

Statistical Methods For Comparing Labelled Graphs

A thesis presented for the degree of
Doctor of Philosophy of the Imperial College London
and the
Diploma of Imperial College (DIC)
by

Da Ruan

Department of Mathematics
Imperial College
180 Queen's Gate, London SW7 2BZ

JULY 22, 2014

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: Da Ruan

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

This thesis is dedicated to my parents.

For their endless love, support and encouragement

Abstract

Due to the availability of the vast amount of graph-structured data generated in various experiment settings (e.g., biological processes, social connections), the need to rapidly identify network structural differences is becoming increasingly prevalent. In many fields, such as bioinformatics, social network analysis and neuroscience, graphs estimated from the same experimental settings are always defined on a fixed set of objects. We formalize such a problem as a labelled graph comparison problem. The main issue in this area, i.e. measuring the distance between graphs, has been extensively studied over the past few decades. Although a large distance value constitutes evidence of difference between graphs, we are more interested in the issue of inferentially justifying whether a distance value as large or larger than the observed distance could have been obtained simply by chance. However, little work has been done to provide the procedures of statistical inference necessary to formally answer this question. Permutation-based inference has been proposed as a theoretically sound approach and a natural way of tackling such a problem. However, the common permutation procedure is computationally expensive, especially for large graphs.

This thesis contributes to the labelled graph comparison problem by addressing three different topics. Firstly, we analyse two labelled graphs by inferentially justifying their independence. A permutation-based testing procedure based on Generalized Hamming Distance (GHD) is proposed. We show rigorously that the permutation distribution is approximately normal for a large network, under three graph models with two different types of edge weights. The statistical significance can be evaluated without the need to resort to computationally expensive permutation procedures. Numerical results suggest the validity of this approximation. With the Topological Overlap edge weight, we suggest that the

GHD test is a more powerful test to identify network differences.

Secondly, we tackle the problem of comparing two large complex networks in which only localized topological differences are assumed. By applying the normal approximation for the GHD test, we propose an algorithm that can effectively detect localised changes in the network structure from two large complex networks. This algorithm is quickly and easily implemented. Simulations and applications suggest that it is a useful tool to detect subtle differences in complex network structures.

Finally, we address the problem of comparing multiple graphs. For this topic, we analyse two different problems that can be interpreted as corresponding to two distinct null hypotheses:

- (i) a set of graphs are mutually independent;
- (ii) graphs in one set are independent of graphs in another set.

Applications for the multiple graphs problem are commonly found in social network analysis (i) or neuroscience (ii). However, little work has been done to inferentially address the problem of comparing multiple networks. We propose two different statistical testing procedures for (i) and (ii), by again using a normality approximation for GHD. We extend the normality of GHD for the two graphs case to multiple cases, for hypotheses (i) and (ii), with two different permutation strategies. We further build a link between the test of group independence to an existing method, namely the Multivariate Exponential Random Graph Permutation model (MERGP). We show that by applying asymptotic normality, the maximum likelihood estimate of MERGP can be analytically derived. Therefore, the original, computationally expensive, inferential procedure of MERGP can be abandoned.

Acknowledgements

I would like to thank my two supervisors Prof. Alastair Young and Prof. Giovanni Montana for their generous donation of time, effort and patience. They both provided a constant source of ideas, advices and encouragements as well as constructive criticisms which drove me to improve my work. Especially, they both kindly helped with the presentation of the thesis.

I feel very glad to study at Imperial College. At Imperial, I would like to thank Dr. Yvonne Ho who helped me a lot in my first year of study. Additionally, the PhD students at Imperial College: Dr. Christopher Minas, Dr. Xiaole Zhang, Dr. Wang Han, Zhana Kuncheva, and Zi Wang for their insights when I need them most. I also would like to thank Dr. John Gibbons and Mrs. Svanidze Rusudan for their help and patience with my PHD submission extension.

Finally, I would like to thank my father, my mother and my girl friend for their support and love.

Da Ruan

Table of contents

Abstract	5
1 Introduction	13
1.1 Motivation	13
1.2 Testing The Independence Between Two Labelled Graphs	15
1.3 Differential Analysis Of Large Complex Networks	17
1.4 Testing The Independence For Multiple Networks	18
1.5 Contributions	20
1.6 Outline Of The Thesis	22
2 Review Of Network Models And Statistical Methods For Comparing Networks	23
2.1 Introduction And Primer On Graph Theory	24
2.2 Graph Models	25
2.2.1 Erdős-Rényi Random graph (ER)	26
2.2.2 Random Geometric graph (RG)	26
2.2.3 Scale Free network (SF)	29
2.3 Permutation Based Procedures For Comparing Two Networks	31
2.3.1 Permutation based non-parametric inference	31
2.3.2 Mean Absolute Difference (MAD)	33
2.3.3 Quadratic Assignment Procedure (QAP)	34
2.4 MCUG: A Model Based Test Of Network Mutual Independence For Multiple Networks	34
2.5 Multivariate Exponential Random Graph Permutation Model: An Inferential Procedure For Comparing Groups Of Networks	36
2.5.1 Exponential random graph model (ERG)	36
2.5.2 Multivariate Exponential Random Graph Permutation (MERGP) model	38
2.6 A Brief Introduction To Mathematical Notations	41
3 GHD: A Permutation Based Test Of Independence	43
3.1 The Generalized Hamming Distance	44

3.2	Statistical Inference For GHD	49
3.2.1	Asymptotic normality under three graph models	53
3.2.2	Normality when using the HD for SF network	53
3.2.3	Normality when using the GHD_TO for SF network	59
3.2.4	Normality when using the HD for RG graph	72
3.2.5	Normality when using the GHD_TO for RG graph	75
3.2.6	Normality when using the HD for ER graph	83
3.2.7	Normality when using the GHD_TO for ER graph	85
3.2.8	A summary of asymptotic results	88
3.3	Simulation Studies	89
3.3.1	Asymptotic property of GHD	89
3.3.2	Power study	92
3.4	Summary	94
4	Differential Analysis Of Large Complex networks	97
4.1	An Algorithm To Detect Differential Subnetwork	98
4.1.1	A remark on subnetwork detection algorithm	99
4.2	Computing $\delta_{\mathcal{V}_K s}$	101
4.2.1	A computationally efficient vector representation of $\delta_{\mathcal{V}_K s}$ for HD	103
4.2.2	Evaluating $\delta_{\mathcal{V}_K s}$ for GHD_TO	104
4.3	Simulation Study	107
4.3.1	Generating networks with localized differences	107
4.3.2	Results	108
4.4	An Application To Co-methylation Networks In Ovarian Cancer	115
4.4.1	Main results	116
4.4.2	Further studies of subnetwork detection algorithm with co-methylation data	119
4.5	Summary	123
5	Comparing Multiple Networks	125
5.1	A Test Of Independence Of Networks Based On Independent Permutations	126
5.1.1	Proof of joint asymptotic normality of $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$	127
5.1.2	A test of independence for $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$	130
5.2	A Test Of Group Independence Based On Common Permutation	131
5.2.1	Proof of asymptotic normality for $\text{GHD}(\mathcal{G}_{\pi}(\mathcal{A}), \mathcal{G}(\mathcal{B}))$	133
5.2.2	A test of group independence based on $\text{GHD}(\mathcal{G}_{\pi}(\mathcal{A}), \mathcal{G}(\mathcal{B}))$	138
5.3	Approximation For Multivariate Exponential Random Graph Permutation model (MERGP)	140
5.3.1	Introduction	140
5.3.2	Derivation of the normalizing constant $\mathcal{X}'_{(\phi)}$	143
5.3.3	Formulation of the MLE $\hat{\phi}$	144

5.3.4	A test of independence between two sets of networks based on the MLE $\hat{\phi}$	145
5.4	Discussion	146
5.4.1	A remark on the two tests	146
5.4.2	Discussion on duplicated samples	147
5.4.3	The null hypothesis of independence and future work	150
5.5	Simulation Study	150
5.5.1	Empirical justification of (5.22)	151
5.5.2	Validation of null sampling distribution for testing of mutual independence	152
5.5.3	Validation of null sampling distribution for testing of group independence	154
5.5.4	Power study	158
5.6	Summary	160
6	Conclusion	161
6.1	Summary Of The Thesis	161
6.2	Future Work	162
	References	174
A	Additional Derivation And Proofs	175
A.1	Closed-form moment computation	175
A.1.1	Notation	175
A.1.2	Calculation of the mean	177
A.1.3	Calculation of the covariance	177
A.2	Proof Of (5.22)	182
B	DNA Methylation Analysis: Further Experimental Results	186
B.1	Power Law Of Co-methylation Networks	187
B.2	Results From GO Analysis	188
B.2.1	Top 10 enriched GO terms and pathways for probe set of differential co-methylated (DC) subnetwork	188
B.2.2	Top 10 enriched GO terms for Differential methylated (DM) probes that also found enriched with the probe set of DC subnetworks	192
B.2.3	Top 5 enriched GO terms and pathways for community $\mathcal{C}_1$	194
B.2.4	Top 5 enriched GO terms and pathways for community $\mathcal{C}_2$	196
B.2.5	Top 5 enriched GO terms and pathways for community $\mathcal{C}_3$	198
B.2.6	Top 5 enriched GO terms and pathways for community $\mathcal{C}_4$	199
B.2.7	Top 5 enriched GO terms and pathways for community $\mathcal{C}_5$	201
B.2.8	Top 5 enriched GO terms and pathways for community $\mathcal{C}_6$	203

List of Publications

Most of the authors own research can be found after Chapter 2 of this thesis. Part of the research presented in Chapter 3 and Chapter 4 of this thesis can also be found in the following publication:

- [91] Ruan, D. Young, A. and Montana, G. Differential analysis of biological networks. *Imperial College London*. 2014.

Which has the abstract:

”Biological processes are often represented in the form of networks. In cancer research, the comparison of gene expression or DNA methylation networks observed in cases and healthy controls can lead to the discovery of biological pathways associated to the disease. As a cancer progresses, its signalling and control networks are subject to some degree of localised re-wiring or re-arrangement. Being able to detect localised changes in network topology induced by the presence or progression of the disease can lead to the discovery of novel molecular diagnostic and prognostic signatures. However currently there is a lack of scalable statistical algorithms for the detection of such interaction patterns and the assessment of their statistical significance. We propose the subnetwork detection algorithm, a computationally efficient statistical procedure for comparing two biological networks defined on the same set of genes, and detecting a subset of genes whose interaction patterns in the two networks are statistically independent. We call this set of nodes a differential subnetwork. The algorithm relies on a statistic, the Generalised Hamming Distance (GHD), which quantifies the topological difference between two labelled networks. We demonstrate that, under the null hypothesis of independence between the networks, the sampling distribution of this statistic is asymptotically normal under the

Erdős-Rényi, random geometric, and scale free random network models. The performance of the subnetwork detection algorithm has been assessed using extensive simulations, and we further discuss an application to detect differential DNA co-methylation networks in ovarian cancer.”

Chapter 1

Introduction

1.1 Motivation

A graph, in general, models a network of relationships between objects. In recent years, due to the increasing availability of graph data generated from different areas of study, the need for rapidly identifying and making inference on graph differences is increasingly prevalent. A large number of applications for comparison of graphs can be found in bioinformatic analysis [44, 77, 78, 111], social network analysis [64, 21, 4, 22, 23] and neuroscience [28, 106, 68].

In bioinformatics analysis, the studies of biological networks inferred from gene expression and DNA methylation (a biochemical process where a methyl group is added to the cytosine or adenine DNA nucleotides) data abound. A number of high-throughput platforms are commonly used to compare genome-wide molecular profiles of large cohorts of normal control subjects and diseased individuals, to understand patterns that differentiate them. For instance, epigenetic studies identify aberrant DNA methylation profiles within CpG islands (regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length) that differ between patients and controls, as changes in DNA methylation levels are believed to be among the earliest and most common alterations in human cancers [101, 11]. Such applications are commonly analysed using a network-based methodology in which each node represents a genomic feature, such as a gene or CpG island, and the presence of an edge between two nodes

indicates the interaction between two corresponding features. To compare a case versus a control, both networks are defined using the same set of labelled nodes. The question then arises whether an observed difference in topological structure (i.e. interaction among set of nodes) between the two networks can be related to the different experimental conditions.

In the field of social network analysis, there is an interest in predicting the structure of relationships among social entities, and analysing the impact of the structure on other social phenomena using a collection of social relations. The analysis of a social network is mainly built around a shared core of concepts and methods for the measurement, representation and analysis of social structure [23]. A classic application in social networks, collected by Krackhardt [65], is to analyse the interpersonal relations among managers in high-technology companies. Networks are estimated by managers status characteristics or structural properties such as age, tenure, membership, relationships or manager/subordinate position. There is a need to understand different relationships among those structural properties (e.g., different age groups, manager versus subordinates, etc.).

In neuroscience studies, the brain is divided into different subregions. Each region is defined as a node and a significant correlation between two parts of the brain (activated or deactivated) forms an edge. Commonly, two samples of brain networks of individuals in different experimental settings are estimated from the same partition of the brain [68]. For the study of brain networks, samples of brain networks for different groups (case/control) are used. It is always necessary to understand that, overall, brain networks from distinct groups act differently.

Literature regarding network comparisons are well established. There are generally two different approaches to address this problem. The first approach for graph comparison studies the similarity of unlabelled or/and unequal sized graphs. Examples of such methods include graph and subgraph isomorphism matching [88, 100], graph edit distance (GED) and their extensions [94, 18, 74, 43], maximum common subgraph-based measures [55, 67, 19, 107] and Graphlet Degree distribution (GDD) [87, 86].

Another class of methods solves the problem of comparing labelled graphs (graph with labels defined on nodes). The labelled graph comparison problem can be informally interpreted as comparing graphs that are estimated from a common set of objects that are defined by a specific experimental setting.

The main method of comparing labelled graphs, i.e. measuring distance, has been extensively studied. Large distance values constitute evidence of a difference between graphs, and it is often of interest to inferentially justify whether a distance value as large or larger than the distance observed could have been obtained by chance only. However, little work has been done to provide the statistical inference that is needed to formally answer this question. Although methods using permutation-based inference (permutation, in the context of graph comparison, refers to randomizing the node labels of a graph, details please refer to Section 2.3.1) have been proposed as a natural way of tackling labelled graph comparisons, the computational cost of such inference procedures, especially when comparing large networks, cannot be ignored.

In the thesis, we propose permutation based inferential procedures to overcome different problems for comparing labelled graphs. The common question that permutation based graph comparison procedures answer is whether two observed networks of relations (or for more than two networks in multiple case) are significantly associated with each other [5, 58, 60]. Usually, the null hypothesis of such a permutation procedure is stated as one of no association between graphs. However, the term no association between graphs is not clear and is generally not formally defined in the literature. In this thesis, we take no association between graphs as meaning *independence* between graphs. Formal interpretation and definition of hypotheses in a permutation based procedure for graph comparison are presented in Section 2.3.1. The research presented focuses on addressing three different aspects of inference for labelled graph comparisons:

- Testing the independence between two labelled graphs (Section 1.2)
- Differential analysis of large complex networks (Section 1.3)
- Testing the independence for multiple graphs (Section 1.4)

In addition, the contributions and outline of this thesis are presented in Section 1.5 and 1.6.

1.2 Testing The Independence Between Two Labelled Graphs

Methods concerning the comparison of two labelled graphs have long been studied. There are two different approaches to this problem. One approach, for instance, Odibat and Reddy

[77, 78], Lord et al. [68], studies node differences between two networks using node-level measures (e.g. node centrality measures).

Another approach, which we will focus on in this thesis, analyses the structural difference between labelled graphs. The central issue for comparing graphs, i.e. measuring distance, has been studied extensively. For instance, distance-based methods with labelled graphs, GED [43, 94], defined as the cost (obtained by using a cost function) of the least expensive sequence of edit operations that are needed to transform one graph into another, can also be used to measure the distance between labelled graphs. GED is a good way of measuring the distance between networks, because it is error-tolerant to noise and distortion [43]. However, the computational cost of GED is expensive, which makes it difficult to do inference studies. Although large distance values could provide evidence of differences in graph topology, we are more interested in the issue of inferentially assessing whether a distance value as large or larger than the observed distance could have been obtained by chance.

However, little work has been done to provide the requisite statistical procedures for evaluating significance for such an assessment. In early efforts, Katz [63] defined a 2 by 2 association table by conformity and concordance of social relations (i.e., adjacency information) for two social networks. Under the null hypothesis of no association between networks, the distribution of the edges of one network should be independent of the distribution of the edges of the other network. By applying Fisher's exact test on the obtained table, the association between graphs can be inferred if the null hypothesis of independence is rejected. The problem with this approach is that it entails the assumption that edges are independent of one another. This assumption is difficult to justify in real applications [80].

For permutation-based procedures, inferences can be made easily. The Mean Absolute Difference (MAD) [20, 106] counts the number of different edges in two graphs. This distance corresponds to the Hamming distance between two networks [49]. The Quadratic Assignment Procedure (QAP) [59] uses edge set product statistics that measure the similarity between networks. Statistical inference for the above two measures is obtained by a standard permutation testing procedure (permuting labels of the network). The idea of the permutation testing procedure is that the indices of conformity for the labels of nodes are constructed by regressing one adjacency matrix with the other, and the permutation distri-

bution of the statistic is employed to estimate the probability of the observed magnitude of a specific index, given the null hypothesis of independence between networks [80]. The permutation-based inference is a theoretically sound approach that does not assume independence among different edges. However, a shortcoming of this inference procedure is that it is computationally expensive.

When comparing graphs, an initial decision has to be made regarding which metric to use to characterise their topologies, so that a comparison can then be carried out. For the above inferential procedures, we used adjacency information, which is commonly used to measure structural similarity [41, 21, 106, 59, 58]. The adjacency information only represents the direct neighbourhood of a network. It sometimes fails to distinguish structural differences (see Section 3.1 for details). For example, the simplest measure of network distance between two labelled networks is the Hamming distance, which uses the adjacency information to quantify the extent of edge differences [28] in the two networks. Although commonly used in the literature, this distance measure is not sufficiently sensitive to capture subtle changes in network structure. Instead of counting only the properties of the immediate neighbourhoods within a network, a more sensitive distance measure should take into account the higher-order neighbourhood structure of the network.

1.3 Differential Analysis Of Large Complex Networks

Another interesting problem is to compare large complex networks in which only localized differences are assumed. For this particular problem, the assumption of independence between case/control networks can be too strong. This is because we commonly deal with problems for which only localised changes in the network structure are expected to be observed. Therefore, the structural changes between networks can be relatively subtle to detect. Very often, we are more interested in the identification of a small sub-component that differentially interacted in the entire network, which is analogous to the detection of localised differences in the mean levels of some node level measures. Applications for such a problem are vast. For example, in cancer studies, signalling of case and control networks are subject to some degree of localised re-wiring or re-arrangement as cancer progresses [8]. Being able to detect localised changes in network topology induced by the presence

or progression of the disease can lead to the discovery of novel molecular diagnostic and prognostic signatures.

The problem of detecting randomly structured biological subnetworks in two-sample comparisons has not been extensively studied. Most of the methods are concerned with finding differentially connected nodes. Especially in differential connectivity analysis, such as [77, 78, 89, 42, 61], differentially connected nodes are individually tested and then a set of significant differentially connected nodes can be identified. However, the problem for these node-based tests is that, although individual nodes are differentially connected, the overall structure that they construct is not necessarily significantly different. Apart from these node comparing methods, there are several approaches that attempt to compare two large networks in modular (subnetwork) levels. For instance, instead of comparing the networks directly, gene modules can first be identified independently in each network, and then compared by assessing the degree of overlap (of nodes) between sets of modules of the different networks [44]. In another study, Zhang and Horvath [111] proposed a method to find "consensus modules" (modules having the same set of nodes for different networks), and further compare those consensus modules to understand the similarity/dissimilarity between them.

1.4 Testing The Independence For Multiple Networks

In neuroscience, there are problems that require the comparison of structural differences of two samples of networks. The development of statistical comparison tools is needed to meet the demand of comparing brain networks across groups of subjects [92]. Despite this need, a large amount of work needs to be done in this area to match its level of importance [106]. Hence, developing a systematic approach to capture the network structures from a group of subjects brain networks has great appeal and would help to fill the gap in the group comparison literature [97].

In neuroscience, connectivity research mainly applies descriptive models based on a specific overall measure of the network, such as graph centrality measures and clustering coefficients [17]. Some inferential studies have employed relatively rudimentary testing procedures, such as the ANOVA used in Meunier et al. [71], or the permutation-based T

test used in Lord et al. [68], to justify group differences based on specific overall graph measures. Such methods evaluate graph measures individually for each network and then combine the information. As argued in Simpson et al. [96], obtaining average measures for each group only provides a global view of the system and does not capture the complex organization within a population of networks. One possible way of assessing the group differences, based on comparing network structures instead of an overall measure, is to first construct a representative *mean* network [1, 71, 103] or *median* network [98] for each group and then compare the representative networks. Here the *mean* or *median* network is estimated from the threshold group *mean* or *median* matrix to obtain a *mean* or *median* network. As a result, the group *mean* or *median* matrix is the average or median across the individual entries of the connectivity matrices. The defect of this method is that for a *mean* network, the outlying functional connectivity values may affect the resulting *mean* network. Also, for both type of networks, these two approaches use the edge-based average/median across the individual entries of the connectivity matrices, and, thus, the topological dependence among edges is ignored within each individual network [97].

In the literature assessing structural differences among multiple graphs, Anderson et al. [4] proposed a flexible model-based testing procedure called the Multivariate Conditional Uniform Graph (MCUG), which compares social networks with graph level measures or structure-based distances/similarities (e.g., graph correlation). Graph level measures indicate a particular property of the network. Examples of graph level measure include network density, cluster index, network centrality measures, etc. According to this procedure, random graphs are generated with fixed graph properties (e.g., size and density). Then the distribution for the measure of interest is computed from those networks generated from the graph models. Statistical significance is assessed based on the above distributions, given the null hypothesis of independence among all networks. Under this procedure, more re-samples may be needed than the standard permutation (of labels) to obtain a reliable p -value.

In another study, Butts [22] proposed a likelihood-based framework called Multivariate Exponential Random Graph Permutations model (MERGP), which is a probabilistic model for the set of label permutations. This procedure is an extension of the Multivariate p^* model of Pattison and Wasserman [79], which is used to model social networks. Unlike the

standard permutation procedure, which can only provide a value for statistics and a p -value, MERGP offers a probability distribution on the set of graph label permutation vectors and an inference for the association between two groups of networks or a set of networks under the null hypothesis that two groups (or a set of networks) are independent. Butts [22] proposed a two-step, pseudo-likelihood-based Markov Chain Monte Carlo (MCMC) procedure to estimate the parameter. However, as stated in Robins et al. [90], the properties of the pseudo-likelihood estimate are not well understood. It is not completely clear from existing research as to when pseudo-likelihood estimates may become acceptable. Note that both methods of comparing groups/set of networks can be reduced to compare two networks.

1.5 Contributions

This thesis contributes to four different issues related to comparing labelled networks. First, it considers the issue of choosing a distance measure between a pair of networks that is able to capture subtle topological differences, while also being easily interpreted and computed. We propose a distance measure, namely the generalised Hamming distance (GHD), which can apply to various edge weight measures. In this work, we suggest an edge weight called topological overlap measure (TO) to apply with the GHD. This TO measure has been studied extensively in biological network analysis, and found to capture biologically meaningful patterns [56]. The measure takes into account higher-order neighbourhood structures and, hence, has more power to detect differential structures, while being amenable to statistical analysis.

Secondly, we discuss how to establish whether an observed discrepancy in network topology, measured using the chosen distance measure, can be deemed non-random, hence statistically significant. Although larger distance values provide stronger evidence in support of a real difference in topology, formal inferences can only be drawn by taking into account the sampling properties of the distance measure. We are therefore confronted with the issue of assessing whether a GHD distance as large or larger than the observed one, given the two networks, could also have been obtained by chance. We address this problem by introducing a test, the GHD test, for assessing the null hypothesis that two paired

networks are independent. We demonstrate that, under the null hypothesis, the sampling distribution of the GHD test statistic approaches a standard normal distribution as the size of the networks increases. Thus, p -values can be obtained without the need to resort to computationally expensive permutation procedures. Two sufficient conditions for asymptotic normality are provided, which can be easily checked given the two observed networks. Furthermore, we show that the two conditions are satisfied in most situations for adjacency information and TO measures, when the networks are assumed to be generated by any of three different random graph models: Erdős-Rényi, Random Geometric, and Scale Free networks. The first two moments of the standardised GHD test statistic can be cheaply computed in closed-form, so that p -values can be obtained routinely. Simulation shows the validity of the normality approximation for GHD with different measures, and, furthermore, it is demonstrated by the simulation that the GHD with TO outperforms other permutation-based methods.

Thirdly, we ask whether it is possible to identify a differential subnetwork, starting from two large networks, so that localised differential behaviour can be unveiled. By using the GHD test, we further propose a simple differential subnetwork detection algorithm, that repeatedly applies the test to smaller and smaller sub-graphs until a minimal network size is reached. At each iteration, one or multiple nodes are removed so that the distance between the paired networks is increased, and a new p -value is computed in closed-form, without requiring repeated permutation testing. We demonstrate that the algorithm generates a sequence of monotonically increasing p -values, which are then adjusted for multiple testing and enable the identification of a differential subnetwork. Extensive simulation experiments have been performed to illustrate the main features of the proposed algorithm for detecting differential subnetworks. We also present an application to DNA co-methylation networks estimated for ovarian cancer patients and healthy subjects. Using this algorithm, we were able to identify two subnetworks that present statistically significant changes in network topologies between cases and controls.

Finally, we extend the procedure for comparing two networks to a more general problem, in which comparing the structural changes between two groups of networks or among a set of networks is considered. The interest then becomes whether there are structural differences between groups or a set of networks given that two groups are independent

or the set of networks are mutually independent. Two tests of independence for multiple networks are proposed, based on a different permutation strategy under the assumption that the permutation distribution is multivariate normal under the null. We also discuss using both strategies in comparing two groups of networks. We further show that using standardized statistics, under the normality assumption, the maximum likelihood estimate of Multivariate Exponential Random Graph Permutation (MERGP) for Butts [22] can be analytically derived, and, hence, the potentially inaccurate and computationally expensive pseudo-likelihood-based MCMC procedure can be replaced.

1.6 Outline Of The Thesis

The thesis is organized as follows. In Chapter 2, relative procedures for comparing networks are reviewed together with three graph models. In Chapter 3, the GHD test is proposed and proofs of normality for two edge weight measures and three network models are presented. Then in Chapter 4, an algorithm for detecting differential subnetworks is proposed, together with applications in bioinformatics regarding the analysis of large biological networks. In Chapter 5, we propose two tests for comparing multiple networks, and then build a link between one proposed method and the MERGP. Finally, in Chapter 6 we draw conclusions and also provide a discussion for future studies.

Chapter 2

Review Of Network Models And Statistical Methods For Comparing Networks

In this chapter, we review some graph comparison methods and useful graph models that are relevant to the methods developed in this thesis and will appear multiple times in the upcoming chapters. The descriptions of the existing graph comparison methods and graph models will offer the reader some details regarding how they are performed and related.

This chapter sets the context of the work in this thesis. We first formally define the problem and then we provide a short review of: a primer on graph theory (Section 2.1), three random graph models, including Erdős-Rényi (ER), Random Geometric (RG), and Scale Free networks (SF) (Section 2.2), two permutation-based network comparison methods for comparing two networks (Section 2.3), e.g., MAD and QAP, two methods for comparing multiple networks: the Multivariate Conditional Uniform Graph model (MCUG) (Section 2.4) and Multivariate Exponential Random Graph Permutation model (MERGP) (Section 2.5). Finally we provide a brief introduction about the mathematic notations we use in this thesis (Section 2.6)

2.1 Introduction And Primer On Graph Theory

In this thesis, we focus on solving the problem of comparing labelled graphs. Before we formally define the problem, we will need a primer on graph theory. The purpose of this section is to define terminology and notation for the remainder of this thesis, and to provide the definitions from graph theory that are useful to interpreting our main results [16].

A graph is a set of nodes connected by edges. More formally, it can be defined as follows

Definition 1. (Graph) A graph is a pair $G = (\mathcal{V}, \mathcal{E})$ of sets of nodes $\mathcal{V}$ and edges $\mathcal{E}$, where each edge connects a pair of nodes, i.e., $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$. In general, $\mathcal{V}(G)$ refers to the set of nodes of graph G , and $\mathcal{E}(G)$ refers to the edges of graph G .

A graph with labels on its nodes is called node-labelled, a graph with labels on edges is called edge-labelled. In this thesis, we consider the case where all the nodes of the network are labelled. Here, in a slight abuse of terminology, we refer to the node-labelled graph as a labelled graph. The size of a graph $G = (\mathcal{V}, \mathcal{E})$ is determined by the number of nodes of a graph written as $\langle \mathcal{V}(G) \rangle$. Here $\langle \mathcal{S} \rangle$ is denoted as the total number of elements in a set $\mathcal{S}$. In this thesis we always denote labels for nodes of a graph to be $\{1, 2, \dots, N\}$, where N is the size of the graph.

Let v_i and v_j be two nodes in $\mathcal{V}$. Then nodes v_i and v_j are **neighbor** or **adjacent** (and $v_i \neq v_j$) if and only if they are connected by an edge.

We say that a graph is **empty** if there are no adjacent nodes in $\mathcal{V}(G)$, and also a graph is considered **complete** if all the nodes in $\mathcal{V}(G)$ are pairwise neighbours.

The adjacency information for all pairs of nodes in a graph is represented by an adjacency matrix.

Definition 2. (Adjacency Matrix) The adjacency matrix $A = (A_{ij})$ (of size $N \times N$) of graph $G = (\mathcal{V}, \mathcal{E})$ is defined by

$$A_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases}$$

The number of neighbours of a node is called node degree.

Definition 3. (Degree of a Node) The degree d_i of a node v_i in $G = (\mathcal{V}, \mathcal{E})$ is the total number of neighbors of v_i ,

$$d_i = \sum_{j=1}^N A_{ij}.$$

To formally define the comparison of labelled graphs, suppose we have two sets of graphs, $\mathcal{G}(\mathcal{A}) = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$ and $\mathcal{G}(\mathcal{B}) = \{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_Q\}$, where $P > 0$ and $Q > 0$ are constants. Let $\mathcal{A}_k, k = 1, \dots, P$ and $\mathcal{B}_l, l = 1, \dots, Q$ each be represented by a graph in $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$, denoted $\mathcal{A}_k = (\mathcal{V}, \mathcal{E}_{\mathcal{A}_k})$ and $\mathcal{B}_l = (\mathcal{V}, \mathcal{E}_{\mathcal{B}_l})$, respectively. Both graphs are defined using a common set of labels $\{1, 2, \dots, N\}$ of nodes, representing features of interest. The respective sets $\mathcal{E}_{\mathcal{A}_k}$ and $\mathcal{E}_{\mathcal{B}_l}$ of edges indicate the connections between the nodes in the two graphs. The matrices $A_k = (A_{kij})$ and $B_l = (B_{lij})$ denote the two $(N \times N)$ adjacency matrices associated with graphs $\mathcal{A}_k$ and $\mathcal{B}_l$, respectively. In a two networks comparison problem, for notational simplicity, we denote the two networks to be $\mathcal{A}$ and $\mathcal{B}$, respectively, while the corresponding adjacency matrices for $\mathcal{A}$ and $\mathcal{B}$ are $A = (A_{ij})$ and $B = (B_{ij})$, respectively. It is of great interest to understand whether there is a tendency for elements that are strongly connected in a network to be similarly connected in the other network [22], or whether there is a correspondence between objects of one (or one set of) event(s) to the same set of objects in another event.

2.2 Graph Models

In this section, we introduce three different types of graph model: Erdős-Rényi Random graph (ER), Geometric Random graph (RG) and Scale Free network model (SF). Node degree distributions for each of the graph models are introduced, and possible graph generating procedures are presented. For both the SF and RG models, we also discuss recent studies and review related applications in which they are involved.

2.2.1 Erdős-Rényi Random graph (ER)

Firstly, we introduce the simplest and well studied Erdős-Rényi [37, 38] random graph model. An ER graph is generated by assuming that the presence or absence of an edge between two particular nodes is independent of that of any other edge. Under this independence assumption, each edge is of equal probability, say p . Supposing the node set $\mathcal{V}$ contains N nodes, then the expected number of edges and expected node degrees are $\frac{pN(N-1)}{2}$ and $p(N-1)$, respectively. Denote the probability distribution of node d_i with node degree k as $P(d_i = k)$. Then for the Erdős-Rényi random graph, $P(d_i = k)$ follows a binomial distribution:

$$P(d_i = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

For large N and small p , the above equation approximates a Poisson distribution with parameter $\lambda_N = p(N-1)$:

$$P(d_i = k) = \frac{e^{-pN} (p(N-1))^k}{k!}. \quad (2.1)$$

For the case when pN tends to infinity, the binomial node degree distribution is approximated by a normal distribution with mean $\mu_N = p(N-1)$ and variance $\sigma_N^2 = (N-1)p(1-p)$. Random graphs generated in such cases do not accurately reflect real-world phenomena Newman et al. [76]. This is caused by the independence assumption among edges as well as nodes. For real world networks, such as gene-to-gene interaction, social networking, the Internet, etc., such independence cannot be assumed.

2.2.2 Random Geometric graph (RG)

A random geometric graph is a widely used graph model in many areas and has many different applications. Such applications involve protein-protein interactions [87, 52], epidemics networks [48, 85], energy consumption networks [10] etc. A random geometric graph can be constructed by an independent and uniform random scattering of N nodes in a region A in m -dimensional space R^m [81, 52]. For a given constant r , an edge is formed

if and only if the distance between two random scattered nodes is less than or equal to r .

More specifically, consider A to be the unit cube in R^m : $[0, 1]^m$, $U_1, U_2, \dots, U_N$ are scatters in $[0, 1]^m$ with each U_i generated independently from the uniform distribution in $[0, 1]^m$. For $\|r\| < 1$, U_i and U_j are connected (denoted as $i \sim j$):

$$i \sim j \quad \text{iff} \quad \|U_i - U_j\| \leq r.$$

Here, $\|\cdot\|$ is a vector norm, commonly an Euclidean norm. The node degree distribution for the above model is not obvious [29], because the node degree distribution is closely associated with the connectivity of the graph. The connectivity is the average number of connections per vertex (i.e., average node degree) denoted as λ_N .

Denote the volume of a m -dimensional sphere with radius $r' = r/2$ as V , Dall and Christensen [29] stated that the connectivity of a random geometric graph is directly related to excluded volume V_{ex} , where $V_{ex} = 2^m V$ [29] and $\lambda_N = NV_{ex}$.

$$V = \frac{\pi^{m/2} r'^m}{\Gamma(\frac{m+2}{2})}.$$

where $\Gamma(\cdot)$ is the gamma function. Now consider a cube containing spheres with the above radius r and volume V centred at nodes. Two nodes are connected if and only if there is an overlap between two spheres that is less than or equal to the distance ($r = 2r'$) between the two nodes. Then the node degree distribution can be approximated by:

$$P(d_i = k) = \frac{e^{-\lambda_N} \lambda_N^k}{k!},$$

which also defines a Poisson distribution.

Except for RG network with fixed size N , there is another procedure that can generate RG network with size $N_{(A)}$ in a fixed region A . Assume the number of nodes in a particular region is assumed to follow a homogeneous Poisson process. This process is illustrated as follows:

- The number of nodes in a region A follows a homogeneous Poisson process in R^m with intensity λ : $N_{(A)} \sim \text{Poisson}(\lambda [A])$, Where $[A]$ denote the volume of A .

- Conditional on the number of nodes $N_{(A)}$, a random geometric graph with A of $N_{(A)}$ nodes and parameter r is generated by following the same steps as previously stated.

If we consider the above node generating process, the node degree distribution can be derived more specifically. Haenggi [47] derived the distribution of the distance for a node and its k^{th} ordered neighbour for a Poisson point process in m dimensional space with intensity λ . This distribution is of the form of a generalized Gamma distribution:

$$f_R(x) = e^{-\lambda c_m x^m} \frac{m(\lambda c_m x^m)^k}{x \Gamma(k)} = e^{-\lambda c_m x^m} x^{km-1} \frac{m(\lambda c_m)^k}{\Gamma(k)},$$

where $c_m x^m$ is the volume of an m -dimensional ball of radius x in m dimensional space. Notice that the probability that the distance between a node and its k^{th} nearest neighbour is greater than r is equivalent to the probability of a node i having a node degree d_i smaller than k . This relation is given by:

$$P(d_i \leq k-1) = P(x > r) = 1 - P(x \leq r).$$

Since we know that distance follows a generalized Gamma distribution, we have

$$\begin{aligned} P(d_i \leq k-1) &= 1 - P(x \leq r) \\ &= 1 - \frac{\gamma(k, \lambda c_m r^m)}{\Gamma(k)} \\ &= \frac{\Gamma(k) - \gamma(k, \lambda c_m r^m)}{\Gamma(k)} \\ &= \frac{\Gamma(k, \lambda c_m r^m)}{\Gamma(k)} \\ &= e^{-(\lambda c_m r^m)} \sum_{i=0}^{k-1} \frac{(\lambda c_m r^m)^i}{i!}. \end{aligned}$$

where $\gamma(\cdot, \cdot)$ and $\Gamma(\cdot, \cdot)$ are a lower incomplete gamma function and an upper incomplete gamma function, respectively. The third step is needed because the summation of the lower and upper incomplete gamma functions is equal to the gamma function: $\Gamma(k) = \gamma(k, \lambda c_m r^m) + \Gamma(k, \lambda c_m r^m)$. Then, the probability $P(d \leq k-1)$ has a Poisson cumulative distribution function with parameter $\lambda_N = \lambda c_m r^m$. For large N , the Poisson distribution

may be approximated by a normal distribution with mean λ_N and variance λ_N .

It is not surprising to see that random geometric graphs have similar node degree distributions as random graphs. This may be because both procedures generate nodes independently and uniformly. In a recent study, Faragó [39] showed that a random geometric graph is of equivalent class to a random graph. That is, for any random geometric graph, there exists an isomorphic random graph, and vice versa.

2.2.3 Scale Free network (SF)

The scale free network is probably the most popular class of network model, that has been applied to all fields of real network study [2, 9, 75, 32, 62, 25, 45, 36] etc. Many empirical results have demonstrated that real complex networks have two main characteristics: the first feature is the so-called small world phenomenon [109]. This phenomenon suggests that real networks tend to contain sub-networks that hold connections between almost any two nodes within them, and that most pairs of nodes have at least one shortest path. The average shortest path $\bar{L}$ is of the order $\log N$. Another characteristic is that nodes in real networks tend to follow a power law node degree distribution that is heavy tailed. Again, let $P(d_i = k)$ be the probability of node i having degree k . We first denote m to be lower cut off of node degree and K to be upper cut off of node degree. Then it follows that

$$P(d_i = k) = ck^{-\alpha}, k = m, m + 1, \dots, K, \quad (2.2)$$

where m and K are lower and upper cut-offs of the node degree, respectively. The lower cut-off m is generally taken to be 1. Also, c is a normalising constant, and α represents the power exponent. It is generally assumed that α is greater than 1.

For different exponent α , we may assume different upper cut-off of K , this is because for large α , the probably of having a node with very large node degree might be ignorable. When $\alpha > 2$, K is generally specified as $K = N^{\frac{1}{\alpha-1}}$. This is because the probability of observing a node with node degree greater than $K = N^{\frac{1}{\alpha-1}}$ is $\frac{1}{N}$ [33, 26, 27],

$$\sum_{i=K}^N \frac{N_i}{N} = \frac{1}{N},$$

Where N_i is the number of nodes in the entire graph with degree i , and the total number of nodes is $\sum_i N_i = N$, . We also specify $K = N - 1$ for $1 < \alpha \leq 2$.

The meaning of scale free lies in the fact that the asymptotic nature of the degree sequence is independent of its size [104]. Cohen and Havlin [26] provided an interesting fact regarding the scale free network. They showed that the scale free network is 'ultra-small', which indicates that the average shortest path for a scale free network is of the order $\log \log N$, which is of a lower order than the small-world network ($\log N$).

Another interesting issue that has been established empirically is the choice of power exponent α for different types of real network. As showed in Table 2.2.3, it shows that biological networks tend to have small exponents (smaller than 2),

Network	Approx Exponent α
Non-Biological network	
Internet	2.1 (in), 2.5 (out)
Citations	3
Actors	2.3
Power grid	4
Phone call	2.1-2.3
Biological network	
Yeast protein-protein net	1.5,1.6,1.7,2.5
E. coli metabolic net	1.7,2.2
Yeast gene expression net	1.4-1.7
Gene functional interactions	1.6

Table 2.1: List of power exponents analysed by Chung [25]

Barabási and Albert [7] provided a process for generating graphs with scale-free properties. They claimed that the scale free property of complex networks arises because the network grows with the addition of nodes and that there is a greater probability of a new node connecting to a node with a high node degree. Hence, they introduced a network generating process concerning the preferential attachment of new incoming nodes. The

incoming node connects to a specific node i with probability $\frac{d_i}{\sum_j d_j}$, where d_i indicates the node degree for node i and $\sum_j d_j$ is the total node degree for the existing network. It has been shown that scale free networks generated from such graph models have the exponent $\alpha = 3$.

2.3 Permutation Based Procedures For Comparing Two Networks

In this section, we will introduce the permutation procedure (section 2.3.1), together with two procedures, the Mean Absolute Difference (MAD) and Quadratic Assignment Procedure (QAP). For these two tests, adjacency information is commonly applied as similarity measures [41, 21, 106, 59, 58]. The adjacency information only represents the direct neighbourhoods of a network and is not capable of identifying structural differences (see section 3.1 for detail). Also, the inference procedures for both tests rely on permutations, and, therefore, they are computationally expensive for large graph comparisons.

2.3.1 Permutation based non-parametric inference

The problem of comparing the structures of labelled networks has been well studied. However, there are not many procedures that have been proposed to evaluate statistical significance. A natural way of evaluating statistical significance, in the context of labelled network comparison, is to use a permutation procedure. The common question that permutation based graph comparison procedures answer is whether two observed networks are significantly related to each other [5, 58, 60]. The null hypothesis of a permutation procedure is commonly presented as one of no association between networks. However, this term, 'no association between networks', is vaguely presented without a formal definition in the literature.

In this thesis, we interpret no association between networks as meaning *independence* between two networks. To formally define independence between two networks, suppose $\mathcal{A}$ and $\mathcal{B}$ are two networks defined on the same node set $\mathcal{V}$ of size N . Then $\mathcal{A}$ and $\mathcal{B}$ are described by sets of random variables, say $\{a_{ij}, 1 \leq i \leq N, 1 \leq j \leq N, i \neq j\}$ and $\{b_{ij}, 1 \leq i \leq N, 1 \leq j \leq N, i \neq j\}$, indicating edge weights (a measure of association

between nodes i and j given a particular network). We say that network $\mathcal{A}$ and network $\mathcal{B}$ are *independent*, if the two random vectors $\{a_{ij}, 1 \leq i \leq N, 1 \leq j \leq N, i \neq j\}$ and $\{b_{ij}, 1 \leq i \leq N, 1 \leq j \leq N, i \neq j\}$ of edge weights are independent. Under the definition, the null hypothesis for the permutation-based procedure is,

$$\mathcal{H}_0 : \text{Two networks are independent.}$$

The permutation based method attempts to compute the probability that a distance as large or larger than the observed value could have been observed by chance, under the null hypothesis assumption that two networks are independent. By treating the test statistic as a random variable with an unknown distribution, this probability may be computed non-parametrically using a standard permutation testing procedure.

A **permutation** is a bijection of a set onto itself. For a label set (for nodes) of size N , denoted by $\mathcal{L} = \{1, 2, \dots, N\}$, there are $N!$ possible permutations collected in a set Π . The permutation procedure, in the context of network comparison, works as follows.

Denote by π a particular index of the labels $\mathcal{L}$ after permutation. The procedure consists of permuting the elements of $\mathcal{L}$ for network $\mathcal{A}$, while keeping the edges unchanged, thus obtaining a permuted network $\mathcal{A}_\pi$. The idea of the permutation (permuting the labels of the nodes $\mathcal{L}$) method is that the statistic for index π is calculated for each of the new pairs of comparison (between $\mathcal{A}_\pi$ and $\mathcal{B}$) and, under the null hypothesis that the two networks are independent of one another, all values of the statistic under permutation are equally likely. If the observed index is extreme in this distribution, then the hypothesis of independence is rejected.

More specifically, given a distance measure (or similarity measure) between two networks defined in terms of edge weights, the null sampling distribution for such a distance (similarity), under independence, can be estimated by random permutation of the node labels in one of the networks. Evidence against independence is provided by extreme values of such a distance measure (similarity measure).

2.3.2 Mean Absolute Difference (MAD)

The MAD statistic sums over all edge differences between two networks:

$$d(\mathcal{A}, \mathcal{B}) = \frac{1}{N(N-1)} \sum_{i \neq j} |a_{ij} - b_{ij}|, \quad (2.3)$$

where edge weights a_{ij} and b_{ij} are given by the adjacency information for the two networks. This measure of distance has been extensively used in the literature [21, 106, 44] to compare networks.

Now we denote $a_{\pi(i)\pi(j)}$ to be a measure of edge weight for nodes i and j after permutation $\pi \in \Pi$. The corresponding permuted MAD value of (2.3) has the form

$$d(\mathcal{A}_\pi, \mathcal{B}) = \frac{1}{N(N-1)} \sum_{i,j} |a_{\pi(i)\pi(j)} - b_{ij}|$$

which is discussed in detail below. This is evaluated for every $\pi \in \Pi$ so as to infer the exact permutation distribution of MAD under the null hypothesis. This distribution, as stated in 2.3.1, can then be utilized to evaluate statistical significance.

The statistic (2.3) measures the edge differences between two networks, hence it is equivalent to the well known Hamming distance [49, 50] between two graphs, which is defined as:

$$H(\mathcal{A}, \mathcal{B}) = \frac{1}{N(N-1)} \sum_{i \neq j} I(A_{ij} \neq B_{ij}).$$

The utilization of Hamming distance in network analysis has not been limited to comparing networks, it has also been used as a measure of distance in network modelling [6].

The Hamming distance is formed by the arithmetic mean of the edge differences. Apart from using the arithmetic mean, da F Costa et al. [28] considered the test statistic to be the geometrical average between coinciding ones and coinciding zeros in an adjacency matrix (i.e. where $A_{ij} = B_{ij}$). Let c_0 and c_1 be the number of coinciding zeros and ones for two networks. Let A_0 , and A_1 be total number of zeros and ones in network $\mathcal{A}$. The statistic

they proposed is

$$gh(\mathcal{A}, \mathcal{B}) = \sqrt{\frac{c_0}{A_0} \frac{c_1}{A_1}}.$$

The permutation test can also be applied to this statistic. However, this measure is not a symmetric measure of distance, as the choice of denominator (i.e., $gh(\mathcal{A}, \mathcal{B}) \neq gh(\mathcal{B}, \mathcal{A})$) sometimes may affect the test result.

2.3.3 Quadratic Assignment Procedure (QAP)

Another popular permutation-based procedure, called the QAP, was introduced by Hubert [59], Hubert and Schultz [58], who adopted the procedure developed by Mantel [70]. Hubert [59] suggested to use edge set comparison product statistics to test for the association between networks. The formulation of the QAP is:

$$Q_{(\mathcal{A}, \mathcal{B})} = \frac{1}{N(N-1)} \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ij},$$

where the edge weights a_{ij} and b_{ij} are given by adjacency information, and the corresponding permuted QAP value has the form

$$Q_{(\mathcal{A}_\pi, \mathcal{B})} = \frac{1}{N(N-1)} \sum_{i=1}^n \sum_{j=1}^n a_{\pi(i)\pi(j)} b_{ij}.$$

This idea was further extended in sociology analysis by Krackhardt [64, 66]. Hubert and Schultz [58] provided a comparison between an approximated normal distribution and a sample cumulative distribution with a sample size of 2500.

2.4 MCUG: A Model Based Test Of Network Mutual Independence For Multiple Networks

Different from the permutation-based approach, Anderson et al. [4] proposed a model-based alternative to the permutation testing procedure for comparing two graphs, which is

known as the Conditional Uniform Graph (CUG) approach. In a more general scenario described in Butts [23], an extension of this approach, called Multivariate Cumulative Uniform Graph model (MCUG), was proposed to analyse the mutual independence between multiple networks. In this procedure, random graphs are generated from a graph model with fixed graph properties, such as size and density, with equal probability. A distribution of the measures of interest is formed with those networks that are randomly generated from particular graph models.

More specifically, let $\mathbf{t} = (t_1, \dots, t_P)$ be a vector (of size P) of real-valued functions and $\mathbf{x}$ be a vector corresponding to an observed value of $\mathbf{t}$. The choice of t_i could be defined by the network density, network size, etc. Let $\mathcal{G}$ be a graph set including all graphs of a particular feature. Then, the CUG probability distribution is given as:

$$P(G = g|\mathbf{t}, \mathbf{x}) = [\langle \{g' \in \mathcal{G} : \mathbf{t}(g') = \mathbf{x}\} \rangle]^{-1} I_{\{g' \in \mathcal{G} : \mathbf{t}(g') = \mathbf{x}\}}(g),$$

where $\langle . \rangle$ as in Section 2.1, denotes the size of a set.

In the multiple network case, let $G_1, G_2, \dots, G_P$ be a set of networks. Then, the multivariate CUG probability distribution is:

$$\begin{aligned} P((G_1, \dots, G_P) = (g_1, \dots, g_P) | \mathbf{t}, \mathbf{x}_1, \dots, \mathbf{x}_P) \\ = (\phi_{\mathbf{t}, \mathbf{x}}^P)^{-1} I_{\phi_{\mathbf{t}, \mathbf{x}}^P}(g_1, g_2, \dots, g_P). \end{aligned}$$

where

$$\phi_{\mathbf{t}, \mathbf{x}}^P = \langle \{(g'_1, \dots, g'_P) \in \mathcal{G}^P : \mathbf{t}(g'_1), \dots, \mathbf{t}(g'_P) = \mathbf{x}_1, \dots, \mathbf{x}_P\} \rangle$$

When comparing networks, we select a statistic t based on $\mathcal{G}^P$. Then, for the one tail p -value of $t(g_1, \dots, g_P)$,

$$P(t(G_1, \dots, G_P) \geq t(g_1, \dots, g_P) | \mathbf{t}, \mathbf{x}_1, \dots, \mathbf{x}_P),$$

is the upper tail and

$$P(t(G_1, \dots, G_P) \leq t(g_1, \dots, g_P) | \mathbf{t}, \mathbf{x}_1, \dots, \mathbf{x}_P),$$

is the lower tail.

