_U(i)$ and j as $d_V(j)$. The subscript U and V indicate the sets of nodes. Let $d_U^{\max} = \max_i\{d_U(i)\}$ and $d_V^{\max} = \max_j\{d_V(j)\}$. We start with the case $c = |S^*_\lambda|/|T^*_\lambda|$. Let

$$(S^*_{\lambda,c}, T^*_{\lambda,c}) = \underset{S\subset U, T\subset V, |S|/|T|=c}{\arg\max} \; d_\lambda(S,T),$$

Then, since all edges in $E[S^*_{\lambda,c}, T^*_{\lambda,c}]$ will be assigned to node i or j which will be counted in $d_U(i)$ or $d_V(j)$,

$$|E[S^*_{\lambda,c}, T^*_{\lambda,c}]| \leq |S^*_{\lambda,c}|d_U^{\max} + |T^*_{\lambda,c}|d_V^{\max} \tag{B.1}$$

Consider a specific way to assign edges. Let the edges are not assigned at the beginning of this algorithm. A node gets assigned when it is removed in the greedy algorithm. In other words, if a node $i \in U$ is deleted at some iteration, assume the subgraph at this iterationis $G[S', T']$, then after this iteration the nodes set changes from $(S', T') \rightarrow (S'/\{i\}, T')$. The edges assigned to this node i is the number of edges removed at this iteration, which is the degree of node i in this iteration. The same for a node removed from nodes set V. Therefore, since the node i or j is deleted for the smallest degree, we will have the degree of i at this iteration is smaller than the average degree :

$$d_U(i) \leq \frac{|E[S', T']|}{|S'|} \text{ and } d_V(i) \leq \frac{|E[S', T']|}{|T'|}.$$

Hence,

$$\begin{aligned}\min(\sqrt{c}d_U(i), \frac{1}{\sqrt{c}}d_V(j)) \leq \sqrt{d_U(i)d_V(j)} &\leq \frac{|E[S', T']|}{\sqrt{|S'||T'|}} \\ &\leq \frac{|E[S', T']|}{(|S'||T'|)^{\lambda}}(mn)^{\lambda - \frac{1}{2}} \leq d_{\lambda}(\tilde{S}_{\lambda}, \tilde{T}_{\lambda})(mn)^{\lambda - \frac{1}{2}}\end{aligned}$$

If $\sqrt{c}d_U(i) \leq \frac{1}{\sqrt{c}}d_V(j)$, then node i is deleted and deleted edges are all assigned to $d_U(i)$. In this case, from the inequality above, $\sqrt{c}d_U(i) \leq d_{\lambda}(\tilde{S}_{\lambda}, \tilde{T}_{\lambda})(mn)^{\lambda - \frac{1}{2}}$. If j is deleted, $\sqrt{c}d_U(i) > \frac{1}{\sqrt{c}}d_V(j)$, then $\frac{1}{\sqrt{c}}d_V(j) \leq d_{\lambda}(\tilde{S}_{\lambda}, \tilde{T}_{\lambda})(mn)^{\lambda - \frac{1}{2}}$. Therefore,

$$\sqrt{c}d_U^{\max} \leq d_{\lambda}(\tilde{S}_{\lambda}, \tilde{T}_{\lambda})(mn)^{\lambda - \frac{1}{2}} \text{ and } \frac{1}{\sqrt{c}}d_V^{\max} \leq d_{\lambda}(\tilde{S}_{\lambda}, \tilde{T}_{\lambda})(mn)^{\lambda - \frac{1}{2}}. \tag{B.2}$$

Thus,

- **Case 1:** for $\lambda > \frac{1}{2}$,

$$\begin{aligned}d_{\lambda}(S^*_{\lambda,c}, T^*_{\lambda,c}) = \frac{|E[S^*_{\lambda,c}, T^*_{\lambda,c}]|}{(|S^*_{\lambda,c}||T^*_{\lambda,c}|)^{\lambda}} \leq \frac{|E[S^*_{\lambda,c}, T^*_{\lambda,c}]|}{\sqrt{|S^*_{\lambda,c}||T^*_{\lambda,c}|}} &\leq \sqrt{c}d_U^{\max} + \frac{1}{\sqrt{c}}d_V^{\max} \\ &\leq 2d_{\lambda}(\tilde{S}_{\lambda}, \tilde{T}_{\lambda})(mn)^{\lambda - \frac{1}{2}}\end{aligned} \tag{B.3}$$

- **Case 2:** For $\lambda < \frac{1}{2}$,

$$\begin{aligned}
d_\lambda(S^*_{\lambda,c}, T^*_{\lambda,c}) &= \frac{|E[S^*_{\lambda,c}, T^*_{\lambda,c}]|}{(|S^*_{\lambda,c}||T^*_{\lambda,c}|)^\lambda} \\
&\leq \frac{[S^*_{\lambda,c}, T^*_{\lambda,c}]|}{\sqrt{|S^*_{\lambda,c}||T^*_{\lambda,c}|}}(mn)^{\frac{1}{2}-\lambda} \leq \left(\sqrt{c}d_U^{\max} + \frac{1}{\sqrt{c}}d_V^{\max}\right)(mn)^{\frac{1}{2}-\lambda} \\
&\leq 2d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)(mn)^{\frac{1}{2}-\lambda} \qquad \text{(B.4)}
\end{aligned}$$

Since (B.2) doesn't depend on the choice of c, taking maximum of the left-hand side over c for equation (3) and (4) gives the result. □

Proof of Theorem 3.1. Equivalently, we consider to maximize a generalized metric:

$$d_\lambda(S, T) = \frac{|\boldsymbol{W}[S, T]|}{(|S||T|)^\lambda}, \qquad \text{(B.5)}$$

with $\lambda \in (0, 1)$.