When dealing with two networks, comparing the difference between two graph measures as an example, if we define t as the absolute difference $t(G_1, G_2) = |f(G_1) - f(G_2)|$, where f is the graph level, such as the centrality measure of the network with upper tail test, it becomes a two tailed test of equality of a particular graph measure. This could be illustrated by assessing whether the absolute difference between graph measures (f) is large compared to the expected distribution of absolute differences [4, 24]. Under the null hypothesis of mutual independence among networks, the probability for all possible networks belonging to the underlying network model is uniform. For this network generating (randomizing) procedure, it entails the assumption that edges in the network are independent of each other, which is difficult to justify in real applications [80]. Notice that, commonly, the numbers of possible networks in $\mathcal{G}$ (or $\mathcal{G}^P$) is far greater than the possible number of node permutations ($N!$). Therefore, this procedure may need to generate more random graphs than a standard permutation approach to obtain a reliable p -value.

Later, in a power study of two graphs comparison methods, CUG (MCUG with two graph comparison) is also used as an alternative method to compare to GHD; for details please refer to Section 3.3.

2.5 Multivariate Exponential Random Graph Permutation Model: An Inferential Procedure For Comparing Groups Of Networks

2.5.1 Exponential random graph model (ERG)

The exponential random graph (ERG) model is a parametric class of probability models for graphs. Suppose a graph $\mathcal{A}$ is observed from a set of possible graphs $\mathcal{G}$ with important features (such as the same node set, overall topology etc.). The graph $\mathcal{A}$ is considered to be a realization of some stochastic process related to $\mathcal{G}$. In general, we do not know what

stochastic process generated the observed network [90]. However, in this study, the graph set $\mathcal{G}$ is taken to be the graph set containing all possible permutation graphs of $\mathcal{A}$.

The idea of the ERG is to evaluate the probability distribution of structural statistics for an observed network with respect to the set of all possible structures of networks. When an extreme structural difference occurs between networks, it is more interesting to parameterize such a structure, rather than perform a dichotomous statistical decision analysis [23]. The first work that used the exponential random graph family as a probability model to find relationships among graphs is Holland and Leinhardt [54]. Frank and Strauss [40], Wasserman and Pattison [108] further extended this idea to test for structural independence among networks. Denote by $\mathbf{t} = (t_1, t_2, \dots, t_k)$ a vector of network structural characteristics, where k is the total number of features being considered. Possible structural characteristics could be edge degrees, number of triangles, etc. Denote by $\vec{\phi} = (\phi_1, \phi_2, \dots, \phi_k)$ a vector of parameters corresponding to structural characteristics of $\mathcal{A}$. The probability mass function can be written as

$$P(G = \mathcal{A}) = \frac{\exp(\vec{\phi}^T \mathbf{t}(\mathcal{A}))}{\sum_{\mathcal{A}' \in \mathcal{G}} \exp(\vec{\phi}^T \mathbf{t}(\mathcal{A}'))}, \quad (2.4)$$

where $\sum_{\mathcal{A}' \in \mathcal{G}} \exp(\vec{\phi}^T \mathbf{t}(\mathcal{A}'))$ is a normalizing constant, obtained by summing over all possible $G \in \mathcal{G}$. Equation (2.4) defines the probability of observing any particular graph $\mathcal{A}$. Notice that $\vec{\phi}^T \mathbf{t}(\mathcal{A})$ is a linear weighted combination of structural statistic values. In the case of a positive element t_i , this probability increases as ϕ_i increases.

For the multivariate case, let $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$ be a realization drawn from a distribution with finite joint support $\mathcal{G}_1 \otimes \mathcal{G}_2 \otimes \dots \otimes \mathcal{G}_P$. Then, the probability mass function of the joint distribution is given by

$$\begin{aligned} & P[(G_1, G_2, \dots, G_P) = (\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P)] \\ &= \frac{\exp(\vec{\phi}^T \mathbf{t}(\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P))}{\sum_{(\mathcal{A}'_1, \mathcal{A}'_2, \dots, \mathcal{A}'_P) \in \mathcal{G}_1 \otimes \mathcal{G}_2 \otimes \dots \otimes \mathcal{G}_P} \exp(\vec{\phi}^T \mathbf{t}(\mathcal{A}'_1, \mathcal{A}'_2, \dots, \mathcal{A}'_P))}. \end{aligned} \quad (2.5)$$

This multivariate exponential random graph model (MERG) is a direct generalization of the ERG model.

The ERG model provides a simple, yet flexible, family of probability distributions that can be used to analyse networks. This model has been extensively studied for analysing complex networks in the social science literature and there has been growing interest in ERG in recent years (for details please read Robins et al. [90] and references therein). However, there is not much work that uses the ERG model for the analysis of biological or brain networks. Saul and Filkov [93] described ERG models and demonstrated their utility in modelling the architecture of biological networks as a function of local structural features. Simpson et al. [96] used ERG to assess how a set of interacting local brain network features gives rise to global structure by creating brain networks that can capture the topological structure of an individual brain network. Simpson et al. [97] further extended this work to produce a brain network that was representative of a group of brain networks. The multivariate version of ERG (MERG) can be used to model multiple networks. Pattison and Wasserman [79], Wasserman and Pattison [108] used MERG to formulate models for multiple social networks by considering different kinds of theoretical claims about social structure. They further interpreted the models by developing structural models for several multiple social networks. Banks and Carley [6] defined two parameters for MERG to analyse the centres and dispersions of some particular graph measures for a group of networks.

2.5.2 Multivariate Exponential Random Graph Permutation (MERGP) model

In this subsection, we will describe a general class of models, termed MERGP, described in Butts [22], that can be used to evaluate the structural interdependence of two groups of networks or graphs.

The MERGP procedure defines an exponential random graph family of distributions on a set of permutation vectors whose sufficient statistics are given by product moments (of general correlation form) of the matrices to be compared [22]. Let $\pi \in \Pi$ be a permutation of $\{1, 2, \dots, N\}$ that indicates an ordering of nodes in the network. Denote by $\mathcal{A}_\pi = \{\mathcal{A}_{1\pi}, \mathcal{A}_{2\pi}, \dots, \mathcal{A}_{P\pi}\}$ the networks obtained by permuting the nodes of each network in $G(\mathcal{A})$ by π . Let the joint support $\mathcal{G}_1 \otimes \mathcal{G}_2 \otimes \dots \otimes \mathcal{G}_P$ contain all $N!$ possible graphs $\mathcal{A}_{k\pi}$ ($\pi \in \Pi$), with each of the $\mathcal{A}_k$ corresponding to the same permutation π , analogous for $\mathcal{B}_\pi = \{\mathcal{B}_{1\pi}, \mathcal{B}_{2\pi}, \dots, \mathcal{B}_{Q\pi}\}$. Define a matrix $\phi \in \mathcal{R}^{P \times Q}$. Let a_{kij}, b_{lij} be graph adjacency

indicators for networks $\mathcal{A}_k$ and $\mathcal{B}_l$. Then, the model is defined on the space of permutations by

$$P(\pi|G(\mathcal{A}), G(\mathcal{B}), \phi) = \frac{\exp(\sum_{k=1}^P \sum_{l=1}^Q \phi_{kl} \sum_{i=1}^N \sum_{j=1}^N a_{k\pi(i)\pi(j)} b_{lij})}{\mathcal{X}_{(\phi)}}, \quad (2.6)$$

where

$$\mathcal{X}_{(\phi)} = \sum_{\pi' \in \Pi} \exp(\sum_{k=1}^P \sum_{l=1}^Q \phi_{kl} \sum_{i=1}^N \sum_{j=1}^N a_{k\pi'(i)\pi'(j)} b_{lij}) \quad (2.7)$$

is a normalizing constant. This formula defines the probability of a particular permutation π , given two observed sets of networks $G(\mathcal{A})$, $G(\mathcal{B})$ and specified parameter ϕ . Under a null hypothesis of independence between networks in groups $G(\mathcal{A})$ and $G(\mathcal{B})$ (i.e. $\phi = 0$), the probability of any permutation is equal to $\frac{1}{N!}$. The notion of parameterizing permutations by a discrete exponential family has been applied for some time [31, 83]. The most common approach [83] parameterizes the permutation in terms of conditional probabilities of winning with N contestants (with probabilities of winning equal to $p_i, i = 1, \dots, N$) and in terms of order statistics [51].

Notice that a positive value of ϕ_{ij} indicates a positive association between graphs $\mathcal{A}_i$ and $\mathcal{B}_j$, while $\phi_{ij} < 0$ indicates a negative association [22]. This can also be intuitively understood by comparing two permutations π_1 and π_2 . For ease of interpretation, consider the case of comparing two networks. Then,

$$\begin{aligned} & \frac{P(\pi_1|\mathcal{A}, \mathcal{B}, \phi)}{P(\pi_2|\mathcal{A}, \mathcal{B}, \phi)} \\ &= \frac{\exp(\phi \sum_{i=1}^N \sum_{j=1}^N a_{\pi_1(i)\pi_1(j)} b_{ij})}{\exp(\phi \sum_{i=1}^N \sum_{j=1}^N a_{\pi_2(i)\pi_2(j)} b_{ij})} \\ &= \exp(\phi \sum_{i=1}^N \sum_{j=1}^N (a_{\pi_1(i)\pi_1(j)} b_{ij} - a_{\pi_2(i)\pi_2(j)} b_{ij})). \end{aligned} \quad (2.8)$$

This ratio indicates that ϕ acts as a weight on the edgewise product moment difference. This means that if two networks $\mathcal{A}$ and $\mathcal{B}$ are positively associated, then a permuted net-

work with more edge overlap will have larger probability (e.g., if $\phi > 0$ and $\mathcal{A}_{\pi_1}$ and $\mathcal{B}_{\pi_1}$ have more edge overlap than $\mathcal{A}_{\pi_2}$ and $\mathcal{B}_{\pi_2}$, then $\frac{P(\pi_1|\mathcal{A},\mathcal{B},\phi)}{P(\pi_2|\mathcal{A},\mathcal{B},\phi)} > 1$). Hence, the parameter ϕ also provides a measure of network association. In Section 5.3, we will show that the maximum likelihood estimates for parameter ϕ are equivalent to the normalized generalized correlation coefficient under the assumption that the permutation distribution of $\sum_{k=1}^P \sum_{l=1}^Q \phi_{kl} \sum_{i=1}^N \sum_{j=1}^N a_{k\pi(i)\pi(j)} b_{lij}$ is normal.

A direct evaluation of $\mathcal{X}_{(\phi)}$ is infeasible, since computing the exact value requires summing over all $N!$ possible permutations. Therefore, calculation of the Maximum Likelihood Estimator (MLE) $\hat{\phi}_{ij}$ requires the use of an approximation. To estimate the MLE, Butts [22] provided a two step procedure involving: (1) Maximum pseudo-likelihood estimation and (2) Markov chain Monte Carlo (MCMC). The maximum likelihood estimate $\hat{\phi}_{ij}$ indicates the strength of correspondence between $\mathcal{A}_i$ and $\mathcal{B}_j$. The statistical significance of observing an extreme structural difference between groups of networks could be assessed by applying log likelihood ratio test (with null: $\phi = 0$).

This inferential procedure consists of two steps. The first step is to perform an initial estimation of ϕ based on a subset of Π . The second step uses the initial estimate of ϕ and further approximates the true MLE using the likelihood ratio. At the first step, Butts [22] used the maximum pseudo-likelihood estimation to obtain a rough estimate of the parameter ϕ (called the Maximum Pseudo-likelihood Estimator, MPLE, and denoted by $\tilde{\phi}$). This estimation is performed using only the set of all one step permutations (such a permutation will, for instance, swap the indices i and j and leave other indices unchanged). In the second step, Butts [22] showed that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \left\{ \exp \left(\sum_{k=1}^P \sum_{l=1}^Q (\tilde{\phi}_{kl} - \phi_{kl}) \sum_{i=1}^N \sum_{j=1}^N a_{k\pi(i)\pi(j)} b_{lij} \right) \right. \\ & \left. \left[\frac{1}{n} \sum_{m=1}^n \exp \left(\sum_{k=1}^P \sum_{l=1}^Q (\tilde{\phi}_{kl} - \phi_{kl}) \sum_{i=1}^N \sum_{j=1}^N a_{k\pi_m(i)\pi_m(j)} b_{lij} \right) \right]^{-1} \right\} \\ &= \frac{P(\pi|\tilde{\phi}, G(\mathcal{A}), G(\mathcal{B}))}{P(\pi|\phi, G(\mathcal{A}), G(\mathcal{B}))}, \end{aligned}$$

where permutations $\pi_1, \pi_2, \dots, \pi_n$ are drawn from the permutation model (2.6) with parameter ϕ . Then, the approximate MLE can be computed by first drawing a sample from the

initial model (e.g., with parameter $\tilde{\phi}$), and then by maximizing the Monte Carlo approximation to $\frac{P(\pi|\tilde{\phi}, G(\mathcal{A}), G(\mathcal{B}))}{P(\pi|\phi, G(\mathcal{A}), G(\mathcal{B}))}$ using standard methods such as Newton-Raphson, etc. This two-step procedure is highly reliant on the MPLE estimate being close to the true MLE with reasonable precision. If the MPLE is not very well estimated, then the convergence of the second step can be slow and will need further iterations. The properties of the pseudo-likelihood estimator are not well understood. It is still not clear from existing research as to when pseudo-likelihood estimates may provide an acceptable approximation to the MLE [90]. Therefore, the procedure just described is potentially unreliable. Also, when implementing multiple comparisons, the computational cost may be substantial [23].

2.6 A Brief Introduction To Mathematical Notations

In this section we will briefly introduce mathematical notation that will be used in this thesis.

Limit is defined as a value that a function or sequence approaches as the input or index approaches some value. More formally, suppose f is a real-valued function and c is a real number,

$$\lim_{x \rightarrow c} f(x) = M,$$

indicates that by making x sufficiently close to c , $f(x)$ can be made to be as close to M as desired.

The **Big O notation** describes the limiting behavior of a function when it tends towards a particular value or in terms of simpler functions (such as N). Suppose $f(x)$ and $g(x)$ are two functions defined on some subset of the real numbers. We say,

$$f(x) = O(g(x)),$$

if and only if there exist some constants c and n such that,

$$|f(x)| \leq c|g(x)| \text{ for all } x \geq n,$$

where $|c|$ denote the absolute value of c . Intuitively, this indicates f does not grow faster

than g .

The **Weak Law of Large Numbers** states that the average of observations converges towards the expected value in probability. Let $X_1, \dots, X_n$ be a sequence of independent and identically distributed random variables, each having a mean $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2 < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1,$$

as $n \rightarrow \infty$ or alternatively,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0,$$

as $n \rightarrow \infty$.

Chapter 3

GHD: A Permutation Based Test Of Independence

In this chapter, we present a test of independence between a pair of networks. Traditional permutation-based tests of independence have two major issues. One is the choice of graph measure (i) that represents the graph topology. Previous permutation-based methods, such as MAD and QAP, commonly apply adjacency information to compare structure [41, 21, 106, 59, 58]. However, the adjacency information only reflects the direct neighbourhoods of a network, which is not sufficiently powerful to capture subtle changes in network structure. Another issue is the computational cost (ii) of the permutation procedure. In practice, to obtain a reasonable precision, the evaluation of the permutation distribution for large graphs is too computationally demanding to be feasible.

In this chapter, we address (i) by proposing a distance measure called Generalized Hamming Distance (GHD), which can apply various edge weights. We introduce an edge weight, i.e., the Topological Overlap (TO) measure, which takes into account higher-order neighbourhood structures, and, hence, has more power to detect differential structures while being amenable to statistical analysis. For (ii), we show that by applying TO and adjacency to the GHD with three random graph models under the null hypothesis of independence, the sampling distribution of the GHD test statistic approaches a standard normal distribution as the size of the networks increases. Thus, p -values can be obtained without resorting to computationally expensive permutation procedures.

The structure of this chapter is as follows. In Section 3.1, we introduce the GHD and TO measures. We show that GHD in combination with TO is a superior measure of topological change than the commonly used Hamming Distance. In Section 3.2, we describe the inferential testing procedure based on GHD, and prove that the sampling distribution of the GHD test statistic approaches a standard normal distribution for TO and adjacency under three random graph models. Simulation results are presented in Section 3.3. Finally, we summarize this chapter in Section 3.4.

3.1 The Generalized Hamming Distance

Once again, we denote two graphs $\mathcal{A} = (\mathcal{V}, \mathcal{E}_A)$ and $\mathcal{B} = (\mathcal{V}, \mathcal{E}_B)$. Both graphs are defined on a common label set $\mathcal{L} = \{1, 2, \dots, N\}$, for node set $\mathcal{V}$, and different edge sets $\mathcal{E}_A$ and $\mathcal{E}_B$, respectively, indicating the connection between the nodes. The matrices $A = (A_{ij})$ and $B = (B_{ij})$ denote the two $(N \times N)$ adjacency matrices associated with the graphs $\mathcal{A}$ and $\mathcal{B}$, respectively. The GHD is a Euclidean-like distance measure, which is defined as

$$\text{GHD}(\mathcal{A}, \mathcal{B}) = \frac{1}{N(N-1)} \sum_{i,j} (a_{ij} - b_{ij})^2, \quad (3.1)$$

where a_{ij} and b_{ij} are edge weights that are extracted from the networks, thereby providing a measure of connectivity between every pair of nodes i and j in $\mathcal{A}$ and $\mathcal{B}$, respectively. Here, $\sum_{i,j}$ denotes summation over distinct i and j . When a_{ij} and b_{ij} are the binary values indicating the presence or absence of an edge, i.e., they are the elements of the adjacency matrices, then $\text{GHD}(\mathcal{A}, \mathcal{B})$ is related to the Hamming distance, i.e. edge difference between networks (see Section 2.3.2 on page 32). In this work, for notational simplicity we refer to $\text{GHD}(\mathcal{A}, \mathcal{B})$ with adjacency information as the Hamming Distance (HD).

The weights we propose here, in contrast to the HD, quantify the TO between any two nodes, by taking into account the local neighbourhood structure around the nodes [56]. A one-step TO [56] between two nodes i and j indicates that they share a direct connection to other nodes. It can be computed directly from the adjacency matrix as follows:

$$a_{ij} = \frac{\sum_{u \neq i,j} A_{iu} A_{uj} + A_{ij}}{\min(\sum_{u \neq i} A_{iu} - A_{ij}, \sum_{u \neq j} A_{iu} - A_{ij}) + 1}, \quad (3.2)$$

when $i \neq j$, and otherwise $a_{ij} = 1$. An analogous set of weights b_{ij} is defined for the second graph. Note that the term $\sum_{u \neq i, j} A_{iu}A_{uj}$ is a count of all node triples $(i - u - j)$ involving node pair (i, j) . Hence, this term measures the connectivity information of (i, j) node pairs plus their common one-step neighbours. Holland and Leinhardt [53] argued that the frequency of occurrence of connections involving three nodes in the observed structure could not be explained by the connection structure of the lower-order two nodes using adjacency alone. Hence, this measure considered patterns of three nodes, which reveal more information than the commonly used adjacency information. The denominator in (3.2) can be written as $\min(d_i, d_j) + 1 - A_{ij}$, where d_i and d_j represent the degrees of nodes i and j , respectively:

$$a_{ij} = \frac{\sum_{u \neq i, j} A_{iu}A_{uj} + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}}. \quad (3.3)$$

Since

$$\sum_{u \neq i, j} A_{iu}A_{uj} + A_{ij} \leq \min\left\{\sum_{u \neq i} A_{iu} = d_i, \sum_{u \neq j} A_{ju} = d_j\right\},$$

a_{ij} is roughly equal to the smaller of the node degrees, such that $0 \leq a_{ij} \leq 1$. Then a large discrepancy between a_{ij} and b_{ij} in (3.1) indicates a topological difference localised around that pair of nodes. The computational complexity of an evaluation of this distance lies in the calculation of the numerator. In the numerator, $\sum_{l \neq i, j} A_{il}A_{jl}$ for all pairs (i, j) can be computed by matrix multiplication $A'A$. The evaluation of $A'A$ requires a computational complexity of around $O(N^3)$. However, in practice, the adjacency matrix A can be stored and manipulated in the form of a sparse matrix. Hence, each element $\sum_{l \neq i, j} A_{il}A_{jl}$ is actually obtained by a sparse matrix-matrix multiplication. For a relatively sparse graph, the computational complexity is therefore actually much smaller than $O(N^3)$.

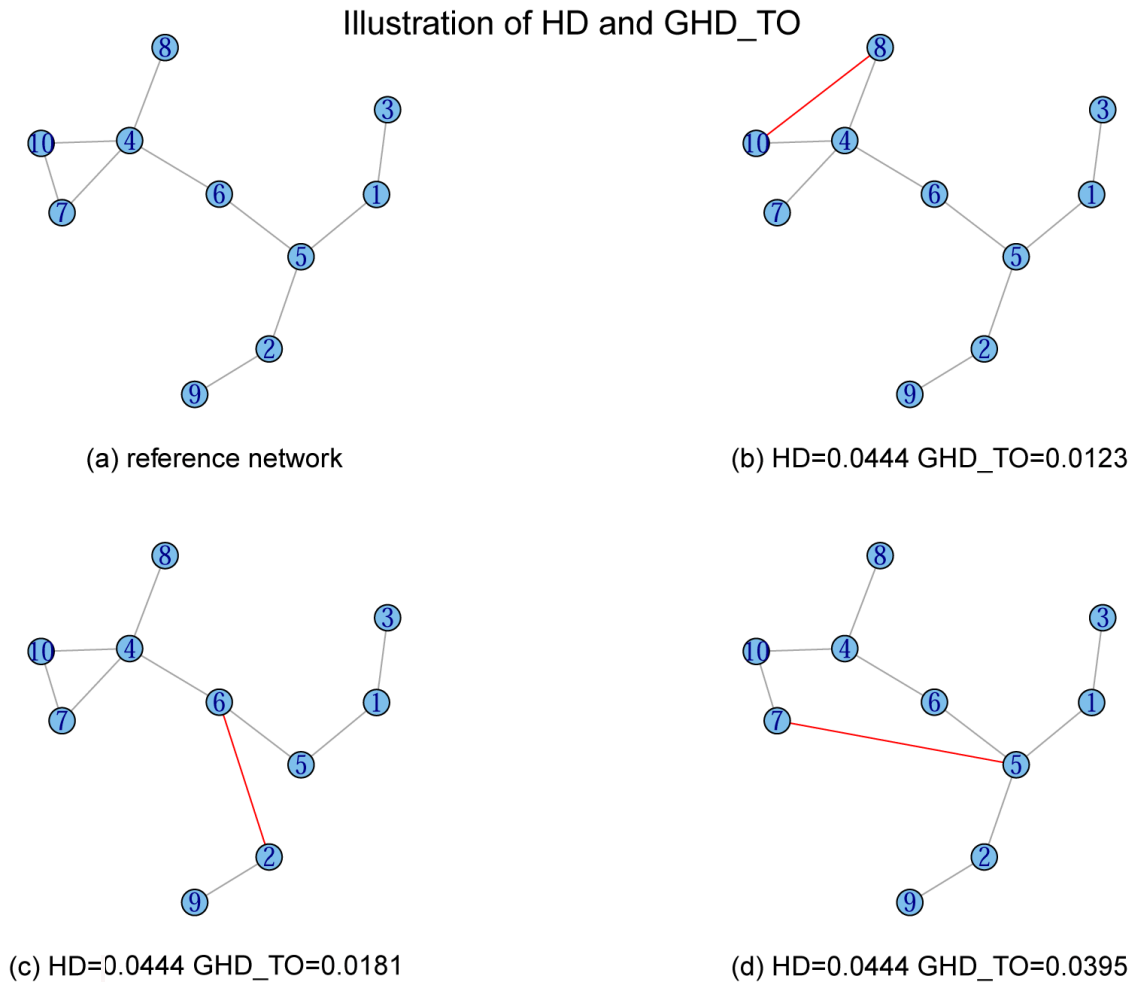


Figure 3.1: Example of HD: GHD with adjacency and GHD_TO, GHD with TO. Networks (b), (c) and (d) are generated from the reference network (a) by a single edge change. Both HD and GHD_TO between the reference network and each modified paired network have been computed in each case

In the literature, the TO measure has been successfully used to detect communities in biological networks, because highly integrated modules are expected to have high TO with their neighbours [56]. There is strong empirical evidence that the TO measure carries biological meaning as a similarity measure for biological networks [111, 3], although it has not been used in the context of deriving a distance between two paired networks. For instance, in gene interaction networks, genes with high topological overlap were found to

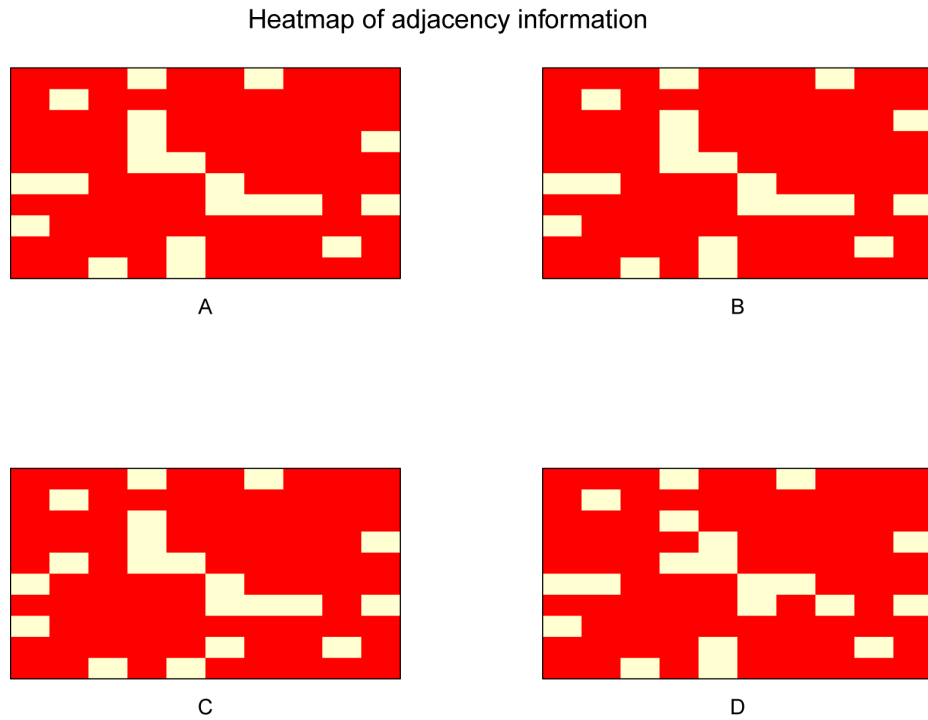


Figure 3.2: Heatmap of with adjacency of Figure 3.1 where all (B) (C) (D) have one grid difference with (A)

have an increased chance of being part of the same biological pathway.

The TO measure relies on the principle that the more neighbours that are shared between two nodes, the tighter their relationship [56]. As such, the GHD distance with the TO measure can detect subtle changes, with higher sensitivity compared to the use of adjacency information. An illustration of this is given in Figure 3.1. Four different networks are shown: (a) is the reference network, and the other three networks are all obtained from (a) by changing an edge (red). Each one of these three networks is compared to (a) using both the HD (with adjacency) and the GHD_TO constructed from the TO measure. The values of the HDs are the same for all three changes. However, the values of the GHD_TO reveal different levels of change. For $(a) \rightarrow (b)$, changes were made to nodes with small node degrees and also at the side portion of the network, which results in only two triangle differences ($4 - 7 - 10$ and $4 - 8 - 10$), and, hence, the GHD_TO exhibited the smallest distance. For a $(a) \rightarrow (c)$ comparison, a change was made to more centred nodes with

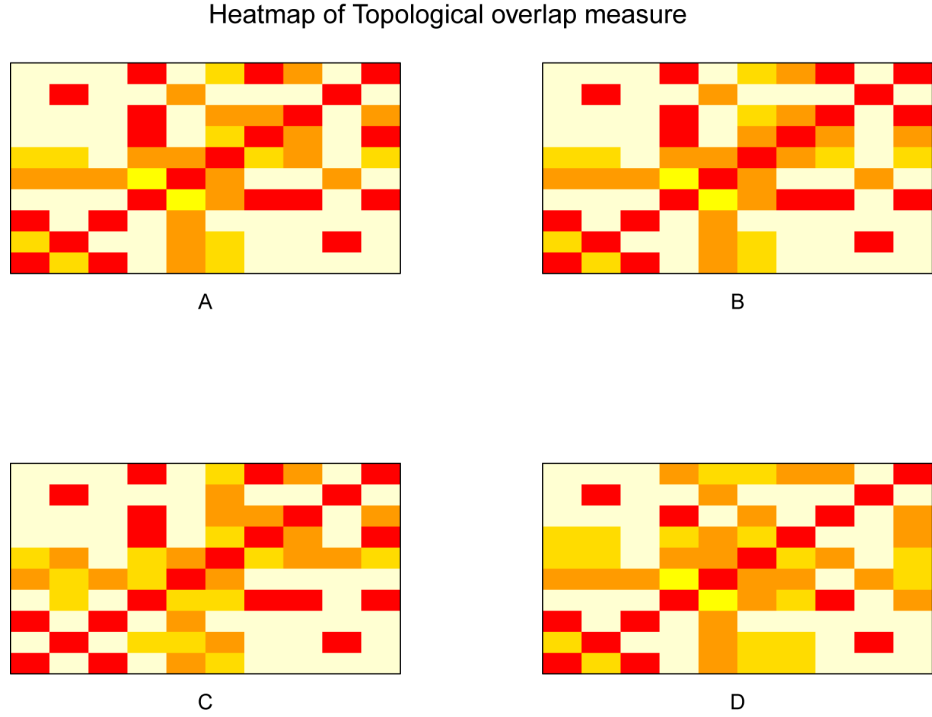


Figure 3.3: Heatmap of TO matrix of Figure 3.1. TO measure response more to different one edge change than adjacency. Heatmap of (A) and (B) are much similar then two others, this is because (A) and (B) are graph isomorphism.

slightly higher node degrees. Although there are six triangle differences, the center-right portion remains a tree, hence the GHD_TO value is only slightly higher than $(a) \rightarrow (b)$. Finally, for $(a) \rightarrow (d)$, a change was made to the nodes with the highest node degrees, and the center portion forms a big loop whose topology is largely changed. This comparison has eight triangle differences and significant topology changes, hence it has the largest distance. Or intuitively, we can observe their edge weight pattern by looking at the heat map of the edge weight matrix. Figure 3.2 and Figure 3.3 show the heat map of the edge weight matrix for the adjacency matrix and TO measure matrix, respectively. For GHD_TO, the color changes are much more diverse than those of the adjacency information. Clearly, the HD is not able to differentiate between changes in topology, whereas the scale of the observed GHD_TO values depends on the amount of change that can be seen. Additional Monte Carlo simulations are presented in Section 3.3.

3.2 Statistical Inference For GHD

Having introduced the GHD, we would like to be able to compute the probability that a distance as extreme or more extreme than the observed GHD value could have been observed by chance under the assumption that the two graphs are independent. By treating the GHD as a random variable with unknown distribution, this probability could in principle be computed non-parametrically using a standard permutation testing procedure. Both graphs $\mathcal{A}$ and $\mathcal{B}$ are defined on a common set of label $\mathcal{L} = \{1, 2, \dots, N\}$. The procedure consists of permuting the elements of $\mathcal{L}$ while keeping the edges unchanged, thus obtaining a permuted network $\mathcal{A}_\pi$. A permutation is a bijection of $\mathcal{L}$ onto itself, and all the $N!$ possible permutations are collected in a set Π . We denote $a_{\pi(i)\pi(j)}$ as a measure of edge weight for nodes i and j after permutation $\pi \in \Pi$. The corresponding permuted GHD value has the form

$$\text{GHD}_\pi(\mathcal{A}_\pi, \mathcal{B}) = \frac{1}{N(N-1)} \sum_{i,j} (a_{\pi(i)\pi(j)} - b_{ij})^2,$$

which is discussed in detail below. This is evaluated for every $\pi \in \Pi$ to infer the exact permutation distribution of GHD under the null hypothesis. This distribution can then be used to evaluate p -values.

Despite the usefulness of this approach, in practice, the evaluation of $\text{GHD}_\pi(\mathcal{A}_\pi, \mathcal{B})$ for large graphs is too computationally demanding to be feasible. Achieving a sufficient coverage probability for a confidence interval of the true mean of the GHD type statistic would require a very large number of permutations. For instance, for a network of size $N = 100$, more than 10 million permutations would be needed to achieve confidence intervals of the same length as that which is achievable by an asymptotic test with 95% coverage probability [73].

In what follows, we propose an alternative approach that removes the need to carry out computationally expensive permutation testing. We demonstrate that under the null hypothesis of independence between the graphs, the exact GHD permutation distribution can be well approximated by a normal distribution, assuming the networks are sufficiently large. First, we observe that the distance can be rewritten in an equivalent form in terms of

a generalised correlation coefficient of the form $\sum_{i,j} a_{ij} b_{ij}$ as follows:

$$\begin{aligned}
 \text{GHD}_\pi(\mathcal{A}, \mathcal{B}) &= \frac{1}{N(N-1)} \sum_{i,j} (a_{\pi(i)\pi(j)} - b_{ij})^2 \\
 &= \frac{1}{N(N-1)} \sum_{i,j} (a_{\pi(i)\pi(j)}^2 + b_{ij}^2) - \frac{2}{N(N-1)} \sum_{i,j} a_{\pi(i)\pi(j)} b_{ij} \\
 &= c - \frac{2}{N(N-1)} \sum_{i,j} a_{\pi(i)\pi(j)} b_{ij}
 \end{aligned} \tag{3.4}$$

where c is a constant that does not change under permutations.

Now we define the mean-centred connectivity weights, a'_{ij} and b'_{ij} of a_{ij} and b_{ij} ,

$$a'_{ij} = a_{ij} - \frac{1}{N(N-1)} \sum_{i,j} a_{ij}, \quad b'_{ij} = b_{ij} - \frac{1}{N(N-1)} \sum_{i,j} b_{ij}.$$

By replacing a_{ij} and b_{ij} by mean-centered edge weights in (3.4),

$$\begin{aligned}
 \text{GHD}_\pi(\mathcal{A}, \mathcal{B}) &= c - \frac{2}{N(N-1)} \sum_{i,j} \left(a'_{\pi(i)\pi(j)} + \frac{\sum_{i,j} a_{\pi(i)\pi(j)}}{N(N-1)} \right) \left(b'_{ij} + \frac{\sum_{i,j} b_{ij}}{N(N-1)} \right) \\
 &= c - \frac{2}{N(N-1)} \sum_{i,j} \left[a'_{\pi(i)\pi(j)} b'_{ij} + b'_{ij} \frac{\sum_{i,j} a_{\pi(i)\pi(j)}}{N(N-1)} + a'_{\pi(i)\pi(j)} \frac{\sum_{i,j} b_{ij}}{N(N-1)} \right. \\
 &\quad \left. + \frac{\sum_{i,j} a_{\pi(i)\pi(j)}}{N(N-1)} \frac{\sum_{i,j} b_{ij}}{N(N-1)} \right] \\
 &= c - \frac{2}{N(N-1)} \sum_{i,j} a'_{\pi(i)\pi(j)} b'_{ij} + \frac{\sum_{i,j} b'_{ij} \sum_{i,j} a_{\pi(i)\pi(j)}}{N(N-1)} \\
 &\quad + \frac{\sum_{i,j} a'_{\pi(i)\pi(j)} \sum_{i,j} b_{ij}}{N(N-1)} + \sum_{i,j} \frac{\sum_{i,j} a_{\pi(i)\pi(j)}}{N(N-1)} \frac{\sum_{i,j} b_{ij}}{N(N-1)}.
 \end{aligned}$$

Since a'_{ij} and b'_{ij} are mean-centered measure, then $\sum_{i,j} a'_{ij} = \sum_{i,j} b'_{ij} = 0$. Also $\frac{\sum_{i,j} a_{\pi(i)\pi(j)}}{N(N-1)}$ and $\frac{\sum_{i,j} b_{ij}}{N(N-1)}$ are constants under permutation. Therefore,

$$\text{GHD}_\pi(\mathcal{A}, \mathcal{B}) = c_1 - \frac{2}{N(N-1)} \sum_{i,j} a'_{\pi(i)\pi(j)} b'_{ij}, \tag{3.5}$$

where c_1 is constant under permutation.

By making use of this alternative representation, we are able to exploit well-known conditions for asymptotic normality that, in practice, can be easily checked empirically. For a generalised correlation coefficient of the form (3.5), Daniels [30] showed that the approximate joint distribution of two quantities of the form (3.5) is normal if conditions for both a'_{ij} 's and b'_{ij} 's hold (see below). From the proof of [30], asymptotic normality holds true if the following conditions are obtained for both a'_{ij} and b'_{ij} (note, a sufficient condition):

$$\begin{aligned} \sum_{i,j} a'_{ij} &= \sum_{i,j} b'_{ij} = 0 \quad \text{and} \\ \sum_{i,j,k} a'_{ij} a'_{ik} &= O(N^3), \quad \sum_{i,j,k} b'_{ij} b'_{ik} = O(N^3). \end{aligned}$$

However, both [82] and [41] point out that the second condition given by [30] is far too strong. Pham et al. [82] and Friedman and Rafsky [41] both provide a weaker version of the above condition. Now the exact permutation distribution is asymptotically normal under two conditions [30, 41, 82]:

$$\sum_{i,j} a'_{ij} = \sum_{i,j} b'_{ij} = 0 \quad \text{and} \tag{3.7a}$$

$$\lim_{N \rightarrow \infty} \frac{[\sum_{ijkl} a'_{ij} a'_{ik} a'_{il}]^2}{[\sum_{ijk} a'_{ij} a'_{ik}]^3} = \lim_{N \rightarrow \infty} \frac{[\sum_{ijkl} a'_{ij} a'_{ik} a'_{il}]^2}{[\sum_{ijk} a'_{ij} a'_{ik}]^3} = 0. \tag{3.7b}$$

Condition (3.7a) follows directly from the definition of a'_{ij} and b'_{ij} as being mean-centred, thus it is always satisfied. Notice that because of the first condition, the substituted a'_{ij} and b'_{ij} must be a mean-centered measure. In a proof of normality for the General Correlation Coefficient (GCC) with a K-minimum spanning tree graph, Friedman and Rafsky [41] directly substituted in the adjacency information to (3.7b), rather than a centred version. Pham et al. [82] stated that centering can cause the ratio of the numerator and denominator to vanish for constant row sums. Additionally, they further provided a condition of exchangeability between raw and centred data:

$$\limsup \frac{(\sum_{ij} a_{ij})^2}{N \sum_i a_i^2} < 1, \tag{3.8}$$

where $a_{i\cdot} = \sum_j a_{ij}$.

For the direct application of an adjacency matrix in [41], Pham et al. [82] point out that some of the K-minimum spanning tree graphs, which violated the condition of exchangeability, will mistakenly accept condition (3.7b) (for details please refer to [82]). Therefore, the proof of normality of the permutation distribution of GCC for K-minimum spanning graphs in Friedman and Rafsky [41] was, to some extent, incomplete.

To gain some insight into the meaning of condition (3.7b) in our context, it is instructive to consider the case where a_{ij} and b_{ij} are elements of the two adjacency matrices, i.e., they indicate the presence of an edge. By defining $a_{i\cdot} = \sum_{j \neq i} a_{ij}$ and $\bar{a} = \frac{1}{N} \sum_i a_{i\cdot}$, we have

$$a'_{i\cdot} = \sum_{j \neq i} a'_{ij} = a_{i\cdot} - \bar{a}, \quad (3.9)$$

and condition (3.7b) can be written in an analogous form as

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (Na'_{i\cdot})^3]^2}{[\sum_i (Na'_{i\cdot})^2]^3} = \lim_{N \rightarrow \infty} \frac{[\sum_i (a_{i\cdot} - \bar{a})^3]^2}{[\sum_i (a_{i\cdot} - \bar{a})^2]^3} = 0. \quad (3.10)$$

It can be observed that when using the adjacency matrix, $a_{i\cdot}$ represents the degree of the i^{th} node. An analogous condition is observed for the quantities b_{ij} . Therefore, checking (3.7b) amounts to computing the degree of each node in the two networks, and assessing the limiting behaviour. When the TO measure is used instead, the coefficient $a_{i\cdot}$ represents the overall TO information at node i , which can also be analyzed using (3.10).

When both (3.7a) and (3.7b) hold true, the permutation distribution of $\text{GHD}_\pi(\mathcal{A}_\pi, \mathcal{B})$ is approximately normal. For convenience, it is easier to standardise the GHD value by mean-centering and normalizing it so that it follows a standard normal distribution asymptotically, namely

$$\frac{\text{GHD}_\pi(\mathcal{A}_\pi, \mathcal{B}) - \mu_\pi}{\sigma_\pi} \sim N(0, 1). \quad (3.11)$$

Here, μ_π and σ_π are the mean and standard deviation of GHD under the exact permutation distribution, respectively. These two moments can be computed explicitly and in closed-form by enumerative combinatoric calculations. Detailed derivations of these two moments can be found in Appendix A.1, and follow similar developments described in the context of

related permutation-based testing procedures [70]. With the formula for the first two exact moments, a corresponding p -value can therefore be efficiently computed from the normal approximation, even for very large networks.

3.2.1 Asymptotic normality under three graph models

It is not uncommon to assume that observed networks follow a particular random graph model. In this section we demonstrate that conditions (3.7a) and (3.7b) always holds true when the networks being compared are assumed to be generated by three common random network models: the Erdős-Rényi (ER), Random Geometric (RG), and Scale Free (SF) models.

In this work, we showed that for ER and RG, (3.10) holds for both HD and GHD_TO. For a SF network, the asymptotic normality property also holds when the power exponent in equation (2.2) (page 29) falls in the range $1 < \alpha < 2$, or $\alpha \geq 3$.

Before starting the proof, consider the case where a_{ij} and b_{ij} are entries of the adjacency matrices. Then (3.10) becomes

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = 0, \quad (3.12)$$

where $\bar{d}$ is the average node degree. In order to evaluate the above limit, we exploit the fact that both the numerator and denominator in (3.12) are powers of centralised empirical moments of the node degree distribution. Our strategy consists of establishing a connection between empirical and population moments, which are available in closed-form under each of the random network models considered, and then study the limiting behaviour.

3.2.2 Normality when using the HD for SF network

Recall the scale free network discussed in Section 2.2.3. The power law node degree distribution (2.2) is,

$$P(d_i = k) = ck^{-\alpha}, \quad k = m, m+1, \dots, K,$$

where m and K are lower and upper cut-offs of the node degree, respectively, c is a normalising constant, and α represents the power exponent. The lower cut-off m is generally taken to be 1. For the upper cut-off K , when $\alpha > 2$, K is generally specified as $K = N^{\frac{1}{\alpha-1}}$ and $K = N - 1$ for $1 < \alpha \leq 2$.

We let $\mu_s = c \sum_{d=1}^K d^{s-\alpha}$ denote the s^{th} theoretical moment of the node degree distribution and $m_s = \frac{1}{N} \sum_{i=1}^N d_i^s$ the corresponding empirical moment. From (3.12), we see that we are only interested in the order of magnitude of the empirical moments. In order to study the limit above we need to characterise the order of m_s , for $s = 1, 2, 3$, as N increases. Our strategy here consists of first characterising the order of μ_s asymptotically, for the first three moments, and then establishing a correspondence with the order of the empirical moments. Later, in the proofs of normality for other graph models and edge weights, we will repeatedly apply this strategy.

We start by examining the order of μ_s , for $s = 1, 2, 3$, in the limit. Since for different values of the power exponent s , the order of μ_s are different, we consider three distinct cases where the order of μ_s can be discussed:

- (a) $s - \alpha + 1 < 0$,
- (b) $s - \alpha + 1 = 0$,
- (c) $s - \alpha + 1 > 0$.

For case (a), the order of μ_s is

$$\sum_{d=1}^K \frac{1}{\alpha - 1} d^{-1} = O(1).$$

For case (b), the order of μ_s is given by the well-known result that

$$\sum_{d=1}^K d^{-1} = O(\ln(K)).$$

Now consider case (c), as $K \rightarrow \infty$. We apply the Euler-Maclaurin formula to obtain an

alternative expression for the sum

$$\sum_{d=1}^K d^{s-\alpha} = K^{s-\alpha+1} + (\alpha - s) \int_1^K \frac{\lfloor x \rfloor}{x^{\alpha-s+1}} dx + O(1),$$

where $\lfloor x \rfloor$ denotes the largest integer that is not greater than x . To compute the order of $\sum_{d=1}^K d^{s-\alpha}$, we need to know which term dominates in order. By applying l'Hospital's rule we have

$$\lim_{K \rightarrow \infty} \frac{s \int_1^K \frac{\lfloor x \rfloor}{x^{\alpha-s+1}} dx}{K^{s-\alpha+1}} = \frac{s}{s - \alpha + 1},$$

a finite constant. It follows that both terms in the expression are of the same order, and therefore μ_s will have the same order as $K^{s-\alpha+1}$. Hence for any $s - \alpha + 1 > 0$, μ_s will have the same order as $K^{s-\alpha+1}$. We then have

$$\mu_s \approx \begin{cases} O(K^{s-\alpha+1}), & s - \alpha + 1 > 0, \\ O(\ln K), & s - \alpha = -1, \\ O(1), & s - \alpha < -1. \end{cases} \quad (3.13)$$

So, for a SF network conditions for asymptotic normality of the permutation distribution may differ according to the value of α .

Consider first the case $1 < \alpha < 2$, for which $K = N - 1$. The order of the theoretical moments is

$$\begin{cases} \mu_1 \approx O(N^{2-\alpha}), \\ \mu_2 \approx O(N^{3-\alpha}), \\ \mu_3 \approx O(N^{4-\alpha}). \end{cases}$$

Now consider the s th empirical moment of the node degree observed from the SF network,

$$m_s = \frac{1}{N} \sum_{i=1}^N d_i^s.$$

We now try to compare the orders of μ_s and m_s in the limit. Specifically, we assess whether the order of each μ_s established above also hold true for the corresponding m_s . This can be verified by checking that

$$\lim_{N \rightarrow \infty} \frac{m_s}{\mu_s} = c_s, \quad s = 1, 2, 3, \quad (3.14)$$

for some positive constant c_s . To show the above limit, we apply the Weak Law of Large Numbers (WLLN). For WLLN to hold, μ_s must be finite. Hence we first transform d_i , such that μ_s after transformation is finite. We show that the orders of the empirical moments m_s , $s = 1, 2, 3$, are the same as those of the theoretical moments. Define $z_{si} = d_i / N^{\frac{s+1-\alpha}{s}}$. The distribution of z_{si} is specified by:

$$P(z_{si} = z) = c' z^{-\alpha}, \quad z = 1/N^{\frac{s+1-\alpha}{s}}, 2/N^{\frac{s+1-\alpha}{s}}, \dots, K/N^{\frac{s+1-\alpha}{s}},$$

where $c' = cN^{\frac{s+1-\alpha}{s}}$. Thus the s^{th} theoretical moment of z_{si} is

$$\mu_{zs} = c' \sum_z z^{s-\alpha} = c' \sum_d \left[\frac{d}{N^{\frac{s+1-\alpha}{s}}} \right]^s d^{-\alpha} = \frac{\mu_s}{N^{s+1-\alpha}},$$

which is finite. Denoting by m_{zs} the s^{th} empirical moment of z_{si} , $i = 1, \dots, N$, we have

$$m_{zs} = \frac{1}{N} \sum_{i=1}^N z_{si}^s = \frac{1}{N} \sum_{i=1}^N \left(\frac{d_i}{N^{\frac{s+1-\alpha}{s}}} \right)^s = \frac{m_s}{N^{s+1-\alpha}}.$$

Now since μ_{zs} is finite and we suppose $z_{s1}, z_{s2}, \dots, z_{sN}$ are N independent, identically distributed random variables, according to the Weak Law of Large Numbers (WLLN), m_{zs} converges to μ_{zs} in probability, and hence we have:

$$1 = \lim_{N \rightarrow \infty} \frac{m_{si}}{\mu_{si}} = \lim_{N \rightarrow \infty} \frac{\frac{m_s}{N^{s+1-\alpha}}}{\frac{\mu_s}{N^{s+1-\alpha}}} = \lim_{N \rightarrow \infty} \frac{m_s}{\mu_s}.$$

This indicates that m_s and μ_s are of the same order. Now we can approximate the order of

the numerator and denominator of condition (3.12):

$$\begin{aligned}
 \sum_i (d_i - \bar{d})^3 &= \sum_i (d_i^3 - 3d_i^2\bar{d} + 3d_i\bar{d}^2 - \bar{d}^3) \\
 &= N(m_3 - 3m_2m_1 + 2m_1^3) \\
 &\approx N(\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3) \\
 &\approx O(N^{4-\alpha+1}),
 \end{aligned}$$

and

$$\sum_i (d_i - \bar{d})^2 = N(m_2 - m_1^2) \approx N(\mu_2 - \mu_1^2) \approx O(N^{3-\alpha+1}).$$

The limit of (3.12) is:

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[N^{4-\alpha+1}]^2}{[N^{3-\alpha+1}]^3} = 0 \quad \text{for } 1 < \alpha < 2.$$

Thus, for $1 < \alpha < 2$ the condition (3.12) holds.

Considering the case $2 \leq \alpha < 3$, with $K = N^{\frac{1}{\alpha-1}}$, we have

$$\begin{cases} \mu_1 \approx O(1), \quad \text{or } O(\log(N)) & \alpha = 2, \\ \mu_2 \approx O(N^{\frac{3-\alpha}{\alpha-1}}), \\ \mu_3 \approx O(N^{\frac{4-\alpha}{\alpha-1}}). \end{cases}$$

Following a similar argument as before, we can show that the empirical moment m_s has the same order of magnitude as the theoretical moment μ_s , and that

$$\begin{aligned}
 \sum_i (d_i - \bar{d})^3 &= N(m_3 - 3m_2m_1 + 2m_1^3) \approx O(N^{\frac{4-\alpha}{\alpha-1}+1}), \\
 \sum_i (d_i - \bar{d})^2 &= N(m_2 - m_1^2) \approx O(N^{\frac{3-\alpha}{\alpha-1}+1}).
 \end{aligned}$$

Hence, for this case $2 \leq \alpha < 3$,

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[N^{\frac{4-\alpha}{\alpha-1}+1}]^2}{[N^{\frac{3-\alpha}{\alpha-1}+1}]^3} = 1, \quad \text{as } N \rightarrow \infty,$$

the condition (3.12) therefore does not hold.

For $\alpha = 3$, the theoretical moments are of order

$$\begin{cases} \mu_1 \approx O(1), \\ \mu_2 \approx O(\log(N)), \\ \mu_3 \approx O(N^{\frac{4-\alpha}{\alpha-1}}), \end{cases}$$

and similarly to before

$$\begin{aligned} \sum_i (d_i - \bar{d})^3 &= N(m_3 - 3m_2m_1 + 2m_1^3) \approx O(N^{\frac{4-\alpha}{\alpha-1}+1}), \\ \sum_i (d_i - \bar{d})^2 &= N(m_2 - m_1^2) \approx O(N \log(N)). \end{aligned}$$

Then

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[N^{\frac{4-3}{3-1}+1}]^2}{[N \ln N]^3} = 0.$$

Therefore condition (3.12) holds for the case $\alpha = 3$.

For $\alpha > 3$, the theoretical moments have orders

$$\begin{cases} \mu_1 \approx O(1), \\ \mu_2 \approx O(1), \\ \mu_3 \approx O(N^{\frac{4-\alpha}{\alpha-1}}) \text{ or } O(\ln N) \text{ or } O(1), \quad \text{for } \alpha < 4, \alpha = 4 \text{ and } \alpha > 4, \end{cases}$$

and

$$\begin{aligned}\sum_i (d_i - \bar{d})^3 &= N(m_3 - 2m_2m_1 + 2m_1^3) \approx O(N^{\frac{3}{\alpha-1}}) \text{ or } O(N \ln N) \text{ or } O(N) \\ \sum_i (d_i - \bar{d})^2 &= N(m_2 - m_1^2) \approx O(N).\end{aligned}$$

Hence the limit is obtained as

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[N^{\frac{3}{\alpha-1}}]^2}{[N]^3} = 0 \text{ for } 4 > \alpha > 3.$$

For $\alpha \geq 4$, the condition also holds:

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[N \ln N]^2}{[N]^3} = 0 \text{ for } \alpha = 4,$$

while

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[N]^2}{[N]^3} = 0 \text{ for } \alpha > 4.$$

In summary, the condition holds true for any $1 < \alpha < 2$, and $\alpha \geq 3$.

3.2.3 Normality when using the GHD_TO for SF network

To verify the limit of (3.10) (in page 52), first we define the TO measure of node i

$$a_{i.} = \sum_j a_{ij} = \sum_j \frac{\sum_{u \neq i,j} A_{iu} A_{uj} + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}}.$$

When verifying the limit of (3.10), we only interested in the order of magnitude of both numerator and denominator. Therefore, in order to understand the order of magnitude, we are going to approximate the expected value of $a_{i.}$ (i.e. $E(a_{i.}|d_i)$) given that the node degree of i^{th} node is d_i . We first find the expected value of $A_{iu} A_{uj}$, given the node degrees:

$E(A_{iu}A_{uj}|d_i, d_j, d_u)$. Because A_{iu} and A_{uj} represent adjacency information, we have

$$A_{iu}A_{uj} = \begin{cases} 1 & \text{with probability } P_{iuj}, \\ 0 & \text{with probability } 1 - P_{iuj}. \end{cases} \quad (3.15)$$

Here P_{iuj} is the conditional probability, given that d_i , d_j and d_u are known, of node u being connected to both node i and node j . Then we have $E(A_{iu}A_{uj}|d_i, d_j, d_u) = P_{iuj}$. To estimate this probability, we make the following two assumptions:

- (I) the probability of a node u being connected to a node i is proportional to the node degrees d_i and d_u . More specifically, this probability corresponds to that of a randomly weighted sampling of d_u objects from a total of $N - 1$ objects without replacement, constrained to include node i .
- (II) given d_i and d_j fixed, the random quantities $A_{iu}A_{uj}$, ($u = 1, 2, \dots, N$, $u \neq i$ and $u \neq j$) are independent, identically distributed random variables.

As a special case of the first assumption, if $d_u = 1$, then the probability of node u being connected to node i is $\frac{d_i}{T-1}$, in terms of the total node degree T , where the total node degree is written as

$$T = \sum_{i=1}^N d_i.$$

Notice that here T depends on N .

P_{iuj} is actually the probability of randomly sampling d_u individuals from a total of $N - 1$ objects without replacement, constrained to include individuals i and j . The detailed combinatoric formulation of such a problem is infeasible. However, for this study, it is not necessary to obtain an exact formulation of P_{iuj} : to check the condition (3.10), it is only necessary to approximate the order of both the numerator and the denominator. Hence, only an approximation of the order of $\sum_u P_{iuj}$ is required.

Now consider node u , with degree d_u , and a set of d_u nodes $\{i, j, v_1, v_2, \dots, v_{d_u-2}\}$ which connect with u .

The probability of selecting a set of consecutive nodes $\{i, j, v_1, v_2, \dots, v_{d_u-2}\}$ is:

$$\left[\frac{d_i}{T - d_u} \frac{d_j}{T - d_u - d_i} \frac{d_{v_1}}{T - d_u - d_i - d_j} \cdots \frac{d_{v_{d_u-2}}}{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-2} d_{v_k}} \right].$$

Then, summing over all possible sets $v_1, v_2, \dots, v_{(d_u-2)}$, we calculate the quantity

$$\begin{aligned} & P_{\{i,j,v_1,v_2,\dots,v_{d_u-2}\}} \\ &= \sum_{v_1 \neq i,j} \sum_{v_2 \neq i,j,v_1} \cdots \sum_{v_{(d_u-2)} \neq i,j,v_1,v_2,\dots} \frac{d_i}{T - d_u} \frac{d_j}{T - d_u - d_i} \frac{d_{v_1}}{T - d_u - d_i - d_j} \cdots \\ & \quad \frac{d_{v_{d_u-2}}}{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-2} d_{v_k}} \\ &= \frac{d_i}{T - d_u} \frac{d_j}{T - d_u - d_i} \sum_{v_1 \neq i,j} \sum_{v_2 \neq i,j,v_1} \cdots \sum_{v_{(d_u-2)} \neq i,j,v_1,v_2,\dots} \frac{d_{v_1}}{T - d_u - d_i - d_j} \cdots \\ & \quad \frac{d_{v_{d_u-2}}}{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-3} d_{v_k}} \\ &= \frac{d_i}{T - d_u} \frac{d_j}{T - d_u - d_i} \sum_{v_1 \neq i,j} \cdots \sum_{v_{d_u-3} \neq i,j,v_1} \frac{d_{v_1}}{T - d_u - d_i - d_j} \cdots \\ & \quad \frac{d_{v_{d_u-3}}}{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-4} d_{v_k}} \sum_{v_{d_u-2} \neq i,j,v_1,v_2,\dots} \frac{d_{v_{d_u-2}}}{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-3} d_{v_k}}. \end{aligned}$$

Notice that

$$\sum_{v_{(d_u-2)} \neq i,j,v_1,v_2,\dots} \frac{d_{v_{d_u-2}}}{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-3} d_{v_k}} = \frac{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-3} d_{v_k}}{T - d_u - d_i - d_j - \sum_{k=1}^{d_u-3} d_{v_k}} = 1.$$

Hence we have

$$P_{\{i,j,v_1,v_2,\dots,v_{d_u-2}\}} = \frac{d_i}{T - d_u} \frac{d_j}{T - d_u - d_i}. \quad (3.16)$$

Also, from the fact that for any node i , $d_i = o(T)$ for the SF network model, we have

$$\lim_{N \rightarrow \infty} \frac{d_i}{T} \rightarrow 0.$$

In fact, for $1 < \alpha < 2$, the total node degree

$$T = \sum_{i=1}^N d_i \approx N \sum_{d=1}^{N-1} d \times d^{-\alpha} = N \sum_{d=1}^{N-1} d^{1-\alpha} \approx O(N^{3-\alpha}),$$

of order $N^{3-\alpha}$, of greater order than $\max\{d_i\} = N-1$. For $\alpha = 2$, $T \sim O(N \ln N)$, which is also of a larger order than $N-1$. For $\alpha > 2$, we consider as before the natural cut-off $K = N^{\frac{1}{\alpha-1}}$, which is also of a smaller order than the total degree $T \sim O(N)$. Therefore, for any k , so that (3.16) can be approximated by,

$$P_{\{i,j,v_1,v_2,\dots,v_{d_u-2}\}} \approx \frac{d_i d_j}{T^2}.$$

Now consider a general selection of consecutive nodes,

$$\pi^* = \{v_1, v_2, \dots, v_s, i, v_{s+1}, \dots, v_t, j, v_{t+1}, \dots, v_{u-2}\}$$

where $1 \leq s \leq t \leq u-2$.

Then the probability of selecting this set of consecutive connections, P_{π^*} , is equal to

$$\begin{aligned} P_{\pi^*} &= \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{d_i}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots \\ &\quad \sum_{d_{v_t}} \frac{d_{v_t}}{T - d_u - d_i - \sum_{k=1}^{t-1} d_{v_k}} \frac{d_j}{T - d_u - d_i - \sum_{k=1}^t d_{v_k}} \\ &\quad \sum_{d_{v_{t+1}}} \frac{d_{v_{t+1}}}{T - d_u - d_i - d_j - \sum_{k=1}^t d_{v_k}} \cdots \sum_{d_{v_{u-2}}} \frac{d_{v_{u-2}}}{T - d_u - d_i - d_j - \sum_{k=1}^{u-3} d_{v_k}} \\ &= \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{d_i}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots \\ &\quad \sum_{d_{v_t}} \frac{d_{v_t}}{T - d_u - d_i - \sum_{k=1}^{t-1} d_{v_k}} \frac{d_j}{T - d_u - d_i - \sum_{k=1}^t d_{v_k}} \\ &\geq \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{d_i}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots \\ &\quad \sum_{d_{v_t}} \frac{d_{v_t}}{T - d_u - d_i - \sum_{k=1}^{t-1} d_{v_k}} \frac{d_j}{T - d_u - d_i} \end{aligned}$$

$$\begin{aligned}
&= \frac{d_j}{T - d_u - d_i} \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \\
&\quad \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{d_i}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots \sum_{d_{v_t}} \frac{d_{v_t}}{T - d_u - d_i - \sum_{k=1}^{t-1} d_{v_k}} \\
&= \frac{d_j}{T - d_u - d_i} \frac{d_i}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_1}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^s d_{v_k}} \\
&\geq \frac{d_j}{T - d_u - d_i} \frac{d_i}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_1}}{T - d_u} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \\
&= \frac{d_i}{T - d_u} \frac{d_j}{T - d_u - d_i} \\
&= P_{\{i,j,v_1,v_2,\dots,v_{d_u-2}\}} \\
&\geq \frac{d_i d_j}{T^2}.
\end{aligned}$$

Hence $P_{\{i,j,v_1,v_2,\dots,v_{d_u-2}\}}$ is the smallest probability among all possible orders of selecting nodes i and j , i.e.

$$P_{\pi^*} \geq P_{\{i,j,v_1,v_2,\dots,v_{d_u-2}\}} \geq \frac{d_i d_j}{T^2} \quad \text{for all } \pi^* \in \Pi$$

Also, for any selection $\pi^* = \{v_1, v_2, \dots, v_s, i, v_{s+1}, \dots, v_t, j, v_{t+1}, \dots, v_{u-2}\}$ of consecutive connected nodes, P_{π^*} can be approximated by a quantity of the form $\frac{d_i d_j}{T^2} f_{\pi^*}(T, d_u)$. For any $d_i = o(T)$,

$$\begin{aligned}
P_{\pi^*} &= \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{d_i}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots \\
&\quad \sum_{d_{v_t}} \frac{d_{v_s}}{T - d_u - d_i - \sum_{k=1}^{t-1} d_{v_k}} \frac{d_j}{T - d_u - d_i - \sum_{k=1}^t d_{v_k}} \\
&\quad \sum_{d_{v_{t+1}}} \frac{d_{v_{t+1}}}{T - d_u - d_i - d_j - \sum_{k=1}^t d_{v_k}} \\
&\quad \cdots \sum_{d_{v_{u-2}}} \frac{d_{v_{u-2}}}{T - d_u - d_i - d_j - \sum_{k=1}^{v_{u-3}} d_{v_k}} \\
&= \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{d_i}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots
\end{aligned}$$

$$\begin{aligned}
& \sum_{d_{v_t}} \frac{d_{v_t}}{T - d_u - d_i - \sum_{k=1}^{t-1} d_{v_k}} \frac{d_j}{T - d_u - d_i - \sum_{k=1}^t d_{v_k}} \\
& \approx \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{d_i}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots \\
& \sum_{d_{v_t}} \frac{d_{v_t}}{T - d_u - \sum_{k=1}^{t-1} d_{v_k}} \frac{d_j}{T - d_u - \sum_{k=1}^t d_{v_k}} \\
& \approx \frac{d_i d_j}{T^2} \sum_{d_{v_1}} \frac{d_{v_1}}{T - d_u} \sum_{d_{v_2}} \frac{d_{v_2}}{T - d_u - d_{v_1}} \cdots \sum_{d_{v_s}} \frac{d_{v_s}}{T - d_u - \sum_{k=1}^{s-1} d_{v_k}} \frac{T}{T - d_u - \sum_{k=1}^s d_{v_k}} \cdots \\
& \sum_{d_{v_t}} \frac{d_{v_t}}{T - d_u - \sum_{k=1}^{t-1} d_{v_k}} \frac{T}{T - d_u - \sum_{k=1}^t d_{v_k}} \\
& = \frac{d_i d_j}{T^2} f_{\pi^*}(T, d_u).
\end{aligned}$$

Since $P_{\pi^*} \geq \frac{d_i d_j}{T^2}$, then

$$f_{\pi^*}(T, d_u) > 1 \quad \text{for all } \pi^* \in \Pi.$$

Finally, we obtain all possible selections for i and j in a set with nodes $\{v_1, v_2, \dots, v_{u-2}, i, j\}$. Considering different ways of placing two balls into an ordered set $\{v_1, v_2, \dots, v_{u-2}\}$, this is equivalent to selecting 2 balls from a set of u balls, which is $\frac{1}{2}d_u(d_u - 1)$. Then, because the order of i and j is interchangeable, there are in total $d_u(d_u - 1)$ different possible rankings of i and j , hence

$$P_{iuj} \approx \frac{d_i d_j}{T^2} \sum_{\pi^*} f_{\pi^*}(T, d_u) = d_u(d_u - 1) \frac{d_i d_j}{T^2} \sum_{\pi^*} \frac{f_{\pi^*}(T, d_u)}{d_u(d_u - 1)}.$$

We calculate a lower bound:

$$P_{iuj} > d_u(d_u - 1) P_{\{i,j,v_1,v_2,\dots,v_{d_u-2}\}} \approx d_u(d_u - 1) \frac{d_i d_j}{T^2}. \quad (3.17)$$

Now, the topological overlap measure to a node i can then be written as

$$a_i = \sum_j \frac{\sum_{u \neq i,j} A_{iu} A_{uj} + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}}$$

$$\begin{aligned}
&\approx \sum_j \frac{\sum_{u \neq i,j} d_u(d_u - 1) \left[\frac{d_i d_j}{T^2} \sum_{\pi^*} \frac{f_{\pi^*}(T, d_u)}{d_u(d_u - 1)} \right] + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
&= \sum_j \frac{\sum_{u \neq i,j} (d_u - 1)(d_u) \left[\frac{d_i d_j}{T^2} \sum_{\pi^*} \frac{f_{\pi^*}(T, d_u)}{(d_u - 1)(d_u)} \right] + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
&= \sum_j \frac{\left[\frac{d_i d_j}{T^2} \sum_{u \neq i,j} \sum_{\pi^*} f_{\pi^*}(T, d_u) \right] + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}}. \tag{3.18}
\end{aligned}$$

The approximation follows by use of WLLN.

Further, the quantities $A_{iu}A_{uj}$, ($u = 1, \dots, N$ and $u \neq i, j$), are assumed to be independent and identically distributed and $0 < E(A_{iu}A_{uj}) \leq 1$ is finite.

Now we discuss about different values of α . First, consider $\alpha > 3$. We denote

$$F(N) = \frac{1}{T^2} \sum_{u \neq i,j} \sum_{\pi^*} f_{\pi^*}(T, d_u).$$

From (3.17), we can conclude a lower bound for $F(N)$,

$$\sum_{u \neq i,j} d_u(d_u - 1)T^2.$$

Applying results for the adjacency case, then

$$\sum_{u \neq i,j} d_u(d_u - 1) = \sum_{u \neq i,j} (d_u)(d_u - 1) = \sum_{u \neq i,j} d_u^2 - d_u = N(m_2 - m_1) \approx N(\mu_2 - \mu_1),$$

where $m_s = \frac{1}{N} \sum_i d_i^s$, $s = 1, 2$ are the empirical moments of d_i and μ_1, μ_2 are the corresponding theoretical moments. Here, we also need to approximate $\sum_{u \neq i,j} d_u(d_u - 1)$ by the theoretical moments of d_i . In the following argument, there are many approximations which are justified by use of WLLN, though we suppress these steps. Similar to the adjacency case, the order of the empirical moment m_s can be approximated by that of the theoretical moment μ_s . Hence

$$\sum_{u \neq i,j} d_u(d_u - 1) \approx c'N.$$

Then, $F(N) > c'N^{\frac{1}{T^2}}$, and for $\alpha > 3$, the order of $F(N)$ is greater than $O(\frac{1}{N})$.

Equation (3.18) gives

$$\begin{aligned}
 a_i &= \sum_j \frac{d_i d_j F(N) + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
 &= \sum_j \frac{d_i d_j F(N)}{\min(d_i, d_j) + 1 - A_{ij}} + \sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
 &= F(N) \sum_j \frac{d_i d_j}{\min(d_i, d_j) + 1 - A_{ij}} + \sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}}. \quad (3.19)
 \end{aligned}$$

Consider the second term of (3.19). This is smaller than d_i , and contains d_i non-zero terms. For small node degree d_i , the expected value of this second term is very close to 1. For d_i large, we first separate the summation into two terms,

$$\begin{aligned}
 &\sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
 &= d_i \frac{1}{d_i} \sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
 &\approx d_i \sum_{d=1}^{N^{\frac{1}{\alpha-1}}} \frac{A_{i|d}}{\min(d_i, d) + 1} cd^{-\alpha} \\
 &= d_i \sum_{d \leq d_i} \frac{A_{i|d}}{d_j + 1} cd^{-\alpha} + d_i \sum_{d_i \leq d} \frac{A_{i|d}}{d_i + 1} cd^{-\alpha}. \quad (3.20)
 \end{aligned}$$

Here we notice that $A_{i|d}$ is representing the expected value of node i having an edge with a node with degree d . More precisely:

$$A_{i|d} = E(A_{ij} | d_j = d).$$

Here in denominator A_{ij} is ignored in the second step, due to the fact that the chance that a node with degree d_i connect to a node with degree d is small, and as node degree increases, A_{ij} is ignorable. Later in this proof we will repeatedly apply this strategy.

The value of the first term in (3.20) can be approximated, for large d_i , as

$$d_i \sum_{d \leq d_i} \frac{A_{i|d}}{d+1} cd^{-\alpha} \approx d_i \sum_{d=1}^{d_i-1} c \frac{1}{d} d^{-\alpha} = d_i \sum_{d=1}^{d_i-1} cd^{-\alpha-1}.$$

The second term of (3.20) is approximated as

$$d_i \sum_{d_i \leq d} \frac{A_{i|d}}{d_i+1} cd^{-\alpha} \approx d_i \sum_{d=d_i}^{N^{\frac{1}{\alpha-1}}} \frac{1}{d_i} cd^{-\alpha} = P(d > d_i).$$

Hence, for d_i large,

$$\sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1} \approx d_i \sum_{d=1}^{d_i-1} cd^{-\alpha-1} + P(d > d_i) = G(d_i),$$

say. Notice that $G(d_i)$ is an increasing function, since

$$G(d_i + 1) - G(d_i) = \sum_{d \leq d_i} cd^{-\alpha-1} + (d_i + 1)cd_i^{-\alpha-1} - cd_i^{-\alpha} > 0.$$

However, according to (3.13), $\sum_{d=1}^{N^{\frac{1}{\alpha-1}}} cd^{-\alpha-1}$, is a constant smaller than 1. Hence, for any $d_i > 0$, the first term in $G(d_i)$ satisfies

$$d_i \geq d_i \sum_{d=1}^{d_i-1} cd^{-\alpha-1} \geq c_\alpha d_i,$$

with $c_\alpha = \frac{1}{\zeta(\alpha)}$. Here $\zeta(\alpha)$ is the ζ function of power α , defined by

$$\zeta(\alpha) = \sum_{i=1}^{\infty} \frac{1}{i^\alpha},$$

which has $\zeta(3) = 1.202$. Also the order of higher moments for second term of $G(d_i)$, $P(d > d_i)$ (this is a function of random d_i) is at most $O(1)$. Therefore, the order of the higher moments of the second term of (3.19) is just the same as the corresponding moments of d_i .

Now, consider the first term of (3.19). This may be further expanded as follows:

$$\begin{aligned}
& F(N) \sum_j \frac{d_i d_j}{\min(d_i, d_j) + 1 - A_{ij}} \\
& \approx NF(N) \sum_{d=1}^{N^{\frac{1}{\alpha-1}}} \frac{d_i d}{\min(d_i, d) + 1} cd^{-\alpha} \quad (\text{by WLLN}) \\
& = NF(N) \left[\sum_{d=1}^{d_i} \frac{d_i d}{d+1} cd^{-\alpha} + \sum_{d=d_i+1}^{N^{\frac{1}{\alpha-1}}} \frac{d_i d}{d_i+1} cd^{-\alpha} \right].
\end{aligned}$$

The first term in brackets may be approximated,

$$\sum_{d=1}^{d_i} \frac{d_i d}{d+1} d^{-\alpha} \approx d_i \sum_{d=1}^{d_i} \frac{d}{d+1} d^{-\alpha},$$

and, as $N \rightarrow \infty$, the second term in brackets may be approximated as

$$\sum_{d=d_i+1}^{N^{\frac{1}{1-\alpha}}} \frac{d_i d}{d_i+1} cd^{-\alpha} \approx \sum_{d=d_i+1}^{N^{\frac{1}{1-\alpha}}} \frac{d_i d}{d_i+1} cd^{-\alpha} = \frac{d_i}{d_i+1} \left(\frac{\zeta(\alpha-1)}{\zeta(\alpha)} - \sum_{d=1}^{d_i} cd^{1-\alpha} \right).$$

The approximations in the above are reasonable, since by using assumption (I) we can calculate that the expected value of A_{ij} is approximately the order of $\frac{d_i d_j}{T}$, for small values of d_i . Hence the first term of (3.19) can be approximated by

$$\begin{aligned}
& F(N) \sum_j \frac{d_i d_j}{\min(d_i, d_j) + 1 - A_{ij}} \\
& \approx NF(N) \left[d_i \sum_{d=1}^{d_i} \frac{d}{d+1} cd^{-\alpha} + \frac{d_i}{d_i+1} \left(\frac{\zeta(\alpha-1)}{\zeta(\alpha)} - \sum_{d=1}^{d_i} cd^{1-\alpha} \right) \right].
\end{aligned}$$

Since we have

$$0 < \frac{d_i}{d_i+1} \left(\frac{\zeta(\alpha-1)}{\zeta(\alpha)} - \sum_{d=1}^{d_i} cd^{1-\alpha} \right) < 1,$$

then the value of higher moments offered by this term is $O(1)$ at most. Notice that

$$d_i \sum_{d=1}^{d_i} \frac{d}{d+1} cd^{-\alpha} \rightarrow d_i \quad \text{as } d_i \rightarrow \infty.$$

Hence we have

$$\frac{cd_i}{2} \leq d_i \sum_{d=1}^{d_i} \frac{d}{d+1} cd^{-\alpha} \leq d_i.$$

Therefore the order of the higher moments in the first term in (3.19) are the same as for the corresponding moments of $NF(N)d_i$.

We know that the order of $F(N)$ is at least greater than $O(\frac{1}{N})$. Then the order of moments in the first term in (3.19) is greater than or equal to those of the second term in (3.19). Hence, the orders of the 2^{nd} and 3^{rd} moments of $a_{i.}$ are the same as those of $NF(N)d_i$. Then,

$$\begin{aligned} \sum_i (a_i - \bar{a})^3 &= N(m_3 - 2m_2m_1 + 2m_1^3) \approx O(N^3 F^3(N)) \sum_i (d_i - \bar{d})^3, \\ \sum_i (a_i - \bar{a})^2 &= N(m_2 - m_1^2) \approx O(N^2 F^2(N)) \sum_i (d_i - \bar{d})^2. \end{aligned}$$

Applying the central moments to (3.10), we have

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (a_{i.} - \bar{a}_{i.})^3]^2}{[\sum_i (a_{i.} - \bar{a}_{i.})^2]^3} = \lim_{N \rightarrow \infty} \frac{[(NF(N))^3 \sum_i (d_i - \bar{d})^3]^2}{(NF(N))^2 \sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = 0.$$

Hence, for $\alpha > 3$ the asymptotic normality holds.

For $\alpha = 3$ and $2 \leq \alpha < 3$ similar steps can be followed, and we have that the orders of the 2^{nd} and 3^{rd} moments of $a_{i.}$ are the same with those of $NF(N)d_i$.

When applying these values for different α to condition (3.10), the conclusion is the same as seen when considering adjacency information. The coefficient $NF(N)$ can be taken out of both denominator and numerator, and cancelled. As with the adjacency case, the condition for asymptotic normality holds for $\alpha = 3$, but for $2 \leq \alpha < 3$ asymptotic normality of the permutation distribution is not guaranteed.

For $1 < \alpha < 2$, (3.18) becomes:

$$\begin{aligned}
a_{i.} &\approx \sum_j \frac{\frac{d_i d_j}{T^2} \left[\sum_u d_u (d_u - 1) \sum_{\pi^*} \frac{f_{\pi^*}(T, d_u)}{d_u (d_u - 1)} \right] + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
&\approx \sum_j \frac{d_i d_j F(N) + A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
&= F(N) \sum_j \frac{d_i d_j}{\min(d_i, d_j) + 1 - A_{ij}} + \sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}} \\
&= F(N) \sum_j \frac{d_i d_j}{\min(d_i, d_j) + 1 - A_{ij}} + \sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}}. \tag{3.21}
\end{aligned}$$

Similarly to before,

$$F(N) = \frac{1}{T^2} \sum_u d_u (d_u - 1) \sum_{\pi^*} \frac{f_{\pi^*}(T, d_u)}{d_u (d_u - 1)},$$

is a function of N and it is greater than

$$\frac{\sum_u d_u (d_u - 1)}{T^2} \approx \frac{c' N^{3-\alpha}}{N^{4-2\alpha}} = c' N^{\alpha-1}. \tag{3.22}$$

The second term

$$\sum_j \frac{A_{ij}}{\min(d_i, d_j) + 1 - A_{ij}}, \tag{3.23}$$

is of order d_i , the same as for the previous case. For the first term:

$$\begin{aligned}
&F(N) \sum_j \frac{d_i d_j}{\min(d_i, d_j) + 1 - A_{ij}} \\
&= NF(N) \left[\sum_{d=1}^{d_i} \frac{d_i d}{d + 1 - A_{i|d}} d^{-\alpha} + \sum_{d=d_i+1}^N \frac{d_i d}{d_i + 1 - A_{i|d}} d^{-\alpha} \right] \\
&\approx NF(N) \left[d_i \sum_{d=1}^{d_i} \frac{d}{d + 1} d^{-\alpha} + \frac{d_i}{d_i + 1} \sum_{d=d_i+1}^N d^{1-\alpha} \right]
\end{aligned}$$

$$\approx NF(N) \left[d_i \sum_{d=1}^{d_i} \frac{d}{d+1} d^{-\alpha} + \frac{d_i}{d_i+1} (N^{2-\alpha} - (d_i+1)^{2-\alpha}) \right].$$

Also, it is easy to see that

$$\frac{c'd_i}{2} \leq d_i \sum_{d=1}^{d_i} \frac{d}{d+1} cd^{-\alpha} \leq d_i,$$

still holds true. This is bounded by $[\frac{cd_i}{2}, d_i]$, therefore the order is still the same as that of d_i .

From (3.22), we know that $F(N)$ is greater than $O(N^{\alpha-1})$, with $\alpha > 1$, therefore the order of (3.23) is always smaller than the order of

$$NF(N) \left[d_i \sum_{d=1}^{d_i} \frac{d}{d+1} d^{-\alpha} \right].$$

For $\alpha > 1$, $N^{3-\alpha}$ is always greater than $(N^{2-\alpha})^2$, the order of the second moment for

$$\frac{d_i}{d_i+1} (N^{2-\alpha} - (d_i+1)^{2-\alpha}),$$

and the same conclusion can be made for the third moment. Hence again the orders of the 2^{nd} and 3^{rd} moments of a_i are the same as the orders of the 2^{nd} and 3^{rd} moments of $NF(N)d_i$:

$$\begin{aligned} \sum_i (a_i - \bar{a})^3 &\approx N(m_3 - 2m_2m_1 + 2m_1^3) \approx O(N^3 F^3(N)) \sum_i (d_i - \bar{d})^3, \\ \sum_i (a_i - \bar{a})^2 &\approx N(m_2 - m_1^2) \approx O(N^2 F^2(N)) \sum_i (d_i - \bar{d})^2. \end{aligned}$$

Then, substituting directly to (3.10),

$$\begin{aligned} &\lim_{N \rightarrow \infty} \frac{[\sum_i (a_i - \bar{a}_i)^3]^2}{[\sum_i (a_i - \bar{a}_i)^2]^3} \\ &= \lim_{N \rightarrow \infty} \frac{[(NF(N))^3 \sum_i (d_i - \bar{d})^3]^2}{(NF(N))^2 \sum_i (d_i - \bar{d})^2]^3} \end{aligned}$$

$$\begin{aligned}
&= \lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} \\
&= \lim_{N \rightarrow \infty} \frac{[N^{4-\alpha}]^2}{[N^{3-\alpha}]^3} \\
&= \lim_{N \rightarrow \infty} N^{\alpha-2} \\
&= 0, \quad 1 < \alpha < 2.
\end{aligned}$$

This justifies that for $1 < \alpha < 2$ asymptotic normality holds.

Hence, we conclude that for a scale free network, for both adjacency and topological overlap measures, a sufficient condition for asymptotic normality of the permutation distribution is that $1 < \alpha < 2$, or $\alpha \geq 3$.

3.2.4 Normality when using the HD for RG graph

For RG graph, the node degree (d_i) distribution can be approximated by a Poisson distribution with parameter λ_N :

$$P(d_i = k) = \frac{e^{-\lambda_N} \lambda_N^k}{k!},$$

where the Poisson parameter λ_N depends on the total number of nodes N , the threshold r and dimension m [29], for details please refer to Section 2.2.2 (in page 26).

To check (3.12) (in page 53), similarly to (3.15), we write the numerator and denominator in terms of, respectively,

$$\begin{aligned}
\sum_i (d_i - \bar{d})^2 &= N(m_2 - m_1^2) \quad \text{and} \\
\sum_i (d_i - \bar{d})^3 &= N(m_3 - 3m_2m_1 + 2m_1^3),
\end{aligned} \tag{3.24}$$

where $m_s = \frac{1}{N} \sum_{i=1}^N d_i^s$ is the s^{th} empirical moment of the node degree distribution. In order to find the limit in (3.12), we approximate the empirical moments by the corresponding theoretical moments.

For finite λ_N , the series of random variables $\{d_i^s, i = 1, \dots, N\}$ are independent, identically distributed, with $E[d_i^s] = \mu_s < \infty$. According to the WLLN, the empirical moments

converge to the corresponding theoretical moments in probability, hence we have

$$\lim_{N \rightarrow \infty} \frac{m_s}{\mu_s} = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{i=1}^N d_i^s}{\sum_{d=1}^{N-1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} d^s} = 1, \quad s = 1, 2, 3,$$

where μ_s is the theoretical moment of the node degree distribution.

Hence we have for the numerator

$$\begin{aligned} \sum_{i=1}^N (d_i - \bar{d})^3 &= N \frac{1}{N} \sum_{i=1}^N (d_i - \bar{d})^3 \\ &= N(m_3 - 3m_2m_1 + 2m_1^3) \\ &\approx N(\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3) \\ &= N\lambda_N \text{ as } N \rightarrow \infty. \end{aligned}$$

Similarly, for the denominator

$$\sum_{i=1}^N (d_i - \bar{d})^2 = N \frac{1}{N} \sum_{i=1}^N (d_i - \bar{d})^2 = N(m_2 - m_1^2) \approx N(\mu_2 - \mu_1^2) = N\lambda_N \text{ as } N \rightarrow \infty.$$

Substituting into (3.12),

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{[N\lambda_N]^2}{[N\lambda_N]^3} = 0, \quad (3.25)$$

provided λ_N is finite and greater than $O(\frac{1}{N})$.

For large enough λ_N , the Poisson distribution can be approximated by a normal distribution with mean λ_N and variance λ_N . Let $Z_i = \frac{d_i}{\lambda_N}$. Then we have

$$Z_i \sim N(1, \frac{1}{\lambda_N}).$$

Denoting by μ'_s the s^{th} theoretical moment of the random variable Z_i , μ'_s is finite. Let

$$m'_s = \frac{1}{N} \sum_{i=1}^N \left[\frac{d_i}{\lambda_N} \right]^s,$$

so that $m_s = \lambda_N^s m'_s$. For the set of independent and identically distributed random variables $Z_1, Z_2, \dots, Z_N$, $E[Z_i^s] = \mu_s$ exists and is finite. Again according to WLLN, m'_s will converge to μ'_s in probability:

$$1 = \lim_{N \rightarrow \infty} \frac{m'_s}{\mu'_s} = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{i=1}^N \left(\frac{d_i}{\lambda_N} \right)^s}{\frac{\mu_s}{\lambda_N^s}} = \lim_{N \rightarrow \infty} \frac{m_s}{\mu_s}, \quad s = 1, 2, 3.$$

This indicates that m_s and μ_s are of the same order. Now we can approximate the order of the numerator and denominator of condition (3.12). The numerator is 0, since

$$\begin{aligned} \sum_{i=1}^N (d_i - \bar{d})^3 &= N(m_3 - 3m_2m_1 + 2m_1^3) \\ &\approx N(\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3) \\ &= N[\mu_1(\mu_1^2 + 3(\mu_2 - \mu_1^2)) - 3\mu_1\mu_2 + 2\mu_1^3] \\ &= 0. \end{aligned}$$

Similarly, the denominator approximates $N\lambda_N^2$:

$$\sum_{i=1}^N (d_i - \bar{d})^2 = N(m_2 - m_1^2) \approx N(\mu_2 - \mu_1^2) = N\sigma_N^2 = N\lambda_N^2 \text{ as } N \rightarrow \infty.$$

Substituting in to (3.12):

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{0}{[N\lambda_N^2]} = 0.$$

Hence the limit of (3.12) also holds for large λ_N . These results suggest that, for λ_N of greater order than $O(\frac{1}{N})$, condition (3.12) is always satisfied for an RG graph. For λ_N smaller than order $O(\frac{1}{N})$, when the condition (3.12) does not hold, the average node degree in the network is smaller than $O(1)$. This fact indicates that in such a case the network is almost empty and hence is of little practical interest.

3.2.5 Normality when using the GHD_TO for RG graph

Similar to the analysis of the topological overlap measure for a scale free network, we approximate the value of a_i given a particular node with degree d_i by first calculating the quantity $E(A_{iu}A_{ju}|d_i, d_j)$. We know that the node degree distribution is Poisson with parameter λ_N . According to the results for the adjacency case, with any λ_N ,

$$\lim_{N \rightarrow \infty} \frac{m_s}{\mu_s} = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{i=1}^N d_i^s}{\sum_{d=1}^{N-1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} d^s} = 1, \quad s = 1, 2, 3.$$

As before, denote by T the total node degree,

$$T = \sum_i d_i = N \frac{1}{N} \sum_i d_i \approx N \mu_N = N \lambda_N.$$

Again, we are only interested in the order of $E(A_{iu}A_{uj}|d_i, d_j)$, hence we can write

$$E(A_{iu}A_{uj}|d_i, d_j) \approx \frac{d_i d_j}{T^2} \frac{d_u(d_u - 1)}{2} \sum_{\pi^*} \frac{f_{\pi^*}(T, d_u)}{d_u(d_u - 1)} = \frac{d_i d_j}{2N^2 \lambda_N^2} \sum_{\pi^*} f_{\pi^*}(T, d_u). \quad (3.26)$$

Then in the expression for the topological overlap measure, the summation $\sum_u A_{iu}A_{uj}$ can further be expressed as:

$$\sum_{u \neq i, j} p_{iuj} \approx \sum_{u \neq i, j} \frac{d_i d_j}{2N^2 \lambda_N^2} \sum_{\pi^*} f_{\pi^*}(T, d_u) \frac{e^{-\lambda_N} \lambda_N^{d_u}}{d_u!} = d_i d_j F(N),$$

where again $F(N)$ is defined as a function of N with order greater than

$$d_u(d_u - 1)P_{\{i, j, v_1, v_2, \dots, v_{d_u-2}\}} = \frac{1}{\lambda_N^2 N} \sum_{u \neq i, j} \frac{1}{N} \frac{d_u(d_u - 1)}{2} \approx \frac{(\lambda_N^2 - 2\lambda_N)}{2\lambda_N^2 N}.$$

where

$$d_u(d_u - 1)P_{\{i, j, v_1, v_2, \dots, v_{d_u-2}\}}$$

is the lower bound for $F(N)$ (see (3.17) in page 64). This result shows that $F(N)$ is at least of order $O(\frac{1}{N})$. Again, the approximation is justified by the WLLN. Plugging into the

topological overlap measure we have:

$$\begin{aligned} E(a_i | d_i) &\approx \sum_{j=1}^N \frac{d_i d_j F(N) + A_{ij}}{\min(d_i, d_j) - A_{ij} + 1} \\ &= F(N) \sum_{j=1}^N \frac{d_i d_j}{\min(d_i, d_j) - A_{ij} + 1} + \sum_{j=1}^N \frac{A_{ij}}{\min(d_i, d_j) - A_{ij} + 1}. \end{aligned} \quad (3.27)$$

The first term in (3.27) can be expanded, as $N \rightarrow \infty$,

$$\begin{aligned} &F(N) \sum_{j=1}^N \frac{d_i d_j}{\min(d_i, d_j) - A_{ij} + 1} \\ &\approx NF(N) \left[\sum_{d=1}^{d_i} \frac{d_i d}{d - A_{i|d} + 1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} + \sum_{d=d_i+1}^N \frac{d_i d}{d_i - A_{i|d} + 1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} \right] \\ &\approx NF(N) \left[\sum_{d=1}^{d_i} \frac{d_i d}{d+1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} + \sum_{d=d_i+1}^{\infty} \frac{d_i d}{d_i+1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} \right] \\ &= NF(N) \left[d_i \sum_{d=1}^{d_i} \frac{d}{d+1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} + \frac{d_i}{d_i+1} \sum_{d=d_i+1}^{\infty} d \frac{e^{-\lambda_N} \lambda_N^d}{d!} \right] \\ &= NF(N) \left[d_i \sum_{d=1}^{d_i} \frac{d}{d+1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} + \frac{d_i \lambda_N}{d_i+1} \sum_{d=d_i+1}^{\infty} \frac{e^{-\lambda_N} \lambda_N^{d-1}}{(d-1)!} \right] \\ &= NF(N) \left[d_i \left(\sum_{d=1}^{d_i} \frac{d+1}{d+1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} - \sum_{d=1}^{d_i} \frac{1}{d+1} \frac{e^{-\lambda_N} \lambda_N^d}{d!} \right) + \frac{d_i \lambda_N}{d_i+1} \sum_{d=d_i}^{\infty} \frac{e^{-\lambda_N} \lambda_N^d}{(d)!} \right] \\ &= NF(N) \left[d_i P(d \leq d_i) - \frac{d_i}{\lambda_N} \sum_{d=1}^{d_i} \frac{e^{-\lambda_N} \lambda_N^{d+1}}{(d+1)!} + \frac{d_i \lambda_N}{d_i+1} P(d \geq d_i) \right] \\ &= NF(N) \left[d_i P(d \leq d_i) - \frac{d_i}{\lambda_N} P(2 \leq d \leq d_i + 1) + \frac{d_i \lambda_N}{d_i+1} (1 - P(d \leq d_i - 1)) \right]. \end{aligned}$$

Then (3.27) can be approximated as

$$\begin{aligned} E(a_i | d_i) &\approx NF(N) \left[d_i P(d \leq d_i) - \frac{d_i}{\lambda_N} P(2 \leq d \leq d_i + 1) + \frac{d_i \lambda_N}{d_i+1} (1 - P(d \leq d_i - 1)) \right] \\ &\quad + \sum_{j=1}^N \frac{A_{ij}}{\min(d_i, d_j) - A_{ij} + 1}. \end{aligned} \quad (3.28)$$

We first show that the order of the s^{th} moment for the second term in (3.28) is at most $O(E(d_i^s))$,

$$\begin{aligned}
& \sum_{j=1}^N \frac{A_{ij}}{\min(d_i, d_j) - A_{ij} + 1} \\
&= \sum_{j'=1}^{d_i} \frac{1}{d_{j'}}, \quad \text{considering terms with } A_{ij} = 1 \text{ only} \\
&\approx d_i \sum_{d \leq d_i} \frac{1}{d} \frac{e^{-\lambda_N} \lambda_N^d}{d!} + d_i \sum_{d_i \leq d} \frac{1}{d_i} \frac{e^{-\lambda_N} \lambda_N^d}{d!} \\
&= d_i \sum_{d \leq d_i} \frac{1}{d} \frac{e^{-\lambda_N} \lambda_N^d}{d!} + P(d > d_i).
\end{aligned}$$

Again, since $0 < P(d > d_i) < 1$, the order of any moment of $P(d > d_i)$ is at most $O(1)$. Also,

$$e^{-\lambda_N} \lambda_N \leq d_i \sum_{d \leq d_i} \frac{1}{d} \frac{e^{-\lambda_N} \lambda_N^d}{d!} \leq d_i. \quad (3.29)$$

Therefore, the order of moment of the second term is at most the same as the order of the corresponding moment of d_i .

We check the order of the first term in (3.28), by discussing different values of λ_N . For bounded λ_N , $NF(N)$ is at least $O(1)$.

Actually the term inside bracket in (3.28) is $O(d_i)$, which in the case of bounded λ_N has finite 2^{nd} and 3^{rd} moments. Therefore, the mean of a_i is at least $O(NF(N))$ and the second moment of a_i is of $O(N^2 F^2(N))$. Then for finite value of λ_N ,

$$E \left[d_i P(d \leq d_i) - \frac{d_i}{\lambda_N} P(2 \leq d \leq d_i + 1) + \frac{d_i \lambda_N}{d_i + 1} (1 - P(d \leq d_i - 1)) \right], \quad (3.30)$$

is $O(1)$. By following a similar procedure, we conclude that the first moment of (3.28) is $O(NF'(N))$, the second moment of (3.28) is $O(N^2 F'(N)^2)$ and the third moment of (3.28) is $O(N^3 F'(N)^3)$. However the order of central moments for (3.28) can not be obtained

easily. Especially for the third central empirical moment,

$$\sum_i (a_{i\cdot} - \bar{a})^3 = N(m_3 - 3m_2m_1 + 2m_1^3) \approx N(\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3),$$

where all three terms, μ_3 , $\mu_2\mu_1$ and μ_1^3 , have the same order of magnitude.

In order to verify the limit of (3.10), we are seeking to verify condition (3.8). An alternative representation of (3.8) (in page 51) is given by

$$\liminf_{N \rightarrow \infty} \frac{\sum_{i=1}^N (a_{i\cdot} - \bar{a})^2}{N\bar{a}^2} > 0. \quad (3.31)$$

If the above inequality holds true, the non-centered measure $a_{i\cdot}$ can directly be applied. Then (3.10) can be replaced by

$$\lim_{n \rightarrow \infty} \frac{[\sum_i (a_{i\cdot})^3]^2}{[\sum_i (a_{i\cdot})^2]^3} = 0. \quad (3.32)$$

For a detailed derivation, please refer to [82].

We can check that the variance of $a_{i\cdot}$ is of $O((NF(N))^2)$ as well. This is because for d_i large,

$$d_i P(d \leq d_i) - \frac{d_i}{\lambda_N} P(2 \leq d \leq d_i + 1) + \frac{d_i \lambda_N}{d_i + 1} (1 - P(d \leq d_i - 1)), \quad (3.33)$$

tends to $d_i - \frac{d_i}{\lambda_N}$, where λ_N is finite. Hence for d_i large enough, this term increases linearly with d_i . When d_i is small, the value,

$$\frac{d_i \lambda_N}{d_i + 1} (1 - P(d \leq d_i - 1))$$

dominates. Also,

$$\frac{d_i \lambda_N}{d_i + 1} (1 - P(d \leq d_i - 1)),$$

decreases as d_i increases. Hence (3.33) offers sufficient variation such that the variance of (3.33) is not degenerate. Hence the variance of $a_{i\cdot}$ is of order at least greater than

$O((NF(N))^2)$. Therefore

$$\liminf_{N \rightarrow \infty} \frac{\sum_{i=1}^N (a_{i\cdot} - \bar{a})^2}{N\bar{a}^2} = \liminf_{N \rightarrow \infty} \frac{NN^2F(N)^2}{N(NF(N))^2} > 0,$$

so we can use condition for non-centralized measures instead of (3.10). We have

$$\begin{aligned} & \lim_{N \rightarrow \infty} \frac{[\sum_i a_{i\cdot}^3]^2}{[\sum_i a_{i\cdot}^2]^3} \\ &= \lim_{N \rightarrow \infty} \frac{[N \frac{1}{N} \sum_i a_{i\cdot}^3]^2}{[N \frac{1}{N} \sum_i a_{i\cdot}^2]^3} \\ &= \lim_{N \rightarrow \infty} \frac{[NE(a_{i\cdot}^3)]^2}{[NE(a_{i\cdot}^2)]^3} \quad \text{by WLLN} \\ &= \lim_{N \rightarrow \infty} \frac{[N(N^3F(N))^3]^2}{[N(N^2F(N))^2]^3} \\ &= \lim_{N \rightarrow \infty} N^{-1} \\ &= 0. \end{aligned}$$

Thus, for bounded λ_N , the asymptotic normality holds.

Now consider λ_N increasing with N . Then (3.28) can be further approximated as

$$\begin{aligned} E(a_{i\cdot}|d_i) &\approx NF(N) \left[d_i P(d \leq d_i) - \frac{d_i}{\lambda_N} P(2 \leq d \leq d_i + 1) + \frac{d_i \lambda_N}{d_i + 1} (1 - P(d \leq d_i - 1)) \right] \\ &\quad + \sum_{j=1}^N \frac{A_{ij}}{\min(d_i, d_j) - A_{ij} + 1} \\ &\approx NF(N) [d_i P(d \leq d_i) + \lambda_N (1 - P(d \leq d_i - 1))] \\ &\quad + d_i \sum_{d \leq d_i} \frac{1}{d} \frac{e^{-\lambda_N} \lambda_N^d}{d!} + P(d > d_i). \end{aligned} \tag{3.34}$$

This is justified because, for λ_N increasing with N ,

$$\frac{d_i}{\lambda_N} P(2 \leq d \leq d_i + 1),$$

is negligible compared to $d_i P(d \leq d_i)$. Still we have that $P(d > d_i)$ is $O(1)$ and

$$\frac{d_i}{\lambda_N} P(2 \leq d \leq d_i + 1) < d_i \sum_{d \leq d_i} \frac{1}{d} \frac{e^{-\lambda_N} \lambda_N^d}{d!} < d_i P(d \leq d_i).$$

Hence the order of moments for

$$d_i \sum_{d \leq d_i} \frac{1}{d} \frac{e^{-\lambda_N} \lambda_N^d}{d!}$$

is at most equal to the order of the corresponding moment of $d_i P(d \leq d_i)$, and therefore at most the same as that of

$$NF(N) d_i P(d \leq d_i).$$

Hence (3.34) can further be reduced to

$$E(a_i | d_i) \approx NF(N) [d_i P(d \leq d_i) + \lambda_N (1 - P(d \leq d_i - 1))]. \quad (3.35)$$

The mean value of $d_i P(d \leq d_i)$ for large N is

$$\begin{aligned} E[d_i P(d \leq d_i)] &= \sum_{d_i=0}^{\infty} \sum_{k=0}^{d_i} \frac{e^{-\lambda_N} \lambda_N^k}{k!} \frac{d_i e^{-\lambda_N} \lambda_N^{d_i}}{d_i!} \\ &= \sum_{k=0}^{\infty} \sum_{d_i=k}^{\infty} \frac{d_i e^{-\lambda_N} \lambda_N^{d_i}}{d_i!} \frac{e^{-\lambda_N} \lambda_N^k}{k!} \\ &= \sum_{k=0}^{\infty} \sum_{d_i=k}^{\infty} \frac{\lambda_N e^{-\lambda_N} \lambda_N^{d_i-1}}{(d_i-1)!} \frac{e^{-\lambda_N} \lambda_N^k}{k!} \\ &= \sum_{k=0}^{\infty} \sum_{d_i=k-1}^{\infty} \frac{\lambda_N e^{-\lambda_N} \lambda_N^{d_i}}{(d_i)!} \frac{e^{-\lambda_N} \lambda_N^k}{k!} \\ &= \lambda_N \sum_{d_i=0}^{\infty} \sum_{k=0}^{d_i+1} \frac{e^{-\lambda_N} \lambda_N^k}{k!} \frac{e^{-\lambda_N} \lambda_N^{d_i}}{d_i!} \\ &= \lambda_N E[P(d \leq d_i + 1)]. \end{aligned}$$

Then, taking the expectation of (3.35), we have

$$\begin{aligned}
E(E(a_{i.}|d_i)) &\approx NF(N)E[d_i P(d \leq d_i) + \lambda_N(1 - P(d \leq d_i - 1))] \\
&= NF(N)E[d_i P(d \leq d_i)] + E[\lambda_N(1 - P(d \leq d_i - 1))] \\
&= NF(N)[\lambda_N + \lambda_N E[P(d \leq d_i + 1)] - \lambda_N E[P(d \leq d_i - 1)]] \\
&= NF(N)[\lambda_N + \lambda_N E[P(d \leq d_i + 1) - P(d \leq d_i - 1)]] \\
&= NF(N)[\lambda_N + \lambda_N E[P(d = d_i + 1) + P(d = d_i)]] \tag{3.36}
\end{aligned}$$

The expectation in the second term can be evaluated based on the fact that for large enough λ_N , the Poisson distribution is approximated by a normal distribution with mean λ_N and variance λ_N . Then, letting $f(x)$ be probability density function of $N(\lambda_N, \lambda_N)$, we have

$$\begin{aligned}
E[P(d = d_i)] &\approx \int_{-\infty}^{\infty} f(x)^2 dx \\
&= \frac{1}{2\pi\lambda_N} \int_{-\infty}^{\infty} e^{-\frac{(x-\lambda_N)^2}{2\lambda_N} - \frac{(x-\lambda_N)^2}{2\lambda_N}} dx \\
&= \frac{1}{2\pi\lambda_N} \int_{-\infty}^{\infty} e^{-\frac{(x-\lambda_N)^2}{\lambda_N}} dx \text{ letting } y = \frac{x - \lambda_N}{\sqrt{\frac{\lambda_N}{2}}} \\
&= \frac{1}{2\pi\lambda_N} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} \sqrt{\frac{\lambda_N}{2}} dy \\
&= \frac{1}{\sqrt{4\pi\lambda_N}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
&= \frac{1}{\sqrt{4\pi\lambda_N}} \\
&\sim O\left(\frac{1}{\sqrt{\lambda_N}}\right).
\end{aligned}$$

Similarly we can show that $E[P(d = d_i + 1)]$ is also of $O(\frac{1}{\lambda_N})$. Then from (3.36)

$$E(a_{i.}) \approx NF(N)[\lambda_N + c\sqrt{\lambda_N}] \approx NF(N)\lambda_N \text{ as } \lambda_N \rightarrow \infty \text{ and } N \rightarrow \infty.$$

Then for $\lambda_N \rightarrow \infty$ and $N \rightarrow \infty$,

$$\begin{aligned}
\sum_{i=1}^N (a_{i.} - \bar{a})^2 &= N \frac{1}{N} \sum_{i=1}^N (a_{i.} - \bar{a})^2 \\
&\approx N \frac{1}{N} \sum_{i=1}^N (a_{i.} - E(a_{i.}))^2 \\
&\approx N \frac{1}{N} \sum_{i=1}^N (a_{i.} - NF(N)\lambda_N)^2 \\
&= NE_{a_{i.}} [(a_{i.} - NF(N)\lambda_N)^2] \\
&= NE_{d_i} [(NF(N))^2 (d_i P(d \leq d_i) + \lambda_N (1 - P(d \leq d_i - 1)) - \lambda_N)^2] \\
&= N(NF(N))^2 E_{d_i} [(d_i P(d \leq d_i) - \lambda_N P(d \leq d_i - 1))^2] \\
&\approx N(NF(N))^2 E_N [(d_i - \lambda_N)^2 P(d \leq d_i)^2] \\
&= N^3 (F(N))^2 \int_{-\infty}^{\infty} (x - \lambda_N)^2 F_x(x)^2 f(x) dx.
\end{aligned}$$

Here, $F_x(x)$ denote the cdf of $N(\lambda_N, \lambda_N)$, which can be further represented in terms of the standard normal cdf: $F_x(x) = \Phi(\frac{x - \lambda_N}{\sqrt{\lambda_N}})$. Here $\Phi(x)$ denotes the standard normal cumulative distribution function.