Then, from similar argument in Lemma 1:

- **Case 1:** For $1/2 \leq \lambda < 1$, we can see $d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda) \geq 1$ and

$$d_\lambda(S^*_\lambda, T^*_\lambda) = \frac{|E[S^*_\lambda, T^*_\lambda]|}{(|S^*_\lambda||T^*_\lambda|)^\lambda} \leq (|S^*_\lambda||T^*_\lambda|)^{1-\lambda}.$$

Assume there exist $0 < a < 1$ with $d_\lambda(S^*_\lambda, T^*_\lambda) \leq 2d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)(mn)^a$. Firstly, if

$|S^*_\lambda||T^*_\lambda| \leq (mn)^a$, $d_\lambda(S^*_\lambda, T^*_\lambda) \leq (mn)^{(1-\lambda)a}$. If $|S^*_\lambda||T^*_\lambda| > (mn)^a$,

$$d_\lambda(S^*_\lambda, T^*_\lambda) \leq (mn)^{-a(\lambda-\frac{1}{2})} \frac{|E[S^*_\lambda, T^*_\lambda]|}{\sqrt{|S^*_\lambda||T^*_\lambda|}} \leq (mn)^{-a(\lambda-\frac{1}{2})} \cdot 2d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)(mn)^{\lambda-\frac{1}{2}}$$
$$= 2d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)(mn)^{(\lambda-\frac{1}{2})(1-a)}.$$

Let $a = 2\lambda - 1$, we have

$$d_\lambda(S^*_\lambda, T^*_\lambda) \leq 2d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)(mn)^{(1-\lambda)(2\lambda-1)}$$

- **Case 2:** For $0 < \lambda < \frac{1}{4}$,

$$d_\lambda(S^*_\lambda, T^*_\lambda) \leq |E[U, V]| \leq \frac{|E[U, V]|}{(|U||V|)^\lambda} \cdot (mn)^\lambda \leq d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)(mn)^\lambda$$

- **Case 3:** For $\frac{1}{4} < \lambda < \frac{1}{2}$, keep the result in Lemma 1.

□

Proof of Theorem 3.2. For simplicity, we assume $m = n$ and $|S_0| = |T_0| = s_0$. We prove the claim in two parts. In part 1, we show that with high probability, the state space searched by Algorithm (1) includes candidates of (S, T) with misarrangement error smaller than n^ϵ. In part 2, we illustrate that the objective function would favor the (S, T) with misarrangement error smaller than n^ϵ with high probability.

Part 1. We denote $\mathcal{C} = \{(S_1, T_1), (S_2, T_2), ..., (S_{mn-1}, T_{mn-1})\}$ be the sequence of biclusters generated by deleting the smallest-degree nodes according to the decision rule in Algorithm 1 when $c = c_0 = \frac{|S_0|}{|T_0|}$ (i.e., under our assumption $c_0 = 1$). Then, we target to prove with high probability $(S_0, T_0) \in \mathcal{C}$.

At the stage k, within the subgraph $G[S_k, T_k]$, the node degrees satisfy:

$$\begin{aligned}
&\text{For } i \in S_k \cap S_0: \quad \deg_X(i) = \sum_{j=1}^{|T_k \cap T_0|} \xi_{ij}^1 + \sum_{j=|T_k\cap T_0|+1}^{|T_k|} \xi_{ij}^0, \\
&\text{For } i' \in S_k \cap S_0^c: \quad \deg_X(i') = \sum_{j=1}^{|T_k|} \xi_{i'j}^0, \\
&\text{For } j \in T_k \cap T_0: \quad \deg_Y(j) = \sum_{i=1}^{|S_k \cap S_0|} \xi_{ij}^1 + \sum_{i=|S_k\cap S_0|+1}^{|S_k|} \xi_{ij}^0, \\
&\text{For } j' \in T_k \cap T_0^c: \quad \deg_Y(j') = \sum_{i=1}^{|S_k|} \xi_{ij'}^0,
\end{aligned}$$

where $\xi_{ij}^1 \sim$ Bernouli(π_1), $\xi_{ij}^0 \sim$ Bernouli(π_0).

From chernoff bound, for $i \in S_k \cap S_0$ and $\delta \in (0, 1)$

$$\mathbb{P}(\deg_X(i) \leq (1-\delta) \left(|T_k \cap T_0|\pi_1 + |T_k \cap T_0^c|\pi_0\right)) \leq \exp\left[-\frac{\delta^2}{2}\left(|T_k \cap T_0|\pi_1 + |T_k \cap T_0^c|\pi_0\right)\right],$$

and for $i' \in S_k \cap S_0^c$,

$$\mathbb{P}(\deg_X(i') \geq (1+\delta)|T_k|\pi_0) \leq \exp\left[-\frac{\delta^2}{2+\delta^2}|T_k|\pi_0\right].$$

Hence, when $\delta > |T_k|^{-\frac{1}{2}+\frac{\epsilon}{2}}$ and $\frac{|T_k \cap T_0|}{|T_k|} \geq 2\frac{\delta}{1-\delta}\frac{\pi_0}{\pi_1 - \pi_0}$, there exist $\epsilon_0(\delta) > 0$,

$$\begin{aligned}
\mathbb{P}(\deg_X(i) - \deg_X(i') \geq \epsilon_0(\delta)) \geq 1 - \exp&\left[-\frac{\delta^2}{2}\left(|T_k \cap T_0|\pi_1 + |T_k \cap T_0^c|\pi_0\right)\right] \\
&- \exp\left[-\frac{\delta^2}{2+\delta^2}|T_k|\pi_0\right] \to 0.
\end{aligned}$$

Similarly, when $\delta > |S_k|^{-\frac{1}{2}+\frac{\epsilon}{2}}$ and $\frac{|S_k \cap S_0|}{|S_k|} \geq 2\frac{\delta}{1-\delta}\frac{\pi_0}{\pi_1-\pi_0}$, there exist $\epsilon_0(\delta) > 0$,

$$\mathbb{P}(\text{deg}_Y(j) - \text{deg}_Y(j') \geq \epsilon_0(\delta)) \geq 1 - \exp\left[-\frac{\delta^2}{2}(|S_k \cap S_0|\pi_1 + (|S_k \cap S_0^c|)\pi_0)\right] \\ - \exp\left[-\frac{\delta^2}{2+\delta^2}|S_k|\pi_0\right] \to 0.$$

In this case, we have

$$\begin{aligned}
&\mathbb{P}(\{\text{deg}_X(i) - \text{deg}_X(i') \geq \epsilon_0(\delta) \text{ and } \text{deg}_Y(j) - \text{deg}_Y(j') \geq \epsilon_0(\delta)\}) \\
&\geq \mathbb{P}(\text{deg}_X(i) - \text{deg}_X(i') \geq \epsilon_0(\delta)) + \mathbb{P}(\text{deg}_Y(j) - \text{deg}_Y(j') \geq \epsilon_0(\delta)) - 1. \\
&\geq 1 - \exp\left[-\frac{\delta^2}{2}(|T_k \cap T_0|\pi_1 + (|T_k \cap T_0^c|)\pi_0)\right] - \exp\left[-\frac{\delta^2}{2+\delta^2}|T_k|\pi_0\right] \\
&\qquad - \exp\left[-\frac{\delta^2}{2}(|S_k \cap S_0^c|\pi_1 + (|S_k \cap S_0^c|)\pi_0)\right] - \exp\left[-\frac{\delta^2}{2+\delta^2}|S_k|\pi_0\right] \\
&:= p_1
\end{aligned}$$

At stage k of, the event of deleting a node outside S_0 and T_0 is:

$$\cap_{i \in S_k \cap S_0} \cap_{i' \in S_k \cap S_0^c} \cap_{j \in T_k \cap T_0} \cap_{j' \in T_k \cap T_0^c} \{\text{deg}_X(i) \geq \text{deg}_X(i') \text{ and } \text{deg}_Y(j) \geq \text{deg}_Y(j')\}$$

Then,

$$\begin{aligned}
&\mathbb{P}\left(\cap_{i \in S_k \cap S_0} \cap_{i' \in S_k \cap S_0^c} \cap_{j \in T_k \cap T_0} \cap_{j' \in T_k \cap T_0^c} \{\text{deg}_X(i) \geq \text{deg}_X(i') \text{ and } \text{deg}_Y(j) \geq \text{deg}_Y(j')\}\right) \\
\geq& 1 - \sum_{i \in S_k \cap S_0} \sum_{i' \in S_k \cap S_0^c} \sum_{j \in T_k \cap T_0} \sum_{j' \in T_k \cap T_0^c} \mathbb{P}(\text{deg}_X(i) < \text{deg}_X(i') \text{ or } \text{deg}_Y(j) < \text{deg}_Y(j')) \\
\geq& 1 - |S_k \cap S_0||S_k \cap S_0^c||T_k \cap T_0||T_k \cap T_0^c|p_1
\end{aligned}$$

Using similar argument,

$$\begin{aligned}
&\mathbb{P}((S_0, T_0) \in \mathcal{C}) \\
=&\mathbb{P}\left(\cap_{k=1}^{n^2-s_0^2} \cap_{i\in S_k\cap S_0} \cap_{i'\in S_k\cap S_0^c} \cap_{j\in T_k\cap T_0} \cap_{j'\in T_k\cap T_0^c}\{\deg_X(i) \geq \deg_X(i') \text{ and } \deg_Y(j) \geq \deg_Y(j')\}\right) \\
\geq& 1-(n^2-s_0^2)\mathbb{P}\left(\cap_{i\in S_k\cap S_0} \cap_{i'\in S_k\cap S_0^c} \cap_{j\in T_k\cap T_0} \cap_{j'\in T_k\cap T_0^c}\{\deg_X(i) \geq \deg_X(i') \text{ and } \deg_Y(j) \geq \deg_Y(j')\right. \\
\to& 1 \text{ as } n \to \infty,
\end{aligned}$$

when $s_0 = O(n^{\frac{1}{2}+\epsilon})$.

Part 2. We first show the population version that the true subgraph $G[S_0, T_0]$ is the global optimal of the objective function under expectation.

Denote $d_\lambda(S,T) = \|\boldsymbol{A}[S,T]\|_1 - \lambda|S||T|$ be the objective function under tuning parameter λ. Let $\boldsymbol{\Delta}(S,T)$ be the matrix of indicator variables: $\{\boldsymbol{\Delta}(S,T)\}_{ij} = I(e_{ij} \in G[S,T])$, and $\boldsymbol{\Delta}_0 \equiv \boldsymbol{\Delta}(S_0,T_0)$ corresponds to the true IGDB.

Then, $\mathbb{E}\boldsymbol{A} = \boldsymbol{P} = \pi_1\boldsymbol{\Delta}_0 + \pi_0(\mathbf{1}_n\mathbf{1}_m^T - \boldsymbol{\Delta}_0)$, and

$$\begin{aligned}
\mathbb{E}d_\lambda(S,T) &= \|\left\{\pi_1\boldsymbol{\Delta}_0 + \pi_0(\mathbf{1}_n\mathbf{1}_m^T - \boldsymbol{\Delta}_0)\right\} \circ \boldsymbol{\Delta}(S,T)\|_1 - \lambda\|\boldsymbol{\Delta}(S,T)\|_1 \\
&= \pi_1\|\boldsymbol{\Delta}_0 \circ \boldsymbol{\Delta}(S,T)\|_1 + \pi_0\|(\mathbf{1}_n\mathbf{1}_m^T - \boldsymbol{\Delta}_0) \circ \boldsymbol{\Delta}(S,T)\|_1 - \lambda\|\boldsymbol{\Delta}(S,T)\|_1 \\
&= (\pi_1 - \lambda)\|\boldsymbol{\Delta}_0 \circ \boldsymbol{\Delta}(S,T)\|_1 + (\pi_0 - \lambda)\|(\mathbf{1}_n\mathbf{1}_m^T - \boldsymbol{\Delta}_0) \circ \boldsymbol{\Delta}(S,T)\|_1.
\end{aligned} \tag{B.6}$$

Hence, for $\lambda \in (\pi_0, \pi_1)$, $S = S_0$ and $T = T_0$ is the unique maximizer of $\mathbb{E}d_\lambda(S,T)$.

We then consider to select from the sequence of subgraphs $\mathcal{C}$ by objective function

$d_\lambda(S;T)$. From Chernouff bounds,

$$\mathbb{P}(\|\boldsymbol{A} \circ \boldsymbol{\Delta}(S,T)\|_1 \geq (1+\delta)\|\boldsymbol{P} \circ \boldsymbol{\Delta}(S,T)\|_1) \leq \exp\left[-\frac{\delta^2}{2+\delta^2}\|\boldsymbol{P} \circ \boldsymbol{\Delta}(S,T)\|_1\right]$$

and

$$\mathbb{P}(\|\boldsymbol{A} \circ \boldsymbol{\Delta}_0\|_1 \leq (1-\delta)\left\{\pi_1\|\boldsymbol{\Delta}_0\|_1\right\}) \leq \exp\left[-\frac{\delta^2}{2}\left\{\pi_1\|\boldsymbol{\Delta}_0\|_1\right\}\right].$$

Then,

$$\begin{aligned}&\mathbb{P}(\|\boldsymbol{A} \circ \boldsymbol{\Delta}(S,T)\|_1 - \|\boldsymbol{A} \circ \boldsymbol{\Delta}_0\|_1 \leq (1+\delta)\|\boldsymbol{P} \circ \boldsymbol{\Delta}(S,T)\|_1 - (1-\delta)\pi_1\|\boldsymbol{\Delta}_0\|_1)\\ \geq& 1 - \exp\left[-\frac{\delta^2}{2+\delta^2}\|\boldsymbol{P} \circ \boldsymbol{\Delta}(S,T)\|_1\right] - \exp\left[-\frac{\delta^2}{2}\left\{\pi_1\|\boldsymbol{\Delta}_0\|_1\right\}\right] := 1-p.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}(d_\lambda(S,T) - d_\lambda(S_0,T_0) \leq \mathbb{E}(d_\lambda(S,T)) - \mathbb{E}(d_\lambda(S_0,T_0))\\ &\qquad\qquad + \delta(\|\boldsymbol{P} \circ \boldsymbol{\Delta}(S,T)\|_1 + \pi_1\|\boldsymbol{\Delta}_0\|_1))\\ \geq& 1-p. \end{aligned} \tag{B.7}$$