Letting $y = \frac{x - \lambda_N}{\sqrt{\lambda_N}}$, after a change of variable we have

$$\begin{aligned}
&N^3 (F(N))^2 \int_{-\infty}^{\infty} (x - \lambda_N)^2 F_x(x)^2 f(x) dx \\
&= \frac{1}{\sqrt{2\pi\lambda_N}} N^3 (F(N))^2 \int_{-\infty}^{\infty} \lambda_N y^2 \Phi(y)^2 e^{-\frac{1}{2}y^2} \sqrt{\lambda_N} dy \\
&= \frac{\lambda_N}{\sqrt{2\pi}} N^3 (F(N))^2 \int_{-\infty}^{\infty} y^2 \Phi(y)^2 e^{-\frac{1}{2}y^2} dy \text{ using integration by parts} \\
&= \frac{\lambda_N}{\sqrt{2\pi}} N^3 (F(N))^2 \left[\int_{-\infty}^{\infty} \Phi^2(y) e^{-\frac{1}{2}y^2} dy + \int_{-\infty}^{\infty} 2y \Phi(y) \frac{e^{-\frac{1}{2}y^2} e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}} dy \right] \\
&\quad - \frac{N\lambda_N}{\sqrt{2\pi}} \left[u \Phi^2 e^{-\frac{1}{2}y^2} \right] \Big|_{-\infty}^{\infty} \\
&= \frac{\lambda_N}{\sqrt{2\pi}} N^3 (F(N))^2 \left[\sqrt{2\pi} \int_0^1 \Phi^2(y) d\Phi(y) + \int_{-\infty}^{\infty} 2y \Phi(y) \frac{e^{-\frac{1}{2}y^2} e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}} dy \right] - 0
\end{aligned}$$

$$\begin{aligned}
&= \frac{\lambda_N}{\sqrt{2\pi}} N^3 (F(N))^2 \left[\frac{\sqrt{2\pi}}{3} + \int_{-\infty}^{\infty} 2y\Phi(y) \frac{e^{-y^2}}{\sqrt{2\pi}} dy \right] \\
&= \frac{\lambda_N}{\sqrt{2\pi}} N^3 (F(N))^2 \left[\frac{\sqrt{2\pi}}{3} + \frac{1}{3\pi} \right] \\
&\approx c_1 N^3 (F(N))^2 \lambda_N, \quad \text{for constant } c_1.
\end{aligned}$$

Similarly,

$$\sum_{i=1}^N (a_i - \bar{a})^3 \approx N N^3 (F(N))^3 \int_{-\infty}^{\infty} (x - \lambda_N)^3 F_x(x)^3 f(x) dx.$$

Again using a change of variable, we have

$$\begin{aligned}
&N^4 (F(N))^3 \int_{-\infty}^{\infty} (x - \lambda_N)^3 F_x(x)^3 f(x) dx \\
&= \frac{1}{\sqrt{2\pi\lambda_N}} N^4 (F(N))^3 \int_{-\infty}^{\infty} \lambda_N^{\frac{3}{2}} y^3 \Phi(y)^3 e^{-\frac{y^2}{2}} \sqrt{\lambda_N} dy \\
&= \frac{\lambda_N^{\frac{3}{2}}}{\sqrt{2\pi}} N^4 (F(N))^3 \int_{-\infty}^{\infty} y^3 \Phi(y)^3 e^{-\frac{y^2}{2}} dy \\
&\approx c_2 N^4 (F(N))^3 \lambda_N^{\frac{3}{2}}, \quad \text{for constant } c_2.
\end{aligned}$$

Hence condition (3.10), for the topological overlap measure and Poisson node degree distribution, is

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{(c_2 N^4 (F(N))^3 \lambda_N^{\frac{3}{2}})^2}{(c_1 N^3 (F(N))^2 \lambda_N)^3} = \lim_{N \rightarrow \infty} \frac{c_2}{c_1 N} = 0.$$

This completes the proof.

3.2.6 Normality when using the HD for ER graph

In order to check (3.12) (in page 53), similarly as the proof for SF and RG, we write the numerator and denominator in a similar form as (3.15), and approximate the empirical moments by the corresponding theoretical moments. For the ER random graph with node degree distribution approximated by a Poisson distribution with finite $\lambda_N = pN$ (> 0), this

is the same case as RG with finite λ_N , hence the limit for (3.25) holds when pN is finite and greater than $O(\frac{1}{N})$.

For the ER random graph with the node degree d following an approximate normal distribution: $N(pN, Np(1-p))$, with $pN \rightarrow \infty$, we have that $\frac{d_i}{N}$ follows a $N(p, \frac{p(1-p)}{N})$ distribution asymptotically. Again, denote by μ'_s the s^{th} theoretical moment of the random variable $\frac{d_i}{N}$, so that μ'_s is finite. Now let $m'_s = \frac{1}{N} \sum_{i=1}^N (\frac{d_i}{N})^s$ so that $m_s = N^s m'_s$. For a series of independent and identically distributed random variables: $Z_1 = \frac{d_1}{N}, Z_2 = \frac{d_2}{N}, \dots, Z_N = \frac{d_N}{N}$, $E[Z_i^s] = \mu_s$ exists and is finite. Again by WLLN we have

$$1 = \lim_{N \rightarrow \infty} \frac{m'_s}{\mu'_s} = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{i=1}^N (\frac{d_i}{N})^s}{\frac{\mu_s}{N^s}} = \lim_{N \rightarrow \infty} \frac{m_s}{\mu_s}, \quad s = 1, 2, 3.$$

Hence m_s and μ_s are of the same order of magnitude. Now approximate the numerator and denominator of condition (3.12). Similar to the case of RG, the numerator is 0:

$$\begin{aligned} \sum_{i=1}^N (d_i - \bar{d})^3 &= N(m_3 - 3m_2m_1 + 2m_1^3) \\ &\approx N(\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3) \\ &= 0. \end{aligned}$$

The denominator is determined by the variance σ_N^2 :

$$\sum_{i=1}^N (d_i - \bar{d})^2 = N(m_2 - m_1^2) \approx N(\mu_2 - \mu_1^2) = N^2 p(1-p) \text{ as } N \rightarrow \infty.$$

Then (3.12) indicates

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (d_i - \bar{d})^3]^2}{[\sum_i (d_i - \bar{d})^2]^3} = \lim_{N \rightarrow \infty} \frac{0}{[N^2 p(1-p)]^3} = 0.$$

Hence the limit for (3.12) also holds for large $\lambda_N = pN$. Therefore, for pN of greater order than $O(\frac{1}{N})$, condition (3.12) is always satisfied for an ER random graph. Again similarly as for the RG case, for λ_N smaller than order $O(\frac{1}{N})$, when the condition (3.12) does not hold, the average node degree in the network is $o(1)$, and it is also (as for RG) of little practical

interest.

3.2.7 Normality when using the GHD_TO for ER graph

The proof for the topological overlap measure with an ER random graph can be divided into two cases. First, consider the case when pN remains finite. Then we know that the node degree distribution of d_i follows a Poisson distribution, with parameter $\lambda = pN$. In this case, the steps will be the same as for a RG graph with λ_N finite. Hence for pN finite, condition (3.10) (in page 52) holds true.

In the second case, when pN is large and tends to infinity, we know that the node degree distribution is approximately normal with mean $\mu_N = Np$ and variance $\sigma_N^2 = Np(1 - p)$. In this case, $E(A_{iu}A_{ju}|d_i, d_j)$ can be approximated as follows. We have

$$E(A_{iu}A_{uj}|d_i, d_j) \approx \frac{d_i d_j}{T^2} \frac{d_u(d_u - 1)}{2} \sum_{\pi^*} \frac{f_{\pi^*}(T, d_u)}{d_u(d_u - 1)} \approx \frac{d_i d_j}{N^4 p^2} \sum_{\pi^*} f_{\pi^*}(T, d_u).$$

Then $\sum_u A_{iu}A_{uj}$ can further be expressed as

$$\sum_{u \neq i, j} A_{iu}A_{uj} \approx \sum_{u \neq i, j} \frac{d_i d_j}{N^4 p^2} \sum_{\pi^*} f_{\pi^*}(T, d_u) \approx d_i d_j F(N),$$

where $F(N)$ is defined as a function of N with order greater than

$$d_u(d_u - 1)P_{\{i, j, v_1, v_2, \dots, v_{d_u-2}\}} = \sum_{u \neq i, j} \frac{1}{N^4 p^2} \frac{d_u(d_u - 1)}{2} \approx \frac{((Np)^2 - 2Np)}{N^3 p^2}.$$

where

$$d_u(d_u - 1)P_{\{i, j, v_1, v_2, \dots, v_{d_u-2}\}}$$

is the lower bound for $F(N)$ (see (3.17) in page 64). For $Np \rightarrow \infty$, $F(N)$ is of order greater than $O(\frac{1}{N})$. Again, the two approximations use the WLLN. Substituting into the definition of the topological overlap measure we have

$$E(a_i|d_i) \approx \sum_{j=1}^N \frac{d_i d_j F(N) + A_{ij}}{\min(d_i, d_j) - A_{ij} + 1}$$

$$= F(N) \sum_{j=1}^N \frac{d_i d_j}{\min(d_i, d_j) - A_{ij} + 1} + \sum_{j=1}^N \frac{A_{ij}}{\min(d_i, d_j) - A_{ij} + 1}.$$

The second term above is smaller than d_i , hence the order of a moment given by this term is at most the same as that of d_i .

For $N \rightarrow \infty$, the first term can be expanded as

$$\begin{aligned} & F(N) \sum_{j=1}^N \frac{d_i d_j}{\min(d_i, d_j) - A_{ij} + 1} \\ & \approx NF(N) \left[\int_1^{d_i} \frac{d_i x}{x+1} f(x) dx + \int_{d_i}^N \frac{d_i x}{d_i+1} f(x) dx \right] \\ & \approx NF(N) \left[\int_1^{d_i} \frac{d_i x}{x+1} f(x) dx + \int_{d_i}^N \frac{d_i x}{d_i+1} f(x) dx \right] \\ & = NF(N) \left[d_i \int_1^{d_i} \frac{x}{x+1} f(x) dx + \frac{d_i}{d_i+1} \int_{d_i}^N x f(x) dx \right] \\ & \approx NF(N) \left[d_i \left(\int_{-\infty}^{d_i} f(x) dx - \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) dx \right) + \frac{d_i}{d_i+1} \int_{d_i}^{\infty} x f(x) dx \right] \\ & = NF(N) \left[d_i \Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) - d_i \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) dx + \frac{d_i}{d_i+1} \int_{d_i}^{\infty} x f(x) dx \right] \\ & = NF(N) \left[d_i \Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) - d_i \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) dx \right. \\ & \quad \left. + \frac{d_i Np}{d_i+1} - \frac{d_i}{d_i+1} \int_{-\infty}^{d_i} x f(x) dx \right]. \end{aligned} \tag{3.37}$$

The last term in formula (3.37) can further calculated as

$$\begin{aligned} \int_{-\infty}^{d_i} x f(x) dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{d_i-\mu}{\sigma}} (\sigma y + \mu) e^{-\frac{y^2}{2}} d(\sigma y), \text{ letting } y = \frac{x-\mu}{\sigma}, \\ &= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\frac{d_i-\mu}{\sigma}} \sigma y e^{-\frac{y^2}{2}} d(y) + \mu \int_{-\infty}^{\frac{d_i-\mu}{\sigma}} e^{-\frac{y^2}{2}} d(y) \right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\frac{d_i-\mu}{\sigma}} e^{-\frac{y^2}{2}} d\left(\frac{y^2}{2}\right) + \mu \Phi\left(\frac{d_i-\mu}{\sigma}\right) \\ &= \frac{\sigma}{\sqrt{2\pi}} \left(-e^{-\frac{y^2}{2}} \right) \Big|_{-\infty}^{\frac{d_i-\mu}{\sigma}} + \mu \Phi\left(\frac{d_i-\mu}{\sigma}\right) \end{aligned}$$

$$\begin{aligned}
&= \mu \Phi \left(\frac{d_i - \mu}{\sigma} \right) - \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{(d_i - \mu)^2}{2\sigma^2}} \\
&= \mu \Phi \left(\frac{d_i - \mu}{\sigma} \right) - \sigma^2 f(d_i) \\
&= Np \Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) - \sigma^2 f(d_i).
\end{aligned}$$

Plugging into (3.37), we have

$$\begin{aligned}
&F(N) \sum_{j=1}^N \frac{d_i d_j}{\min(d_i, d_j) - A_{ij} + 1} \\
&\approx NF(N) \left[d_i \Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) - d_i \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) d(x) + \frac{d_i Np}{d_i + 1} \right. \\
&\quad \left. - \frac{d_i}{d_i + 1} \left(\mu \Phi \left(\frac{d_i - \mu}{\sigma} \right) - \sigma^2 f(d_i) \right) \right] \\
&= NF(N) \left[\frac{d_i Np}{d_i + 1} + \Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) \left(d_i - \frac{d_i Np}{d_i + 1} \right) - d_i \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) d(x) \right. \\
&\quad \left. - \frac{d_i \sigma^2}{d_i + 1} f(d_i) \right].
\end{aligned}$$

Clearly, the mean value of $E(a_{i.}|d_i)$ with respect to d_i is of order $O(NF(N)Np)$, the second moment of $E(a_{i.}|d_i)$ is of order $O(N^2F(N)^2(Np)^2)$, and the third moment is of order $O(N^3(F(N)Np)^3)$.

Again we check for condition (3.31). Let

$$C(d_i) = \frac{F(N) \sum_{j=1}^N \frac{d_i d_j}{\min(d_i, d_j) - A_{ij} + 1}}{NF(N)Np}.$$

Then

$$\begin{aligned}
C(d_i) &= \frac{d_i}{d_i + 1} + \Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) \left(\frac{d_i}{Np} - \frac{d_i}{d_i + 1} \right) - \frac{d_i}{Np} \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) d(x) \\
&\quad - \frac{d_i \sigma^2}{(d_i + 1)Np} f(d_i)
\end{aligned}$$

$$\approx \frac{d_i}{d_i + 1} \left[1 - \Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) - \frac{(1-p)}{\sqrt{2\pi}} e^{-\frac{(d_i - \mu)^2}{2\sigma^2}} \right] \\ - \frac{d_i}{Np} \left[\Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) - \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) d(x) \right].$$

Since the second moment of $E(a_i | d_i)$ with respect to d_i is of order $O(N^2 F(N)^2 (Np)^2)$, then $C(d_i)$ is of order $O(1)$ at most. For large value of d_i

$$\frac{d_i}{Np} \left[\Phi \left(\frac{d_i - Np}{\sqrt{Np(1-p)}} \right) - \int_{-\infty}^{d_i} \frac{1}{x+1} f(x) d(x) \right]$$

dominates and has value close to $\frac{d_i}{Np}$. For small d_i , the value is close to $\frac{d_i}{d_i+1}$. Again, there is enough variation for $C(d_i)$, and hence we can easily check (3.31) holds. Therefore the condition involving non-centralized moments can be applied:

$$\begin{aligned} & \lim_{N \rightarrow \infty} \frac{[\sum_i a_i^3]^2}{[\sum_i a_i^2]^3} \\ &= \lim_{N \rightarrow \infty} \frac{[N \frac{1}{N} \sum_i a_i^3]^2}{[N \frac{1}{N} \sum_i a_i^2]^3} \\ &\approx \lim_{N \rightarrow \infty} \frac{[NE(a_i^3)]^2}{[NE(a_i^2)]^3} \quad \text{by WLLN} \\ &= \lim_{N \rightarrow \infty} \frac{[NN^3(F(N)Np)^3]^2}{[NN^2F(N)^2(Np)^2]^3} \\ &= \lim_{N \rightarrow \infty} N^{-1} \\ &= 0. \end{aligned}$$

The proof is therefore completed.

3.2.8 A summary of asymptotic results

In this section, we discussed and verified asymptotic normality, for two edge weights, adjacency information (HD) and the topological overlap measure (GHD_TO), for three different graph models: the SF network model (2.2.3), the RG random graph model (2.2.2) and the ER random graph model (2.2.1). Here we list a summary of results:

- For ER and RG random graph models, for both HD and GHD_TO, asymptotic normality holds true for non-degenerate average node degree (the order of average node degree is greater than $O(\frac{1}{N})$);
- For SF networks, with both HD and GHD_TO, asymptotic normality holds true for $1 < \alpha < 2$ and $\alpha \geq 3$.

3.3 Simulation Studies

In this section, we report on a number of simulation experiments carried out to study the properties of the GHD and compare it to a number of alternative approaches.

3.3.1 Asymptotic property of GHD

In the simulations, we mainly used two random graph models: the ER random graph model and the RG graph model. For an ER random graph, a $N \times N$ probability matrix was generated with each element independently uniformly distributed on $[0,1]$. A threshold p was used such that all the values greater than p are set to correspond to no edge and all values smaller than p are defined as specifying an edge. The procedure provides an ER random graph with size N and edge probability p . For an RG random graph, we uniformly scattered points on an $R = 2$ dimensional square $[0, 1]^2$, where each point corresponds to a node in the graph. Then an edge was introduced between a pair of nodes if the Euclidean distance between the corresponding two-dimensional points is smaller than a pre-determined threshold d [81]. The SF graph model was used in the power study. For detailed properties of the three graph models, please refer to Section 2.2.

In the first simulation, we assessed the asymptotic null sampling distribution of the GHD statistic. We randomly generated paired networks $\mathcal{A}$ and $\mathcal{B}$ independently from ER and RG with sizes $N = 50, 100$ and 250 with parameters p and d , which guarantees that all the networks we generated have a similar density. In each case, we simulated 100,000 random pairs of graphs and obtained the corresponding asymptotic p -values for testing independence, with three different edge weights. Additionally, we visualized parts of the results by applying quantile-quantile plots for those p -values against uniform $[0, 1]$.

In Section 3.2, we claimed that two sufficient conditions for asymptotic normality can be easily checked given two observed networks for any possible edge weight measures. Hence, in this simulation, apart from the two rigorously proved edge weights, adjacency and TO, we also applied the normal approximation to the permutation distribution for edge weight $\frac{1}{sp}$, where sp refer to the shortest path between two nodes. Figure 3.4 and Figure 3.5 are the QQ plots, for all three edge weights with different network sizes, for two graph models. In the plot, green points show all the observed p -values while the red line in each figure indicate the reference line $y = x$. These figures show that for both ER and RG networks of size greater than 100, the empirical moments of this distribution agree well with the theoretical moments of a uniform distribution with all three measures. Therefore, these results validate the permutation distributions of the GHD for all three measures, which are well approximated by a normal distribution.

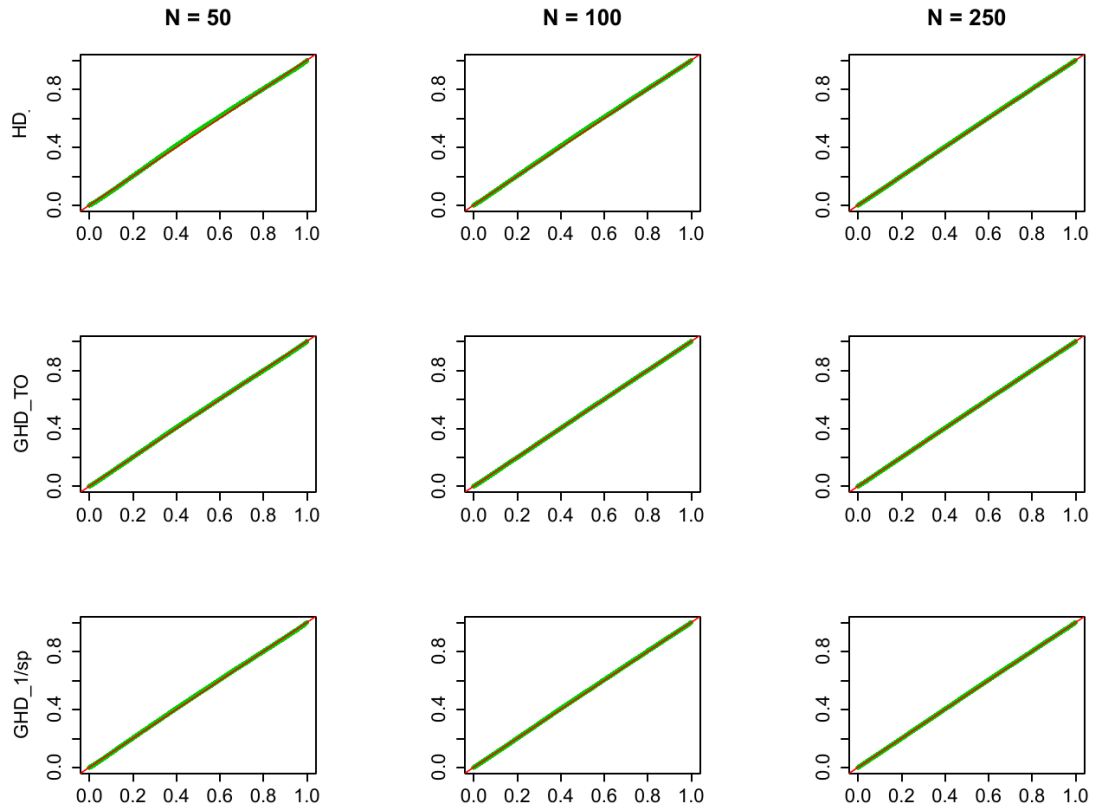


Figure 3.4: 9 QQ plots for p -values obtained from three different types of tests (HD , GHD_TO , $GHD_{-1/sp}$) with three different network sizes (50,100,250) against uniform $[0,1]$. For each one of the plots, 100,000 pairs of ER networks were generated. P-values were obtained using asymptotic distributions. In the plot, green points show all the p-values while red line indicate the reference line $y = x$. For the ER random graph model for a network size at least 100, the normal approximation performs well.

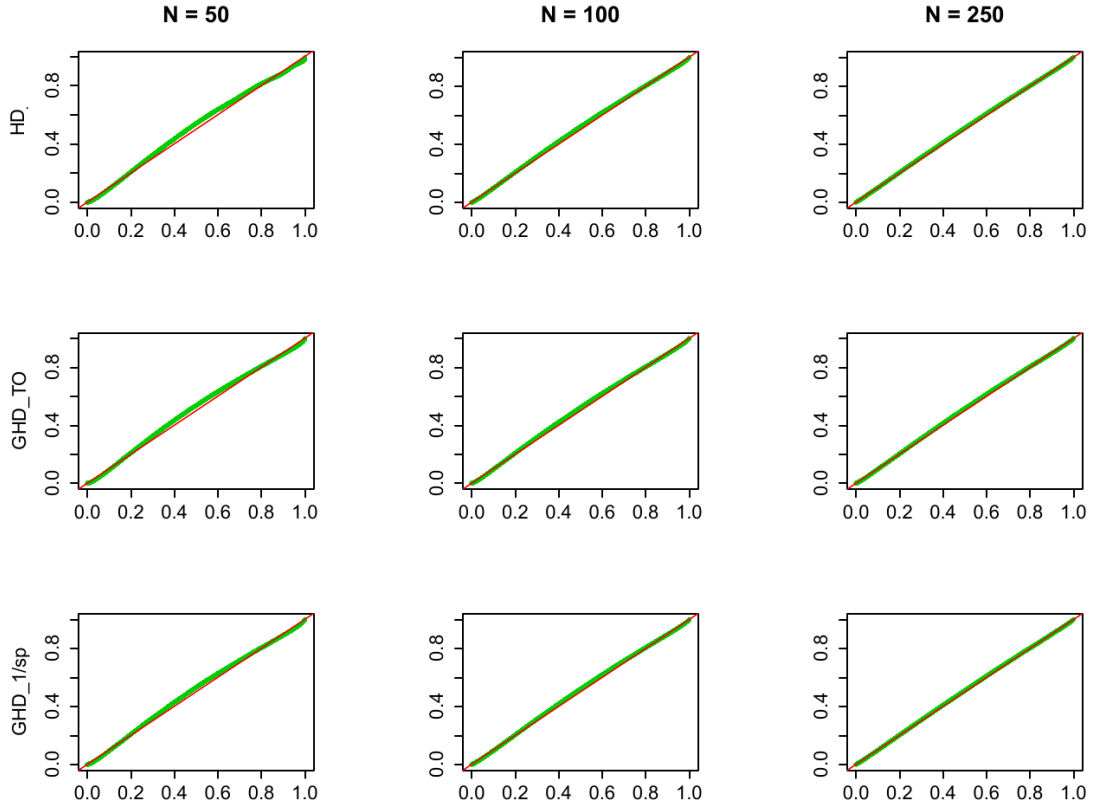


Figure 3.5: 9 QQ plots for p -values obtained from three different types of tests (HD , GHD_TO , $GHD_{-1/sp}$) with three different network sizes (50,100,250) against uniform $[0,1]$. For each one of the plots, 100,000 pairs of RG networks were generated. P-values were obtained using asymptotic distributions. In the plot, green points show all the p -values while red line indicate the reference line $y = x$. For the network size greater than 100, the normality assumption works well.

3.3.2 Power study

In the second study, we compared the power of the GHD against a competing method: the QAP with adjacency information. For the MAD procedure with adjacency, it is equivalent to QAP with adjacency. This due to the fact that QAP measures the total edge overlaps which is equals to the total number of edges sums for two networks (known constant) minus the total edge differences (MAD with adjacency). Therefore, we only consider QAP

as the competing method. For detailed formulations and procedures please refer to Section 2.3 (page 31).

In this study, three graph models, ER, RG and SF, were considered. Random ER and RG networks were generated in the same manner as in Section 3.3.1. For the random SF network, we first generated two connected nodes as the initial connected network. Then nodes were added one by one following the preferential attachment rule: the incoming nodes connect to a specific node i with probability $\frac{d_i}{\sum_j d_j}$ (please refer to Section 2.2.3, in page 31).

For this experiment, we needed to generate paired networks with a pre-specified degree of topological similarity. We present a strategy to generate two associated networks as follows. We first generated $\mathcal{A}$ with size 100 from the network models stated above. The paired network $\mathcal{B}$ was obtained by first making an exact copy of $\mathcal{A}$. Next, a fixed number of edges were randomly shuffled such that $\mathcal{A}$ and $\mathcal{B}$ differs by a fixed percentage of edges. This procedure guarantees a roughly fixed percentage of noise (percentage edge difference) between $\mathcal{A}$ and $\mathcal{B}$. We denoted this fixed percentage of noise as p_m . As p_m increases, the dissimilarity between $\mathcal{A}$ and $\mathcal{B}$ also increases. For each given value of p_m , we generated 1,000 pairs of networks, applied the competing tests, and evaluated the proportion of tests that rejected the null hypothesis of independence at a 5% significance level. P-values were adjusted for multiple testing by controlling the false discovery rate [13]. Here, the normality of the permutation distribution was assumed for edge weight $\frac{1}{sp}$ with all graph models. For graphs generated under the RG, ER and SF models, Figures 3.7, 3.6 and 3.8 show that the GHD_TO perform better than other methods and edge weights for all the graph models. $\text{GHD}_{\frac{1}{sp}}$ is also better than other existing methods, except for GHD_TO. In summary, the GHD_TO has more power to detect subtle differences between networks. It is also very interesting to see that among three different graph models, GHD_TO perform the best for ER graphs. This may due to the fact that ER networks have simpler structures than networks generated from two other graph models. For ER networks, since edges are independently generated, commonly, the path between nodes (if exists) are short and uniformly presence. Knowing that the TO measure we used in this thesis is the TO with step one, hence the TO with step one may best interpret those local structure with only short range paths. Therefore, it has the best power for ER networks.

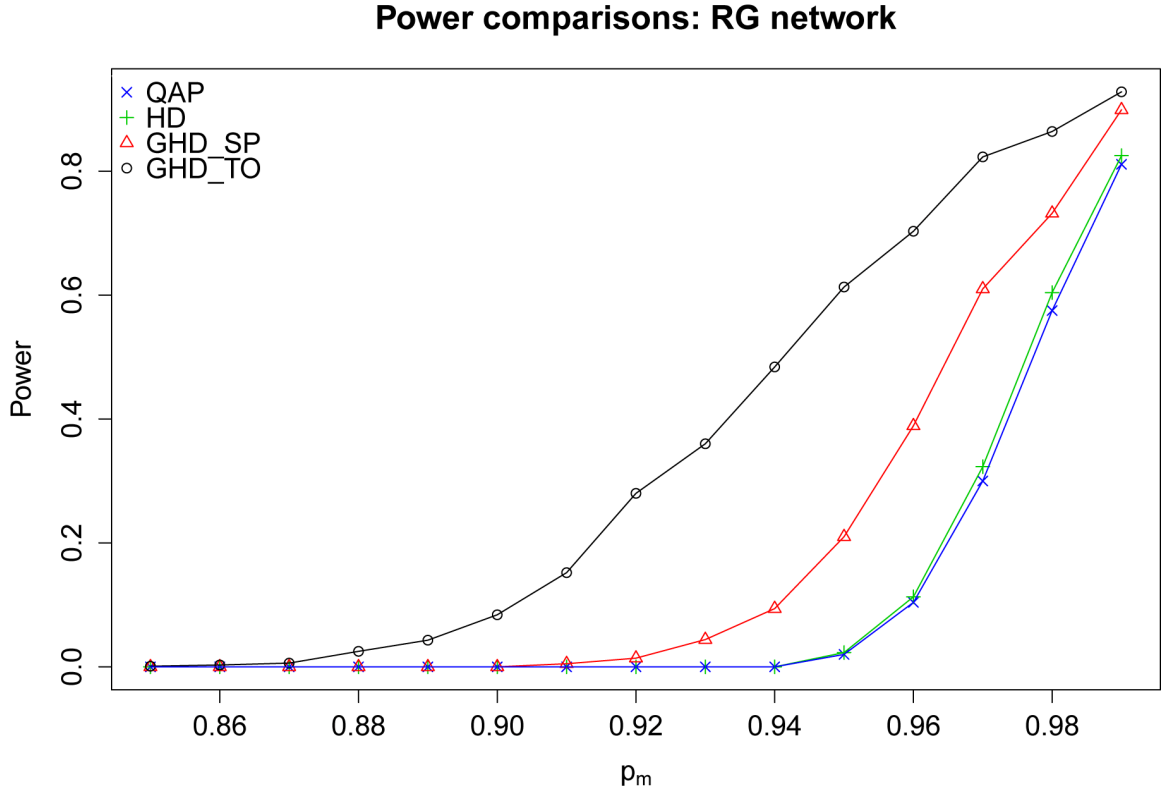


Figure 3.6: Power study using random geometric 2D graphs of size 100. The GHD with TO measure is more sensitive to detect true topological structure compared to the other alternatives.

3.4 Summary

In this chapter, we addressed the problem of comparing two graphs. We proposed a distance measure, namely the GHD, and suggested that the TO edge weight be applied with the GHD. Then we showed that for adjacency and TO measures, the sampling distribution under permutation of the GHD test statistic is approximated by a standard normal distribution, as the size of the networks increases, with three different random graph models. The numerical results showed that GHD_TO is superior to competing methods in detecting subtle differences between networks.

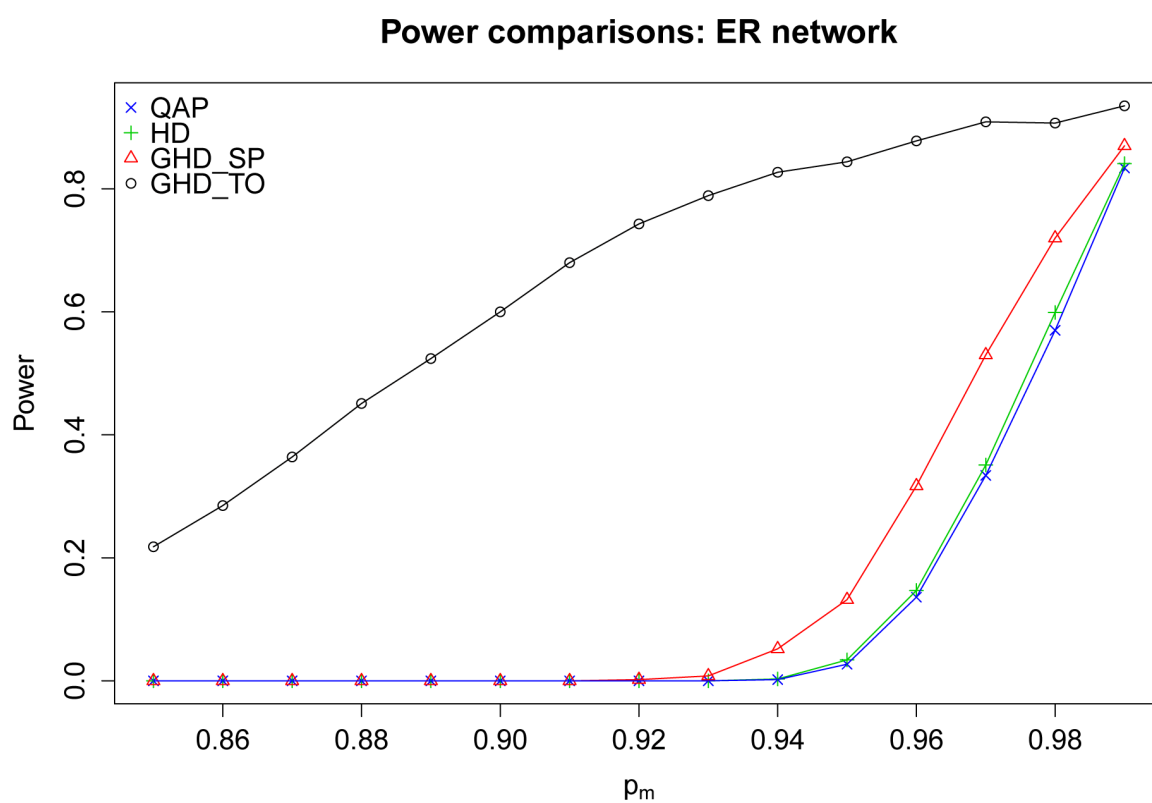


Figure 3.7: Power study using ER random graphs of size 100. The GHD with TO measure is more sensitive to detect true topological structure compared to the other alternatives.

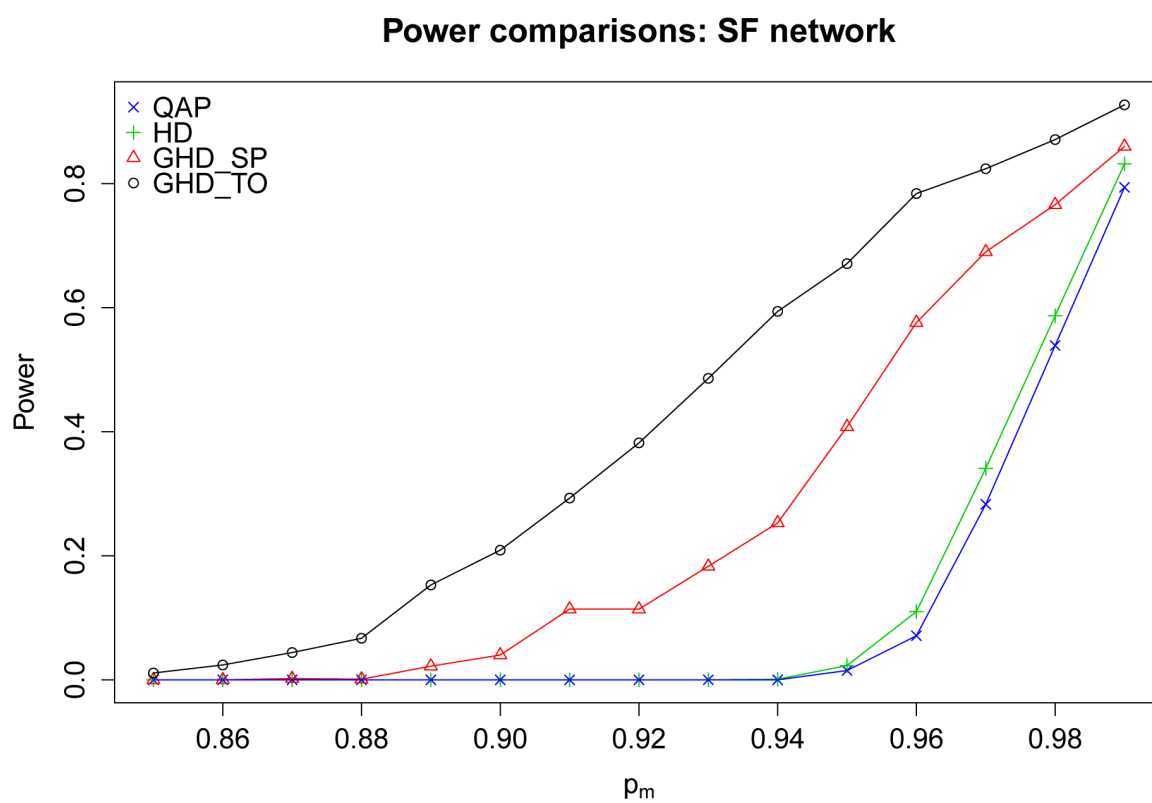


Figure 3.8: Power study using SF networks of size 100. Again the GHD with TO is more sensitive to detect true topological structure compared to the other alternatives.

Chapter 4

Differential Analysis Of Large Complex networks

When comparing large complex networks, the null hypothesis of independence can be too strong. This is because in large complex networks only localised differences in the network structure (i.e. interactions among a set of nodes) are commonly expected to be observed, and when comparing the structural differences of the entire network, such localised differences may not substantially affect the entire structure. Hence, we are more interested in the identification of a differential structured subnetwork, which is analogous to the detection of localised differences in the mean levels of some graph measures.

In this chapter, we address the problem of comparing two large complex networks. We propose a simple differential subnetwork detection algorithm that takes advantages of the GHD test we proposed in Chapter 3. Using the GHD test, we detect effectively an optimal differential subnetwork using the asymptotic property of the GHD. Extensive simulation experiments are then performed to illustrate features of the proposed algorithm. We also apply the procedure to DNA co-methylation networks, estimated for ovarian cancer patients and healthy subjects. We report on the subnetwork that presents statistically significant changes in network topology and further investigate its biological meaning.

This chapter is structured as follows. In Section 4.1, we introduce the proposed subnetwork detection algorithm. In Section 4.2, we provide a computationally efficient way of applying the algorithm to adjacency and TO measures. Simulation results are presented in

Section 4.3. In Section 4.4, we analyse a DNA co-methylation network for ovarian cancer. Finally, in Section 4.5, we present our conclusions.

4.1 An Algorithm To Detect Differential Subnetwork

We assume two large complex structured graphs $\mathcal{A}$ and $\mathcal{B}$ defined on a common node set $\mathcal{V}$ of size N . In some studies it is expected that only the local structures of $\mathcal{A}$ and $\mathcal{B}$ have been perturbed by the different experimental conditions, while the other parts of the networks are expected to remain roughly unchanged. For instance, when comparing gene expression interaction networks between case and control subjects, we would normally expect only a confined subset of genes to present altered interaction patterns. Hence, in such cases, these local altered interaction patterns are not sufficiently influential to alter the judgement that their topologies are similar globally.

Our objective here is to assess whether a connectivity pattern involving a set of nodes in one network is informative of the link structure amongst the same nodes in the paired network, and to assess the statistical significance of this finding in a computationally efficient manner. When this set of nodes exists, we call it a *differential subnetwork*.

To define it formally, we consider the task of detecting a subset $\mathcal{V}^* \subset \mathcal{V}$ for which there is no sufficient evidence to reject the null hypothesis that the corresponding networks $\mathcal{A}^*(\mathcal{V}^*, \mathcal{E}_{A^*})$ and $\mathcal{B}^*(\mathcal{V}^*, \mathcal{E}_{B^*})$ are independent. In this section, we describe an algorithm for the detection of such an optimal subset $\mathcal{V}^*$, if it exists, by making use of the GHD test.

First, we indicate by $\mathcal{V}_K$ a subset of $\mathcal{V}$ of size $K \leq N$, and define a centralised GHD value computed by comparing the two networks $\mathcal{A} = (\mathcal{V}_K, \mathcal{E}_A)$ and $\mathcal{B} = (\mathcal{V}_K, \mathcal{E}_B)$,

$$\Delta_{\mathcal{V}_K} = K(K-1)(\text{GHD}(\mathcal{A}(\mathcal{V}_K, \mathcal{E}_A), \mathcal{B}(\mathcal{V}_K, \mathcal{E}_B)) - \mu_{\mathcal{V}_K}), \quad (4.1)$$

where $\mu_{\mathcal{V}_K}$ is the mean of the permutation distribution for node set $\mathcal{V}_K$. We also define $\Delta_{\mathcal{V}_K|s}$ to be the centralised GHD value computed by comparing the networks after removing node s from a network of size K so that

$$\delta_{\mathcal{V}_K|s} = \Delta_{\mathcal{V}_K|s} - \Delta_{\mathcal{V}_K},$$

quantifies the influence that node s has on the centred GHD test when comparing two subnetworks defined on set $\mathcal{V}_K$.

We propose an iterative procedure that removes a node or set of nodes at each step, and generates a sequence of node sets of increasingly smaller size,

$$\mathcal{V}_N \supset \mathcal{V}_{N-1} \supset \dots \supset \mathcal{V}_{N_{\min}}$$

where $N_{\min} < N$ is a constant indicating the smallest expected size of the subnetwork. For each node set, the two corresponding subnetworks are compared by the GHD test, and a p -value is computed, as described previously. This procedure produces a monotonically decreasing sequence of p -values, which is expected to feature a large peak corresponding to the size of the differential subnetwork.

The algorithm works as follows. Initially, we start with node set $\mathcal{V}_N$, and obtain a p -value p_N . Obtain $\mathcal{V}_{N-1}$ by removing the node (or nodes) that correspond to the largest $\delta_{\mathcal{V}_K|s}$ with $\delta_{\mathcal{V}_K|s} > 0$ and compute the corresponding p -value, i.e. p_{N-1} . The process is repeated: given node set $\mathcal{V}_k$, a node or nodes are removed to obtain $\mathcal{V}_{k-1}$. At each stage, compare the two subgraphs to obtain a p -value, i.e. p_{k-1} . At the end of this procedure, the set of p -values $\{p_N, p_{N-1}, \dots, p_{N_{\min}}\}$ is adjusted for multiple testing, for instance by controlling the false discovery rate [14], and a corresponding set of adjusted p -values $\{p^*_N, p^*_{N-1}, \dots, p^*_{N_{\min}}\}$ is obtained. To select the optimal subgraph with significance level α , find the largest K , with $N > K > N_{\min}$, such that $p^*_K > \alpha$. At each step, the p -value for two graphs with a particular node set is evaluated. Hence, the algorithm takes advantage of the fact that p -values are computed quickly in closed-form.

4.1.1 A remark on subnetwork detection algorithm

The subnetwork detection algorithm is an iterative procedure used to detect a differential subnetwork, by constructing a sequence of subnetworks of decreasing size, the sequence being chosen to maximise a difference score (between the two networks) at each step. Suppose we have two observed networks following some random graph model of the type considered in the previous Chapter. At each stage, two networks of smaller size are obtained. Then as the procedure works through the sequence, because of the selection method used,

it is natural to be concerned that the subnetworks, say at a particular stage K , are going to be further apart than for random realisations of the assumed graph model with K nodes, and that this therefore might potentially invalidate the assumption about the normality of the null distribution.

Clearly, at each stage it would be inferentially correct to consider the null distribution of the statistic conditional on having removed the specified node having the extreme value of $\delta_{\mathcal{V}_K|s}$ at the previous step. In the context of this non-parametric analysis, in general, we can only assume that permutation will again be a reasonable way to approximate the appropriate null distribution at each stage. The sequence of p -values will certainly be dependent, but the evidence appears to be convincing that the assemblage of p -values constructed over the series of tests, suitably (Benjamini and Yekutieli [14]) corrected for multiple testing, does give reasonable properties.

The assumption we invoked above is, to some extent, reasonable. Recall that in Chapter 3, we demonstrated sufficient conditions, e.g. in the form of condition (3.7a) and (3.7b), for asymptotic normality of the null sampling distribution. An equivalent form, condition (3.10), may be written as follows,

$$\lim_{N \rightarrow \infty} \frac{[\sum_i (a_{i\cdot} - \bar{a})^3]^2}{[\sum_i (a_{i\cdot} - \bar{a})^2]^3} = 0,$$

where $a_{i\cdot} = \sum_{j \neq i} a_{ij}$ is the node level measure summing over edge weights a_{ij} , in a network $\mathcal{A}$, that involves node i . An analogous condition applies to the edge weights b_{ij} of network $\mathcal{B}$. From the above condition, we notice that the asymptotic normality of the null sampling distribution holds only if both a_{ij} and b_{ij} satisfy the sufficient condition separately (not jointly). Therefore, to validate the assumption that the null sampling distribution is still properly approximated by normal at each stage, we only need to guarantee that at each stage, condition (3.10) holds for both subnetworks. At each step, although there are only slight change in a particular network, as the procedure goes, the subnetwork we obtain may differ greatly from the original random graph model we assumed. However, the procedure is not restricted by requiring that the random graph model assumed for all descendent subnetworks has to follow exactly the same random graph model as the original network. At a particular stage, it is entirely possible that the obtained subnetwork follows a particular

random graph model having, say, a slightly different parameter to subnetworks obtained in the previous stages.

The above situation can possibly be found when comparing two networks following the RG graph model. Intuitively, at any stage, for any node we removed from a RG network, the remaining subnetwork is still a RG network (with RG pattern, i.e. nodes are connected to each other if and only if their Euclidean distance is smaller than r , say). Recall the second procedure in Section 2.2.2, used for generating RG networks with random size of network $N_{(A)}$ in a fixed region A . At least at those stages when K is still large, it is reasonable to assume that the obtained subnetwork roughly follows a RG network model, with slightly smaller node intensity in the same region A . A similar argument can possibly be applied to ER networks as well.

For scale free networks, the issue becomes quite complicated and cannot easily be discussed. However, we know that the null sampling distribution is approximately normal if the ratio in condition (3.10), denoted as the U score,

$$U = \frac{[\sum_i (a_{i\cdot} - \bar{a})^3]^2}{[\sum_i (a_{i\cdot} - \bar{a})^2]^3},$$

is reasonably small compared to the size of the network. Therefore, conservatively, one can compute the U score at each step to justify if the asymptotic normal approximation is appropriate. Notice that the subnetwork detection algorithm with U score adapted at each step can be generalized to detect differential subnetworks for any pair of observed networks: it is not restricted to the graph model assumption we made initially. In a simulation study, we present plots showing U scores at each stage during the subnetwork detection algorithm for both RG and ER random graph models. The results support the argument we have made in the previous paragraph.

4.2 Computing $\delta_{\mathcal{V}_K|s}$

In this section, we formulate $\delta_{\mathcal{V}_K|s}$, and discuss issues related to computing $\delta_{\mathcal{V}_K|s}$. First, $\Delta_{\mathcal{V}_K}$ can be written as

$$\begin{aligned}
\Delta_{\mathcal{V}_K} &= K(K-1) \left[\frac{1}{K(K-1)} \sum_{i,j} (a_{ij} - b_{ij})^2 - E_{\Pi} \left(\frac{1}{K(K-1)} \sum_{i,j} (a_{\pi(i)\pi(j)} - b_{ij})^2 \right) \right] \\
&= \left[\sum_{i,j} a_{ij}^2 + \sum_{i,j} b_{ij}^2 - 2 \sum_{i,j} a_{ij} b_{ij} - \left(\sum_{i,j} a_{ij}^2 + \sum_{i,j} b_{ij}^2 - 2 \sum_{i,j} b_{ij} E_{\Pi}(a_{\pi(i)\pi(j)}) \right) \right] \\
&= 2 \left(\sum_{i,j} b_{ij} E_{\Pi}(a_{\pi(i)\pi(j)}) - \sum_{i,j} a_{ij} b_{ij} \right) \\
&= 2 \left(\frac{\sum_{i,j} a_{ij} \sum_{i,j} b_{ij}}{K(K-1)} - \sum_{i,j} a_{ij} b_{ij} \right),
\end{aligned}$$

where a_{ij} and b_{ij} are edge weights, π is a permutation of vector $1, \dots, N$, and $\pi \in \Pi$. Denote $a_{ij|s}$ and $b_{ij|s}$ to be edge weight measures obtained from networks after removing node s , then, equivalently, we can formulate $\Delta_{\mathcal{V}_K|s}$ as

$$\Delta_{\mathcal{V}_K|s} = 2 \left(\frac{\sum_{i,j \neq s} a_{ij|s} \sum_{i,j \neq s} b_{ij|s}}{(K-1)(K-2)} - \sum_{i,j \neq s} a_{ij|s} b_{ij|s} \right).$$

Therefore $\delta_{\mathcal{V}_K|s}$ can then be expressed by edge weight matrices:

$$\begin{aligned}
\delta_{\mathcal{V}_K|s} &= \Delta_{\mathcal{V}_K|s} - \Delta_{\mathcal{V}_K} \\
&= 2 \left(\frac{\sum_{i,j \neq s} a_{ij|s} \sum_{i,j \neq s} b_{ij|s}}{(K-1)(K-2)} - \sum_{i,j \neq s} a_{ij|s} b_{ij|s} \right. \\
&\quad \left. - \frac{\sum_{i,j} a_{ij} \sum_{i,j} b_{ij}}{K(K-1)} + \sum_{i,j} a_{ij} b_{ij} \right). \tag{4.2}
\end{aligned}$$

Given networks $\mathcal{A}$ and $\mathcal{B}$, this measure can be calculated accordingly during the procedure. The computational efficiency of this measure $\delta_{\mathcal{V}_K|s}$ is crucial. This is because during the subnetwork detection, procedure $\delta_{\mathcal{V}_K|s}$ will be computed repeatedly at each step. In the following sections, we propose strategies to compute $\delta_{\mathcal{V}_K|s}$ for HD and GHD_TO.