Based on the proof of part 1, we only need to discuss the sets $\{(S,T) : S \supset S_0, T \supset T_0\}$ and $\{(S,T) : S \subset S_0, T \subset T_0\}$. For $S_0 \subset S$ and $T_0 \subset T$, assume $|S/S_0| + |T/T_0| \geq (|S_0| \wedge |T_0|)^\epsilon$. Then we observe the largest value of $\mathbb{E}(d_\lambda(S,T)) - \mathbb{E}(d_\lambda(S_0,T_0))$ when

$(|S|, |T|) = (s_0 + s_0^{\epsilon}/2, s_0 + s_0^{\epsilon}/2)$. In this case,

$$\mathbb{E}(d_\lambda(S,T)) - \mathbb{E}(d_\lambda(S_0,T_0)) = (\pi_0 - \lambda)\left(s_0^{1+\epsilon} - \frac{1}{4}s_0^{2\epsilon}\right),$$

while

$$\|\boldsymbol{P} \circ \boldsymbol{\Delta}(S,T)\|_1 + \pi_1\|\boldsymbol{\Delta}_0\|_1 = 2\pi_1 s_0^2 + \pi_0\left(s_0^{1+\epsilon} + \frac{1}{4}s_0^{2\epsilon}\right).$$

Hence, from (B.7), when δ is order $s_0^{-1+\frac{\epsilon}{2}}$ and $\lambda > \pi_0$, $d_\lambda(S_0,T_0) > d_\lambda(S,T)$ for any $S \supset S_0$ and $T \supset T_0$ with high probability.

On the other hand, when $S \subset S_0, T \subset T_0$, for simplicity, we assume $m = n$ and $|S_0| = |T_0| = s_0$, then the largest value of $\mathbb{E}(d_\lambda(S,T)) - \mathbb{E}(d_\lambda(S_0,T_0))$ is observed when $(|S|, |T|) = (s_0 - s_0^{\epsilon}/2, s_0 - s_0^{\epsilon}/2)$. In this case,

$$\mathbb{E}(d_\lambda(S,T)) - \mathbb{E}(d_\lambda(S_0,T_0)) = -(\pi_1 - \lambda)\left(s_0^{1+\epsilon} - \frac{1}{4}s_0^{2\epsilon}\right),$$

and

$$\|\boldsymbol{P} \circ \boldsymbol{\Delta}(S,T)\|_1 + \pi_1\|\boldsymbol{\Delta}_0\|_1 = \pi_1\left(2s_0^2 - s_0^{1+\epsilon} + \frac{1}{4}s_0^{2\epsilon}\right).$$

Similarly, when δ is order $s_0^{-1+\frac{\epsilon}{2}}$ and $\lambda < \pi_1$, $d_\lambda(S_0,T_0) > d_\lambda(S,T)$ for any $S \subset S_0$ and $T \subset T_0$ with high probability.

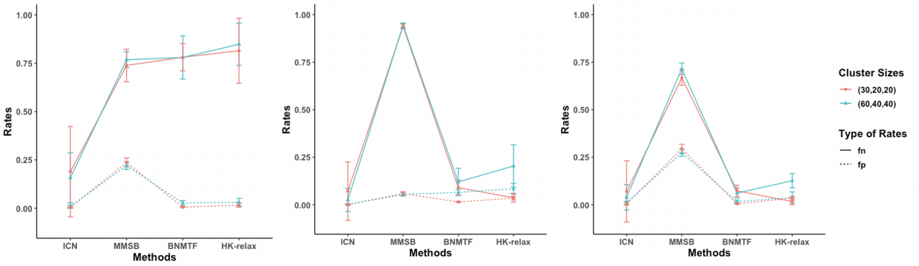
□

Appendix C: Supplemental for Chapter 4

C.1 Additional Simulation Results

C.1.1 Illustrative plots of simulation results

To conveniently display the difference of FPR and FNR from the ICN and competing methods, we draw the plots indicating the mean values and SDs as error bars as follows:



(a) (b) (c)

Figure C.1: Plots of results in Table 1 of main context. Estimated edge-level FPR, FNR and standard errors for $C_0 = 3$ and $n = 200$ based on different detection methods. The FPR and FNR are calculated separately for (a) between-community edges, (b) within-community edges, and (c) overall edges.

C.1.2 Additional simulation settings

We include additional simulation results from different settings. Table C.1 shows the edge-level FPR and FNR for the synthetic data with two positively interconnected communities under generating mechanism described in Simulation Results in Chapter 4. Two non-trivial communities are considered in this case, such that the community sizes and the number of singleton nodes are $(60, 40; 100)$ and $(30, 20; 150)$, respectively. The matrix A is generated from f_1, f_2 and f_0 with mean values 0.8, 0.5 and 0. 40% of edges between two communities are true connecting edges.

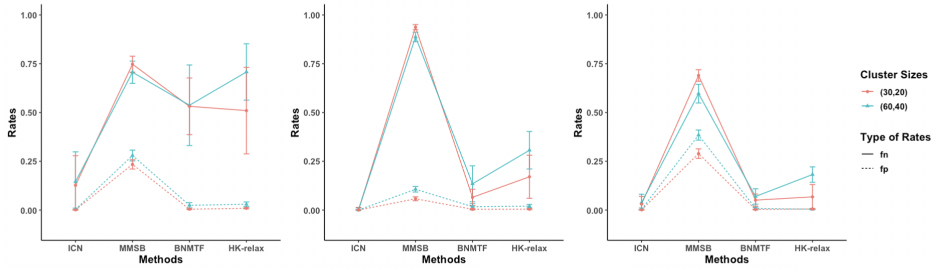
The corresponding figure is displayed as follows:

C.1.3 Simulation results for community-level inference

We also evaluate the power of testing interconnected community from Algorithm 5 in Step 2 of our method in Chapter 4. Despite the testing of interconnection is not the ultimate goal, it is the backbone of our method and critically affects the accuracy of the identification for connecting edges. In addition to the experimental settings described

Table C.1: Estimated edge-level FPR, FNR and standard errors for $C_0 = 2$ and $n = 200$ with positive interconnection based on different detection methods. The FPR and FNR are calculated separately for between-community edges, within-community edges, and overall edges.