4.2.1 A computationally efficient vector representation of $\delta_{\mathcal{V}_K|s}$ for HD

First, we denote the adjacency matrices of graphs $\mathcal{A}$ and $\mathcal{B}$ containing a vertex set $\mathcal{V}_K$ with size K to be $A_{\mathcal{V}_K}$ and $B_{\mathcal{V}_K}$, respectively. For adjacency information, after removing a particular node s , the remaining inputs in $A_{\mathcal{V}_K|s}$ and $B_{\mathcal{V}_K|s}$ (either 0 or 1) are not changed, where $A_{\mathcal{V}_K|s}$ and $B_{\mathcal{V}_K|s}$ denote the remaining adjacency matrices after removing node s . We first introduce some notation. Denote $a_{i\cdot} = \sum_j a_{ij}$, $b_{i\cdot} = \sum_j b_{ij}$ and $ab_{i\cdot} = \sum_j a_{ij}b_{ij}$.

Here, in a slight abuse of notation, let a_{ij} have the same subscript (i, j) as $a_{ij|s}$ in both matrices $A_{\mathcal{V}_K}$ and $A_{\mathcal{V}_K|s}$. Therefore, $A_{\mathcal{V}_K|s}$ can be seen as matrix $A_{\mathcal{V}_K}$ with a 0 value for all the elements involving subscript s (i.e., a_{is} , a_{si} , $i = 1, \dots, K$ are all 0). Now that we have adjacency information, we know that $a_{ij|s} = a_{ij}$ for all $i \neq s$ and $j \neq s$, hence,

$$\sum_{i,j \neq s} a_{ij|s} = \sum_{i,j} a_{ij} - a_{s\cdot} - a_{\cdot s} + a_{ss} = \sum_{i,j} a_{ij} - 2a_{s\cdot},$$

and similarly for $\sum_{i,j} b_{ij|s}$ and $\sum_{i,j} a_{ij|s}b_{ij|s}$,

$$\begin{aligned} \sum_{i,j \neq s} b_{ij|s} &= \sum_{i,j} b_{ij} - 2b_{s\cdot}, \\ \sum_{i,j \neq s} a_{ij|s}b_{ij|s} &= \sum_{i,j} a_{ij}b_{ij} - 2ab_{s\cdot}. \end{aligned}$$

Substituting in to (4.2), we have,

$$\begin{aligned} \delta_{\mathcal{V}_K|s} &= 2 \left(\frac{\sum_{i,j \neq s} a_{ij|s} \sum_{i,j \neq s} b_{ij|s}}{(K-1)(K-2)} - \sum_{i,j \neq s} a_{ij|s}b_{ij|s} - \frac{\sum_{i,j} a_{ij} \sum_{i,j} b_{ij}}{K(K-1)} + \sum_{i,j} a_{ij}b_{ij} \right) \\ &= 2 \left[\frac{(K \sum_{i,j \neq s} a_{ij|s} \sum_{i,j \neq s} b_{ij|s} - (K-2) \sum_{i,j} a_{ij} \sum_{i,j} b_{ij})}{K(K-1)(K-2)} \right. \\ &\quad \left. + \sum_{i,j} a_{ij}b_{ij} - \sum_{i,j \neq s} a_{ij|s}b_{ij|s} \right] \\ &= 2 \left[\frac{(K \sum_{i,j \neq s} a_{ij|s} \sum_{i,j \neq s} b_{ij|s} - K \sum_{i,j} a_{ij} \sum_{i,j} b_{ij} + 2 \sum_{i,j} a_{ij} \sum_{i,j} b_{ij})}{K(K-1)(K-2)} + ab_{s\cdot} \right] \\ &= 2 \left[\frac{[K((\sum_{i,j} a_{ij} - 2a_{s\cdot})(\sum_{i,j} b_{ij} - 2b_{s\cdot}) - \sum_{i,j} a_{ij} \sum_{i,j} b_{ij}) + 2 \sum_{i,j} a_{ij} \sum_{i,j} b_{ij}]}{K(K-1)(K-2)} \right. \\ &\quad \left. + ab_{s\cdot} \right] \end{aligned}$$

$$= 2 \left[\frac{[K(4a_s \cdot b_s - 2b_s \cdot \sum_{i,j} a_{ij} - 2a_s \cdot \sum_{i,j} b_{ij}) + 2 \sum_{i,j} a_{ij} \sum_{i,j} b_{ij}]}{K(K-1)(K-2)} + ab_s \right]. \quad (4.3)$$

In formula (4.3), evaluating $\delta_{\mathcal{V}_K|s}$ is only related to the original adjacency matrix with size K . When evaluating $\delta_{\mathcal{V}_K|s}$, $s = 1, \dots, K$, information from an initial matrix of size K is used. Denote the vector

$$\vec{\delta}_{\mathcal{V}_K} = (\delta_{\mathcal{V}_K|1}, \delta_{\mathcal{V}_K|2}, \dots, \delta_{\mathcal{V}_K|K})^T,$$

containing the influence on the centred GHD for each node.

Again we denote

$$\begin{aligned} \vec{a}_K &= (a_{1\cdot}, a_{2\cdot}, \dots, a_{K\cdot})^T, \\ \vec{b}_K &= (b_{1\cdot}, b_{2\cdot}, \dots, b_{K\cdot})^T, \\ \vec{ab}_K &= (ab_{1\cdot}, ab_{2\cdot}, \dots, ab_{K\cdot})^T. \end{aligned}$$

Then applying these quantities to (4.3), $\vec{\delta}_{\mathcal{V}_K}$ can be further formulated as

$$\vec{\delta}_{\mathcal{V}_K} = 2 \left[\frac{(4\vec{a}_K \cdot \vec{b}_K) - 2 \sum_{i,j} a_{ij} \vec{b}_K - 2 \sum_{i,j} b_{ij} \vec{a}_K}{((K-1)(K-2))} + \frac{2 \sum_{i,j} a_{ij} \sum_{i,j} b_{ij}}{K(K-1)(K-2)} + \vec{ab}_K \right], \quad (4.4)$$

where $\cdot$ in $\vec{a}_K \cdot \vec{b}_K$ denotes element-wise multiplication. We notice from (4.4) that when evaluating $\vec{\delta}_{\mathcal{V}_K}$, $\sum_{i,j} a_{ij}$ and $\sum_{i,j} b_{ij}$ are only required to be computed once. Additionally, all vectors $\vec{a}_K$, $\vec{b}_K$ and $\vec{ab}_K$ can be effectively evaluated by the column (or row) sum of matrices A , B and $A \cdot B$. Thus, $\delta_{\mathcal{V}_K|s}$ for adjacency information can be computed effectively using (4.4).

4.2.2 Evaluating $\delta_{\mathcal{V}_K|s}$ for GHD_TO

In this section, we present an iterative way of computing the TO matrix. In a similar way as for the adjacency case, we first analyse the relation between $\sum_{i,j} a_{ij}$ and $\sum_{i,j \neq s} a_{ij|s}$. Recall that in Chapter 3 we introduced the TO measure (3.2), for which we further represent

the numerator in (3.2) in a matrix form,

$$\begin{aligned} a_{ij} &= \frac{\sum_{l \neq i, j} A_{il} A_{lj} + A_{ij}}{\min(\sum_{l \neq i} A_{il} - A_{ij}, \sum_{l \neq j} A_{jl} - A_{ij}) + 1} \\ &= \frac{(A'A)_{ij} + A_{ij}}{\min(\sum_{l \neq i} A_{il} - A_{ij}, \sum_{l \neq j} A_{jl} - A_{ij}) + 1}, \end{aligned}$$

where A_{ij} denotes the adjacency information for nodes i and j in adjacency matrix A for graph $\mathcal{A}$, and $(A'A)_{ij}$ denotes the $(i, j)^{\text{th}}$ element of matrix $A'A$.

We first denote $T_{\mathcal{V}_K}$ to be the TO matrix and $T_{\mathcal{V}_K|s}$ to be the related TO matrix after removing a node s from $\mathcal{V}_K$, with elements $T_{\mathcal{V}_K}$ and $T_{\mathcal{V}_K|s}$ to be denoted by a_{ij} and $a_{ij|s}$, respectively. For the TO measure, we observe that when removing a node s from a set of nodes $\mathcal{V}_K$, the value $a_{ij|s}$, $i \neq s$ and $j \neq s$, is not necessarily equal to a_{ij} . More specifically, when removing node s from $\mathcal{V}_K$, the value of $a_{ij|s}$ may be different from a_{ij} if any of the nodes i or j are connected to s . This is because if a node i is connected to s , the denominator of (3.2) becomes

$$\min\left(\sum_{l \neq i} A_{il} - A_{ij} - 1, \sum_{l \neq j} A_{jl} - A_{ij}\right) + 1.$$

Hence, when

$$\sum_{l \neq i} A_{il} - 1 \leq \sum_{l \neq j} A_{lj},$$

the value of the denominator changes. Because of this property, it is difficult to obtain a representation of the form $\vec{\delta}_{\mathcal{V}_K}$ for the TO measure that is similar to that obtained for adjacency information. However, for the TO measure, we can reduce the computational complexity by proposing a recursive formulation for the TO matrix when nodes are removed iteratively.

Now suppose a node s is removed from $\mathcal{V}_K$. We denote $d_i = \sum_{l \neq i} A_{il}$ to be the node degree for node i . Here, similarly to the adjacency case, we slightly abuse the notation again. We let a_{ij} have the same subscript (i, j) as $a_{ij|s}$ in both matrices $T_{\mathcal{V}_K}$ and $T_{\mathcal{V}_K|s}$. From equation (3.2), there are two simple facts:

1. The denominator of $a_{ij|s}$ may differ from the denominator of a_{ij} if nodes i or j connect to s .
2. The numerator of $a_{ij|s}$ is equal to the numerator of $a_{ij} - 1$ if both nodes i and j

connect to s .

Expressing this in detail, we have the following cases. If both i and j connect to s , then

$$\begin{aligned}
 a_{ij|s} &= \frac{(A'A)_{ij} + A_{ij} - 1}{\min(\sum_{l \neq i} A_{il} - A_{ij} - 1, \sum_{l \neq j} A_{jl} - A_{ij} - 1) + 1} \\
 &= \frac{[(A'A)_{ij} + A_{ij}] - 1}{\min(d_i - 1, d_j - 1) - A_{ij} + 1} \\
 &= \frac{(A'A)_{ij} + A_{ij}}{\min(d_i - 1, d_j - 1) - A_{ij} + 1} - \frac{1}{\min(d_i - 1, d_j - 1) - A_{ij} + 1} \\
 &= a_{ij} \frac{\min(d_i, d_j) - A_{ij} + 1}{\min(d_i - 1, d_j - 1) - A_{ij} + 1} - \frac{1}{\min(d_i - 1, d_j - 1) - A_{ij} + 1}.
 \end{aligned}$$

If i is connected to s , but j is not connected to s , then

$$\begin{aligned}
 a_{ij|s} &= \frac{(A'A)_{ij} + A_{ij}}{\min(\sum_{l \neq i} A_{il} - A_{ij} - 1, \sum_{l \neq j} A_{jl} - A_{ij}) + 1} \\
 &= \frac{(A'A)_{ij} + A_{ij}}{\min(d_i - 1, d_j) - A_{ij} + 1} \\
 &= a_{ij} \frac{\min(d_i, d_j) - A_{ij} + 1}{\min(d_i - 1, d_j) - A_{ij} + 1}.
 \end{aligned}$$

The analogous situation is true if j is connected to s but i is not connected to s . Hence, given matrix $T_{\mathcal{V}_K}$ and node degree $d_i, i = 1, \dots, K$, $a_{ij|s}$ can be computed accordingly.

Suppose the node s has node degree d_s . Denote the set of nodes that s connects to as $\{v_1, v_2, \dots, v_{d_s}\}$. Since $a_{ij|s}$ may differ from a_{ij} if and only if nodes i or j connect to s , the matrix $T_{\mathcal{V}_K|s}$ can be computed by only changing those rows and columns involving $\{v_1, v_2, \dots, v_{d_s}\}$ in $T_{\mathcal{V}_K}$ (also ignoring the s^{th} column and row). In Chapter 3, we discussed that the computational complexity of evaluation of $T_{\mathcal{V}_K}$ corresponds to the computational complexity of calculating $A'A$, which for a matrix of size K is of $O(K^3)$. However, using the above recursive formulation for all the elements involving nodes $\{v_1, v_2, \dots, v_{d_s}\}$, there are only Kd_s elements that need to be re-calculated, while the rest of the elements in $T_{\mathcal{V}_K}$ remain unchanged. Therefore, by knowing $T_{\mathcal{V}_K}$, the computational complexity of applying the recursive formulation is $O(Kd_s)$, where $0 < d_s < K - 1$. Hence, for small d_s , it is possible to compute $T_{\mathcal{V}_K|s}$ in an approximately linear time, $O(K)$. Similarly, we can also obtain the TO matrix after removing node s for b_{ij} . Then $\delta_{\mathcal{V}_K|s}$ is able to be computed

by plugging in a_{ij} , b_{ij} , $a_{ij|s}$ and $b_{ij|s}$ to (4.2). Nevertheless, the computational cost of calculating $\vec{\delta}_{\mathcal{V}_K}$ with the TO measure is much larger than that of the adjacency approach.

Notice that the procedure of computing $T_{\mathcal{V}_K|s}$ can also be applied at each step to iteratively evaluate the TO matrix, given the index of node(s) that was deleted in the last iteration.

4.3 Simulation Study

In this simulation study, we conduct an investigation to assess the behaviour of the differential sub-network detection algorithm. We simulate a pair of networks that contain subnetworks that are independent from each other. We apply the subnetwork detection algorithm for both HD and GHD_TO to the simulated networks and further report on their performance. Part of the simulation results can also be found in [91]

4.3.1 Generating networks with localized differences

In this study, we mainly used two different graph models, ER and RG. We wanted to simulate two networks $\mathcal{A}$ and $\mathcal{B}$, with network $\mathcal{B}$ very "similar" in most parts to $\mathcal{A}$, but with a localized sub-component in $\mathcal{B}$ independent from the same component in $\mathcal{A}$. We denoted the set of all nodes by $\mathcal{V}$, and the set of nodes belongs to the independent subnetwork $\mathcal{V}^*$.

For the ER random graph model, we first generated a network $\mathcal{A}$ with node set $\mathcal{V}$ of size 1,000. Secondly, we generated another network $\mathcal{B}$ that has $p_m = 25\%$ of its edges 'rewired' to $\mathcal{A}$ (definition of p_m please refer to Section 3.3.2 page 92). We then generated a subnetwork $\mathcal{B}_{sub}$ of size 200 independently from the same ER random graph model and replaced the subnetwork consisting of the first 200 nodes (i.e. $\mathcal{V}^*$) in $\mathcal{B}$ with $\mathcal{B}_{sub}$. Hence, the resulting network $\mathcal{B}$ consists of a subnetwork of 200 nodes independent of $\mathcal{A}$, but with the remaining parts of the two networks differing by 25%.

For the RG random graph model, we proposed a network generating procedure to generate networks with RG "noises". We reported on an experiment involving 2D RG networks $\mathcal{A}$ and $\mathcal{B}$ of size 1,000. $\mathcal{A}$ was generated first, as before. For each node $i \in \mathcal{V}$ with coordinates (x_i, y_i) in $\mathcal{A}$, the corresponding coordinates for that node in $\mathcal{B}$ were ob-

tained as $(x_i + \epsilon_{ix}, y_i + \epsilon_{iy})$, where ϵ_{ix} and ϵ_{iy} are independent random variables drawn from a $N(0, \sigma^2)$ distribution. Once all the coordinates for the nodes in $\mathcal{B}$ were obtained, the Euclidean distances between all pairs of nodes were calculated and the edges of $\mathcal{B}$ were generated by thresholding. In this setting, the variance parameter σ controls the extent to which the networks differ. This parameter was chosen to roughly obtain $p_m \approx 25\%$. Finally, in each network, we introduced independent subnetworks, which we denote by $\hat{\mathcal{A}}^*$ and $\hat{\mathcal{B}}^*$, by randomly selecting a subset $\mathcal{V}^* \subset \mathcal{V}$ of size 200, and replacing the corresponding subgraphs with independent 2D RG networks.

4.3.2 Results

In the first simulation, we justify the argument we made under the discussion in Section 4.1.1. We applied the network generating procedure as stated above to generate paired networks of size $N = 1000$ in each case. The subnetwork detection algorithm was applied and U scores are computed at each stage, by using the corresponding subnetwork. For both network models, ER and RG with HD, the procedure was repeated 5 times. Each point in Figure 4.1 is an average of 5 U -scores. The evidence in Figure 4.1 seems to support the argument that by assuming ER and RG random graph models, the assumed asymptotic normality of the null sampling distribution, at each stage, is appropriate for both large and small subnetwork size (e.g. the average U -score is smaller than $\frac{1}{N}$ even when the size of the subnetwork is small).

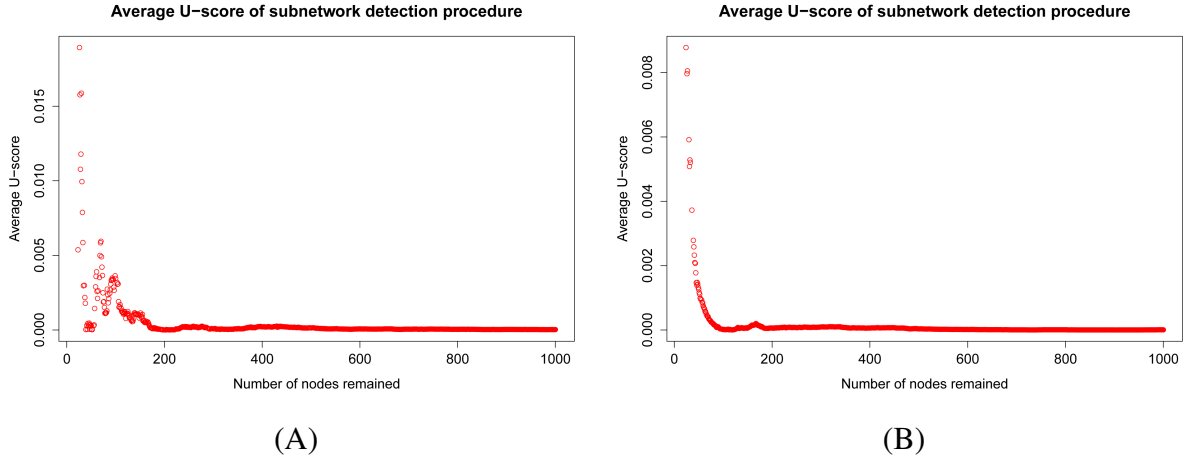


Figure 4.1: (A) and (B) are trend of average U -score with network size decrease during the subnetwork detection procedure for HD with ER (A) and RG (B) random graphs. Each point in both plots is computed as an average of 5 U -scores, obtained from 5 independent trials. These two plots clearly show that the score is very small, and therefore that the asymptotic normal assumption for the null sampling distribution at each stage is appropriate.

In the second simulation, the subnetwork detection algorithm of Section 4.1 was applied for both graph models, which generate a sequence of p -values. These p -values are shown in Figure 4.2 and Figure 4.3. Red circles represent an adjusted p -value smaller than 1%. It can be seen that the algorithm rejects the null hypothesis of independence when all the nodes are included in the comparison, and that the null hypothesis of independence is rejected as soon as the algorithm reached a subnetwork size of around 200.

Figure 4.4 and Figure 4.5 show an illustration of simulated paired networks. To better visualize both networks, we simulated $\mathcal{A}$ and $\mathcal{B}$ with a smaller size. RG networks of size 500 (with node set $\mathcal{V}$) were generated (with $p_m \approx 25\%$) with a subnetwork of size 100 for $\mathcal{V}^*$. Figure 4.4 shows both networks overlapped by true differential subnetworks $\hat{\mathcal{A}}^*$ and $\hat{\mathcal{B}}^*$ (with $\mathcal{V}^*$ of size 100 with edges in red and nodes in green), and Figure 4.5 shows both networks overlapped by subnetworks $\mathcal{A}^*$ and $\mathcal{B}^*$ (with edges in red and nodes in green) that were detected using the subnetwork detection algorithm with GHD_TO (with statistical significance = 10%). These two plots show that the subnetwork detection algorithm identifies well the differential component of $\mathcal{A}$ and $\mathcal{B}$.

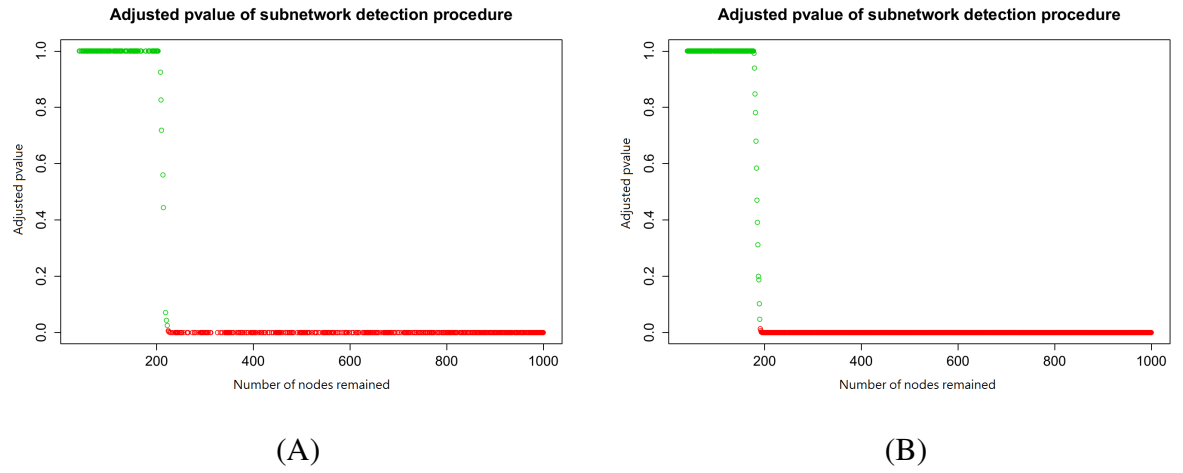


Figure 4.2: (A) and (B) are trend of adjusted p -values with network size decrease during subnetwork detection procedure for HD (A) and GHD.TO (B), with RG graphs. There is no evidence to reject the null hypothesis of independence for a differential subnetwork of around 200 and smaller.

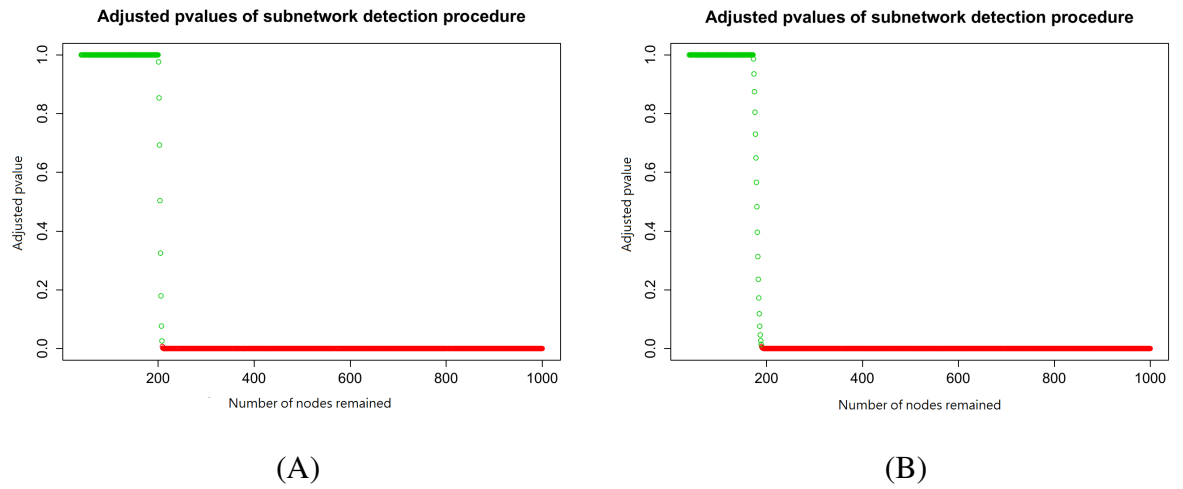


Figure 4.3: (A) and (B) are trend of adjusted p -values with network size decrease during subnetwork detection procedure for HD (A) and GHD.TO (B), with ER graphs. Again there is no evidence to reject the null hypothesis of independence for a differential subnetwork of size around 200 or smaller.

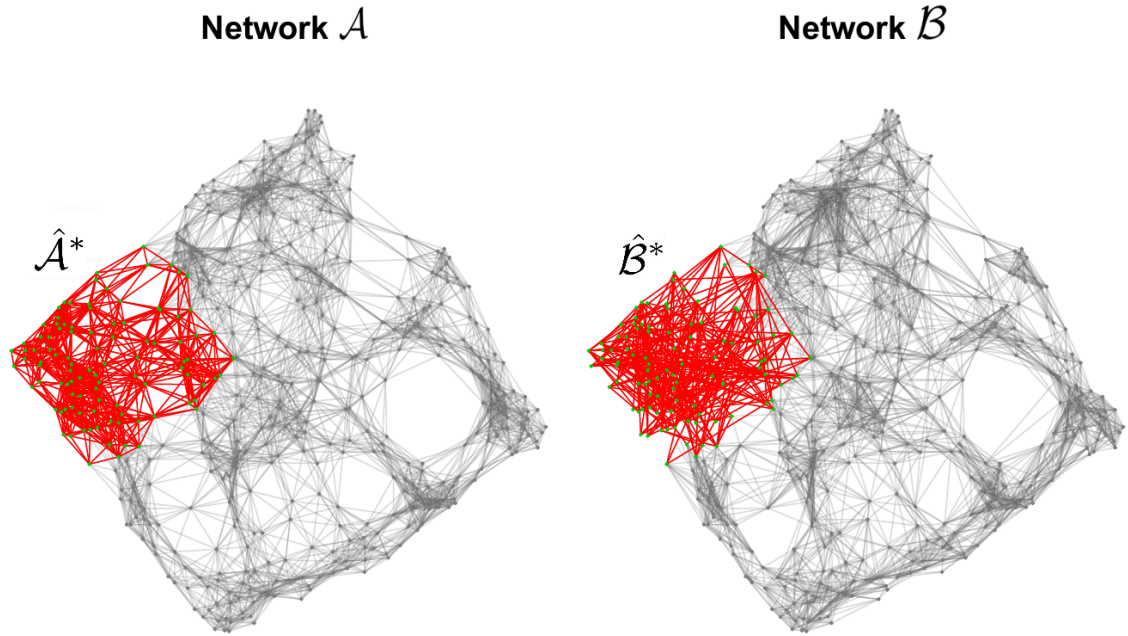


Figure 4.4: Example of the subnetwork detection procedure applied to 2D random geometric networks. Both simulated networks have 500 nodes. For a subset of 100 nodes the corresponding subnetworks are independently generated ($\mathcal{A}$ and $\mathcal{B}$). The TRUE differential subnetwork $\hat{\mathcal{A}}^*$, $\hat{\mathcal{B}}^*$ are highlighted with edges in red, and nodes in green.

In a further simulation, we carried out an investigation to assess the behaviour of the differential subnetwork detection algorithm under repeated trials. We generated 200 pairs of $\mathcal{A}$ and $\mathcal{B}$ for each graph model, with $N = 1000$ and a differential subnetwork of size 200. We then counted the number of times a particular node was selected in the resulting differential subnetwork (with significance 5%). The selection probability (number of times selected divided by 200) was calculated for each node index in each case and all the results are shown in Figures 4.6 and 4.7. These results show that all nodes belonging to the independent part of the network have high selection probabilities in all four cases. Especially

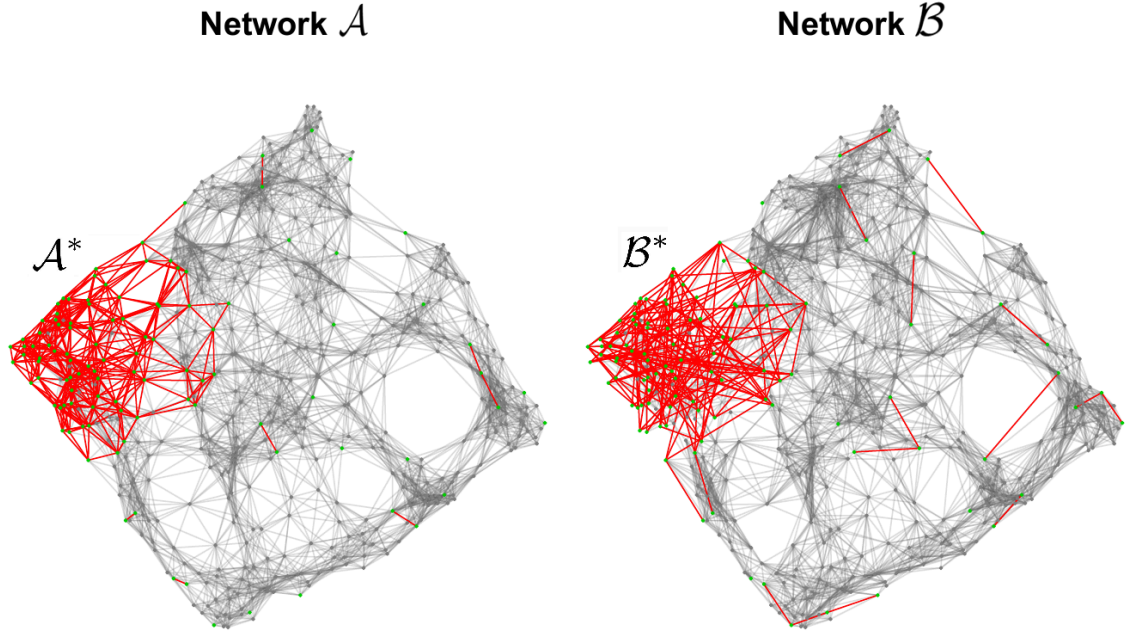


Figure 4.5: Example of subnetwork detection procedure. Subnetwork detection algorithm is applied with GHD_TO. This plot shows two networks $\mathcal{A}$ and $\mathcal{B}$ overlapped by the detected differential subnetwork (with red edges and green node) . The algorithm correctly identifies the differential subnetwork with only a few false positives . The detected differential subnetwork $\mathcal{A}^*$, $\mathcal{B}^*$ are highlighted with edges in red, and nodes in green.

in the cases of the ER graph model, the selection probability of those nodes is close to 1, with both HD and GHD_TO. Here we define the following:

- True Positive (TP), a node that is correctly classified as belonging to the differential subnetwork;
- False Negative (FN), a node that belongs to the subnetwork but has not been detected by the algorithm;
- False Positive (FP), a node that does not belong to the subnetwork but has been mistakenly classified to a differential subnetwork.

Then, as shown in Figures 4.6 and 4.7, there are more FPs and FNs in the RG graphs than the ER graphs. Therefore, the selections are more stable with the ER graph than RG graph, which shows localized patterns (see Figure 4.4).

To investigate the performance of the subnetwork detection algorithm in detecting differences between two networks, we further carried out a repeated trial to compare the GHD_TO and HD models with the RG graph model, as parameterised by σ_m , as this increases. Then for each σ_m value, 100 paired RG networks with 1000 nodes and an independent subnetwork of size 200 were generated as before. We then computed the approximate p_m for each σ_m . In Table 4.1, we report the sensitivity or true positive rate (TPR), computed as $TP/(TP+FN)$, and the specificity (SPC), computed as $TN/(FP+TN)$. These two measures show that the subnetwork detection algorithms for both the GHD and the HD have high TPRs and SPCs for small and moderate noise levels (p_m less than 23%). As can be observed, the subnetwork detection algorithm with GHD_TO maintains high specificity up to moderately high noise levels, and the TO weights yield superior performance across all values of p_m , especially for very high noise levels. We further report a F_β measure that summarises the performance in a single metric. F_β is a measure of test accuracy in binary classification [105]. It defined as:

$$F_\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2FN + FP}$$

The F_β ($\beta = 0.5$) measure of the subnetwork detection algorithm with GHD_TO is always greater than that of the HD, which shows that the GHD_TO is a better procedure for classifying differential subcomponents from a large network.

	p_m	0.055	0.11	0.23	0.54	0.79	0.95
TO	TPR	0.897	0.889	0.855	0.627	0.570	0.789
	SPC	0.987	0.984	0.974	0.912	0.768	0.439
	$F_{0.5}$	0.935	0.924	0.885	0.638	0.408	0.300
ADJ	TPR	0.914	0.904	0.872	0.725	0.712	0.862
	SPC	0.978	0.971	0.956	0.843	0.567	0.201
	$F_{0.5}$	0.912	0.890	0.840	0.566	0.330	0.250

Table 4.1: Sensitivity and specificity of subnetwork detection algorithms for 2D RG networks using both topological overlap (TO) and adjacency (ADJ) weights. For each approximated p_m , we report on average performance across 100 simulations.

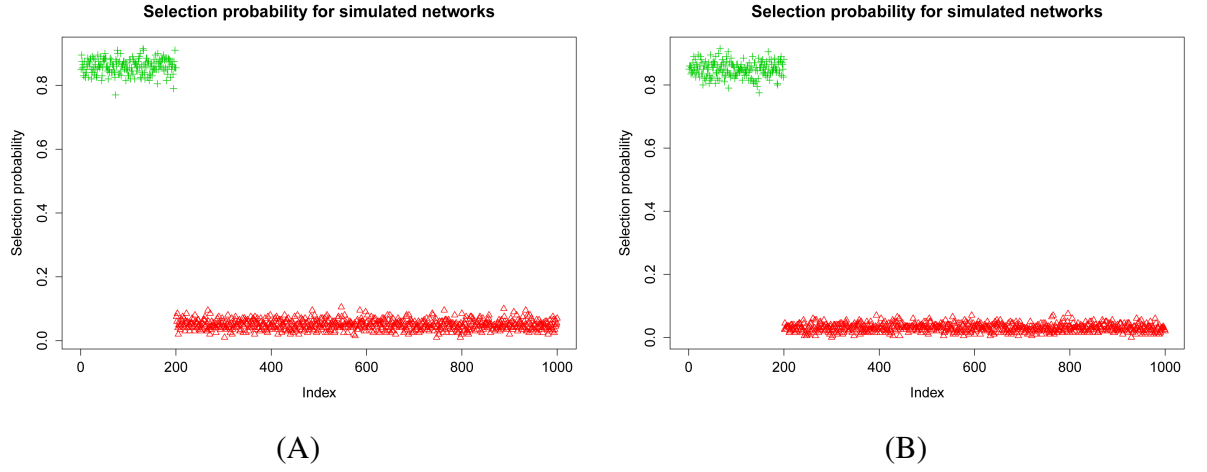


Figure 4.6: (A) and (B) are selection probability of each node with simulated RG networks for HD (A) and GHD.TO (B). Two networks of size 1000 are generated with a subnetwork of size 200 (index from 1 to 200 in green color). This procedure is repeated 200 times and finds 200 pairs of significantly (level 5%) differential subnetworks. The selection probability for each node is computed as the total number of times a particular node is selected in the 200 runs, divided by 200.

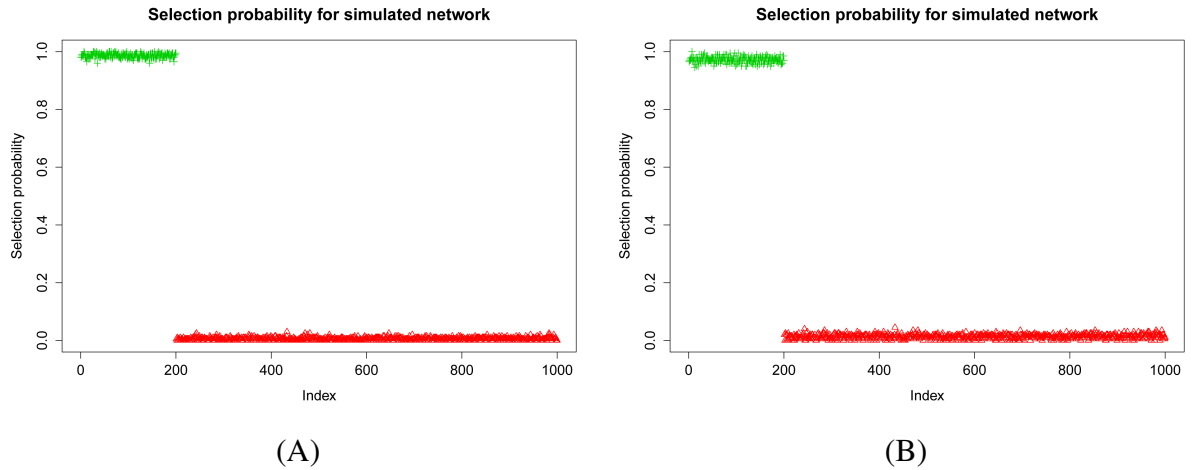


Figure 4.7: (A) and (B) are selection probability of each node with simulated ER networks for Adjacency(A) and TO (B). Two networks of size 1000 are generated with a subnetwork of size 200 (index from 1 to 200 in green color). This procedure is repeated 200 times and finds 200 pairs of significantly (level 5%) differential subnetworks. The selection probability for each node is computed as the total number of times a particular node is selected in the 200 runs, divided by 200.

4.4 An Application To Co-methylation Networks In Ovarian Cancer

As an illustration of the proposed methodology, we applied our algorithms to a case-control epigenetic study of ovarian cancer. The dataset for this study was originally presented in Teschendorff et al. [99]. Methylation profiles for 27,578 CpG islands, were obtained from whole blood samples in 540 women, of which 266 were samples taken from post-menopausal women with ovarian cancer and 274 were from age-matched healthy controls. Here CpG islands are regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length. In their original study, Teschendorff et al. [99] looked at the effects of age on DNA methylation. They also used linear regression to analyse the association between age and methylation beta value for each CpG site. Using the same data, [57] carried out a weighted correlation network analysis, based on the TO measure, with the aim of detecting consensus modules for different age groups.

In our analysis, we set out to compare control and case DNA co-methylation networks in search of a differential subnetwork. Part of the results can also be found in [91]

Raw data files were downloaded from GEO (reposition number GSE19711). DNA methylation profiles were obtained from an Illumina Infinium 27k Human DNA methylation Beadchip v1.2. The raw data were pre-processed using the `lumi` package in R [34]. After color balance adjustment, background level correction and simple scaling normalization of the raw intensities (both methylated and unmethylated intensities, details please refer to [34]), we then compute the methylation beta matrix for both patients and controls by using these intensities. The methylation beta value, denote as g_i , for i^{th} CpG site is defined as [35],

$$g_i = \frac{\max(y_{i,1}, 0)}{\max(y_{i,0}, 0) + \max(y_{i,1}, 0)}$$

where $y_{i,1}$ and $y_{i,0}$ are the intensities measured by the i^{th} methylated and unmethylated probes, respectively. Principal Components Analysis applied to the methylation beta value was used to detect and remove extreme outliers for each group (patient and control), details please refer to [99]. After quality control, 243 control samples and 215 case samples remained for further analysis. The networks were inferred by using each probe as a node. Following [57], an adjacency measure was computed as:

$$\omega_{ij} = \left| \frac{1 + \text{corr}(g_i, g_j)}{2} \right|^b$$

where $\text{corr}(g_i, g_j)$ denotes the Pearson correlation coefficient between methylation beta values observed at the i^{th} and j^{th} CpG sites. The power exponent b was set to a default value of 12 (i.e. correlation ≈ 0.75) to place more emphasis on higher positive correlations. Two nodes were linked in the network if ω_{ij} was higher than 0.2, so that the presence of an edge indicates a strong correlation. The number of edges in the control and case networks were 48,224 and 75,913, respectively. Both networks roughly follow a scale free model (See Appendix, Figure B.1).

4.4.1 Main results

We then applied a differential subnetwork detection algorithm with the GHD_TO for the above case and control networks. At a significance level of 5%, and after correction for multiple testing, we detected a subnetwork of size 1,642, with 1,954 edges in the control

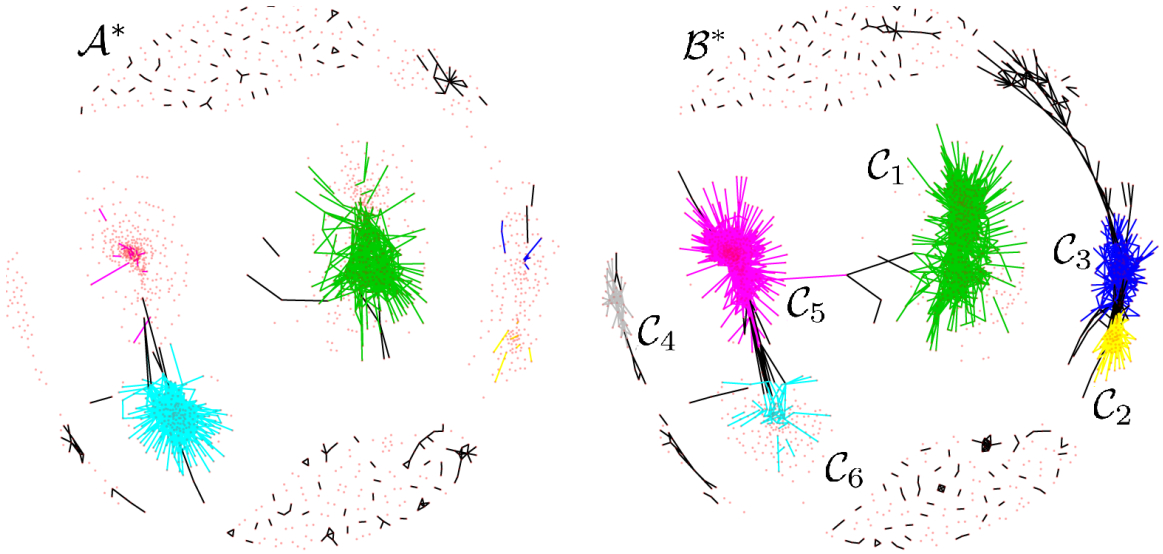


Figure 4.8: DNA co-methylation networks: differential subnetworks $\mathcal{A}^*$ (controls) and $\mathcal{B}^*$ (cases) detected by subnetwork detection algorithm. Six main communities within the subnetworks are characterised by a much higher network density in cancer patients compared to healthy controls. Differential methylation is mostly concentrated in $\mathcal{C}_2$, $\mathcal{C}_3$, $\mathcal{C}_5$ and $\mathcal{C}_6$ (See also Table 4.2 and Figure 4.9).

network $\mathcal{A}^*$ and 12,556 edges in the case network $\mathcal{B}^*$. The Walktrap algorithm implemented in the R statistical package `igraph` [84] was then used to identify communities in these two subnetworks, and the results are presented in Figure 4.8. We found that the densities of the six largest communities, which are denoted $\mathcal{C}_1, \dots, \mathcal{C}_6$, differed quite substantially between networks. In almost all communities, the density was much higher in $\mathcal{A}^*$, with the exception of $\mathcal{C}_5$, in which the density was higher in $\mathcal{B}^*$. This was also the second largest community.

To gain insight into the biological meaning of the subnetworks and the communities within them, we used the R package `GOSTat` [12] to identify enriched Gene Ontology (GO) terms within three broad categories: Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF). At a 5% significance level, the hypergeometric test detected 762 BP, 73 CC and 154 MF statistically significant terms enriched in the subnetworks. The top three BP terms were response to stimulus, cellular response to stimulus and response to chemical stimulus; the top three CC terms were related to plasma membranepart, cell surface, and cell part and the top three MF terms were protein binding,

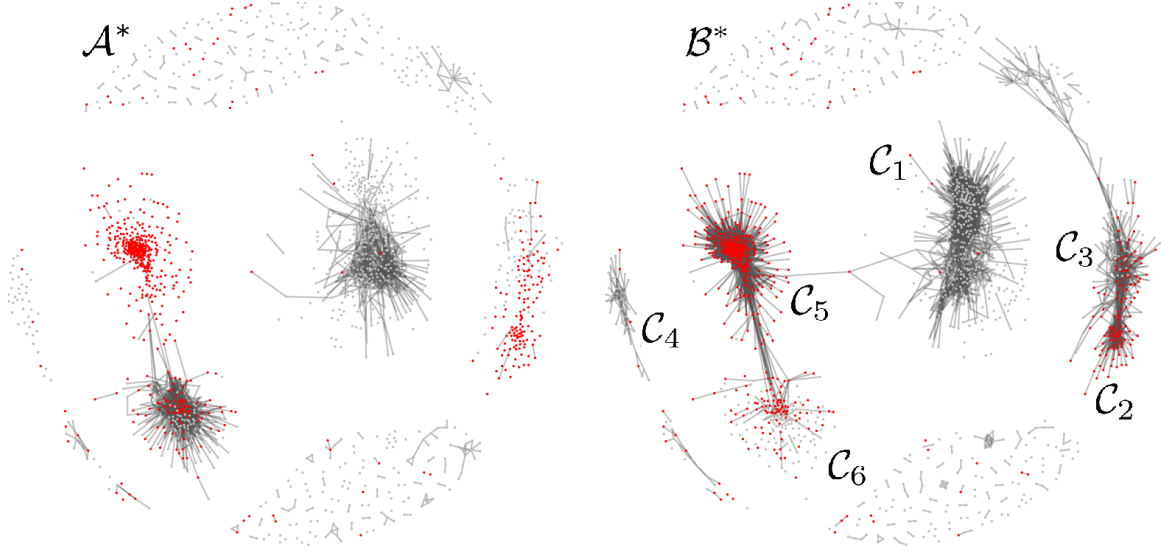


Figure 4.9: Visualization of the distribution of differential methylated probes (red) in differential subnetworks detected by subnetwork detection algorithm in the DNA co-methylation networks.

collagen binding and RNA polymerase II transcription cofactor activity. Furthermore, we carried out a pathway enrichment analysis to identify any significantly enriched KEGG pathways. Again at a 5% significance level, 12 pathways were found to be enriched, including haematopoietic cell lineage, acute myeloid leukemia, and regulation of action cytoskeleton. Tables B.1 to B.4 in the Appendix provide a list of the top 10 enriched (ranked by p -value) GO terms and KEGG terms.

To characterise the biological functions potentially involved in each of the six communities, a similar GO and KEGG enrichment analysis was then repeated independently for each community. The distribution of significantly enriched GO terms and pathways across communities are found in Table 4.2, where m_i represents the number of probes. For the largest community, C_1 , the main biological processes were system development, anatomical structure development, multicellular organismal development, and multicellular organismal process. More interestingly, for community C_5 , the main biological processes were related to interleukin-3, a protein associated with breast and ovarian cancer [69]. The level of interleukin-3 was used as a technical indicator to study chemotherapy for advanced ovarian cancer by [15]. Detailed results can be found in Tables B.8 to B.31 in the Appendix.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_6$	subtotal	overall
m_i	418	66	109	34	347	200	1174	1642
q_i	4	66	54	1	338	97	560	620
R_i	.181	.013	.012	0	.002	23.4	.145	.156
BP	320	25	38	22	236	54	568	762
CC	26	1	6	2	21	0	49	73
MF	54	4	15	3	43	27	125	154
KEGG	5	0	1	1	0	1	8	12

Table 4.2: DNA co-methylation networks: summary of the number of probes (m_i), density ratio between control and case subnetworks (R_i , $\frac{\text{control}}{\text{case}}$ density ratio), number of differentially methylated probes (q_i) and distribution of enriched GO terms and KEGG pathways in 6 communities belonging to the DC subnetworks detected by the subnetwork detection algorithm (See also Figure 4.8). Replicated GO terms and pathways involved in different communities have been excluded in the subtotal.

To compare the probes belonging to the differential subnetworks detected by the algorithm with probes showing statistically significant changes in mean methylation levels, a standard differential analysis was carried out using the two-sample SAM statistic, as implemented in the R package `samr`. After Benjamini and Hochberg [13] correction for multiple testing, 2,770 probes were found to be differentially methylated (DM) at a 5% significance level. Of these, 620 were also found in the differential subnetworks, 90% of which were concentrated in communities $\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_5$ and $\mathcal{C}_6$. Table 4.2 reports the number q_i of DM probes belonging to community $\mathcal{C}_i$, for $i = 1, \dots, 6$, and Figure 4.9 shows the distribution of DM probes in the subnetworks. The GO analysis revealed that 439 BP, 103 MF and 42 CC terms were enriched in the 620 probes, of which 277, 71, and 19, respectively, overlapped with the terms enriched in the differential subnetworks. The top-ranked overlapped GO terms can be found in Tables B.5 to B.7 in the Appendix.

4.4.2 Further studies of subnetwork detection algorithm with co-methylation data

In Section 4.4.1, by applying the algorithm to the full network, we obtained a differential co-methylated network (DC) that contains a set of 1642 probes. We then applied differential methylation (DM) analysis, which resulted in a set of 2770 DM probes. The probe

overlap between the DC and DM sets was 620, which is large.

We then attempted to understand whether the overlapping set of probes was obtained by chance. From the previous study, we knew that a set of 620 DC and DM probes was found to be differentially methylated. The question arose whether we could obtain the same set of probes by identifying differentially connected probes in the network.

We first constructed the network for DM probes using the same method as for the complete probe set. The resulting control network had 6,376 edges (density 0.00166) and the case network had 32,534 edges (0.00848). The two networks had 4612 overlapping edges in total.