			ICN	MMSB	BNMTF	HK-relax
(60, 40)	Between	FPR	0.004 (0.006)	0.280 (0.028)	0.024 (0.013)	0.030 (0.012)
		FNR	0.146 (0.153)	0.706 (0.057)	0.537 (0.207)	0.708 (0.145)
	Within	FPR	0.001 (0.002)	0.107 (0.014)	0.017 (0.016)	0.020 (0.009)
		FNR	0.002 (0.009)	0.887 (0.024)	0.134 (0.092)	0.307 (0.096)
	Overall	FPR	0.004 (0.006)	0.384 (0.026)	0.008 (0.007)	0.004 (0.002)
		FNR	0.038 (0.043)	0.597 (0.048)	0.070 (0.040)	0.182 (0.039)
(30, 20)	Between	FPR	0.001 (0.002)	0.234 (0.024)	0.004 (0.002)	0.009 (0.003)
		FNR	0.127 (0.152)	0.748 (0.041)	0.532 (0.145)	0.510 (0.222)
	Within	FPR	0.001 (0.001)	0.058 (0.010)	0.004 (0.003)	0.004 (0.003)
		FNR	0.001 (0.012)	0.937 (0.013)	0.065 (0.042)	0.171 (0.110)
	Overall	FPR	0.001 (0.002)	0.289 (0.025)	0.003 (0.002)	0.006 (0.003)
		FNR	0.031 (0.038)	0.6898 (0.030)	0.051 (0.031)	0.068 (0.063)



(a) (b) (c)

Figure C.2: Plots of results in Table C.1. Estimated edge-level FPR, FNR and standard errors for $C_0 = 2$ and $n = 200$ with positive interconnection based on different detection methods. The FPR and FNR are calculated separately for (a) between-community edges, (b) within-community edges, and (c) overall edges.

in Simulation Results of Chapter 4 and A.1.2 of the Appendix, we further consider an ICN structure with two negatively interconnected communities, where the matrix A is generated based on f_1, f_2 and f_0 from normal with mean 0.8, -0.6, 0 and standard deviation 0.1.

The results are summarized in Table C.1. We evaluate the accuracy via community-level false negative rate (FNR) and false positive rate (FPR) for KL test among replications. Note, the community-level FPR is not calculated for the settings with $C_0 = 2$ since we assume that the two communities are connected.

In general, we have a high power to detect the truely connected communities with a well-controlled type I error. Both FPR and FNR decrease as the sample size n increases. The networks with two positively interconnected communities have extremely small FNRs, which is signficantly higher than communities with negative interconnections. A possible reason is that the strength of negative connections is constrained by the positive semi-definiteness of the covariance matrix.

Table C.2: Estimated community-level FPR, FNR and standard errors for testing interconnections using Algorithm 1 in main context. The true correlation matrices are considered to have three structures with varied number of non-singleton communities: $C_0 = 3$ with both positive and negative interconnections, $C_0 = 2$ with positive, and $C_0 = 2$ with negative interconnections. Each structure is constructed according to two different cluster sizes. The random samples are generated from the multivariate normal distribution with specified true correlation matrices of varied sample sizes: 100, 200 and 1000.

			$n = 100$	$n = 200$	$n = 1000$
$C_0 = 3$	(60, 40, 40)	FPR	0.0660 (0.0111)	0.0160 (0.0056)	0.0020 (0.0020)
		FNR	0.0220 (0.0066)	0.0090 (0.0042)	0.0100 (0.0044)
	(30, 20, 20)	FPR	0.1140 (0.0142)	0.0760 (0.0119)	0.0000 (0.0000)
		FNR	0.0800 (0.0121)	0.0270 (0.0070)	0.0170 (0.0058)
$C_0 = 2$ (positive)	(60, 40)	FNR	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
	(30, 20)	FNR	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
$C_0 = 2$ (negative)	(60, 40)	FNR	0.0200 (0.0020)	0.0240 (0.0068)	0.0220 (0.0066)
	(30, 20)	FNR	0.0440 (0.0092)	0.0480 (0.0095)	0.0340 (0.0081)

C.1.4 Simulation results with large networks

Furthermore, to illustrate the performance of ICN in a data with a similar number of nodes as real data example, we set $p = 10,000$ nodes, and three communities with sizes: $(2000, 1500, 1500)$ and $(1500, 1000, 1000)$. All other settings are the same as Simulation Results in Chapter 4. We replicate the experiments for 100 times. Since the competing methods under these settings suffer from computational burdens (e.g., MMSB needs more than 6 hours for real data with similar number of nodes), we only reported the results from ICN in the following Table C.1.4.

Table C.3: Estimated edge-level FPR, FNR and standard errors for $C_0 = 3$, $p = 10,000$ and $n = 200$ with positive interconnection based on different detection methods. The FPR and FNR are calculated separately for between-community edges, within-community edges, and overall edges.

		FPR	FNR
(2000, 1500, 1500)	Between	0.0105 (0.0119)	0.1861 (0.2306)
	Within	0.0144 (0.0256)	0.0605 (0.0693)
	Overall	0.0244 (0.0359)	0.0724 (0.0949)
(1500, 1000, 1000)	Between	0.0052 (0.0053)	0.1837 (0.2333)
	Within	0.0055 (0.0113)	0.0614 (0.0574)
	Overall	0.0098 (0.0151)	0.0688 (0.0880)

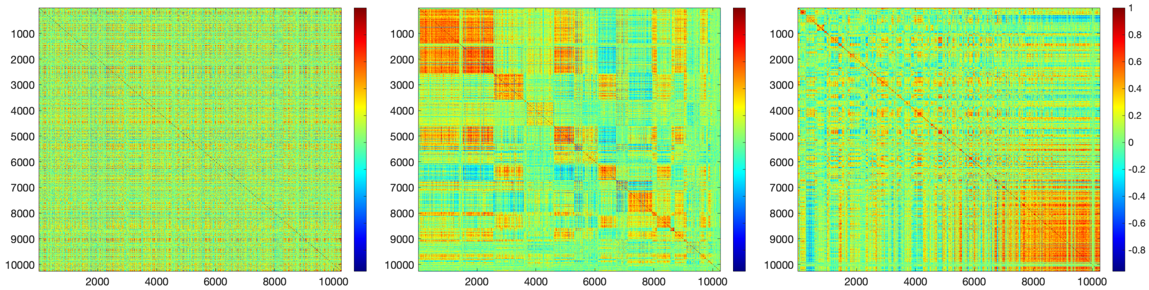
C.2 Addition Data Results

C.2.1 Data results from competing overlapping community methods

We implement the competing methods on the real data example, and the detection results are displayed in the following Figure C.3. We set 17 communities for MMSB method. The results from MMSB yield a balancing partition of all genes into 17 communities, which does not lead to any highly connected communities (i.e., Figure C.3 (a)). BNMTF can identify several blocks, however, it seems to false positively assign week correlations into communities yield false positive results. The HK-relax method detects a number of too small communities (e.g., ranging from a few nodes to a few hundred of nodes), which does not effectively reveal the latent community structure.