The differential subnetwork detection algorithm with the TO edge weights was applied. The differential subnetwork was selected based on the first corrected p -value that was found to be greater than 5%. The resulting differential co-methylated subnetwork for DM networks was of size 715. Figure 4.10 shows the overlap among the three probe sets, DC, DM and Differential Co-methylated subnetwork for DM (DCDM). The result showed that the overlap between the DC and DCDM sets was 597. Therefore, we obtained roughly the same set of overlapped probes (597 out of 620) that were differentially connected within the DM networks.

To further assess the stability of the result, given that the networks were inferred from the data by estimating the correlation coefficients, we carried out a sub-sampling procedure to the DM probe sets (with size 2770). Each time, we randomly selected a sub-sample (with reallocation) comprising 50% of all control samples and 50% of all case samples, estimated both networks, and applied the differential subnetwork detection algorithm. We then kept track of which nodes belonged to the differential subnetwork over 500 sub-samples, and built a probability distribution over the selected nodes.

Figure 4.11 shows the selection probabilities associated with each probe, ranked in decreasing order. The green cross points are those that were initially selected by the algorithm using the entire data set. This plot shows that most of the probes selected from the full sample have a high selection probability.

For adjacency information, we applied a similar analysis to the DM probe set. The size of the subnetwork obtained was 614 for adjacency information for the full data set. For the stability analysis, we repeated the analysis for 1000 subsamples and kept track of which

Venn Diagram for three probe sets

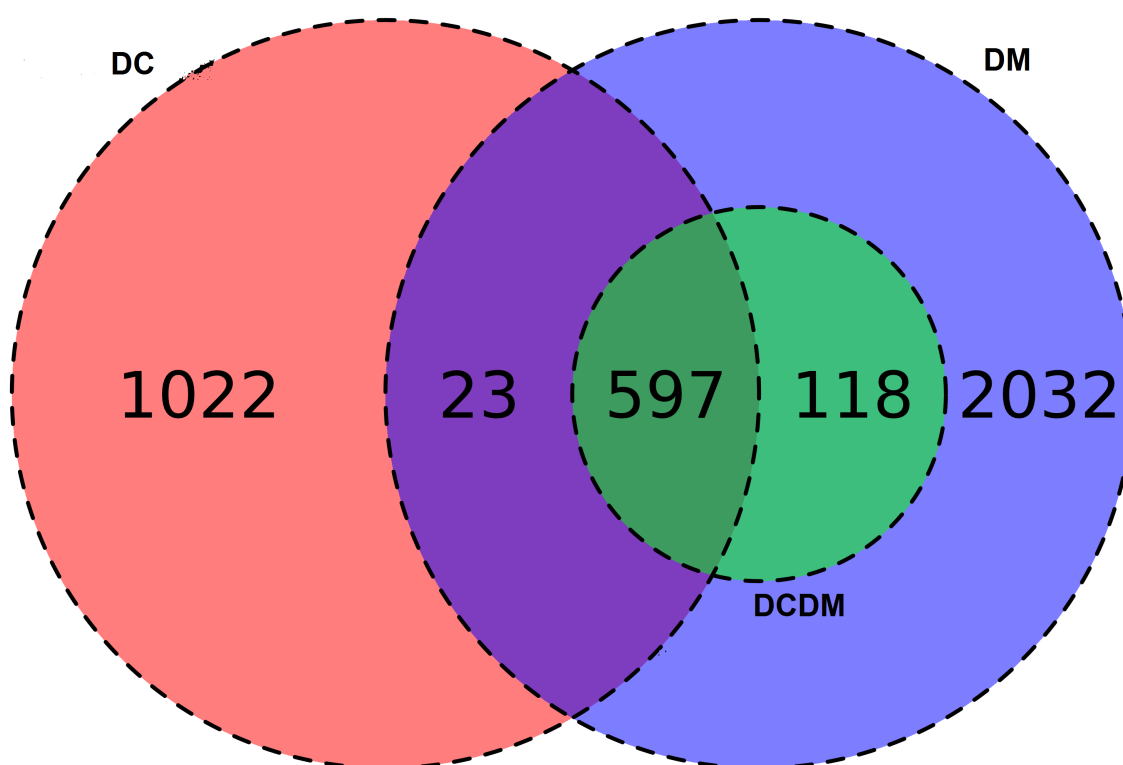


Figure 4.10: Venn Diagram for probe overlap among 3 probe sets

nodes were selected as part of the differential subnetwork. We also built a probability distribution over the selected nodes for adjacency.

Again, Figure 4.12 provides a similar argument as that for the TO measure, in which most of the probes selected from the full sample have a high selection probability.

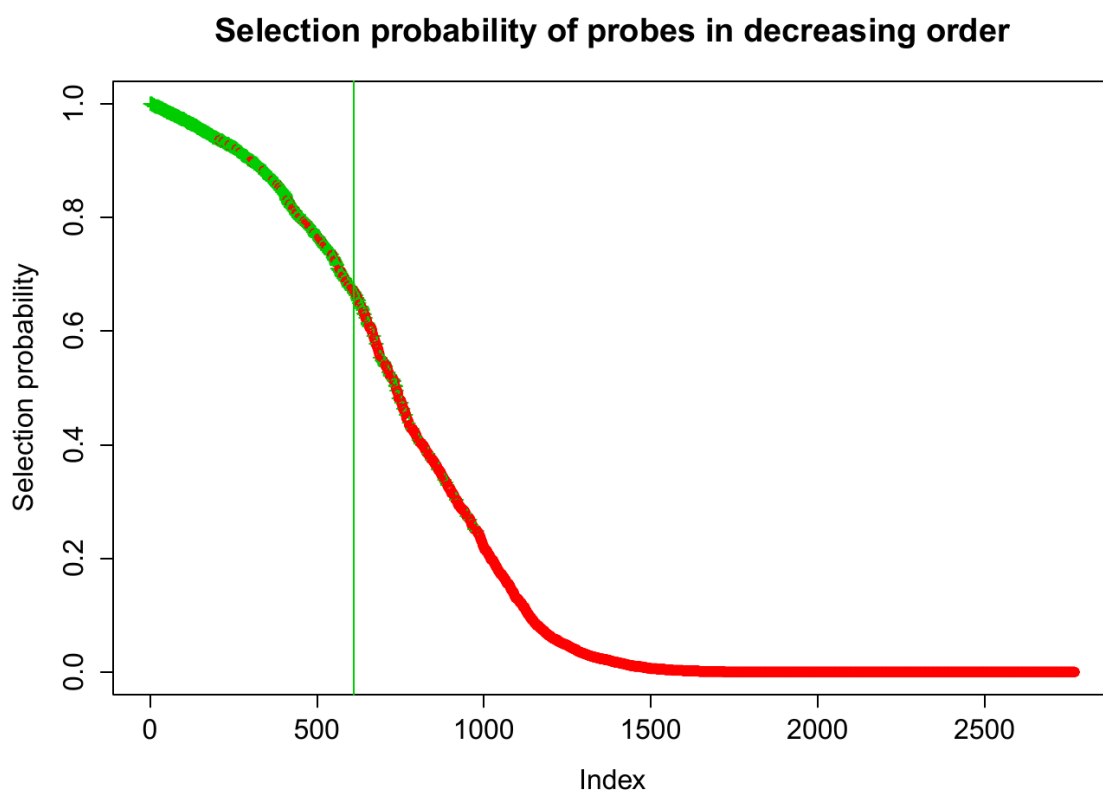


Figure 4.11: DNA co-methylation study: the probe selection probabilities are obtained by repeating the entire subnetwork detection procedure with GHD-TO on 500 subsamples with size $\frac{n}{2}$ (without replacement) where n is the total number of samples. The green cross refer to the nodes selected by the algorithm using the full sample of size n .

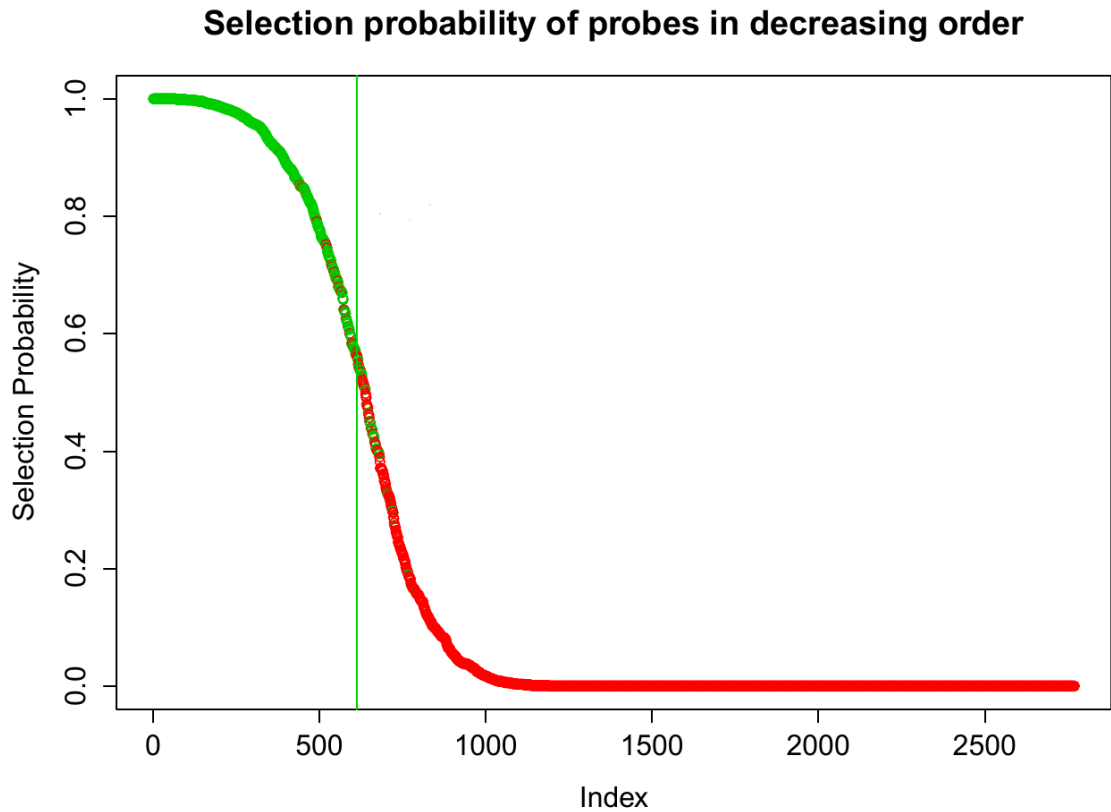


Figure 4.12: DNA co-methylation study: the probe selection probabilities are obtained by repeating the entire subnetwork detection procedure with HD on 1000 subsamples with size $\frac{n}{2}$ (without replacement), where n is the total number of samples. The green cross refer to the nodes selected by the algorithm using the full sample of size n .

4.5 Summary

In this chapter, we proposed a subnetwork detection algorithm for comparing two large complex networks with localized difference. Simulation results showed that it effectively detected a differential subnetwork using the asymptotic property of the GHD. We applied this algorithm to a co-methylation network in ovarian cancer and successfully detected a differential subnetwork. In addition, we applied a GO enrichment analysis and a path-

way enrichment analysis to explore the biological meaning of the differential subnetworks obtained by the subnetwork detection algorithm.

Chapter 5

Comparing Multiple Networks

In this section, we address the problem of comparing multiple networks. We analyse two different problems that can be interpreted with two distinct null hypotheses:

- (i) If a set of graphs are mutually independent,
- (ii) If graphs in one set are independent of graphs in the other set.

Applications for the above problem are commonly found in social network analysis (i) or neuroscience (ii). Especially in neuroscience, the development of statistical comparison tools is needed to meet the demands of comparing brain networks across groups of subjects [92]. The connectivity research in this area mainly applies descriptive models, and most existing inferential studies have employed relatively rudimentary testing procedures, such as ANOVA [71] or the permutation-based T test [68], which use graph level measures and, therefore, fail to capture the complex organization within a population of networks. To compare the mean or median representatives of two groups of networks, edge-based averaging/median across the individual entries of the connectivity matrices ignores the dependency structure within each individual network [97]. MERGP, which takes into account the connectivity patterns for all the individuals, also suffers from two main shortcomings: the reliability of the maximum likelihood estimation and computational issues when comparing large numbers of graphs. These two shortcomings of MERGP can be easily handled by applying our proposed test. We further build a link between the test of group independence and MERGP, by exploiting asymptotic normality. However, in practice, the null hypothesis

of independence for comparing multiple networks can be too strong. Hence this chapter is aimed at developing and mathematically justifying an inferential procedure for the test of independence, rather than being focussed on practical applications. The utilization for the proposed test of independence is set as a future work.

The structure of this chapter is as follows. In Section 5.1, we propose a test of mutual independence among a set of networks. In Section 5.2, we propose a test of independence between two groups of networks. In Section 5.3, we show that using standardized statistics, under the normality assumption, the maximum likelihood estimate of MERGP can be analytically derived and the inferential procedure of MERGP can be replaced by the test for comparing groups we have proposed. In Section 5.4, we discuss issues that relate to the properties and assumptions of the test. We also remark on the limitations of this work and discuss directions for future study. In Section 5.6, we present our conclusions.

Suppose we have two sets of graphs, $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$, containing P and Q graphs, respectively, on a common set of nodes with labels $\mathcal{L} = (1, \dots, N)$. Let the adjacency matrices of $\mathcal{A}_k, k = 1, \dots, P$ and $\mathcal{B}_l, l = 1, \dots, Q$ be A_k and B_l . The values in the adjacency matrices A_k and B_l corresponding to edge (v_i, v_j) are denoted by A_{kij} and B_{lij} , respectively. Again, we denote all the $N!$ possible permutations collected in a set by Π , with π being a permutation of label set $\mathcal{L}$.

5.1 A Test Of Independence Of Networks Based On Independent Permutations

In this section, we address the problem of testing mutual independence among a set of graphs $\mathcal{G}(\mathcal{A})$. Such a problem has been extensively studied in social network studies [79, 80, 4].

Consider a pair-wise comparison of a set of graphs in $\mathcal{G}(\mathcal{A})$. Let GHD_{ij} be the GHD test statistic for comparison of $\mathcal{A}_k$ with $\mathcal{A}_l$, where $k \neq l$:

$$\text{GHD}_{kl} = \frac{1}{N(N-1)} \sum_{i,j} (a_{kij} - a_{lij})^2, \quad (5.1)$$

where a_{kij} and a_{lij} are edge weights extracted from the two networks, thereby providing a

measure of connectivity between every pair of nodes i and j in $\mathcal{A}_k$ and $\mathcal{A}_l$. The similarity measures we apply here are the same as those used when comparing only two networks. Denote by

$$\text{GHD}(\mathcal{G}(\mathcal{A})) = (\text{GHD}_{12}, \text{GHD}_{13}, \dots, \text{GHD}_{kl}, \dots, \text{GHD}_{(P-1)P})^T, \\ \text{for } 1 \leq k < l \leq P,$$

the vector containing all statistics appropriate to pair-wise comparisons within the group. Now, we denote by $a_{k\pi_{kl}(i)\pi_{kl}(j)}$ the between-node measure for nodes i and j after permutation by $\pi_{kl} \in \Pi$. The corresponding permuted value of GHD_{kl} , under a generic permutation π_{kl} , is

$$\text{GHD}_{kl\pi_{kl}} = \frac{1}{N(N-1)} \sum_{i,j} (a_{k\pi_{kl}(i)\pi_{kl}(j)} - a_{lij})^2. \quad (5.2)$$

This may be evaluated for every $\pi_{kl} \in \Pi$ to infer the exact permutation distribution of GHD_{kl} under the null hypothesis of independence of the two networks.

Then consider a vector $\text{GHD}_{\pi_{kl}}$, obtained by considering all possible between-pair comparisons:

$$\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A})) = (\text{GHD}_{12\pi_{12}}, \text{GHD}_{13\pi_{13}}, \dots, \text{GHD}_{(P-1)P\pi_{(P-1)P}})^T. \quad (5.3)$$

In the following sections, under the assumption that each $\text{GHD}_{kl\pi_{kl}}$ is approximately normally distributed, we show that the components of $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$ are joint normally distributed under independent permutation. The proof of asymptotic normality of the distribution of each individual $\text{GHD}_{kl\pi_{kl}}$ can be found in Chapter 3. We then propose a chi-square-based procedure to test the null hypothesis that the networks in $G(\mathcal{A})$ are independently structured.

5.1.1 Proof of joint asymptotic normality of $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$

Suppose that the permutation distribution of each $\text{GHD}_{kl\pi_{kl}}$ is (as showed in Chapter 3) asymptotically normal. We further assume that (a.i) for different components $\text{GHD}_{kl\pi_{kl}}$,

the permutations π_{kl} used in the individual comparisons are independently generated from Π .

Therefore, to show that the joint distribution of (5.3) is asymptotically normal, we only need to show that given any (k, l) and (k', l') with $(k, l) \neq (k', l')$, the random variables $\text{GHD}_{kl\pi_{kl}}$ and $\text{GHD}_{k'l'\pi_{k'l'}}$ are asymptotically independent under (a.i).

$$\begin{aligned}
 & \text{Cov}(\text{GHD}_{kl\pi_{kl}}, \text{GHD}_{k'l'\pi_{k'l'}}) \\
 = & \text{Cov} \left(\frac{1}{N(N-1)} \sum_{i,j} (a_{k\pi_{kl}(i)\pi_{kl}(j)} - a_{tij})^2, \frac{1}{N(N-1)} \sum_{i,j} (a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} - a_{l'ij})^2 \right) \\
 = & \left(\frac{1}{N(N-1)} \right)^2 \text{Cov} \left(\sum_{i,j} a_{k\pi_{kl}(i)\pi_{kl}(j)}^2 + \sum_{i,j} a_{lij}^2 - 2 \sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)} a_{lij} \right. \\
 & \left. , \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)}^2 + \sum_{i,j} a_{l'ij}^2 - 2 \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right) \\
 = & \frac{1}{N^2(N-1)^2} \text{Cov} \left(c_1 - 2 \sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)} a_{lij}, c_2 - 2 \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right) \\
 = & \frac{2}{N^2(N-1)^2} \text{Cov} \left(\sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)} a_{lij}, \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right) \\
 = & \frac{2}{N^2(N-1)^2} \left\{ E \left[\sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)} a_{lij} \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right] \right. \\
 & \left. - E \left[\sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)} a_{lij} \right] E \left[\sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right] \right\} \\
 = & \frac{2}{N^2(N-1)^2} \left\{ E \left[\sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)} a_{lij} \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right] \right. \\
 & \left. - \sum_{i,j} a_{lij} E \left[\sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)} \right] \sum_{i,j} a_{l'ij} E \left[\sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} \right] \right\}, \tag{5.4}
 \end{aligned}$$

where c_1 and c_2 denote constants. The third equal sign follows because $\sum_{i,j} a_{kl\pi_{kl}(i)\pi_{kl}(j)}^2$ and $\sum_{i,j} a_{lij}^2$ are second moments, and, hence, they are constants under permutation; the same is true for $\sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)}^2$ and $\sum_{i,j} a_{l'ij}^2$. The first term in equation (5.4) can be

further derived as

$$\begin{aligned}
& E \left[\sum_{i,j} a_{k\pi_{kl}(i)\pi_{kl}(j)} a_{lij} \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right] \\
&= E \left[\sum_{i,j} \sum_{s,t} a_{k\pi_{kl}(i)\pi_{kl}(j)} a_{lij} a_{k'\pi_{k'l'}(s)\pi_{k'l'}(t)} a_{l'st} \right] \\
&= \sum_{i,j} \sum_{s,t} a_{lij} a_{l'st} E[a_{k\pi_{kl}(i)\pi_{kl}(j)} a_{k'\pi_{k'l'}(s)\pi_{k'l'}(t)}] \\
&= \sum_{i,j} \sum_{s,t} a_{lij} a_{l'st} E[a_{k\pi_{kl}(i)\pi_{kl}(j)}] E[a_{k'\pi_{k'l'}(s)\pi_{k'l'}(t)}] \\
&= \sum_{i,j} \sum_{s,t} E[a_{k\pi_{kl}(i)\pi_{kl}(j)}] a_{lij} E[a_{k'\pi_{k'l'}(s)\pi_{k'l'}(t)}] a_{l'st} \\
&= \sum_{i,j} E[a_{k\pi_{kl}(i)\pi_{kl}(j)}] a_{lij} \sum_{s,t} E[a_{k'\pi_{k'l'}(s)\pi_{k'l'}(t)}] a_{l'st}. \tag{5.5}
\end{aligned}$$

Applying (5.5) to (5.4), we have

$$\begin{aligned}
& Cov(\mathbf{GHD}_{kl\pi_{kl}}, \mathbf{GHD}_{k'l'\pi_{k'l'}}) \\
&= \left(\frac{2}{N(N-1)} \right)^2 \left\{ E \left[\sum_{i,j} a_{k\pi_{kl}(i)\pi_{kl}(j)} a_{lij} \sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} a_{l'ij} \right] \right. \\
&\quad \left. - \sum_{i,j} a_{lij} E \left[\sum_{i,j} a_{k\pi_{kl}(i)\pi_{kl}(j)} \right] \sum_{i,j} a_{l'ij} E \left[\sum_{i,j} a_{k'\pi_{k'l'}(i)\pi_{k'l'}(j)} \right] \right\} \\
&= \left(\frac{2}{N(N-1)} \right)^2 \left\{ \sum_{i,j} E[a_{t\pi_t(i)\pi_t(j)}] b_{tij} \sum_{s,t} E[a_{t'\pi_{t'}(s)\pi_{t'}(t)}] b_{t'st} \right. \\
&\quad \left. - \sum_{i,j} E[a_{k\pi_{kl}(i)\pi_{kl}(j)}] a_{lij} \sum_{s,t} E[a_{k'\pi_{k'l'}(s)\pi_{k'l'}(t)}] a_{l'st} \right\} \\
&= 0.
\end{aligned}$$

Thus, for any comparison, (k, l) and (k', l') and $\{(k, l) \neq (k', l')\}$, $\mathbf{GHD}_{kl\pi_{kl}}$ and $\mathbf{GHD}_{k'l'\pi_{k'l'}}$ are independent under the assumptions. This completes the proof.

5.1.2 A test of independence for $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$

Under the null hypothesis of independence that is represented by independent permutation, among the set of networks, $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$ is approximately jointly normally distributed. The first two moments of the permutation distribution for $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$ can be derived precisely and in closed-form. For a detailed formulation, please refer to the Appendix A.1.

Now denote the mean and covariance matrix of $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$ to be $\vec{\mu}_1$ and Σ_1 , respectively, and also denote the mean and standard deviation of the individual components $\text{GHD}_{kl\pi_{kl}}$ by $\mu_1(kl)$ and σ_{kl} . Then we have the mean vector

$$\vec{\mu}_1 = (\mu_1(12), \mu_1(13), \dots, \mu_1(kl), \dots, \mu_1((P-1)P))^T,$$

and covariance matrix

$$\Sigma_1 = \begin{pmatrix} \sigma_{12}^2 & 0 & \dots & 0 \\ 0 & \sigma_{13}^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_{(P-1)P}^2 \end{pmatrix}.$$

Standardizing the vector $\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A}))$ with respect to the mean and variance matrix of its asymptotic distribution, we define

$$S_{\pi_{kl}} = \Sigma_1^{-\frac{1}{2}}(\text{GHD}(\mathcal{G}_{\pi_{kl}}(\mathcal{A})) - \vec{\mu}_1). \quad (5.6)$$

Under the null hypothesis of independence of the set of networks, $S_{\pi_{kl}} \sim N(0, I)$, approximately. The observed value of the standardized score is

$$S_1 = \Sigma_1^{-\frac{1}{2}}(\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1).$$

Then we form a test of the null hypothesis of independence between the networks. Since $S_{\pi_{kl}} \sim N(0, I)$, asymptotically, $S_{\pi_{kl}}^T S_{\pi_{kl}}$ follows an approximate chi-squared distribution

with degrees of freedom $\frac{(P-1)P}{2}$ under the null hypothesis of independence:

$$C_{1\pi} = S_{\pi_{kl}}^T S_{\pi_{kl}} \sim \chi_T^2,$$

where χ_T^2 denotes the chi-squared distribution with degrees of freedom $T = \frac{(P-1)P}{2}$. More specifically, with the observed statistic S_1 value, and letting $C_1 = S_1^T S_1$, we have

$$\begin{aligned} C_1 &= S_1^T S_1 \\ &= [\Sigma_1^{-\frac{1}{2}}(\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1)]^T [\Sigma_1^{-\frac{1}{2}}(\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1)] \\ &= (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1)^T (\Sigma_1^{-\frac{1}{2}})^T \Sigma_1^{-\frac{1}{2}} (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1) \\ &= (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1)^T \Sigma_1^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1). \end{aligned}$$

Large values of C_1 constitute evidence against the null hypothesis of independence. Therefore, for any significance level α , we reject the null hypothesis of independence of the networks if

$$C_1 = (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1)^T \Sigma_1^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1) < \chi_{T,\alpha}^2,$$

where $\chi_{T,\alpha}^2$ is the α quantile of the chi-squared distribution with degrees of freedom T .

5.2 A Test Of Group Independence Based On Common Permutation

In this section, we address the problem of comparing the independence of two groups of graphs. We first consider a pair-wise comparison of graphs between two groups $G(\mathcal{A})$ and $G(\mathcal{B})$. Let GHD_{kl} be the GHD test statistic appropriate for comparing the graphs $\mathcal{A}_k$ and $\mathcal{B}_l$, so that

$$\text{GHD}_{kl} = \frac{1}{N(N-1)} \sum_{i,j} (a_{kij} - b_{lij})^2, \quad (5.7)$$

where a_{kij} and b_{lij} , as before, are edge weights extracted from the networks, thereby providing a measure of connectivity between every pair of nodes i and j in $\mathcal{A}_k$ and $\mathcal{B}_l$, respec-

tively. We also denote by

$$\begin{aligned} \text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) &= (\text{GHD}_{11}, \text{GHD}_{12}, \dots, \text{GHD}_{kl}, \dots, \text{GHD}_{PQ})^T \\ k &= 1, \dots, P \text{ and } l = 1, \dots, Q, \end{aligned}$$

a vector containing all pair-wise comparisons between the two groups of networks. To simplify the notation, we first replace the double subscript $\{k, l\}$ in (5.7) by a single subscript.

Consider a one-to-one mapping between a double subscript $\{k, l\}$ to a single subscript

$$f : \{k, l\} \rightarrow \{(k-1)P + l\}, \text{ for any } 1 \leq k \leq P, 1 \leq l \leq Q.$$

Intuitively, this relabels:

$$\begin{aligned} \text{From : } & \text{GHD}_{11}, \dots, \text{GHD}_{1Q}, \text{GHD}_{21}, \dots, \text{GHD}_{PQ} \\ \text{To : } & \text{GHD}_1, \dots, \text{GHD}_Q, \text{GHD}_{Q+1}, \dots, \text{GHD}_T \end{aligned}$$

where $T = PQ$ is the total number of comparisons in $\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}))$. Then, the above vector can be written as

$$\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) = (\text{GHD}_1, \text{GHD}_2, \dots, \text{GHD}_t, \dots, \text{GHD}_T)^T, \quad t = 1, \dots, T. \quad (5.8)$$

Similarly, and with a slight abuse of notation, under a particular permutation π , we write

$$\text{GHD}_{t\pi} = \frac{1}{N(N-1)} \sum_{i,j} (a_{t\pi(i)\pi(j)} - b_{tij})^2, \quad t = 1, \dots, T.$$

Then we consider the permutation distribution of (5.8), given that the permutation index is the same for all component comparisons $\text{GHD}_{t\pi}$, $t = 1, 2, \dots, T$.

Define $\text{GHD}(G_\pi(\mathcal{A}), G(\mathcal{B}))$ to be the value of the statistic $\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}))$ obtained under the common permutation π , and further denote

$$\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B})) = (\text{GHD}_{1\pi}, \text{GHD}_{2\pi}, \dots, \text{GHD}_{T\pi})^T. \quad (5.9)$$

In what follows, we show, under the assumption that the distribution under permutation for each $\text{GHD}_{t\pi}$ is asymptotically normal, that the distribution of the vector $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$ is approximately jointly normal. Again, a chi-squared-based procedure is proposed to test the null hypothesis that the two groups of networks $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$ are independent.

5.2.1 Proof of asymptotic normality for $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$

Again, suppose the permutation distribution of $\text{GHD}_{t\pi} = \frac{1}{N(N-1)} \sum_{i,j} (a_{t\pi(i)\pi(j)} - b_{tij})^2$ is asymptotically normal, for all $t = 1, \dots, T$. We further assumes (b.i) that within both groups $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$, networks are neither empty or complete, and they are unique, so that in particular there is no duplicated network in either group $\mathcal{G}(\mathcal{A})$ or $\mathcal{G}(\mathcal{B})$.

Applying to $\text{GHD}_{t\pi}$ results described previously for the case of just two networks, we know that the permutation distribution is asymptotically normal for three graph models with two types of edge weights.

Further, Daniels [30] has shown that the joint distribution for any two such statistics of this type tends to the bivariate normal distribution for large N , so for any $t \neq t'$, the joint distribution of $\text{GHD}_{t\pi}$ and $\text{GHD}_{t'\pi}$ is asymptotically bivariate normal.

It is well known that the joint distribution of a collection of independent normal random variables is normal and that any jointly normally distributed random variables can be constructed by nonsingular linear transformation of independent normal random variables [46]. Hence, proving the joint normality of the quantities $\text{GHD}_{t\pi}$ is now equivalent to showing that there exists a nonsingular linear transformation generating them from independent normal random variables.

To formulate this idea precisely, let $z_1, z_2, \dots, z_T$ be a set of independent normal random variables and $\mathbf{Z} = (z_1, z_2, \dots, z_T)$, where $T = PQ$. Then:

$\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$ is jointly normally distributed if $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B})) = L\mathbf{Z} + \mu$ for some non-singular matrix L .

Here μ is a constant vector of length T . Notice that if $z_1, z_2, \dots, z_T$ are independent, identically distributed normal random variables with $z_t \sim N(0, 1)$, then the matrix L is the lower triangular matrix obtained by the Cholesky decomposition of the covariance matrix, i.e. $\Sigma = LL'$ where Σ is the covariance matrix of $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$. We know that

matrix L is non-singular if Σ is positive definite.

Also, Σ is positive definite if the determinants of all its principal minors are positive, $\det(\Sigma_k) > 0$ for $1 \leq k \leq T$.

Now we will use mathematical induction to show that the determinants for all principal minors of Σ are actually positive under the two assumptions.

$k = 1$: by the assumption (b.ii), that all graphs in both groups $\mathcal{G}_A$ and $\mathcal{G}_B$ are non-empty and not complete, the variance of each $\text{GHD}_{t\pi}$, $t = 1, \dots, T$ is greater than 0.

$k = 2$: by the assumption (b.ii), we know also that for any $t_1 \neq t_2$, the two values $\text{GHD}_{t_1\pi}$ and $\text{GHD}_{t_2\pi}$ cannot be the same. Further, by the assumption (b.i) and Daniels [30], the joint distribution of $\text{GHD}_{t_1\pi}$ and $\text{GHD}_{t_2\pi}$ ($t_1 \neq t_2$) is bivariate normal. Hence, any 2 by 2 minor is positive definite. Therefore, the determinant of any minor of size 2 is positive.

Assume that for $k = s - 1$, $3 \leq s \leq T - 1$, we have that for any set of $s - 1$ random variables

$$(\text{GHD}_{t_1\pi}, \text{GHD}_{t_2\pi}, \dots, \text{GHD}_{t_{s-1}\pi})^T,$$

the covariance matrix Σ_{s-1} is positive definite.

Now consider $k = s$, and any subset of s variables from $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$:

$$\text{GHD}_{st\pi} = (\text{GHD}_{t_1\pi}, \text{GHD}_{t_2\pi}, \dots, \text{GHD}_{t_s\pi})^T.$$

From the previous arguments, all Σ_{s-j} ($j = 1, \dots, s - 1$) are positive definite. We know that a covariance matrix is at least a positive semi-definite matrix. It is hence equivalent to show that the determinant of Σ_s is greater than 0. Without loss of generality, we let $\text{GHD}_{1\pi}$ be the new variable added to a vector of $s - 1$ variables. Then writing Σ_s in the form of a partitioned matrix, let

$$\Sigma_s = \begin{bmatrix} \sigma_{11}^2 & \vec{\sigma}_{1s-1}^T \\ \vec{\sigma}_{1s-1} & \Sigma_{s-1} \end{bmatrix},$$

where σ_{11}^2 is the variance of $\text{GHD}_{1\pi}$. According to the assumption (b.i), we know that $\sigma_{11}^2 > 0$. Also, $\vec{\sigma}_{1s-1} = (\sigma_{1t_1}, \sigma_{1t_2}, \dots, \sigma_{1t_{s-1}})^T$ denotes the column vector containing all the covariances between $\text{GHD}_{1\pi}$ and $(\text{GHD}_{t_1\pi}, \text{GHD}_{t_2\pi}, \dots, \text{GHD}_{t_{s-1}\pi})^T$. Expanding $\det(\Sigma_s)$

with respect to its first row and first column, we have

$$\det(\Sigma_s) = \sigma_{11}^2 \det(\Sigma_{s-1}) + \sum_{i=2}^n (-1)^i \sigma_{1i} \sum_{j=2}^n (-1)^{j-1} \sigma_{j1} \det(\Sigma_{s-1}^{ij}), \quad (5.10)$$

where the Σ_{s-1}^{ij} denote minors formed from Σ_{s-1} by deleting row i and column j . Because Σ_{s-1} is positive definite, (i.e. $\det(\Sigma_{s-1}) > 0$), we know that Σ_{s-1}^{-1} exists, hence we have that the $(i, j)^{th}$ element of Σ_{s-1}^{-1} is given by

$$(\Sigma_{s-1}^{-1})_{ij} = \frac{(-1)^{i+j} \det(\Sigma_{s-1}^{ij})}{\det(\Sigma_{s-1})}.$$

Substituting in (5.10),

$$\det(\Sigma_s) = \sigma_{11}^2 \det(\Sigma_{s-1}) - \det(\Sigma_{s-1}) \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \vec{\sigma}_{1s-1}. \quad (5.11)$$

Consider a vector $\vec{x} = (x_0, \vec{x}_1^T)$ with $\vec{x}_1^T = (x_1, x_2, \dots, x_{s-1})$ of length $s - 1$ and,

$$x_0^2 + \vec{x}_1^T \vec{x}_1 = 1.$$

Define the quadratic form which will verify the positive definite property:

$$\begin{aligned} \Theta &= \vec{x} \Sigma_s \vec{x}^T \\ &= (x_0, \vec{x}_1^T) \Sigma_s (x_0, \vec{x}_1^T)^T \\ &= x_0^2 \sigma_{11}^2 + 2x_0 \vec{\sigma}_{1s-1}^T \vec{x}_1 + \vec{x}_1^T \Sigma_{s-1} \vec{x}_1. \end{aligned} \quad (5.12)$$

By (5.11) we can re-express σ_{11}^2 as

$$\sigma_{11}^2 = \frac{\det(\Sigma_s) + \det(\Sigma_{s-1}) \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \vec{\sigma}_{1s-1}}{\det(\Sigma_{s-1})}.$$

Substituting this formula into (5.12), we have

$$\Theta = x_0^2 \sigma_{11}^2 + 2x_0 \vec{\sigma}_{1s-1}^T \vec{x}_1 + \vec{x}_1^T \Sigma_{s-1} \vec{x}_1$$

$$\begin{aligned}
&= x_0^2 \frac{\det(\Sigma_s) + \det(\Sigma_{s-1}) \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \vec{\sigma}_{1s-1}}{\det(\Sigma_{s-1})} + 2x_0 \vec{\sigma}_{1s-1}^T \vec{x}_1 + \vec{x}_1^T \Sigma_{s-1} \vec{x}_1 \\
&= x_0^2 \frac{\det(\Sigma_s)}{\det(\Sigma_{s-1})} + x_0^2 \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \vec{\sigma}_{1s-1} + 2x_0 \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \Sigma_{s-1} \vec{x}_1 \\
&\quad + \vec{x}_1^T \Sigma_{s-1} \Sigma_{s-1}^{-1} \Sigma_{s-1} \vec{x}_1 \\
&= x_0^2 \frac{\det(\Sigma_s)}{\det(\Sigma_{s-1})} + (x_0 \vec{\sigma}_{1s-1} + \Sigma_{s-1} \vec{x}_1^T)^T \Sigma_s^{-1} (x_0 \vec{\sigma}_{1s-1} + \Sigma_{s-1} \vec{x}_1^T). \quad (5.13)
\end{aligned}$$

Since Σ_{s-1} is positive definite, we know that Σ_{s-1}^{-1} is also positive definite. The second term in Θ is non-negative, and it is strictly positive if $x_0 = 0$. This is because when $x_0 = 0$ we must have the $\|\vec{x}_1\| > 0$, where $\|\cdot\|$ denotes the Euclidean norm of a vector. Therefore the second term becomes $\vec{x}_1^T \Sigma_{s-1} \vec{x}_1 > 0$.

For the case when $x_0 \neq 0$, since both the first and second terms in (5.13) are non-negative, if $\Theta = 0$, we can infer that both terms are 0. Examining the first term we have

$$x_0^2 \frac{\det(\Sigma_s)}{\det(\Sigma_{s-1})} = 0 \implies \det(\Sigma_s) = 0.$$

Similarly, examining the second term gives

$$\begin{aligned}
&(x_0 \vec{\sigma}_{1s-1} + \Sigma_{s-1} \vec{x}_1^T)^T \Sigma_s^{-1} (x_0 \vec{\sigma}_{1s-1} + \Sigma_{s-1} \vec{x}_1^T) = 0 \\
\implies &(x_0 \vec{\sigma}_{1s-1} + \Sigma_{s-1} \vec{x}_1^T) = \vec{0} \\
\implies &\vec{\sigma}_{1s-1} = -\frac{1}{x_0} \Sigma_{s-1} \vec{x}_1^T.
\end{aligned}$$

This first implication $\implies$ is true since Σ_{s-1}^{-1} is positive definite. Substituting into (5.11) we have

$$\begin{aligned}
&\det(\Sigma_s) = \sigma_{11}^2 \det(\Sigma_{s-1}) - \det(\Sigma_{s-1}) \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \vec{\sigma}_{1s-1} \\
\implies &0 = \sigma_{11}^2 \det(\Sigma_{s-1}) - \det(\Sigma_{s-1}) \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \vec{\sigma}_{1s-1} \\
\implies &\sigma_{11}^2 = \vec{\sigma}_{1s-1}^T \Sigma_{s-1}^{-1} \vec{\sigma}_{1s-1} \\
\implies &\sigma_{11}^2 = \frac{1}{x_0} (\Sigma_{s-1} \vec{x}_1^T)^T \Sigma_{s-1}^{-1} \frac{1}{x_0} \Sigma_{s-1} \vec{x}_1^T \\
\implies &\sigma_{11}^2 = \frac{1}{x_0^2} \vec{x}_1^T \Sigma_{s-1} \vec{x}_1^T
\end{aligned}$$

$$\implies \text{Var}(\text{GHD}_{1\pi}) = \text{Var}\left(\frac{\vec{x}_1}{x_0} \text{GHD}_{st\pi}\right).$$

Letting $\vec{\omega}_{s-1}^T = \frac{\vec{x}_1}{x_0} = \left(\frac{x_1}{x_0}, \frac{x_2}{x_0}, \dots, \frac{x_{s-1}}{x_0}\right)$, we have

$$\text{Var}(\text{GHD}_{1\pi}) = \text{Var}(\vec{\omega}_{s-1}^T \text{GHD}_{st\pi}).$$

Therefore verifying the condition is equivalent to showing that the random variable $\text{GHD}_{1\pi}$ can be expressed as a linear combination of $\text{GHD}_{st\pi}$, so,

$$\text{GHD}_{1\pi} = \vec{\omega}_{s-1}^T \text{GHD}_{st\pi} \iff \text{GHD}_{1\pi} = \sum_{k=1}^s \omega_k \text{GHD}_{t_k\pi}. \quad (5.14)$$

Hence our analysis is now equivalent to showing that equation (5.14) holds for any permutation $\pi \in \Pi$. Recall our assumptions that all the networks are neither empty or complete, and also unique. Since all graphs are non-empty and complete, values of $\text{GHD}_{k\pi}$, $1 \leq k \leq T$ under different $\pi \in \Pi$ are not deterministic, therefore $\text{GHD}_{1\pi}$ is not fixed or 0. Also we can conclude from the unique sample assumption that $\text{GHD}_{1\pi}$, $\text{GHD}_{t_1\pi}$, $\text{GHD}_{t_2\pi}$, .. $\text{GHD}_{t_{s-1}\pi}$ are unique comparisons, which implies that $\text{GHD}_{k\pi} = \text{GHD}_{s\pi}$ cannot be true for all permutations π , for any pair of comparisons k and s , $1 \leq k, s \leq T$. Hence we know that $\vec{\omega}_{s-1}$ cannot be a zero vector.

Now, consider any permutation π_1 in Π ,

$$\begin{aligned} \text{GHD}_{1\pi_1} &= \sum_{k=1}^s \omega_k \text{GHD}_{t_k\pi_1} \\ \iff \sum_{i,j} (a_{1\pi_1(i)\pi_1(j)} - b_{1ij})^2 &= \sum_{k=1}^{s-1} \sum_{i,j} \omega_k (a_{t_k\pi_1(i)\pi_1(j)} - b_{t_kij})^2. \end{aligned}$$

There are $N!$ different permutations, therefore quantities $a_{k\pi_1(i)\pi_1(j)}$ s, $b_{k\pi_1(i)\pi_1(j)}$ and ω_k would need to be found to satisfy all $N!$ different linear equations. We know that under the null hypothesis of independence between the two groups of networks, networks in $\mathcal{G}(\mathcal{A})$ are independent of networks in $\mathcal{G}(\mathcal{B})$. Hence $a_{k\pi_1(i)\pi_1(j)}$ and $b_{k\pi_1(i)\pi_1(j)}$ are independently distributed, and it is therefore unlikely that the solution of the system linear equations for different permutations, which are $N!$ in number, are equivalent. More formally, we note the

fact that the total number of comparisons T is a fixed constant and the asymptotic regime considers the size of the network N tending to infinity. For a set of $N!$ linear equations, there are only $sN(N-1) + (s-1)$ variables (all elements in $2s$ N by N matrices plus $\vec{\omega}_{s-1}$). Hence there is an inadequate numbers of variables to be able to find a solution for a linear system of equations of size $N!$.

Therefore when $x_0 \neq 0$, Θ is strictly greater than 0 and Σ_s is positive definite, asymptotically. Then for any subset with size s of variables from $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$, say, $(\text{GHD}_{t_1\pi}, \text{GHD}_{t_2\pi}, \dots, \text{GHD}_{t_s\pi})'$, also has positive definite covariance matrix.

Therefore, we can conclude that, asymptotically, Σ is positive definite and the lower triangular matrix L is unique and non-singular. Hence the asymptotic joint permutation distribution may be approximated as multivariate normal. This completes the proof.

5.2.2 A test of group independence based on $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$

Again, under the null hypothesis of independence, the joint distribution of all between-group comparisons, $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$, is approximately multivariate normal. The first two moments of the permutation distribution for $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$, can be computed explicitly. For a detailed evaluation of the moments of the permutation sampling distribution of $\text{GHD}_{t\pi}$, please refer to the Appendix A.1. Now re-denote the mean and covariance matrix of $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$ to be $\vec{\mu}_2$ and Σ_2 , respectively. Also, denote the mean and standard deviation of $\text{GHD}_{t\pi(t)}$ to be $\mu_2(t)$ and σ_{tt} , and denote the covariance between $\text{GHD}_{t\pi(t)}$ and $\text{GHD}_{t'\pi(t)}$ to be $\sigma_{tt'}$. Then we have

$$\vec{\mu}_2 = (\mu_2(1), \mu_2(2), \dots, \mu_2(T))^T,$$

and

$$\Sigma_2 = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1T} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2T} \\ \dots & \dots & \dots & \dots \\ \sigma_{T1} & \sigma_{T2} & \dots & \sigma_{TT}^2 \end{pmatrix}_{T \times T},$$

where $\sigma_{tt'}^2 = E[(\text{GHD}_{t\pi} - \mu_t)(\text{GHD}_{t'\pi} - \mu_{t'})]$. Again, we standardize $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$. Denote by L_T the lower triangular matrix obtained by Cholesky decomposition of the covariance matrix, so that $\Sigma_2 = L_T L_T^T$. Let S_π be the normalized version of $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}))$, given by

$$S_\pi = L_T^{-1}(\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2), \quad (5.15)$$

and with the elements of S_π independent and standard normal under the null hypothesis of independence of the two groups, $S_\pi \sim N(0, I)$. The observed data value is denoted as

$$S_2 = L_T^{-1}(\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2).$$

Since $S_\pi \sim N(0, I)$, $S_\pi^T S_\pi$ follows a chi-squared distribution with degree of freedom T under the null hypothesis,

$$C_{2\pi} = S_\pi^T S_\pi \sim \chi_T^2.$$

Then, with the observed statistic S_2 , letting $C_2 = S_2^T S_2$ we have

$$\begin{aligned} C_2 &= S_2^T S_2 \\ &= [L_T^{-1}(\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2)]^T [L_T^{-1}(\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2)] \\ &= (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2)^T (L_T^{-1})^T (L_T^{-1}) (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2) \\ &= (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2)^T \Sigma_2^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2). \end{aligned}$$

Therefore, for any significance level α , we reject the null hypothesis of group independence if

$$C_2 = (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2)^T \Sigma_2^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2) < \chi_{T,\alpha}^2,$$

where $\chi_{T,\alpha}^2$ is the α quantile of the chi-squared distribution with degrees of freedom T .

5.3 Approximation For Multivariate Exponential Random Graph Permutation model (MERGP)

Recall that in Section 2.5, we introduced an MERGP model that was proposed by [22]. The formula (2.6) is

$$P(\pi|G(\mathcal{A}), G(\mathcal{B}), \phi) = \frac{\exp(\sum_{k=1}^P \sum_{l=1}^Q \phi_{kl} \sum_{i=1}^N \sum_{j=1}^N a_{k\pi(i)\pi(j)} b_{lij})}{\mathcal{X}_{(\phi)}},$$

where

$$\mathcal{X}_{(\phi)} = \sum_{\pi' \in \Pi} \exp(\sum_{k=1}^P \sum_{l=1}^Q \phi_{kl} \sum_{i=1}^N \sum_{j=1}^N a_{k\pi'(i)\pi'(j)} b_{lij}),$$

is the normalizing constant. Detailed explanation of the formula (2.6) can be found in Section 2.5, start from page 36.

We argue that the maximum likelihood estimator (MLE) obtained using the algorithm that Butts [22] proposed is potentially unreliable and, more importantly, when implementing multiple comparisons with large sample sizes (in both groups), the computational cost may be substantial [23]. In this section, we manage to analytically derive an explicit formula for the MLE using standard statistics, under the assumption we made in Section 5.2.1. This formula enable us to compute the MLE more effectively in close form.

5.3.1 Introduction

In this section, we use the asymptotic normal distribution under permutation for the product moment

$$\Gamma_{\pi\phi} = \sum_{k=1}^P \sum_{l=1}^Q \phi_{kl} \sum_{i=1}^N \sum_{j=1}^N a_{k\pi(i)\pi(j)} b_{lij}. \quad (5.16)$$

Then, under the hypothesis of network independence, the distribution under permutation of the quantity $\sum_{i,j} a_{\pi(i)\pi(j)} b_{ij}$ is approximately normal as N approaches infinity. If we can prove that the quantity $\Gamma_{\pi\phi}$ is well approximated by a normal distribution, then we

can argue that the MLE $\hat{\phi}$ can be evaluated analytically. In the following sections, we will provide a proof of normality for (5.16) and, further, we will provide an analytical solution for the MLE, given that a standardized measure is used. We build a link between MERGP and the standardized measure and the test of group independence proposed in Section 5.2. Using this result, the procedure to obtain a MLE of MERGP for analysing groups of networks is simplified and, more importantly, the estimate we obtained is possibly a more accurately approximation of the true MLE. We first show that the quantity (5.16) is approximately normally distributed, given that the sampling distribution for each component $\sum_{i=1}^N \sum_{j=1}^N a_{k\pi(i)\pi(j)} b_{lij}$ is approximately normal.

Before providing the derivation, to simplify the presentation, we will re-write (5.16) in terms of a vector product. Following similar steps as in Section 5.2, we first replace the double subscript $\{i, j\}$ in (5.16) by a single subscript, and then express the product moment in the form of a vector product. So consider

$$\Gamma_{N(\phi)} = \sum_{i^*=1}^T \phi_{i^*} \sum_{k=1}^N \sum_{s=1}^N a_{i^*ks} b_{i^*ks}, \quad (5.17)$$

where $T = PQ$ is the total number of comparisons. Now, denote the contribution from the i^{th} between-group comparison to be $\gamma_{(i)} = \sum_{k=1}^N \sum_{s=1}^N a_{i^*ks} b_{i^*ks}$, and define two 1 by T vectors $\vec{\gamma}_N = (\gamma_{(1)}, \gamma_{(2)}, \dots, \gamma_{(T)})^T$ and $\vec{\phi} = (\phi_{1^*}, \phi_{2^*}, \dots, \phi_T)^T$. Then the above triple summation can be expressed as a scalar product:

$$\Gamma_{N(\phi)} = \vec{\phi}^T \vec{\gamma}_N.$$

For a given permutation vector π , let

$$\vec{\gamma}_{N\pi} = (\gamma_{N(1)}^\pi, \gamma_{N(2)}^\pi, \dots, \gamma_{N(T)}^\pi)^T,$$

where

$$\gamma_{N(i)}^\pi = \sum_{k=1}^N \sum_{s=1}^N a_{i\pi(k)\pi(s)} b_{i^*ks}.$$

Then the version of $\Gamma_{N(\phi)}$ under permutation π can be written as

$$\Gamma_{N(\phi)}(\pi) = \vec{\phi}^T \gamma_{N\pi}. \quad (5.18)$$

Knowing that the vector (5.9) is jointly normally distributed and also according to (3.5) (in page 50), we can conclude that, under the assumption of independent permutations, $\gamma_{N\pi}$ is also jointly normal. Since ϕ is constant, $\Gamma_{N(\phi)}(\pi)$ is a linear combination of jointly normal random variates, hence it is normally distributed.

Let $\vec{\mu}_3$ and Σ_3 be the corresponding mean vector and variance covariance matrix for $\gamma_{N\pi}$, so that

$$\vec{\mu}_3 = (\mu_3(1), \mu_3(2), \dots, \mu_3(T))^T$$

with $\mu_3(i) = E[\gamma_{N(i)}^\pi]$ and

$$\Sigma_3 = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1T} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2T} \\ \dots & \dots & \dots & \dots \\ \sigma_{T1} & \sigma_{T2} & \dots & \sigma_{TT}^2 \end{pmatrix}_{T \times T}$$

where $\sigma_{ij}^2 = E[(\gamma_{N(i)}^\pi - \mu_3(i))(\gamma_{N(j)}^\pi - \mu_3(j))]$.

For different graph models, values in $\gamma_{N\pi}$ will increase with N in different manners. For some models, values of $\gamma_{N\pi}$ may approach infinity as $N \rightarrow \infty$. Therefore, we use a normalized version of $\gamma_{N\pi}$ instead. Let L_3 be the lower triangular matrix obtained by Cholesky decomposition of the covariance matrix, i.e., $\Sigma_3 = L_3 L_3^T$. By standardizing $\gamma_{N\pi}$, we obtain

$$Z_{N\pi} = L_3^{-1}(\gamma_{N\pi} - \vec{\mu}_3), \quad (5.19)$$

and with the elements in $Z_{N\pi} = (Z_{N(1)}^\pi, Z_{N(2)}^\pi, \dots, Z_{N(T)}^\pi)$ being independent standard normal. We apply (5.15) to (2.6), instead of $\gamma_{N\pi}$. Then the MERGP with the normalized

measure just defined may be written as

$$P(\pi|\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \phi) = \frac{\exp(\vec{\phi}^T Z_{N\pi})}{\sum_{\pi' \in \Pi} \exp(\vec{\phi}^T \cdot Z_{N\pi'})}, \quad (5.20)$$

with the normalizing constant as in (2.7) defined by

$$\mathcal{X}'_{(\phi)} = \sum_{\pi' \in \Pi} \exp(\vec{\phi}^T Z_{N\pi'}). \quad (5.21)$$

In Section 5.3.2, we will first provide an approximate formulation for the normalizing constant $\mathcal{X}'_{(\phi)}$. Given that $Z_{N\pi}$ is multivariate joint normal with mean $\vec{0}$ and covariance matrix I , this normalizing constant $\mathcal{X}'_{(\phi)}$ can be evaluated. Then, in Section 5.3.3, we use the normalizing constant to derive an approximation to the maximum likelihood estimator $\hat{\phi}_{ij}$. Finally, in Section 5.3.4, based on the maximum likelihood estimator, we will propose a chi-square-based test of group difference.

5.3.2 Derivation of the normalizing constant $\mathcal{X}'_{(\phi)}$

We may express the normalizing constant $\mathcal{X}'_{(\phi)}$ as follows:

$$\mathcal{X}'_{(\phi)} = \sum_{\pi \in \Pi} \exp(\vec{\phi}^T Z_{N\pi})$$

From the previous section, we know that $Z_{N\pi} = (Z_{N(1)}^\pi, Z_{N(2)}^\pi, \dots, Z_{N(T)}^\pi)^T$ is (approximately) jointly normally distributed. The final summation is over all possible permutations, with each permutation having equal probability $\frac{1}{N!}$ under null hypothesis of independence.

Let $Z_{N\pi}$ be an observation from the permutation distribution of Z_N , with permutation label π . Then we have:

$$\begin{aligned} \mathcal{X}'_{(\phi)} &= N! \times \frac{\sum_{\pi \in \Pi} \exp(\vec{\phi}^T Z_{N\pi})}{N!} \\ &= N! \times E_{\Pi}[\exp(\vec{\phi}^T Z_{N\pi})] \\ &\approx N! \times E[\exp(\vec{\phi}^T N(0, I))] \\ &= N! \times \exp\left\{\frac{1}{2} \vec{\phi}^T \vec{\phi}\right\}. \end{aligned} \quad (5.22)$$

Note that here E_{Π} denotes the expectation over uniform permutation.

For the approximation of the third step, we know from the previous proofs that as N increases, $Z_{N\pi}^{\rightarrow}$ converges to a multivariate standard normal distribution.

For any random variable X with probability density function f , the expectation $E_f[g(X)]$ may be approximated by $E[g(X_N)]$, if $X_N \rightarrow X$ in distribution for any bounded function g . However, a function of the form $g(x) = \exp(\phi x)$ is an exponential function, and is, therefore, not bounded. Zapała [110] extends the convergence of $E(g(X_N)) \rightarrow E_f(g(X))$ for bounded function g to a set of continuous $\mathcal{S}_{\infty}$ -almost uniform integrable functions g . Using this result, we can show that the third approximation step is justified, given that $Z_{N(i)}^{\pi}$ is a series of random variables of matrix permutations and that the limiting distribution $Z_{\infty(i)}^{\pi}$ is standard normal as $N \rightarrow \infty$. The notation and proof required are complicated, thus we provide a proof in the Appendix A.2 (the interested reader can refer to the Appendix and Zapała [110] for a detailed proof, notation and explanation of the mathematical terms).

5.3.3 Formulation of the MLE $\hat{\phi}$

In this subsection, we obtain analytically the MLE $\hat{\phi}$. First substituting (5.22) into (5.20) we obtain

$$P(\pi|\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \phi) = \frac{\exp(\vec{\phi}^T Z_{N\pi}^{\rightarrow})}{\sum_{\pi' \in \Pi} \exp(\vec{\phi}^T \cdot Z_{N\pi'}^{\rightarrow})} \approx \frac{\exp(\vec{\phi}^T Z_{N\pi}^{\rightarrow})}{N! \times \exp\{\frac{1}{2}\vec{\phi}^T \vec{\phi}\}}. \quad (5.23)$$

Taking the log of (5.23):

$$\begin{aligned} & \log(P(\pi|\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \phi)) \\ & \approx \log \left\{ \frac{\exp(\vec{\phi}^T Z_{N\pi}^{\rightarrow})}{N! \times \exp\{\frac{1}{2}\vec{\phi}^T \vec{\phi}\}} \right\} \\ & = \vec{\phi}^T Z_{N\pi}^{\rightarrow} - \frac{1}{2}\vec{\phi}^T \vec{\phi} - \log(N!) \\ & \sim \vec{\phi}^T Z_{N\pi}^{\rightarrow} - \frac{1}{2}\vec{\phi}^T \vec{\phi}. \end{aligned} \quad (5.24)$$

Now to maximize the log likelihood function (5.24), we take the first derivative of (5.24) with respect to parameter $\vec{\phi}$,

$$\frac{d(\log(P(\pi|\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \phi)))}{d\vec{\phi}} = Z_{N\pi} - \vec{\phi}.$$

Then to obtain the MLE, we set the first derivative to zero, obtaining

$$Z_{N\pi} = \vec{\phi} \implies \hat{\phi} = L_3^{-1}(\gamma_{N\pi} - \vec{\mu}_3). \quad (5.25)$$

Here we should notice that since L_3 is a lower triangular matrix, the inverse L_3^{-1} is also a lower triangular matrix.

5.3.4 A test of independence between two sets of networks based on the MLE $\hat{\phi}$

From the discussion in the previous subsection, we know that $\hat{\phi}$ may be considered as a measure of network association. Values of $\hat{\phi}$ close to 0 indicate that the two samples of the networks are structurally independent, and large values of $||\hat{\phi}||$ indicate that the two groups are strongly associated. In this section, we test the null hypothesis of independence between the two groups of networks, expressed as

$$H_0 : \phi = 0 \quad \text{versus} \quad H_1 : \phi \neq 0.$$

From (5.25), we know that

$$\vec{\phi} = Z_{N\pi} = L_3^{-1}(\gamma_{N\pi} - \vec{\mu}_3).$$

Since $\vec{\phi} = Z_{N\pi}$, asymptotically, under the null hypothesis of group independence, $\vec{\phi}$ follows a joint normal distribution $N(0, I)$. Then a chi-squared based test is proposed as follows. We have $\vec{\phi}^T \vec{\phi} \sim \chi_T^2$, chi-squared with degrees of freedom T . Hence, given the

observed MLE, $\hat{\phi}$,

$$\begin{aligned}\hat{\phi}'\hat{\phi} &= [L_3^{-1}(\gamma_{N\pi} - \vec{\mu}_3)]^T [L_3^{-1}(\gamma_{N\pi} - \vec{\mu}_3)] \\ &= (\gamma_{N\pi} - \vec{\mu}_3)^T L_3^{-T} L_3^{-1} (\gamma_{N\pi} - \vec{\mu}_3) \\ &= (\gamma_{N\pi} - \vec{\mu}_3)^T \Sigma_3^{-1} (\gamma_{N\pi} - \vec{\mu}_3).\end{aligned}$$

Therefore, for any significance level α , we reject the null hypothesis of group independence if

$$C_3 = (\gamma_N - \vec{\mu}_3)^T \Sigma_3^{-1} (\gamma_N - \vec{\mu}_3) < \chi_{T,\alpha}^2.$$

Since the covariances matrices used in C_3 and C_2 are computed based on the same permutation procedure and according to (3.5), the covariance matrices Σ_2 and Σ_3 are similar (actually $\Sigma_2 = 4\Sigma_3$). It is not difficult to derive, also using (3.5), that $2(\gamma_{N\pi} - \vec{\mu}_3) = (\text{GHD}(G_\pi(\mathcal{A}), G(\mathcal{B})) - \vec{\mu}_2)$. Hence, under the normality assumption, the testing procedure for (5.23) can be simplified using the test of group independence we proposed previously.

5.4 Discussion

5.4.1 A remark on the two tests

In Sections 5.1 and 5.2, we proposed two tests of independence relying on different permutation schemes. The difference between the two test statistics,

$$\begin{aligned}C_1 &= (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1)^T \Sigma_1^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A})) - \vec{\mu}_1), \\ C_2 &= (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2)^T \Sigma_2^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2),\end{aligned}$$

arises from the different covariance matrices resulting from the different permutation strategies. Notice that the equations for C_1 and C_2 "look" quite similar. Given the assumption of pair-wise independence (i.e. assumption (a.ii)) among all permutations in C_1 , Σ_1 is diagonal. A question may arise whether C_1 , by applying all pair-wise between-group comparisons, can be used to compare group independence.

Consider all the pair-wise group comparisons $\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}))$ under independent permutation. Then the comparing groups version of C_1 , denoted as C_{1g} , in this case becomes

$$C_{1g} = (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2)^T \Sigma_{1g}^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2).$$

where Σ_{1g} is also a diagonal matrix containing all the diagonal elements of Σ_2 . If C_{1g} is valid, it provides a computationally more effective way of evaluating the group difference. Since the covariance matrix Σ_{1g} is diagonal, then only T elements (the quantities σ_{tt}) need to be computed, while for Σ_2 , evaluation of all $\frac{T(T-1)}{2}$ covariances $\sigma_{tt'}$ is needed.

Intuitively, the permutation strategy used by C_2 is a more sensible way of testing the group independence due to the fact that, in the second procedure, the permutations π used for all the between-group comparisons are the same. This randomizing procedure retains the node associations among different networks within a group, and, hence, makes more sense in regard to the problem because we generally assume that networks within each group are associated with each other.

However, for C_{1g} , under the assumption of independence among all comparisons, the within-group topological dependence among networks is ignored. For networks with complex topological similarity within group of networks, the covariance structure among different comparisons may not be ignorable. In a simulation study (Section 5.5.3), we empirically justified that when comparing two independent groups of networks with C_{1g} , even for networks with only localized topological structural similarities (RG networks), C_{1g} will fail. Therefore, C_{1g} is not a suitable test of group independence.

5.4.2 Discussion on duplicated samples

Recall the assumption we have made during the proof of normality, that all graphs in the two groups $\mathcal{G}_A$ and $\mathcal{G}_B$ are non-empty and unique, i.e., there is no duplicated network in either group $\mathcal{G}_A$ or $\mathcal{G}_B$.

This assumption guaranteed that the covariance matrix Σ_2 we obtained for an asymp-

totically multivariate normal random vector,

$$\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B})) = (\text{GHD}_{1\pi}, \text{GHD}_{2\pi}, \dots, \text{GHD}_{T\pi})^T,$$

is of full rank and, hence, positive definite.

For real applications, it is possible that there may be replicates within each group. Denote by $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P$ and $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_Q$ the collection of all unique samples. Denote by $\omega_{\mathcal{A}t}$ the number of replications that exist of network $\mathcal{A}_t$, and by $\omega_{\mathcal{B}t}$ the number of replications of network $\mathcal{B}_t$. Then the statistic used for the comparison between $\mathcal{A}_t$ and $\mathcal{B}_t$ becomes

$$\begin{aligned} \text{GHD}_{t\pi} &= \frac{1}{N(N-1)} \sum_{i,j} \omega_{\mathcal{A}t} \omega_{\mathcal{B}t} (a_{t\pi(i)\pi(j)} - b_{tij})^2 \\ &= \frac{\omega_{\mathcal{A}t} \omega_{\mathcal{B}t}}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N (a_{t\pi(i)\pi(j)} - b_{tij})^2. \end{aligned}$$

Now, writing $\omega_t = \omega_{\mathcal{A}t} \omega_{\mathcal{B}t}$, let $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega)$ be defined by

$$\begin{aligned} \text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) &= (\omega_1 \text{GHD}_{1\pi}, \omega_2 \text{GHD}_{2\pi}, \dots, \omega_T \text{GHD}_{T\pi})^T \\ &= \text{diag}(\omega_1, \omega_2, \dots, \omega_T) \text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B})), \end{aligned} \quad (5.26)$$

where $\text{diag}(\omega_1, \omega_2, \dots, \omega_T)$ denotes a $T \times T$ diagonal matrix with diagonal entries $\omega_1, \omega_2, \dots, \omega_T$, and all other values equal to zero:

$$\text{diag}(\omega_1, \omega_2, \dots, \omega_T) = \begin{bmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \omega_T \end{bmatrix}.$$

For finite and non-zero numbers of replicates $\infty > \omega_{\mathcal{A}t} > 0$ and $\infty > \omega_{\mathcal{B}t} > 0$, similar procedures as before can be applied to (5.26), and a chi-squared-based test formulated. Again, we standardize $\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega)$, in a similar way as done in (5.15). We

know from (5.15) that

$$\begin{aligned}
S_\pi &= L_T^{-1}(\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B})) - \vec{\mu}_2) \\
&= L_T^{-1}(\text{diag}(\omega_1, \omega_2, \dots, \omega_T)^{-1} \text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \vec{\mu}_2) \\
&= [\text{diag}(\omega_1, \omega_2, \dots, \omega_T) L_T]^{-1} [\text{GHD}(\mathcal{G}_\pi(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}],
\end{aligned}$$

where $\vec{\mu}_{2\omega} = (\omega_1 \mu_2(1), \omega_2 \mu_2(2), \dots, \omega_T \mu_2(T))$ and L is the lower triangular matrix corresponding to the Cholesky decomposition of the covariance matrix Σ_2 . Notice here that Σ_2 contains only those unique comparisons, and, hence, is still invertible. The distribution of S_π under the null hypothesis of group independence is that of independent, identically distributed normal variables. Again, a chi-squared test can be applied.

Denote the observed test statistic by C_ω , with

$$\begin{aligned}
C_\omega &= [(\text{diag}(\omega_1, \omega_2, \dots, \omega_T) L_T)^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega})]^T \\
&\quad (\text{diag}(\omega_1, \omega_2, \dots, \omega_T) L_T)^{-1} (\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}) \\
&= [\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}]^T (L_T^{-1})^T \text{diag}^{-1}(\omega_1^2, \omega_2^2, \dots, \omega_T^2) \\
&\quad L_T^{-1} [\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}] \\
&= [\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}]^T \text{diag}^{-1}(\omega_1^2, \omega_2^2, \dots, \omega_T^2) \\
&\quad \Sigma_2^{-1} [\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}].
\end{aligned}$$

Then with a specified significance level α , we reject the null hypothesis of group independence if

$$\begin{aligned}
C_\omega &= [\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}]^T \text{diag}^{-1}(\omega_1^2, \omega_2^2, \dots, \omega_T^2) \\
&\quad \Sigma_2^{-1} [\text{GHD}(\mathcal{G}(\mathcal{A}), \mathcal{G}(\mathcal{B}), \omega) - \mu_{2\omega}] \\
&< \chi_{T, \alpha}^2.
\end{aligned}$$

where $\chi_{T, \alpha}^2$ is the α quantile of the chi-squared distribution with degrees of freedom T .

5.4.3 The null hypothesis of independence and future work

In this chapter, two tests of independence for multiple networks were proposed based on a different permutation strategy under the assumption that the permutation distribution is multivariate normal under the null. However, in practice, the null hypothesis of independence for comparing groups can be too strong. Commonly in real applications, only localised differences in the network structure of between-group individuals are expected to be observed. For instance, for traumatic brain injury [95], although there is evidence of disruption and abnormalities of functional connectivity for the traumatic brain injury group, the difference is not sufficiently decisive to conclude structural independence of the entire structure of between-group individuals. Therefore, the interest of this chapter focuses on the theoretical property and inferential aspect of the test of independence rather than practical applications. However, in future study, we will further develop, as well as utilise, the proposed test of independence such that it can be used for analysing real networks. Similar to the GHD that is applied in subnetwork detection algorithms, one possible utilisation of the proposed tests of independence would be detecting differential subnetworks for two groups or a set of networks. However, for this possible extension, much work needs to be done to resolve problems such as computation efficiency, the choice of simply evaluated differential scores for each node, etc.

5.5 Simulation Study

In this section, we report on several simulation studies of the tests we proposed for multiple networks. In the first simulation study (Section 5.5.1), we empirically verify the convergence of $E_\pi[\exp(X_N)]$ to $E_f[\exp(X)]$, given that $X_N \rightarrow X$ in (5.22). In the second (Section 5.5.2) and third simulation (Section 5.5.3) studies, we validate the null sampling distribution for $C_{1\pi}$ and $C_{2\pi}$. In Section 5.5.3, we also empirically show that C_{1g} is not a valid test of group independence. In Section 5.5.4, we present a power study, for two edge weights, of the proposed tests, which shows that TO is more powerful than adjacency to detect structure changes.

5.5.1 Empirical justification of (5.22)

In Section 5.3, we showed that the normalizing constant $\mathcal{X}'(\phi)$ can be approximated by equation (5.22) under the null hypothesis that two groups of networks are independent. In this simulation, we attempted to empirically validate the approximation. We first generated independent groups of networks as follows: two independent networks $\mathcal{A}$ and $\mathcal{B}$ of size N were generated from the ER graph model with a common parameter p . Then, P networks of group $\mathcal{G}(\mathcal{A})$ i.e. $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$ from $\mathcal{A}$ and Q networks of group $\mathcal{G}(\mathcal{B})$ i.e. $\{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_Q\}$ from $\mathcal{B}$, were generated with a common percentage of noise $p_m = 0.5$; detailed procedures for generating associated networks are described in Section 3.3.2 (in page 92). Therefore, two groups of independent networks were obtained with each of the groups containing P and Q associated networks. Then the total number of comparisons $T = PQ$. We then computed the corresponding mean vector $\vec{\mu}_3$ and covariance matrix Σ_3 of $\gamma_{N\pi}$, where $\gamma_{N\pi}$ contains all the pair-wise between-group comparisons for $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$ with permutation π .

By (5.21), we knew that for large N , $Z_{N\pi}$ was approximately a vector of I.I.D standard normal random variables, i.e., $Z_{N\pi} \sim_d N(0, I)$ as N increases. For any non-zero constant vector $\vec{\phi}$, $\vec{\phi}^T Z_{N\pi}$ represents a linear combination of standard normal random variables. Therefore, without loss of generality, we let $\vec{\phi} = (1, 1, \dots, 1)^T$ ($1 \times T$ vector). Then it is equivalent to show that

$$\frac{1}{N!} \mathcal{X}(\vec{\phi}) = E_{\Pi}[\exp(\vec{\phi}^T Z_{N\pi})] \approx \exp\left\{\frac{1}{2} \vec{\phi}^T I \vec{\phi}\right\} = \exp\left(\frac{T}{2}\right).$$

In this simulation, for each N , the value of $E_{\Pi}[\exp(\vec{\phi}_1^T Z_{N\pi})]$ was approximated by the mean of $n = 1,000,000$ random permutations,

$$E_{\Pi}[\exp(\vec{\phi}_1^T Z_{N\pi})] \approx \frac{1}{n} \sum_{i=1}^n \exp(\vec{\phi}_1^T Z_{N\pi_i}).$$

We considered two comparisons of groups of different sizes. For the first comparison, $P_1 = Q_1 = 2$ had $T = P_1 Q_1 = 4$ comparisons. Another comparison consisted of $P_2 = Q_2 = 3$ individuals in each group, which has a total of $T = 9$ comparisons. Figure 5.1 shows the

ratio,

$$\frac{E_{\Pi}[\exp(\vec{\phi}_1^T \vec{Z}_{N\pi})]}{E[\exp(\frac{T}{2})]},$$

for two comparisons of different network sizes N . It was known from Section 3.2 that the normal approximation improves as N increases, therefore, we expected the ratio to converge to 1 as the size of the network increases. Figure 5.1 roughly shows this trend for both comparisons. The reason that not all ratios follow that trend may be because of the number of permutations $n = 1,000,000$, which for large N is still insufficient. Figure 5.1 also shows that the ratio for $T = 4$ converged faster to 1 than the ratio for $T = 9$.

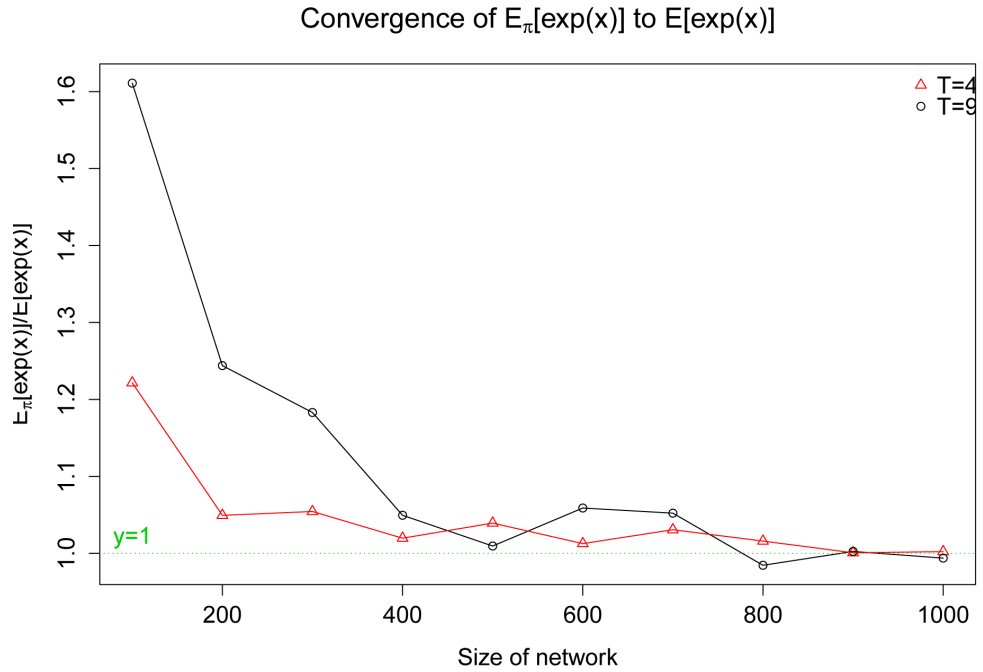


Figure 5.1: Empirical justification of (5.22) for two comparisons of $T = 4$ and $T = 9$. Convergence can be observed for both size of the network (N).

5.5.2 Validation of null sampling distribution for testing of mutual independence

To assess the asymptotic null sampling distribution of the multiple network independence test statistic C_1 , we randomly generated a set of networks $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_{10}\}$ of 10 individual

networks independently from ER and RG with sizes $N = 100, 200$ and 400 with parameters p and d , which guarantee that all the networks we generated have similar densities. In each case, we simulated 10,000 sets of networks and obtained the corresponding asymptotic p -values for testing mutual independence with two different edge weights.

Additionally, we visualised the results by applying quantile-quantile plots for the p -values against the uniform $[0, 1]$ distribution. Figures 5.2 and 5.3 show that for both ER and RG random networks, approximation of the null sampling distribution by a chi-squared distribution is valid.

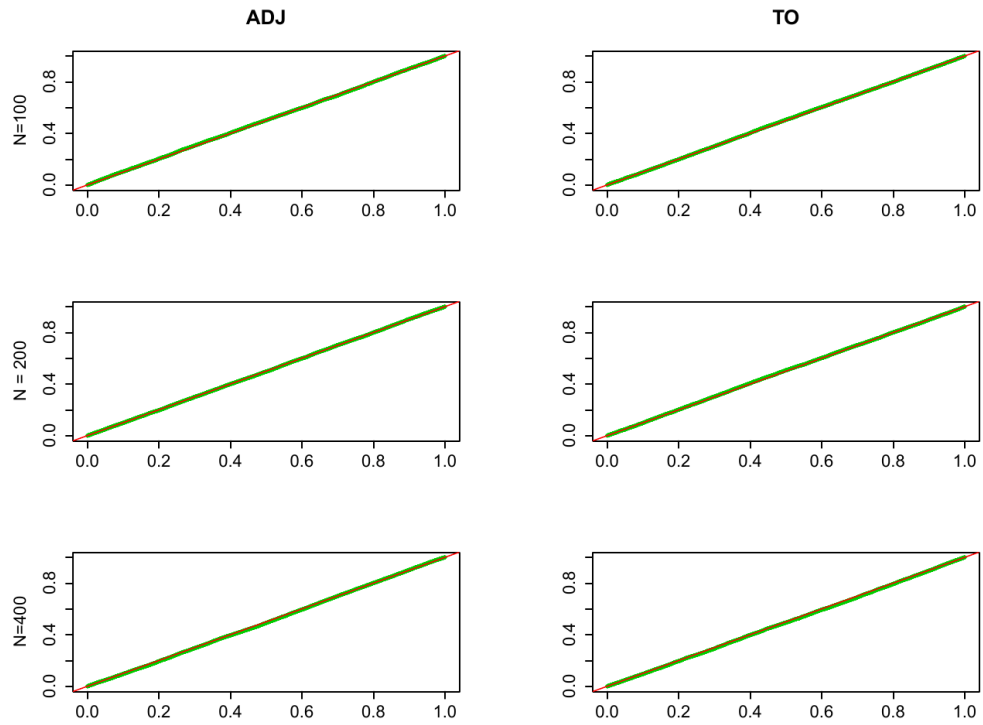


Figure 5.2: QQ plots for p -values obtained from mutual independence test C_1 against uniform $[0,1]$ distribution, for ER random graphs with adjacency information and TO. For each plot, 10,000 sets of networks were generated from the ER random graph model. P -values were estimated by the proposed chi-squared test. The plot shows that with the ER random graph model, the null sampling distribution of $C_{1\pi}$ is well approximated by a chi squared distribution for all N .

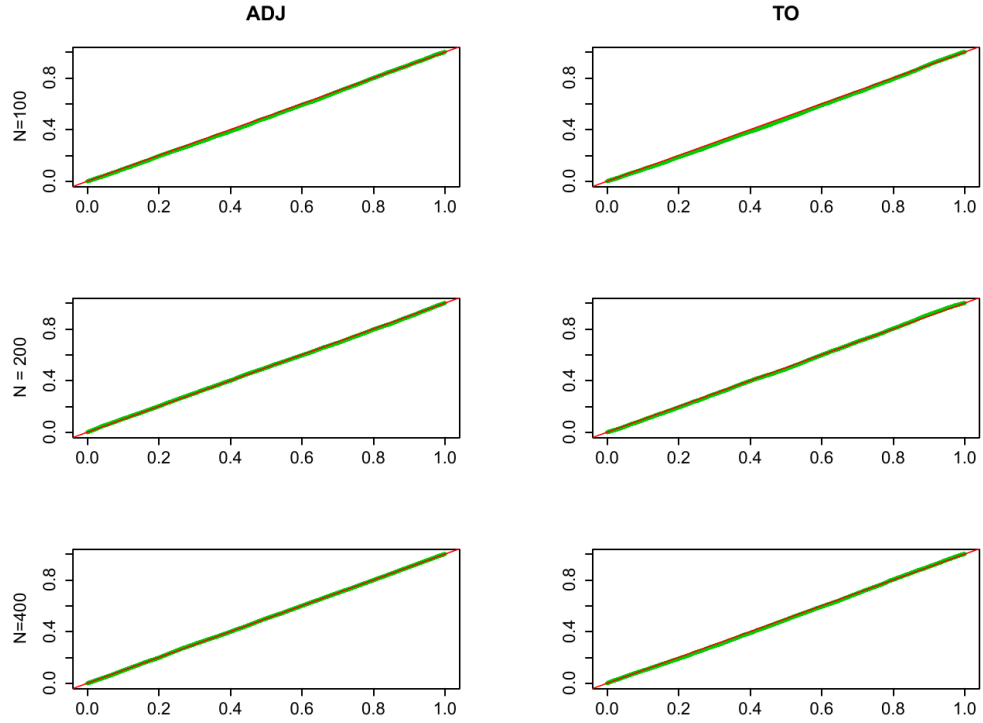


Figure 5.3: QQ plots for p-values obtained from mutual independence test C_1 against uniform $[0,1]$ distribution for RG random graphs with adjacency information and TO. For each plot, 10,000 sets of networks were generated from the RG random graph model. P-values were estimated by the proposed chi-squared test. The plot show that with RG random graph model, when N is greater than 100, approximation of the null sampling distribution of $C_{1\pi}$ by the chi squared distribution is acceptable.

5.5.3 Validation of null sampling distribution for testing of group independence

In the third simulation, we attempted to verify the null sampling distribution of $C_{2\pi}$. We used two network models to investigate this problem. The first graph model we used was the ER random graph model. Network generating procedures that have been used to generate two independent groups of ER networks were the same as in Section 5.5.1. For RG networks, similarly, we first generated two independent RG networks $\mathcal{A}$ and $\mathcal{B}$ with the same parameter d . Additionally, similar to previous chapters, the sizes of the network we used were $N = 100, 200, 400$. Then, $P = 10$ networks of group $\mathcal{G}(\mathcal{A})$ with $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$

from $\mathcal{A}$ and $Q = 10$ networks of group $\mathcal{G}(\mathcal{B})$ with $\{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_Q\}$ from $\mathcal{B}$, were generated with a common percentage of noise $p_m = 0.5$; see Section 3.3.2 for details. 10,000 sets of networks were generated for each case and the asymptotic p -values obtained were used for testing group independence, again with two different edge weights. Again, we used quantile-quantile plots to verify whether the p -values obtained were uniformly distributed. Figure 5.4 and Figure 5.5 show that for both graph models, asymptotic p -values match the uniform quantile for $N > 100$. For the ER graph model, this is also true for $N = 100$.

Apart from verifying the properties of statistic C_2 , we also wanted to justify whether C_{1g} provides a valid test of group independence. We applied a similar procedure as for C_2 with $N = 200$. Figure 5.6 shows that the quantiles of asymptotic p -values obtained from C_{1g} match the theoretical quantiles of the uniform distribution for ER networks. This may be because for ER networks, the topological structure for the network is simple and there is no edge dependence for each network, and, therefore, the covariance structure among different comparisons can be ignored.

However, the results for RG networks fail to match the uniform quantiles. Therefore, C_{1g} is not valid with the RG network. This result justifies the argument we made in 5.4.1, that if complex topological structure similarity exists for networks that are in the same group, the covariance structure among different comparisons should be taken into consideration.

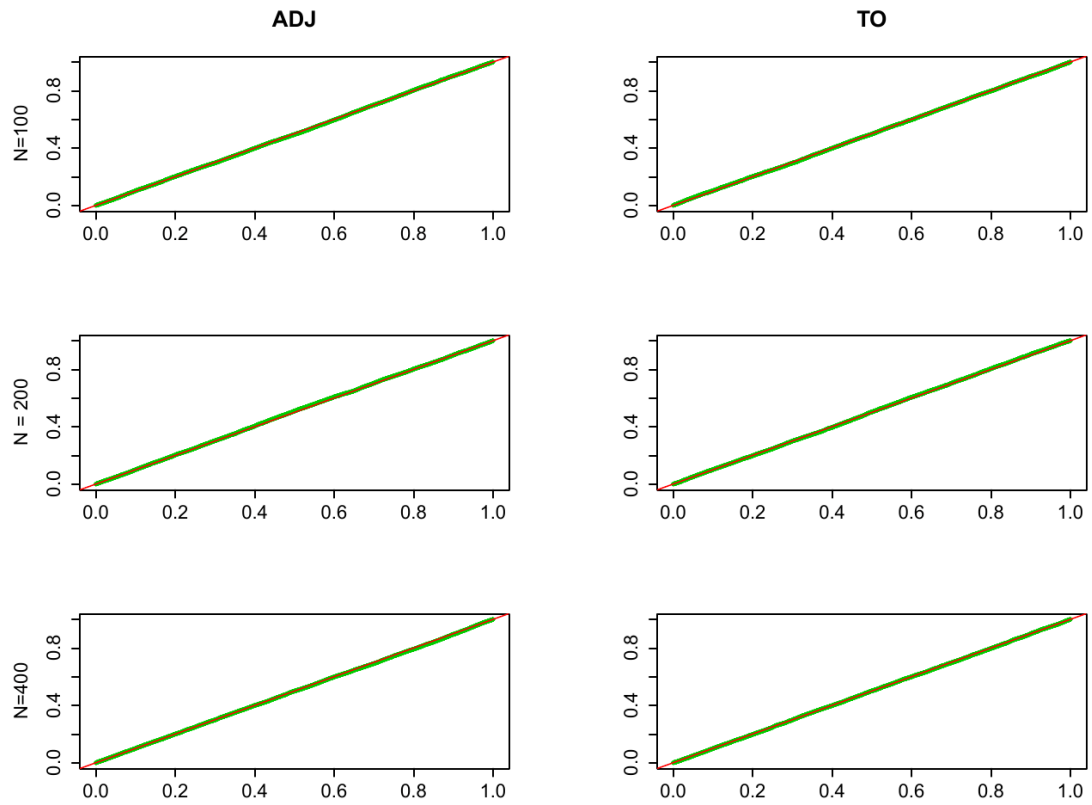


Figure 5.4: QQ plots for p-values obtained from group independence test C_2 against uniform $[0,1]$ distribution for ER random graphs with adjacency information and TO. For each plot, 10,000 sets of networks were generated from the ER random graph model. P-values were estimated by the proposed chi-squared test. The plot shows that with the ER random graph model, when N is greater than 100, approximation of the null sampling distribution of $C_{2\pi}$ by the chi squared distribution is acceptable.

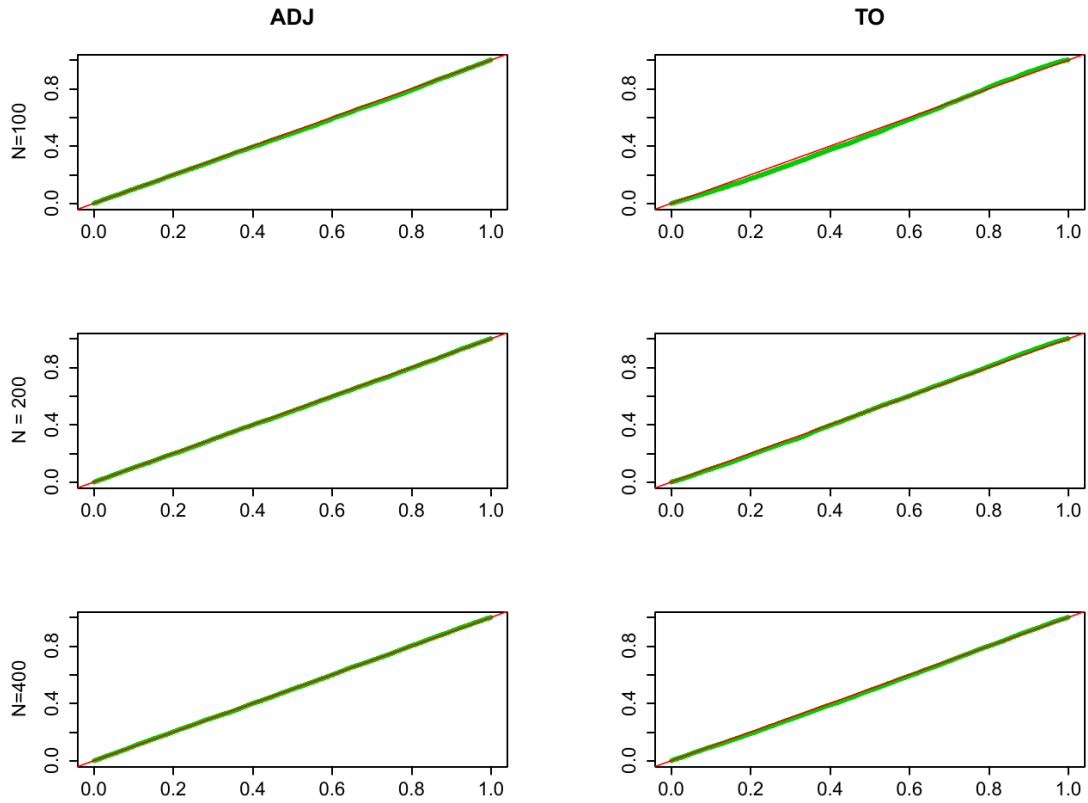


Figure 5.5: QQ plots for p-values obtained from group independence test C_2 against uniform $[0,1]$ distribution for RG random graphs with adjacency information and TO. For each plot, 10,000 sets of networks were generated from the RG random graph model. P-values were estimated by the proposed chi-squared test. The plot shows that with RG random graph model, when N is greater than 100, approximation of the null sampling distribution of $C_{2\pi}$ by the chi squared distribution is acceptable.

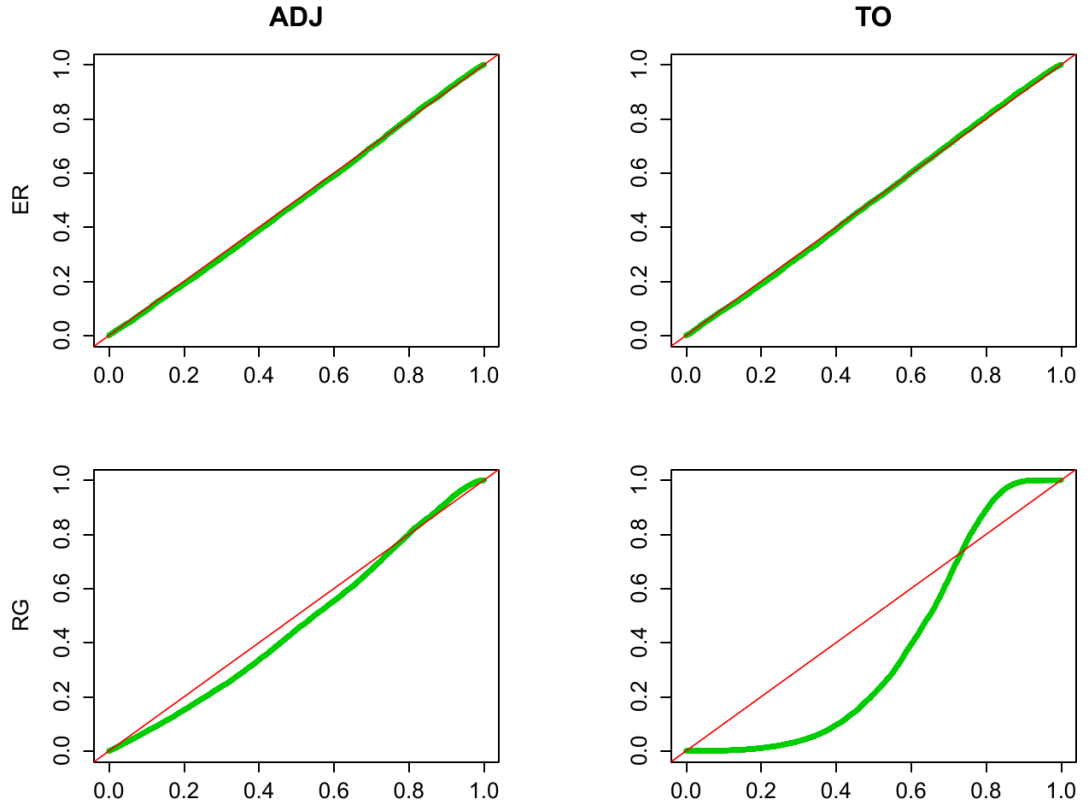


Figure 5.6: QQ plots for asymptotic p-values obtained from group independence test C_{1g} against the uniform $[0,1]$ distribution, with $N = 200$ for ER and RG network models with two edge weight measures. This plots shows that C_{1g} works for the ER random graph model, but it fails for the RG network.

5.5.4 Power study

In the power study, we compared the power when using two measures, adjacency and TO, for both tests.

For the test of mutual independence among networks (C_1), we needed to generate a set of networks with a pre-specified degree of topological similarity. We first generated a network $\mathcal{A}$ from an ER or RG model with size $N = 200$, with parameters p or d . Then, we generated $P = 10$ networks $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$ from $\mathcal{A}$ with a common percentage of noise p_m . As p_m increases, the dissimilarity among $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$ also increases. For each given value of p_m , we generated 1,000 sets of networks, then applied the mutual

independence test with the two measures, and evaluated the proportion of tests that rejected the null hypothesis of independence at a 5% significance level. P-values were also adjusted for multiple testing by controlling the false discovery rate [13]. Figure 5.7 shows that TO has more power to detect subtle differences among a set of networks.

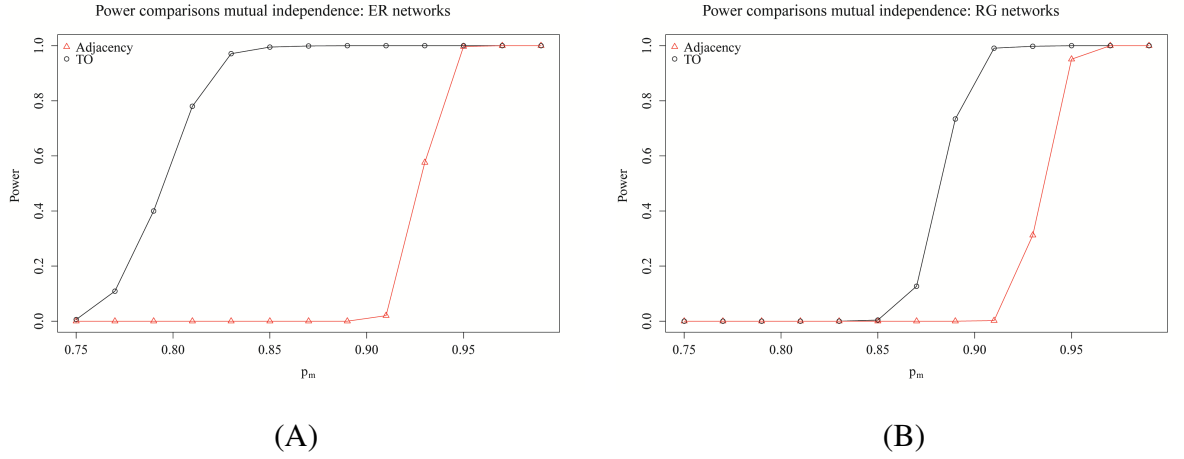


Figure 5.7: (A) and (B) are power of test of mutual independence between a set of networks for ER (A) and RG (B) with adjacency and TO measures. As the noise level p_m increases, the TO has more power to detect true structural changes than adjacency, for both graph models.

For the group independence test (C_2), we needed to generate two groups of networks with an "average" pre-specified degree of topological similarity. We first generated a network $\mathcal{A}$ from the ER or RG model with size $N = 200$ with parameters p or d . Similarly as in Section 3.3.2, we generated a paired network $\mathcal{B}$ by first making an exact copy of $\mathcal{A}$. Then, we randomly shuffled a fixed number of edges, such that $\mathcal{A}$ and $\mathcal{B}$ differed with a level of noise p'_m . Then, we generated $P = 10$ networks $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$ from $\mathcal{A}$ with a common percentage of noise p_m . Then, $P = 10$ networks of group $\mathcal{G}(\mathcal{A})$ with $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_P\}$ from $\mathcal{A}$ and $Q = 10$ networks of group $\mathcal{G}(\mathcal{B})$ with $\{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_Q\}$ from $\mathcal{B}$, were generated with a common percentage of noise $p'_m = 0.5$. Here we denote by p'_m the within group level of noise. Therefore, as p_m increases, the dissimilarity between group $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$ also increases. Again for each given value of p_m , we generated 1,000 paired groups of networks, and then applied the group independence test with the two measures, and evaluated the proportion of tests that rejected the null hypothesis of independence at a 5% significance

level. P-values were again adjusted for multiple testing by controlling the false discovery rate [13]. Figure 5.7 shows that TO is more powerful to detect subtle differences between groups of networks.

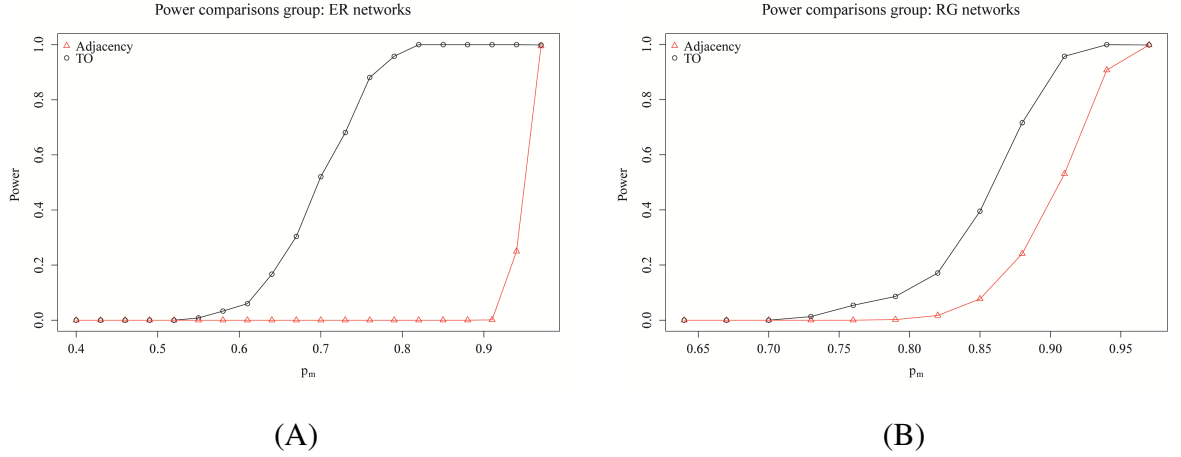


Figure 5.8: (A) and (B) are power of test of group independence between a set of networks for ER (A) and RG (B) with adjacency and TO measures. As the noise level p_m increases, the TO has more power to detect true structural changes than adjacency, for both graph models.

5.6 Summary

In this chapter, we considered the problem of comparing multiple networks. We proposed two different statistical testing procedures for two different problems. We also built a link between the test of group independence with MERGP and analytically derived the maximum likelihood estimate of MERGP using a normality assumption. Simulation studies validated both tests. The TO measure always has more power than adjacency to detect structure changes. Additionally, we further discussed the limitation of the independence assumption and pointed out directions for future studies.

Chapter 6

Conclusion

The central topic of this thesis is the comparison of labelled graphs. Our purpose is to push forward the inferential procedures for different network comparison problems and to provide them with a rigorous theoretical validation. The emphasis of the thesis is on statistical inference for testing independence between networks or sets of networks. The thesis included contributions to inferential procedures for comparing labelled graphs. In this chapter, the main conclusions of the preceding chapters are briefly summarized and suggestions for future research directions are also discussed.

6.1 Summary Of The Thesis

In Chapter 3, we studied the problem of testing the independence between two graphs. We proposed a distance measure, namely the GHD, and suggested that the TO can better characterise local topologies than, say, adjacency information. We then showed that for adjacency and TO measures, as the size of the networks increases, the null sampling distribution of the GHD test statistic may be approximated by a standard normal distribution, under three different random graph models. Therefore, p -values could be obtained without the need to resort to computationally expensive permutation procedures. Simulation results showed that the GHD_TO statistic, which uses the TO, takes into account higher-order neighbourhood structures, and, hence, has more power to detect differential structures. In Chapter 4, we proposed a statistical procedure that is capable of detecting

localized difference in two large complex networks. This procedure can detect a differential subnetwork effectively, by utilising the asymptotic property of the GHD statistic. This procedure was then applied to a co-methylation network in ovarian cancer. GO and KEGG pathway enrichment analysis was applied to obtain insights into the biological meaning of the differential subnetwork detected. In Chapter 5, the problem of comparing multiple networks was studied. We proposed two different inferential procedures for two different problems. We further established a connection between the test of group independence and MERGP, and showed that the maximum likelihood estimate of MERGP can be analytically derived under the null hypothesis. We discussed the limitation of such independence assumptions being too strong, and we also discussed directions for future studies.

6.2 Future Work

For the main materials in this thesis, there are several directions that can be followed for future research. Such extensions are now briefly discussed.

In Chapter 3, we mainly focused on comparing a pair of labelled graphs with two edge weights, TO and adjacency. Although we showed by simulation that the TO measure, together with the GHD statistic, performs better than use of adjacency information and $\frac{1}{sp}$, the question may arise whether there exists a better measure to characteristic graph topology. The answer is yes. For example, the General Topological Overlap Measure (GTOM) with step 2 is a more powerful measure to detect structural differences than the TO (TO is a special case of GTOM with step 1 [56]). However, if we use complex edge weight measures, the normality of the sampling distribution under permutation may become too difficult to be proved or may even be violated. Therefore, to avoid doing permutations, one possible method is to compute the exact third moment of the permutation distribution and further approximate the permutation distribution under the null hypothesis using the Pearson type III distribution [72].

In Chapter 4, although we proposed an algorithm to detect differential subnetworks by sequentially removing nodes from large networks, this algorithm has some potential drawbacks. In Section 4.3, simulation results showed that this algorithm generated some False Positives (FP) that had very small node degrees. One possible way of reducing such FPs

is to reconstruct $\delta_{\mathcal{V}_K|s}$ and add a term to penalize nodes with small node degrees in the remaining subnetwork. Penalizing nodes with small degrees is straightforward, however, such modification may result in loss of the monotonic trend in the p -value, which is not desirable. Another way of improving the subnetwork detection algorithm is that, instead of tracking the differential score $\delta_{\mathcal{V}_K|s}$ for each node, we track the differential score $\delta_{\mathcal{V}_K|\mathcal{V}_s}$ for a small group of nodes $\mathcal{V}_s$ containing small numbers of nodes (e.g. 2 or 3). These small groups of nodes might be picked based on the fact that two nodes are connected or three nodes form a clique of size three. To detect a differential subnetwork, removing a small group of nodes with the most topological impact each time is, to some extent, more sensible than removing a node with the most node-level impact. However, the computational complexity may become a serious issue, in such case, for networks with relatively high densities.

In Chapter 5, we discussed the limitations of the independence hypothesis, and suggested a way to utilise the test of independence. In practice, it is of great interest to detect differential subnetworks between two groups of individuals. However, to detect subnetworks for the two-group case, the computation efficiency is a notable issue when the sample size of both groups are large. A computationally effective algorithm for this case needs to be developed in the future.