C.2.2 Pathway analysis

We include the demonstrative results from pathway analysis in following Figure C.4. We identified pathways enriched with the union of genes in each pair of modules



(a) MMSB (b) BNMTF (c) HK-relax

Figure C.3: The results of real data example from competing methods

(e.g. $\hat{V}_1 \cup \hat{V}_2$ for community 1 and 2, $\hat{V}_7 \cup \hat{V}_8$ community 7 and 8 in leukemia example) using both Fisher's exact test p-values and number of overlapping genes as cutoffs. Then, for each community pair (e.g., community 1 and 2), we conducted pathway analysis separately on three parts: the unique parts of each module (e.g., $\hat{V}_1 \setminus \hat{V}_1^*$ for community 1, e.g., $\hat{V}_2 \setminus \hat{V}_2^*$ for community 2), and the interconnecting part (e.g., $\hat{V}_1^* \cup \hat{V}_2^*$). We display $-\log_{10}$(p-value) of the Fisher's exact test in three parts of each community pair for the top 10 KEGG and Reactome pathways sorted by the p-value from the union set in Figure C.4 (a)(b). We further show the topology plot of MAPK signaling pathway for KEGG with genes from community 1 and 2 highlighted in two different colors in Figure C.4 (c).

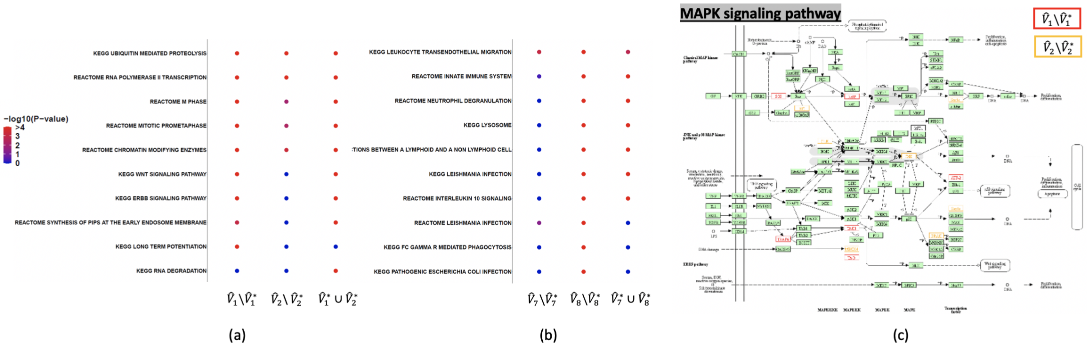


Figure C.4: Pathway enrichment patterns of the three parts identified by ICN for top 10 KEGG and Reactome pathways. (a) Results for community 1-2 pair; (b) Results for community 7-8 pair. Y-axis indicates the pathways and the color reflects the significance level of Fisher's exact tests with red indicating more significant, blue indicating not significant. (c) Topology plot of selected MAPK signaling pathway with the genes from community 1 and 2 highlighted by two different colors.

Bibliography

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

[2] 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.

[3] Martin Rosvall and Carl T Bergstrom. Maps of information flow reveal community structure in complex networks. *arXiv preprint physics.soc-ph/0707.0609*, 2007.

[4] Imola K Fodor. A survey of dimension reduction techniques. Technical report, Citeseer, 2002.

[5] Hujun Yin. Nonlinear dimensionality reduction and data visualization: a review. *International Journal of Automation and Computing*, 4(3):294–303, 2007.

[6] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[7] JingYuan Liu, Wei Zhong, and RunZe Li. A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58(10):1–22, 2015.

[8] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[9] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67 (2):301–320, 2005.

[10] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[11] Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

[12] Fan Li and Nancy R Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American statistical association*, 105(491):1202–1214, 2010.

[13] 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)*, 57(1):289–300, 1995.

[14] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[15] Bradley Efron. Local false discovery rates, 2005.

[16] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[17] Peter J Bickel, Elizaveta Levina, et al. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[18] Peter J Bickel, Elizaveta Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[19] Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

[20] T Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[21] Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

[22] John D Storey et al. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[23] Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[24] Steven D Forman, Jonathan D Cohen, Mark Fitzgerald, William F Eddy, Mark A Mintun, and Douglas C Noll. Improved assessment of significant activation in functional magnetic resonance imaging (fmri): Use of a cluster-size threshold. *Magnetic Resonance in medicine*, 33(5):636–647, 1995.

[25] Satoru Hayasaka and Thomas E Nichols. Validating cluster size inference: random field and permutation methods. *Neuroimage*, 20(4):2343–2356, 2003.

[26] Satoru Hayasaka and Thomas E Nichols. Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage*, 23(1):54–63, 2004.

[27] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[28] Hui Zhang, Thomas E Nichols, and Timothy D Johnson. Cluster mass inference via random field theory. *Neuroimage*, 44(1):51–61, 2009.

[29] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1): 1–25, 2002.

[30] Andrew Zalesky, Alex Fornito, and Edward T Bullmore. Network-based statistic: identifying differences in brain networks. *Neuroimage*, 53(4):1197–1207, 2010.

[31] Thomas E Nichols. Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage*, 62(2):811–815, 2012.

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

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

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

[35] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[36] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[37] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[38] Qiong Wu, Tianzhou Ma, Qingzhi Liu, Donald K Milton, Yuan Zhang, and Shuo Chen. Icn: extracting interconnected communities in gene co-expression networks. *Bioinformatics*, 2021.

[39] Alex Fornito, Andrew Zalesky, Christos Pantelis, and Edward T Bullmore. Schizophrenia, neuroimaging and connectomics. *Neuroimage*, 62(4):2296–2314, 2012.

[40] Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

[41] Xuefei Cao, Björn Sandstede, and Xi Luo. A functional data method for causal dynamic network modeling of task-related fmri. *Frontiers in neuroscience*, 13, 2019.

[42] Gang Chen, Paul-Christian Bürkner, Paul A Taylor, Zhihao Li, Lijun Yin, Daniel R Glen, Joshua Kinnison, Robert W Cox, and Luiz Pessoa. An integrative bayesian approach to matrix-based analysis in neuroimaging. *Human brain mapping*, 2019.

[43] Edward T Bullmore and Danielle S Bassett. Brain graphs: graphical models of the human brain connectome. *Annual review of clinical psychology*, 7:113–140, 2011.

[44] F DuBois Bowman, Lijun Zhang, Gordana Derado, and Shuo Chen. Determining functional connectivity using fmri data with diffusion-based anatomical weighting. *NeuroImage*, 62(3):1769–1779, 2012.

[45] Shuo Chen, F DuBois Bowman, and Yishi Xing. Detecting and testing altered brain connectivity networks with k-partite network topology. *Computational Statistics & Data Analysis*, 141:109–122, 2020.

[46] Joshua T Vogelstein, William Gray Roncal, R Jacob Vogelstein, and Carey E Priebe. Graph classification using signal-subgraphs: Applications in statistical connectomics. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1539–1551, 2012.

[47] Daniele Durante, David B Dunson, et al. Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, 13(1):29–58, 2018.

[48] Debarghya Ghoshdastidar and Ulrike von Luxburg. Practical methods for graph two-sample testing. In *Advances in Neural Information Processing Systems*, pages 3019–3028, 2018.

[49] Cedric E Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, Eric D Kolaczyk, et al. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2):725–750, 2017.

[50] Ixavier A Higgins, Suprateek Kundu, Ki Sueng Choi, Helen S Mayberg, and Ying Guo. A difference degree test for comparing brain networks. *Human brain mapping*, 40(15):4518–4536, 2019.

[51] Suprateek Kundu, Jin Ming, Jordan Pierce, Jennifer McDowell, and Ying Guo. Estimating dynamic brain functional networks using multi-subject fmri data. *NeuroImage*, 183:635–649, 2018.

[52] Joshua Lukemire, Suprateek Kundu, Giuseppe Pagnoni, and Ying Guo. Bayesian joint modeling of multiple brain functional networks. *arXiv preprint arXiv:1708.02123*, 2017.

[53] Amanda F Mejia, Mary Beth Nebel, Yikai Wang, Brian S Caffo, and Ying Guo. Template independent component analysis: Targeted and reliable estimation of subject-level brain networks using big data population priors. *Journal of the American Statistical Association*, pages 1–27, 2019.

[54] Sean L Simpson, Mohsen Bahrami, and Paul J Laurienti. A mixed-modeling framework for analyzing multitask whole-brain network data. *Network Neuroscience*, 3(2):307–324, 2019.

[55] Yin Xia and Lexin Li. Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics*, 73(3):780–791, 2017.

[56] Yin Xia and Lexin Li. Matrix graph hypothesis testing and application in brain connectivity alternation detection. *Statistica Sinica, to appear*, 2018.

[57] Victor E Lee, Ning Ruan, Ruoming Jin, and Charu Aggarwal. A survey of algorithms for dense subgraph discovery. In *Managing and Mining Graph Data*, pages 303–336. Springer, 2010.

[58] Andrew V Goldberg. *Finding a maximum density subgraph*. University of California Berkeley, CA, 1984.

[59] Yuichi Asahiro, Kazuo Iwama, Hisao Tamaki, and Takeshi Tokuyama. Greedily finding a dense subgraph. In *Scandinavian Workshop on Algorithm Theory*, pages 136–148. Springer, 1996.

[60] Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 84–95. Springer, 2000.

[61] Charalampos Tsourakakis, Francesco Bonchi, Aristides Gionis, Francesco Gullo, and Maria Tsiarli. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 104–112. ACM, 2013.

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

[63] Yunpeng Zhao, Elizaveta Levina, Ji Zhu, et al. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

[64] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783, 2017.

[65] Tian Ge, Jianfeng Feng, Derrek P Hibar, Paul M Thompson, and Thomas E Nichols. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *Neuroimage*, 63(2):858–873, 2012.

[66] Andrew Zalesky, Alex Fornito, and Edward T Bullmore. Network-based statistic: identifying differences in brain networks. *Neuroimage*, 53(4):1197–1207, 2010.

[67] Shuo Chen, Yishi Xing, Jian Kang, Peter Kochunov, and L Elliot Hong. Bayesian modeling of dependence in brain connectivity data. *Biostatistics*, 21(2):269–286, 2020.

[68] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[69] Shuo Chen, Jian Kang, Yishi Xing, Yunpeng Zhao, and Donald K Milton. Estimating large covariance matrix with network topology for high-dimensional biomedical data. *Computational Statistics & Data Analysis*, 127:82–95, 2018.

[70] Shuo Chen, Jian Kang, Yishi Xing, and Guoqing Wang. A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks. *Human brain mapping*, 36(12):5196–5206, 2015.

[71] Bhim M Adhikari, L Elliot Hong, Hemalatha Sampath, Joshua Chiappelli, Neda Jahanshad, Paul M Thompson, Laura M Rowland, Vince D Calhoun, Xiaoming Du, Shuo Chen, et al. Functional network connectivity impairments and core cognitive deficits in schizophrenia. *Human brain mapping*, 40(16):4593–4605, 2019.

[72] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R Laird, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, 26(8):3508–3526, 2016.

[73] Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

[74] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[75] Siyi Li, Na Hu, Wenjing Zhang, Bo Tao, Jing Dai, Yao Gong, Youguo Tan, Duanfang Cai, and Su Lui. Dysconnectivity of multiple brain networks in schizophrenia: a meta-analysis of resting-state functional connectivity. *Frontiers in psychiatry*, 10:482, 2019.

[76] Lucina Q Uddin. Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, 16(1):55–61, 2015.

[77] Lena Palaniyappan, Thomas P White, and Peter F Liddle. The concept of salience network dysfunction in schizophrenia: from neuroimaging observations to therapeutic opportunities. *Current topics in medicinal chemistry*, 12(21):2324–2338, 2012.

[78] Tian Ge, Gunter Schumann, and Jianfeng Feng. Imaging genetics—towards discovery neuroscience. *Quantitative Biology*, 1(4):227–245, 2013.

[79] Jingyu Liu and Vince D Calhoun. A review of multivariate analyses in imaging genetics. *Frontiers in neuroinformatics*, 8:29, 2014.

[80] Hongtu Zhu, Zakaria Khondker, Zhaohua Lu, and Joseph G Ibrahim. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, 109(507):977–990, 2014.

[81] Farouk S Nathoo, Linglong Kong, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. A review of statistical methods in imaging genetics. *Canadian Journal of Statistics*, 47(1):108–131, 2019.

[82] Meiyan Huang, Thomas Nichols, Chao Huang, Yang Yu, Zhaohua Lu, Rebecca C Knickmeyer, Qianjin Feng, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Fvgwas: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage*, 118:613–627, 2015.

[83] Chao Huang, Paul Thompson, Yalin Wang, Yang Yu, Jingwen Zhang, Dehan Kong, Rivka R Colen, Rebecca C Knickmeyer, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Fgwas: Functional genome wide association analysis. *NeuroImage*, 159:107–121, 2017.

[84] Derrek P Hibar, Jason L Stein, Omid Kohannim, Neda Jahanshad, Andrew J Saykin, Li Shen, Sungeun Kim, Nathan Pankratz, Tatiana Foroud, Matthew J Huentelman, et al. Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage*, 56(4):1875–1891, 2011.

[85] Jason L Stein, Xue Hua, Suh Lee, April J Ho, Alex D Leow, Arthur W Toga, Andrew J Saykin, Li Shen, Tatiana Foroud, Nathan Pankratz, et al. Voxelwise genome-wide association study (vgwas). *neuroimage*, 53(3):1160–1174, 2010.

[86] Tian Ge, Thomas E Nichols, Debashis Ghosh, Elizabeth C Mormino, Jordan W Smoller, Mert R Sabuncu, Alzheimer's Disease Neuroimaging Initiative, et al. A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application. *NeuroImage*, 109:505–514, 2015.

[87] Hua Wang, Feiping Nie, Heng Huang, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2012.

[88] Hua Wang, Feiping Nie, Heng Huang, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012.

[89] Maria Vounou, Thomas E Nichols, Giovanni Montana, Alzheimer's Disease Neuroimaging Initiative, et al. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage*, 53(3):1147–1159, 2010.

[90] Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

[91] Samir Khuller and Barna Saha. On finding dense subgraphs. In *International Colloquium on Automata, Languages, and Programming*, pages 597–608. Springer, 2009.

[92] Yizong Cheng and George M Church. Biclustering of expression data. in intelligent systems for molecular biology, 2000.

[93] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[94] Peter Kochunov, Laura M Rowland, Els Fieremans, Jelle Veraart, Neda Jahanshad, George Eskandar, Xiaoming Du, Florian Muellerklein, Anya Savransky, Dinesh Shukla, et al. Diffusion-weighted imaging uncovers likely sources of processing-speed deficits in schizophrenia. *Proceedings of the National Academy of Sciences*, 113(47):13504–13509, 2016.

[95] 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.

[96] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):328, 2012.

[97] Marc RJ Carlson, Bin Zhang, Zixing Fang, Paul S Mischel, Steve Horvath, and Stanley F Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC genomics*, 7(1):40, 2006.

[98] Steve Horvath and Jun Dong. Geometric interpretation of gene coexpression network analysis. *PLoS comput biol*, 4(8):e1000117, 2008.

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

[100] Yunpeng Zhao. A survey on theoretical advances of community detection in networks. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5):e1403, 2017.

[101] Muhammad Aqib Javed, Muhammad Shahzad Younis, Siddique Latif, Junaid Qadir, and Adeel Baig. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, 108:87–111, 2018.

[102] Manju Saxena, Scott Williams, Kjetil Taskén, and Tomas Mustelin. Crosstalk between camp-dependent kinase and map kinase through a protein tyrosine phosphatase. *Nature cell biology*, 1(5):305–310, 1999.

[103] Philip JS Stork and John M Schmitt. Crosstalk between camp and map kinase signaling in the regulation of cell proliferation. *Trends in cell biology*, 12(6):258–266, 2002.

[104] Jianhua Ruan, Angela K Dean, and Weixiong Zhang. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC systems biology*, 4(1):8, 2010.

[105] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[106] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[107] Yu Zhang and Dit-Yan Yeung. Overlapping community detection via bounded nonnegative matrix tri-factorization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 606–614, 2012.

[108] Joyce Jiyoung Whang, David F Gleich, and Inderjit S Dhillon. Overlapping community detection using neighborhood-inflated seed expansion. *IEEE Transactions on Knowledge and Data Engineering*, 28(5):1272–1284, 2016.

[109] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting overlapping communities in networks using spectral methods. *SIAM Journal on Mathematics of Data Science*, 2(2):265–283, 2020.

[110] Wei Zhao, Peter Langfelder, Tova Fuller, Jun Dong, Ai Li, and Steve Hovarth. Weighted gene coexpression network analysis: state of the art. *Journal of biopharmaceutical statistics*, 20(2):281–300, 2010.

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

[112] Wei Pan, Jizhen Lin, and Chap T Le. Model-based cluster analysis of microarray gene-expression data. *Genome biology*, 3(2):research0009–1, 2002.

[113] Yanni Zhu, Xiaotong Shen, and Wei Pan. Network-based support vector machine for classification of microarray samples. *BMC bioinformatics*, 10(S1):S21, 2009.

[114] Juliane Schäfer and Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[115] Wei Pan, Jizhen Lin, and Chap T Le. A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & integrative genomics*, 3(3):117–124, 2003.

[116] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[117] Ian Gallagher, Anna Bertiger, Carey Priebe, and Patrick Rubin-Delanchy. Spectral clustering in the weighted stochastic block model. *arXiv preprint arXiv:1910.05534*, 2019.

[118] Rajiv Gandhi, Samir Khuller, and Aravind Srinivasan. Approximation algorithms for partial covering problems. In *International Colloquium on Automata, Languages, and Programming*, pages 225–236. Springer, 2001.

[119] Refael Hassin and Asaf Levin. A better-than-greedy approximation algorithm for the minimum set cover problem. *SIAM Journal on Computing*, 35(1):189–200, 2005.

[120] Jochen Könemann, Ojas Parekh, and Danny Segev. A unified approach to approximating partial covering problems. *Algorithmica*, 59(4):489–509, 2011.

[121] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[122] Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368 (22):2059–2074, 2013.

[123] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[124] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1): D649–D655, 2018.

[125] Weijun Luo and Cory Brouwer. Pathview: an r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, 2013.

[126] Michele Milella, Steven M Kornblau, Zeev Estrov, Bing Z Carter, Hélène Lapillonne, David Harris, Marina Konopleva, Shourong Zhao, Elihu Estey, Michael Andreeff, et al. Therapeutic targeting of the mek/mapk signal transduction module in acute myeloid leukemia. *The Journal of clinical investigation*, 108(6): 851–859, 2001.

[127] Ryan M Teague and Justin Kline. Immune evasion in acute myeloid leukemia: current concepts and future directions. *Journal for immunotherapy of cancer*, 1(1): 1–11, 2013.

[128] Davide Bedognetti, Jessica Roelands, Julie Decock, Ena Wang, and Wouter Hendrickx. The mapk hypothesis: immune-regulatory effects of mapk-pathway genetic dysregulations and implications for breast cancer immunotherapy. *Emerging Topics in Life Sciences*, 1(5):429–445, 2017.

[129] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9 (Sep):1981–2014, 2008.

[130] Kyle Kloster and David F Gleich. Heat kernel based community detection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1386–1395, 2014.

[131] Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

[132] David S Choi, Patrick J Wolfe, and Edoardo M Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

[133] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[134] Yubao Wu, Ruoming Jin, Xiaofeng Zhu, and Xiang Zhang. Finding dense and connected subgraphs in dual networks. In *2015 IEEE 31st International Conference on Data Engineering*, pages 915–926. IEEE, 2015.