References

- [1] S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E.D. Bullmore. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, 26(1):63–72, 2006.
- [2] R. Albert, H. Jeong, and A-L. Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.
- [3] J.D. Allen, Y. Xie, M. Chen, L. Girard, and G.H. Xiao. Comparing statistical methods for constructing large scale gene networks. *PLoS One*, 7(1):e29348, 2012.
- [4] B.S. Anderson, C. Butts, and K. Carley. The interaction of size and density with graph-level indices. *Social Networks*, 21(3):239–268, 1999.
- [5] Frank B Baker and Lawrence J Hubert. The analysis of social interaction data a nonparametric technique. *Sociological Methods & Research*, 9(3):339–361, 1981.
- [6] D. Banks and K. Carley. Metric inference for social networks. *Journal of classification*, 11(1):121–149, 1994.
- [7] A-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
- [8] A-L. Barabási and Z.N. Oltvai. Network biology: understanding the cell’s functional organization. *Nature reviews. Genetics*, 5(2):101–113, February 2004.
- [9] A.L. Barabási and J. Frangos. *Linked: The New Science Of Networks Science Of Networks*. Basic Books, 2002.

- [10] S. Barbarossa and G. Scutari. Decentralized maximum-likelihood estimation for sensor networks composed of nonlinearly coupled dynamical systems. *Signal Processing, IEEE Transactions on*, 55(7):3456–3470, 2007.
- [11] T.E. Bartlett, S.C. Olhede, and A. Zaikin. A DNA methylation network interaction measure, and detection of network oncomarkers. *PLoS ONE*, 9(1):e84573, 2014.
- [12] T. Beißbarth and T.P. Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.
- [13] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 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*, 1:1165–1188, 2001.
- [15] B. Biesma, P.H. Willemse, N.H. Mulder, D.T. Sleijfer, J.A. Gietema, R. Mull, P.C. Limburg, J. Bouma, E. Vellenga, and E.G. de Vries. Effects of interleukin-3 after chemotherapy for advanced ovarian cancer. *Blood*, 80(5):1141–1148, 1992.
- [16] J.A. Bondy and U.S.R. Murty. *Graph Theory*. Springer, 2008.
- [17] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.
- [18] H. Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18(8):689–694, 1997.
- [19] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *Pattern recognition letters*, 19(3):255–259, 1998.
- [20] C. Butts and K. Carley. Multivariate methods for interstructural analysis. CASOS working paper. Carnegie Mellon University, Pittsburgh, PA: Center for the Computational Analysis of Social and Organization Systems, 2001.

- [21] C. Butts and K.M. Carley. Canonical labeling to facilitate graph comparison. Technical report, Carnegie Mellon University, Institute for Complex Engineered Systems, 1998.
- [22] C.T. Butts. Permutation models for relational data. *Sociological Methodology*, 37(1):257–281, 2007.
- [23] C.T. Butts. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11(1):13–41, 2008.
- [24] C.T. Butts and K.M. Carley. Some simple algorithms for structural comparison. *Computational & Mathematical Organization Theory*, 11(4):291–305, 2005.
- [25] Y.B. Chung. Career counseling with lesbian, gay, bisexual, and transgendered persons: The next decade. *The Career Development Quarterly*, 52(1):78–86, 2003.
- [26] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5):058701, 2003.
- [27] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Physical Review Letters*, 86(16):3682–3685, 2001.
- [28] L. da F Costa, M. Kaiser, and C.C. Hilgetag. Predicting the connectivity of primate cortical networks from topological and spatial node properties. *BMC Systems Biology*, 1(1):16, 2007.
- [29] J. Dall and M. Christensen. Random geometric graphs. *Physical Review E*, 66(1):016121, 2002.
- [30] H.E. Daniels. The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 33(2):129–135, 1944.
- [31] M.H. DeGroot, P.I. Feder, and P.K. Goel. Matchmaking. *The Annals of Mathematical Statistics*, 42(2):578–593, 1971.
- [32] Z. Dezső and A.L. Barabási. Halting viruses in scale-free networks. *Physical Review E*, 65(5):055103, 2002.

- [33] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Size-dependent degree distribution of a scale-free growing network. *Physical Review E*, 63(6):062101, 2001.
- [34] P. Du, W.A. Kibbe, and S.M. Lin. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–1548, 2008.
- [35] P. Du, X. Zhang, C.C. Huang, N. Jafari, W. Kibbe, L.f. Hou, and S. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.
- [36] V.M. Eguiluz, D.R. Chialvo, G.A. Cecchi, M. Baliki, and A.V. Apkarian. Scale-free brain functional networks. *Phys. Rev. Lett.*, 94:018102, Jan 2005.
- [37] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.
- [38] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.
- [39] A. Faragó. On the structure of classes of random graphs. *Chicago Journal of Theoretical Computer Science*, 1:1–15, 2010.
- [40] O. Frank and D. Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.
- [41] J.H. Friedman and L.C. Rafsky. Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics*, 11:377–391, 1983.
- [42] T.F. Fuller, A. Ghazalpour, J.E. Aten, T.A. Drake, A.J. Lusis, and S. Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18(6):463–472, 2007.
- [43] X.B. Gao, B. Xiao, D.C. Tao, and X.L. Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129, 2010.
- [44] R. Gill, S. Datta, and S. Datta. A statistical framework for differential network analysis from microarray data. *BMC bioinformatics*, 11(1):95, 2010.

- [45] K-I Goh, B Kahng, and D Kim. Universal behavior of load distribution in scale-free networks. *Physical Review Letters*, 87(27):278701, 2001.
- [46] A. Gut. *An intermediate course in probability*, volume 1. Springer, 1995.
- [47] M. Haenggi. On distances in uniformly random networks. *Information Theory, IEEE Transactions on*, 51(10):3584–3586, 2005.
- [48] M. Haenggi, J.G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *Selected Areas in Communications, IEEE Journal on*, 27(7):1029–1046, 2009.
- [49] R.W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.
- [50] R.W. Hamming. The unreasonable effectiveness of mathematics. *American Mathematical Monthly*, 87(2):81–90, 1980.
- [51] R. J. Henery. Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):pp. 86–91, 1981.
- [52] D.J. Higham, M. Rašajski, and N. Pržulj. Fitting a geometric graph to a protein–protein interaction network. *Bioinformatics*, 24(8):1093–1099, 2008.
- [53] P.W. Holland and S. Leinhardt. The statistical analysis of local structure in social networks. *Sociological Methodology*, 6:1C45, 1975.
- [54] P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50, 1981.
- [55] R. Horaud and T. Skordas. Stereo correspondence through feature grouping and maximal cliques. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(11):1168–1180, 1989.
- [56] S. Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer Verlag, 2011.

- [57] S. Horvath, Y.f. Zhang, P. Langfelder, R.S. Kahn, M.P.M. Boks, K. van Eijk, L.H. van den Berg, and R.A. Ophoff. Aging effects on dna methylation modules in human brain and blood tissue. *Genome Biol*, 13(10):R97, 2012.
- [58] L. Hubert and J. Schultz. Maximum likelihood paired-comparison ranking and quadratic assignment. *Biometrika*, 62(3):655–659, 1975.
- [59] L.J. Hubert. *Assignment methods in combinatorial data analysis*, volume 1. Marcel Dekker, 1987.
- [60] L.J. Hubert and R.G. Golledge. A heuristic method for the comparison of related structures. *Journal of Mathematical Psychology*, 23(3):214–226, 1981.
- [61] N.J. Hudson, A. Reverter, and B.P. Dalrymple. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS computational biology*, 5(5):e1000382, 2009.
- [62] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- [63] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.
- [64] D. Krackhardt. Qap partialling as a test of spuriousness. *Social Networks*, 9(2):171–186, 1987.
- [65] D. Krackhardt. Cognitive social structures. *Social Networks*, 9(2):109–134, 1987.
- [66] D. Krackhardt. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks*, 10(4):359–381, 1988.
- [67] R. Levinson. Pattern associativity and the retrieval of semantic networks. *Computers & Mathematics with Applications*, 23(6-9):573–600, 1992.
- [68] L.D. Lord, P. Allen, P. Expert, O. Howes, R. Lambiotte, P. McGuire, S.K. Bose, S. Hyde, and F.E. Turkheimer. Characterization of the anterior cingulate’s role in the at-risk mental state using graph theory. *Neuroimage*, 56:1531–39, 2011.

- [69] M.H. Mangi and A.C. Newland. Interleukin-3 in hematology and oncology: current state of knowledge and future directions. *Cytokines, cellular & molecular therapy*, 5(2):87–95, 1999.
- [70] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209, 1967.
- [71] D. Meunier, S. Achard, A. Morcom, and E.D. Bullmore. Age-related changes in modular organization of human brain functional networks. *Neuroimage*, 44(3):715–723, 2009.
- [72] P.W. Mielke and K.J. Berry. *Permutation methods: a distance function approach*. Springer, 2007.
- [73] C. Minas, E. Curry, and G. Montana. A distance-based test of association between paired heterogeneous genomic data. *Bioinformatics*, 29(20):2555–2563, 2013. doi: 10.1093/bioinformatics/btt450.
- [74] R. Myers, RC Wison, and E.R. Hancock. Bayesian graph edit distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(6):628–635, 2000.
- [75] M.E.J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
- [76] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.
- [77] O. Odibat and C.K. Reddy. Ranking differential genes in co-expression networks. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 350–354. ACM, 2011.
- [78] O. Odibat and C.K. Reddy. Ranking differential hubs in gene co-expression networks. *Journal of Bioinformatics and Computational Biology*, 10(1):1240002, 2012.
- [79] P. Pattison and S. Wasserman. Logit models and logistic regressions for social networks: Ii. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193, 1999.

- [80] P. Pattison, S. Wasserman, G. Robins, and A.M. Kanfer. Statistical evaluation of algebraic constraints for social networks. *Journal of mathematical psychology*, 44(4):536–568, 2000.
- [81] M. Penrose. *Random geometric graphs*, volume 5. Oxford University Press Oxford, 2003.
- [82] D.T. Pham, J. Möcks, and L. Sroka. Asymptotic normality of double-indexed linear permutation statistics. *Annals of the Institute of Statistical Mathematics*, 41(3):415–427, 1989.
- [83] R.L. Plackett. The analysis of permutations. *Applied Statistics*, 24:193–202, 1975.
- [84] P. Pons and M. Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218, 2006.
- [85] V.M. Preciado and A. Jadbabaie. Spectral analysis of virus spreading in random geometric networks. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 4802–4807. IEEE, 2009.
- [86] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
- [87] N. Przulj, D.G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.
- [88] R.C. Read and D.G. Corneil. The graph isomorphism disease. *Journal of Graph Theory*, 1(4):339–363, 1977.
- [89] A. Reverter, A. Ingham, S.A. Lehnert, S.H. Tan, Y. Wang, A. Ratnakumar, and B.P. Dalrymple. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics*, 22(19):2396–2404, 2006.

- [90] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph models for social networks. *Social networks*, 29(2):173–191, 2007.
- [91] D. Ruan, A. Young, and Montana. G. Differential analysis of biological networks. Technical report, Imperial College London, 2014.
- [92] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.
- [93] Z.M. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604–2611, 2007.
- [94] L.G. Shapiro and R.M. Haralick. Structural descriptions and inexact matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1(5):504–519, 1981.
- [95] D.J. Sharp, C.F. Beckmann, R. Greenwood, K.M. Kinnunen, V. Bonnelle, X. De Boissezon, J.H. Powell, S.J. Counsell, M.C. Patel, and R. Leech. Default mode network functional and structural connectivity after traumatic brain injury. *Brain*, 134(8):2233–2247, 2011.
- [96] S.L. Simpson, S. Hayasaka, and P.J. Laurienti. Exponential random graph modeling for complex brain networks. *PLoS One*, 6(5):e20039, 2011.
- [97] S.L. Simpson, M.N. Moussa, and P.J. Laurienti. An exponential random graph modeling approach to creating group-based representative whole-brain connectivity networks. *Neuroimage*, 60(2):1117–1126, 2012.
- [98] M. Song, Y. Liu, Y. Zhou, K. Wang, C.S. Yu, and T.Z. Jiang. Default network and intelligence difference. *Autonomous Mental Development, IEEE Transactions on*, 1(2):101–109, 2009.
- [99] A.E. Teschendorff, U. Menon, A. Gentry-Maharaj, S.J. Ramus, D.J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C.G. Bell, A.P. Maxwell, et al. Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research*, 20(4):440–446, 2010.

- [100] J.R. Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)*, 23(1):31–42, 1976.
- [101] T. Vaissière, R.J.J Hung, D. Zaridze, A. Moukeria, C. Cuenin, V. Fasolo, G. Ferro, A. Paliwal, P. Hainaut, P. Brennan, Jörg Tost, Paolo Boffetta, and Zdenko Herceg. Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors. *Cancer research*, 69(1):243–252, January 2009.
- [102] N. Vakhania, V. Tarieladze, and S. Chobanyan. *Probability distributions on Banach spaces*, volume 14. Springer, 1987.
- [103] M. Valencia, M.A Pastor, M.A. Fernández-Seara, J. Artieda, J. Martinerie, and M. Chavez. Complex modular structure of large-scale brain networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(2):023119, 2009.
- [104] R. van der Hofstad. *Random graphs and complex networks*, 2013. Lecture notes, available online at: <http://www.win.tue.nl/~rhofstad/NotesRGCN2011.pdf>.
- [105] C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 0408709294.
- [106] B.C.M. van Wijk, C.J. Stam, and A. Daffertshofer. Comparing brain networks of different size and connectivity density using graph theory. *PLoS One*, 5(10):e13701, 2010.
- [107] W.D. Wallis, P. Shoubridge, M. Kraetz, and D. Ray. Graph distances using graph union. *Pattern Recognition Letters*, 22(6):701–704, 2001.
- [108] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, 61(3):401–425, 1996.
- [109] D.J. Watts and S.H. Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

-
- [110] A.M. Zapała. Unbounded mappings and weak convergence of measures. *Statistics & Probability Letters*, 78(6):698–706, 2008.
- [111] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1): 1128, 2005.

Appendix A

Additional Derivation And Proofs

A.1 Closed-form moment computation

A.1.1 Notation

In this section we provide a derivation of first two moments for the permutation distribution of $\text{GHD}_\pi(\mathcal{A}_\pi, \mathcal{B})$. First we introduce some additional notation. Suppose we have two pairs of networks, $(\mathcal{A}_1, \mathcal{B}_1)$ and $(\mathcal{A}_2, \mathcal{B}_2)$. On applying the GHD statistic for both comparisons, we have

$$\begin{aligned}\text{GHD}_1 &= \frac{1}{N(N-1)} \sum_{i,j} (a_{1ij} - b_{1ij})^2 \\ \text{GHD}_2 &= \frac{1}{N(N-1)} \sum_{i,j} (a_{2ij} - b_{2ij})^2,\end{aligned}$$

where a_{1ij} , b_{1ij} , a_{2ij} and b_{2ij} are edge weight measures extracted from networks $\mathcal{A}_1$, $\mathcal{B}_1$, $\mathcal{A}_2$ and $\mathcal{B}_2$ respectively. Given a permutation $\pi \in \Pi$, we denote the GHD values after node permutation of both $\mathcal{A}_1$ and $\mathcal{A}_2$ by π to be

$$\begin{aligned}\text{GHD}_{1\pi} &= \frac{1}{N(N-1)} \sum_{i,j} (a_{1\pi(i)\pi(j)} - b_{1ij})^2 \\ \text{GHD}_{2\pi} &= \frac{1}{N(N-1)} \sum_{i,j} (a_{2\pi(i)\pi(j)} - b_{2ij})^2.\end{aligned}$$

When we are interested in comparing only two networks, $\mathcal{A}$ and $\mathcal{B}$, we let $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{A}$ and $\mathcal{B}_1 = \mathcal{B}_2 = \mathcal{B}$, and hence $\text{GHD}(\mathcal{A}, \mathcal{B}) = \text{GHD}_1 = \text{GHD}_2$. We also introduce the following notation that will be used for the computation of the GHD moments:

$$\begin{aligned}
 {}^a S_t &= \sum_{i=1}^N \sum_{j=1}^N a_{tij} \text{ for } t = 1, 2 \\
 {}^b S_t &= \sum_{i=1}^N \sum_{j=1}^N b_{tij} \text{ for } t = 1, 2 \\
 {}^a S_t^2 &= \sum_{i=1}^N \sum_{j=1}^N a_{tij}^2 \text{ for } t = 1, 2 \\
 {}^b S_t^2 &= \sum_{i=1}^N \sum_{j=1}^N b_{tij}^2 \text{ for } t = 1, 2 \\
 {}^a S &= \sum_{i=1}^N \sum_{j=1}^N a_{1ij} a_{2ij} \\
 {}^b S &= \sum_{i=1}^N \sum_{j=1}^N b_{1ij} b_{2ij} \\
 {}^a S^T &= \sum_{i=1}^N \sum_{j=1}^N a_{1ij} a_{2ji} \\
 {}^b S^T &= \sum_{i=1}^N \sum_{j=1}^N b_{1ij} b_{2ji}
 \end{aligned}$$

and

$$\begin{aligned}
 {}^1 A &= \sum_{i=1}^N a_{1i\cdot} \sum_{i=1}^N a_{2i\cdot} \\
 {}^2 A &= \sum_{i=1}^N a_{1\cdot i} \sum_{i=1}^N a_{2i\cdot} \\
 {}^3 A &= \sum_{i=1}^N a_{1i\cdot} \sum_{i=1}^N a_{2\cdot i}
 \end{aligned}$$

$$\begin{aligned}
 {}^4A &= \sum_{i=1}^N a_{1.i} \sum_{i=1}^N a_{2.i} \\
 {}^1B &= \sum_{i=1}^N b_{1i.} \sum_{i=1}^N b_{2i.} \\
 {}^2B &= \sum_{i=1}^N b_{1.i} \sum_{i=1}^N b_{2i.} \\
 {}^3B &= \sum_{i=1}^N b_{1i.} \sum_{i=1}^N b_{2.i} \\
 {}^4B &= \sum_{i=1}^N b_{1.i} \sum_{i=1}^N b_{2.i}.
 \end{aligned}$$

A.1.2 Calculation of the mean

The expected value is

$$\begin{aligned}
 E_{\Pi}[\text{GHD}_{t\pi}] &= E_{\Pi}\left[\frac{1}{N(N-1)} \sum_{i,j} (a_{t\pi(i)\pi(j)} - b_{tij})^2\right] \\
 &= \frac{1}{N(N-1)} \sum_{i,j} E_{\Pi}[(a_{t\pi(i)\pi(j)}^2 + b_{tij}^2 - 2a_{t\pi(i)\pi(j)}b_{tij})] \\
 &= \frac{{}^bS_t^2}{N(N-1)} + \frac{{}^aS_t^2}{N(N-1)} - \frac{2 \times {}^aS_t {}^bk_t}{N^2(N-1)^2}.
 \end{aligned}$$

for $t = 1$ and $t = 2$.

A.1.3 Calculation of the covariance

The covariance between $\text{GHD}_{1\pi}$ and $\text{GHD}_{2\pi}$:

$$\begin{aligned}
 &Cov(\text{GHD}_{1\pi}, \text{GHD}_{2\pi}) \\
 &= \left(\frac{1}{N(N-1)}\right)^2 Cov\left[\sum_{i,j} (a_{1\pi(i)\pi(j)} - b_{1ij})^2, \sum_{i,j} (a_{2\pi(i)\pi(j)} - b_{2ij})^2\right] \\
 &= \left(\frac{1}{N(N-1)}\right)^2 Cov\left[\sum_{i,j} (a_{1\pi(i)\pi(j)}^2 + b_{1ij}^2 - 2a_{1\pi(i)\pi(j)}b_{1ij}), \sum_{i,j} (a_{2\pi(i)\pi(j)}^2 + b_{2ij}^2 - 2a_{2\pi(i)\pi(j)}b_{2ij})\right],
 \end{aligned}$$

$$\begin{aligned}
& \sum_{i,j} (a_{2\pi(i)\pi(j)}^2 + b_{2ij}^2 - 2a_{2\pi(i)\pi(j)}b_{2ij}) \\
&= \left(\frac{1}{N(N-1)}\right)^2 \text{Cov}\left[\sum_{i,j} a_{1\pi(i)\pi(j)}b_{1ij}, \sum_{i,j} a_{2\pi(i)\pi(j)}b_{2ij}\right] \\
&= \frac{4}{(N(N-1))^2} \text{Cov}\left[\sum_{i,j} a_{1\pi(i)\pi(j)}b_{1ij}, \sum_{i,j} a_{2\pi(i)\pi(j)}b_{2ij}\right]. \tag{A.1}
\end{aligned}$$

Denote the two quantities in (A.1) by $\tau_{1\pi}$ and $\tau_{2\pi}$, respectively, so that

$$\begin{aligned}
\tau_{1\pi} &= \sum_{i,j} a_{1\pi(i)\pi(j)}b_{1ij}, \\
\tau_{2\pi} &= \sum_{i,j} a_{2\pi(i)\pi(j)}b_{2ij}.
\end{aligned}$$

Then the covariance between $\tau_{1\pi}$ and $\tau_{2\pi}$ can be written as

$$\begin{aligned}
\text{Cov}(\tau_{1\pi}, \tau_{2\pi}) &= E_{\Pi}[\tau_{1\pi}\tau_{2\pi}] - E_{\Pi}[\tau_{1\pi}]E_{\Pi}[\tau_{2\pi}] \\
&= E_{\Pi}[\tau_{1\pi}\tau_{2\pi}] - \mu_1\mu_2,
\end{aligned}$$

where μ_1 and μ_2 are the means of $\tau_{1\pi}$ and $\tau_{2\pi}$ given by

$$\begin{aligned}
\mu_t &= E_{\Pi}[\tau_{t\pi}] \quad t = 1, 2 \\
&= E_{\Pi}\left[\sum_{i,j} a_{t\pi(i)\pi(j)}b_{tij}\right] \\
&= \sum_{i,j} b_{tij}E[a_{t\pi(i)\pi(j)}] \\
&= \frac{1}{N(N-1)} {}^bS_t {}^aS_t.
\end{aligned}$$

In order to calculate $E_{\Pi}[\tau_{1\pi}\tau_{2\pi}]$, we observe directly that

$$E_{\Pi}[\tau_{1\pi}\tau_{2\pi}] = E_{\Pi}\left[\sum_{i=1}^N \sum_{j=1}^N a_{1\pi(i)\pi(j)}b_{1ij} \sum_{k=1}^N \sum_{l=1}^N a_{2\pi(k)\pi(l)}b_{2kl}\right]$$

$$\begin{aligned}
 &= E_{\Pi} \left[\sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \sum_{l=1}^N a_{1\pi(i)\pi(j)} b_{1ij} a_{2\pi(k)\pi(l)} b_{2kl} \right] \\
 &= \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \sum_{l=1}^N b_{1ij} b_{2kl} E_{\Pi} [a_{1\pi(i)\pi(j)} a_{2\pi(k)\pi(l)}].
 \end{aligned}$$

The evaluation of $E_{\Pi}(a_{1\pi(i)\pi(j)} a_{2\pi(k)\pi(l)})$ involves 7 distinct cases, depending on the four subscripts, which we consider separately:

1. Two tied subscripts 1: $(i, j) = (k, l)$
2. Two tied subscripts 2: $(i, j) = (l, k)$
3. One tied subscript 1: $i = k$
4. One tied subscript 2: $i = l$
5. One tied subscript 3: $j = k$
6. One tied subscript 4: $j = l$
7. Subscripts are all distinct: $i \neq j \neq k \neq l$.

Case 1: Two tied subscripts $(i, j) = (k, l)$:

$$E_{\Pi}[a_{1\pi(i)\pi(j)} a_{2\pi(i)\pi(j)}] = \frac{{}^a S}{N(N-1)}.$$

Case 2: Two tied subscripts $(i, j) = (l, k)$:

$$E_{\Pi}[a_{1\pi(i)\pi(j)} a_{2\pi(j)\pi(i)}] = \frac{{}^a S^T}{N(N-1)}.$$

Case 3: One tied subscript 1: $i = k$

$$\begin{aligned}
 &E_{\Pi}[a_{1\pi(i)\pi(j)} a_{2\pi(i)\pi(s)}] \\
 &= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j \sum_l a_{1ij} a_{2il}
 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j a_{1ij} (a_{2i\cdot} - a_{2ij}) \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_l (a_{1ij} a_{2i\cdot} - a_{1ij} a_{2ij}) \\
&= \frac{1}{N(N-1)(N-2)} ({}^1A - {}^aS) \\
&:= {}^aC_1,
\end{aligned}$$

Case 4: One tied subscript 1: $i = l$

$$\begin{aligned}
&E_{\Pi}[a_{1\pi(i)\pi(j)} a_{2\pi(k)\pi(i)}] \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j \sum_k a_{1ij} a_{2ki} \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j a_{1ij} (a_{2\cdot i} - a_{2ji}) \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j (a_{1ij} a_{2\cdot i} - a_{1ij} a_{2ji}) \\
&= \frac{1}{N(N-1)(N-2)} ({}^2A - {}^aS^T) \\
&:= {}^aC_2.
\end{aligned}$$

Case 5: One tied subscript 1: $j = k$

$$\begin{aligned}
&E_{\Pi}[a_{1\pi(i)\pi(j)} a_{2\pi(j)\pi(l)}] \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j \sum_l a_{1ij} a_{2jl} \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j a_{1ij} (a_{2j\cdot} - a_{2ji}) \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j (a_{1ij} a_{2j\cdot} - a_{1ij} a_{2ji}) \\
&= \frac{1}{N(N-1)(N-2)} \sum_i \sum_j (a_{1ji} a_{2i\cdot} - a_{1ji} a_{2ij}) \\
&= \frac{1}{N(N-1)(N-2)} ({}^3A - {}^aS^T) \\
&:= {}^aC_3.
\end{aligned}$$

Case 6: One tied subscript 1: $j = l$

$$\begin{aligned}
 & E_{\Pi}[a_{1\pi(i)\pi(j)}a_{2\pi(k)\pi(j)}] \\
 = & \frac{1}{N(N-1)(N-2)} \sum_i \sum_j \sum_k a_{1ij}a_{2kj} \\
 = & \frac{1}{N(N-1)(N-2)} \sum_i \sum_j a_{1ij}(a_{2\cdot j} - a_{2ij}) \\
 = & \frac{1}{N(N-1)(N-2)} \sum_i \sum_j (a_{1ij}a_{2\cdot j} - a_{1ij}a_{2ij}) \\
 = & \frac{1}{N(N-1)(N-2)} \sum_i \sum_j (a_{1ij}a_{2\cdot j} - a_{1ji}a_{2ji}) \\
 = & \frac{1}{N(N-1)(N-2)} (^4A - {}^aS) \\
 := & {}^aC_4
 \end{aligned}$$

Case 7: Subscripts are all distinct: $i \neq j \neq k \neq l$

$$\begin{aligned}
 & E_{\Pi}[a_{1\pi(i)\pi(j)}a_{2\pi(k)\pi(l)}] \\
 = & \frac{1}{N(N-1)(N-2)(N-3)} [\sum_i \sum_j \sum_k \sum_l a_{1ij}a_{2kl}] \\
 = & \frac{1}{N(N-1)(N-2)(N-3)} \sum_i \sum_j \sum_k (a_{1ij}(a_{2k\cdot} - a_{2kj} - a_{2ki})) \\
 = & \frac{1}{N(N-1)(N-2)(N-3)} \sum_i \sum_j \sum_k (a_{1ij}a_{2k\cdot} - a_{1ij}a_{2kj} - a_{1ij}a_{2ki}) \\
 = & \frac{1}{N(N-1)(N-2)(N-3)} \sum_i \sum_j (a_{1ij}(a_{2\cdot\cdot} - a_{2i\cdot} - a_{2j\cdot}) - \sum_k a_{1ij}a_{2kj} - \sum_k a_{1ij}a_{2ki}) \\
 = & \frac{1}{N(N-1)(N-2)(N-3)} \sum_i \sum_j (a_{1ij}({}^aS_2 - a_{2i\cdot} - a_{2j\cdot})) - {}^aC_2 - {}^aC_4 \\
 = & \frac{1}{N(N-1)(N-2)(N-3)} ({}^aS_1 {}^aS_2 - \sum_i \sum_j a_{1ij}(a_{2i\cdot} + a_{2j\cdot}) - {}^aC_2 - {}^aC_4) \\
 = & \frac{1}{N(N-1)(N-2)(N-3)} ({}^aS_1 {}^aS_2 - {}^1A - {}^3A - {}^2A - {}^4A + {}^aS + {}^aS^T) \\
 := & {}^aD
 \end{aligned}$$

Hence

$$\begin{aligned}
 E_{\Pi}[\tau_{1\pi}\tau_{2\pi}] &= \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \sum_{l=1}^N b_{1ij}b_{2kl} E_{\Pi}[a_{1\pi(i)\pi(j)}a_{2\pi(k)\pi(l)}] \\
 &= \frac{1}{N(N-1)}({}^aS^bS + {}^aS^{Tb}S^T) + \sum_{w=1}^4 {}^aC_w{}^bC_w + {}^aD^bD.
 \end{aligned}$$

Then

$$\begin{aligned}
 Cov(\tau_{1\pi}, \tau_{2\pi}) &= E_{\Pi}[\tau_{1\pi}\tau_{2\pi}] - E_{\Pi}[\tau_{1\pi}]E_{\Pi}[\tau_{2\pi}] \\
 &= \frac{({}^aS^bS + {}^aS^{Tb}S^T)}{N(N-1)} + \sum_{w=1}^4 {}^aC_w{}^bC_w + {}^aD^bD - \mu_1\mu_2. \quad (A.2)
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 &Cov(\mathbf{GHD}_{1\pi}, \mathbf{GHD}_{2\pi}) \\
 &= \frac{4}{(n(n-1))^2} Cov[\sum_{ij} a_{1\pi(i)\pi(j)}b_{1ij}, \sum_{ij} a_{2\pi(i)\pi(j)}b_{2ij}] \\
 &= \frac{4}{(n(n-1))^2} Cov(\tau_{1\pi}, \tau_{2\pi}) \\
 &= \frac{4}{(n(n-1))^2} [\frac{({}^aS^bS + {}^aS^{Tb}S^T)}{N(N-1)} + \sum_{w=1}^4 {}^aC_w{}^bC_w + {}^aD^bD - \mu_1\mu_2].
 \end{aligned}$$

A.2 Proof Of (5.22)

In this section we sketch a proof of (5.22) (in page 143). First we know that $\vec{Z}_{N\pi} = (Z_{N(1)}^{\pi}, Z_{N(2)}^{\pi}, \dots, Z_{N(T)}^{\pi})'$ is approximately joint normal. And denote $\vec{Z}_{\infty\pi}$ be standard normal distribution with mean 0 and covariance matrix I . $\vec{Z}_{\infty\pi}$ is the limiting condition of $\vec{Z}_{N\pi}$:

$$\vec{Z}_{N\pi} \rightarrow_d \vec{Z}_{\infty\pi} \quad \text{as } N \rightarrow \infty$$

Where $\rightarrow_d$ means converge in distribution. We want to show that given the above convergence hold, given $g(x) = \exp \phi x$ an exponential function, the following convergence

hold:

$$E(g(\vec{Z}_{N\pi})) \rightarrow_d E(g(\vec{Z}_{\infty\pi})) \quad \text{as } N \rightarrow \infty.$$

This above converge for bounded function $g(x)$ is always true, when applying unbounded function $g(X)$ such as exponential function, Zapala [110] have showed that for any continuous, S_∞ -almost uniform integrable (S_∞ -a.u integrable) functions $g : X \rightarrow \bar{R}$,

$$\lim_{s \in S} \int_X g(X) f_s(X) dX = \int_X g(X) f(X) dX, \quad (\text{A.3})$$

if and only if $\mu_s \rightarrow \mu_\infty$.

Where μ_∞ is a τ -smooth finite Borel measure in the space X . Here, $S_\infty = (\mu_s, s \in S_\infty) \subset \mathcal{U}_+$ denotes a net of finite, non-negative Borel measures in a completely regular topological space X indexed by a index set $S_\infty = S \cup \infty$. For this particular problem, we knew that $\vec{Z}_{N\pi} \rightarrow N(0, I)$ as N increases. Then the series of random variables $\mathcal{Z}_\infty = (\vec{Z}_{N\pi}, N = 1, \dots, \infty)$ are finite and non-negative on R^T , Also $\vec{Z}_{\infty\pi}$ which is $N(0, I)$, is τ -smooth finite [102] in R^T . Hence we only need to show that $g(X) = \exp(\phi X)$ is $\mathcal{Z}_\infty$ -a.u integrable. By definition of Zapala [110], $g(X)$ is $\mathcal{Z}_\infty$ -a.u integrable if

$$\bigwedge_{\epsilon} \bigvee_{0 < C_s < \infty} \bigvee_{z_\epsilon \in Z} \bigwedge_{s \succeq s_\epsilon, z \in S_\infty} \int_{x \in X: \|g(x)\| > C_\epsilon} \|g(x)\| f_s(x) dx < \epsilon, \quad (\text{A.4})$$

First for $\vec{Z}_{\infty\pi}$, we know that $f_\infty(x)$ is the joint density of T i.i.d normals. We have:

$$\begin{aligned} & \int_{\|x\| > C_\epsilon} \exp(\vec{\phi}' x) f(x) dx \\ &= \exp(\vec{\phi}' \vec{\phi}) \int_{\|y + \frac{1}{2} \vec{\phi}\| > C_\epsilon} f_\infty(y) dy \quad y = x - \frac{1}{2} \vec{\phi} \\ &\doteq \exp(\vec{\phi}' \vec{\phi}) (1 - \Phi_T(h(C_\epsilon))). \end{aligned}$$

Here we define $\int_{\|y + \frac{1}{2} \vec{\phi}\| > C_\epsilon} f_\infty(y) dy = \Phi_T(h(C_\epsilon))$, Φ_T joint c.d.f of T i.i.d normals and $h(C_\epsilon)$ is the solution of $\|y + \frac{1}{2} \vec{\phi}\| > C_\epsilon$. $\Phi_T(h(C_\epsilon))$ is a increasing function of C_ϵ . For any $\epsilon > 0$, for fixed constant vector $\vec{\phi}$, $\frac{\epsilon}{\exp(\vec{\phi}' \vec{\phi})}$ is a constant greater than 0. For fixed number T ,

there must exist a finite constant C_ϵ such that:

$$(1 - \Phi_T(h(C_\epsilon))) < \frac{\epsilon}{\exp(\vec{\phi}'\vec{\phi})},$$

Therefore $g(X) = \exp(\phi X)$ is a.u integrable with respect to $Z_{\infty\pi}$. Let c_∞ be the constant s.t. (A.4) is smaller than ϵ .

Now lets look at $\vec{Z}_{N\pi}$. Given a network of size N , for any $\epsilon > 0$, there exist $0 < C_\epsilon < \infty$, we have for the left part of (A.4):

$$\int_{\|x\| > C_\epsilon} \exp(\vec{\phi}'x) f_N(x) dx < \epsilon,$$

For each $\gamma_N^\pi(i)$ in $Z_{N(i)}^\pi$, the value of γ_N^π is actually bounded. From the formulation, we can derive the upper bound for adjacency information

$$\begin{aligned} \gamma_N^\pi(i) &= \sum_{k=1}^N \sum_{s=1}^N a_{i*\pi(k)\pi(s)} b_{i*ks} \\ &\leq \min\left(\sum_{k=1}^N a_{i*\pi(k)\pi(s)}, \sum_{k=1}^N b_{i*\pi(k)\pi(s)}\right) \\ &< N(N-1). \end{aligned}$$

Similarly for the TO measure, we know that the value of TO measure is in the interval $[0, 1]$, hence $\gamma_N^\pi(i)$ for TO measure is also bounded by $N(N-1)$. Then the value of $Z_{N(i)}^\pi = L^{-1}((0, 0, \dots, \gamma_N^\pi(i), \dots, 0)' - (0, 0, \dots, \mu_N(i), \dots, 0))$ is also bounded, since σ_{ii}^{-1} and $\mu_N(i)$ are value that corresponds to N , now lets denote the maximum of $Z_{N(i)}^\pi$ as c_b (e.g. $c_b = \max_i \{Z_{N(i)}^\pi\}$). Then for any $\epsilon > 0$, and $C_\epsilon = T c_b$:

$$\begin{aligned} &\int_{\|x\| > C_\epsilon} \exp(\vec{\phi}'x) f_N(x) dx \\ &= \int_{\|x\| > T c_b} \exp(\vec{\phi}'x) f_N(x) dx \\ &= \int_{\text{at least one } Z_{N(i)}^\pi > c_b} \exp(\vec{\phi}'x) f_N(x) dx \\ &= 0 \end{aligned}$$

$$< \epsilon.$$

Thus we can see for fixed N , there is always a constant c_b bounding $Z_{N(i)}^\pi$, such that the integration (A.4) is 0. Hence for finite N and limiting condition of $\vec{Z}_{\infty\pi}$, (A.4) holds true. Then for those large enough N , notice the fact that the bound of $\gamma_N^\pi(i)$ is an increasing function w.r.t N . Hence the bound of $Z_N^\pi(i)$ is increasing till infinity, however also both tails of $Z_N^\pi(i)$ becoming more and more flat and tends to the tails of $N(0, I)$ as N increases to infinity. For any large $N \rightarrow \infty$, for a large enough $c_N < \infty$, the tail of $f_N(x)$ will always tend to lower than $f_\infty(x)$ as $\|x\|$ increases. Hence for any N , $f_N(x) \leq f_\infty(x)$ as $\|x\|$ tends to infinity. Therefore, for any $\epsilon > 0$, exist constant $c_{\max} > c_\infty$, and for any N large, we can intuitively have:

$$\begin{aligned} & \int_{\|x\| > c_{\max}} \exp(\vec{\phi}'x) f_N(x) dx \\ & \leq \int_{\|x\| > c_{\max}} \exp(\vec{\phi}'x) f_\infty(x) dx \\ & \leq \epsilon \end{aligned}$$

In the later chapter, monte carlo simulation have been done to verify the $\mathcal{Z}_\infty$ a.u integrable property of $g(X) = \exp(\phi X)$. We have thus justified (A.4).

Appendix B

DNA Methylation Analysis: Further Experimental Results

B.1 Power Law Of Co-methylation Networks

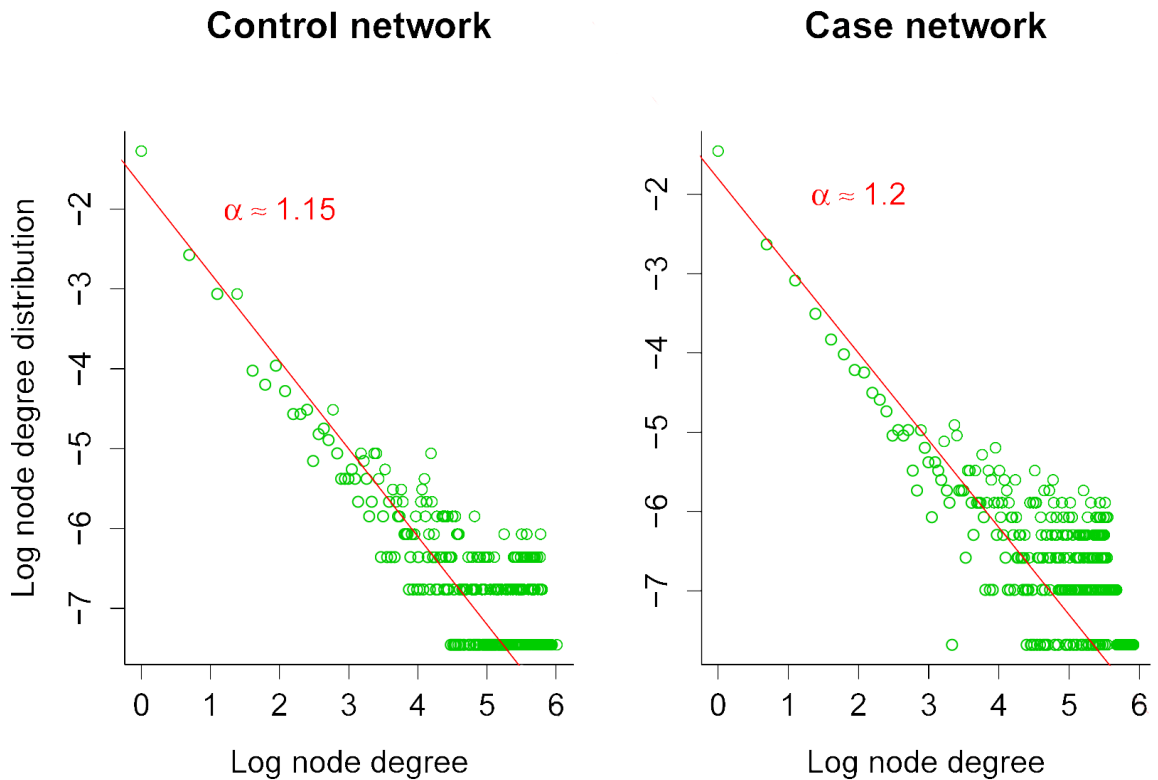


Figure B.1: Check power law property of co-methylation control and case networks. Singular nodes are excluded from both control and case networks. Both plots shows both co-methylation networks roughly follows scale free network models with exponent at around 1.15 to 1.2. Therefore condition for asymptotic normality holds and GHD test can be applied.

B.2 Results From GO Analysis

B.2.1 Top 10 enriched GO terms and pathways for probe set of differential co-methylated (DC) subnetwork

	GO ID	GO:BP TERM	P-value
1	GO:0050896	Response to stimulus	4.01e-09
2	GO:0051716	Cellular response to stimulus	1.44e-08
3	GO:0042221	Response to chemical stimulus	3.29e-08
4	GO:0006950	Response to stress	4.75e-07
5	GO:0048731	System development	1.06e-06
6	GO:0070887	Cellular response to chemical stimulus	1.24e-06
8	GO:0007165	Signal transduction	1.73e-06
7	GO:0007154	Cell communication	2.45e-06
9	GO:0031325	Positive regulation of cellular metabolic process	4.09e-06
10	GO:0048856	Anatomical structure development	4.52e-06

Table B.1: List of Top 10 significant GO terms of BP (Biological Process) for probes in DC subnetwork. There are 762 significant GO terms that related to 1642 selected probes. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0044459	Plasma membrane part	1.19e-05
2	GO:0009986	Cell surface	1.80e-05
3	GO:0044464	Cell part	1.91e-05
4	GO:0071944	Cell periphery	1.92e-05
5	GO:0005623	Cell	1.95e-05
6	GO:0005886	Plasma membrane	3.29e-05
7	GO:0000228	Nuclear chromosome	9.10e-05
8	GO:0044454	Nuclear chromosome part	1.16e-04
9	GO:0005694	Chromosome	3.04e-04
10	GO:0000790	Nuclear chromatin	4.57e-04

Table B.2: List of Top 10 significant GO terms of CC (Cellular Component) for probes in DC subnetwork. There are 73 significant GO terms that related to 1642 selected probes. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0005515	protein binding	9.12e-06
2	GO:0005518	collagen binding	8.44e-05
3	GO:0001104	RNA polymerase II transcription cofactor activity	9.85e-05
4	GO:0004871	signal transducer activity	6.60e-04
5	GO:0060089	molecular transducer activity	6.60e-03
6	GO:0001076	RNA polymerase II transcription factor binding transcription factor activity	6.67e-03
7	GO:0015238	drug transmembrane transporter activity	1.75e-03
8	GO:0042578	phosphoric ester hydrolase activity	1.98e-03
9	GO:0032403	protein complex binding	2.53e-03
10	GO:0044212	transcription regulatory region DNA binding	2.58e-03

Table B.3: List of Top 10 significant GO terms of MF (Molecular Function) for probes in DC subnetwork. There are 154 significant GO terms that related to 1642 selected probes. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	KEGG ID	KEGG pathway	P-value
1	04640	Hemtopoietic cell lineage	2.82e-03
2	05221	Acute myeloid leukemia	6.14e-03
3	04810	Regulation of actin cytoskeleton	1.06e-02
4	04664	Fc epsilon RI signaling pathway	1.14e-02
5	04662	B cell receptor signaling pathway	1.25e-02
6	05144	Malaria	1.38e-02
7	04910	Insulin signaling pathway	1.86e-02
8	04010	MAPK signaling pathway	2.30e-02
9	04964	Proximal tubule bicarbonate reclamation	2.47e-02
10	00603	Glycosphingolipid biosynthesis - globo series	3.29e-02
11	05214	Glioma	4.01e-02
12	05310	Asthma	4.30e-02

Table B.4: List of 12 significant KEGG pathways for probes in DC subnetwork contain 1642 probes. KEGG terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

B.2.2 Top 10 enriched GO terms for Differential methylated (DM) probes that also found enriched with the probe set of DC subnetworks

	GO ID	GO:BP TERM	P-value
1	GO:0006950	Response to stress	5.02e-05
2	GO:0006477	Protein sulfation	9.74e-05
3	GO:0006955	Immune response	1.62e-04
4	GO:0002376	Immune system process	2.10e-04
5	GO:0006952	Defense response	2.13e-04
6	GO:0032672	Regulation of interleukin-3 production	2.96e-04
7	GO:0042223	Interleukin-3 biosynthetic process	2.96e-04
8	GO:0045399	Regulation of interleukin-3 biosynthetic process	2.96e-04
9	GO:0045401	Positive regulation of interleukin-3 biosynthetic process	2.96e-04
10	GO:0033008	Positive regulation of mast cell activation involved in immune response	3.94e-04

Table B.5: List of Top 10 significant enriched GO terms of BP for differential methylated probes that also found enriched in the GO analysis for probe set of DC subnetwork. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0005839	Proteasome core complex	2.96e-03
2	GO:0072357	PTW/PP1 phosphatase complex	3.84e-03
3	GO:0019773	Proteasome core complex, alpha-subunit complex	5.32e-03
4	GO:0005576	Extracellular region	7.29e-03
5	GO:0044464	Cell part	1.24e-02
6	GO:0005623	Cell	1.25e-02
7	GO:0044459	Plasma membrane part	1.47e-02
8	GO:0044297	Cell body	1.59e-02
9	GO:0005586	Collagen type III	1.64e-02
10	GO:0005887	Integral to plasma membrane	1.98e-02

Table B.6: List of Top 10 significant enriched GO terms of CC for differential methylated probes that also found enriched in GO analysis for probe set of DC subnetwork. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0030331	Estrogen receptor binding	7.12e-05
2	GO:0035258	Steroid hormone receptor binding	5.16e-04
3	GO:0019763	Immunoglobulin receptor activity	8.79e004
4	GO:0019767	IgE receptor activity	8.7e-04
5	GO:0004867	Serine-type endopeptidase inhibitor activity	1.71e-03
6	GO:0001104	RNA polymerase II transcription cofactor activity	2.51e-03
7	GO:0001517	N-acetylglucosamine 6-O-sulfotransferase activity	2.86e-03
8	GO:0005345	Purine nucleobase transmembrane transporter activity	2.86e-03
9	GO:0015347	Sodium-independent organic anion transmembrane transporter activity	2.86e-03
10	GO:0019863	IgE binding	2.86e-03

Table B.7: List of Top 10 significant enriched GO terms of MF for differential methylated probes that also found enriched in GO analysis for probe set of DC subnetwork. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

B.2.3 Top 5 enriched GO terms and pathways for community $\mathcal{C}_1$

	GO ID	GO:BP TERM	P-value
1	GO:0048731	System development	2.13e-07
2	GO:0048856	Anatomical structure development	1.602e-06
3	GO:0007275	Multicellular organismal development	2.73e-06
4	GO:0032501	Multicellular organismal process	9.59e-06
5	GO:0007399	Nervous system development	1.47-05

Table B.8: List of Top 5 significant enriched GO terms of BP for community $\mathcal{C}_1$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0000228	Nuclear chromosome	3.94e-06
2	GO:0044454	Nuclear chromosome part	2.73e-05
3	GO:0000790	Nuclear chromatin	3.04e-05
4	GO:0005694	Chromosome	1.30e-04
5	GO:0044424	Intracellular part	1.40e-04

Table B.9: List of Top 5 significant enriched GO terms of CC for community $\mathcal{C}_1$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0005515	Protein binding	1.05e-05
2	GO:0004708	MAP kinase kinase activity	2.52e-04
3	GO:0001012	RNA polymerase II regulatory region DNA binding	6.23e-04
4	GO:0044212	Transcription regulatory region DNA binding	1.00e-03
5	GO:0000978	RNA polymerase II core promoter proximal region DNA binding	1.05e-03

Table B.10: List of Top 5 significant enriched GO terms of MF for community $\mathcal{C}_1$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	KEGG ID	KEGG pathway	P-value
1	04964	Proximal tubule bicarbonate reclamation	2.72e-03
2	05218	Melanoma	1.14e-02
3	04010	MAPK signaling pathway	1.19e-02
4	05219	Bladder cancer	1.69e-02
5	04012	ErbB signaling pathway	2.15e-02

Table B.11: List of Top 5 significant enriched KEGG terms of community $\mathcal{C}_1$. KEGG terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

B.2.4 Top 5 enriched GO terms and pathways for community $\mathcal{C}_2$

	GO ID	GO:BP TERM	P-value
1	GO:0007165	Signal transduction	1.12e-04
2	GO:0051716	Cellular response to stimulus	1.93e-04
3	GO:0023052	Signaling	4.43e-04
4	GO:0007154	Cell communication	5.89e-04
5	GO:0007166	Cell surface receptor signaling pathway	2.45e-03

Table B.12: List of Top 5 significant enriched GO terms of BP for community $\mathcal{C}_2$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0032584	Growth cone membrane	6.35e-03
2	GO:0000177	Cytoplasmic exosome (RNase complex)	7.94e-03
3	GO:0031527	Filopodium membrane	1.74e-02
4	GO:0009897	External side of plasma membrane	2.67e-02
5	GO:0000178	Exosome (RNase complex)	2.67e-02

Table B.13: List of Top 5 significant enriched GO terms of CC for community $\mathcal{C}_2$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0000257	Nitrilase activity	1.74e-03
2	GO:0004997	Thyrotropin-releasing hormone receptor activity	1.74e-03
3	GO:0016815	Hydrolase activity, acting on carbon-nitrogen bonds,	1.74e-03
4	GO:0004937	Alpha1-adrenergic receptor activity	5.22e-03
5	GO:0044323	Retinoic acid-responsive element binding	5.22e-03

Table B.14: List of Top 5 significant enriched GO terms of MF for community $\mathcal{C}_2$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	KEGG ID	KEGG pathway	P-value
1	04020	Calcium signaling pathway	3.85e-03
2	04970	Salivary secretion	1.14e-02
3	04080	Neuroactive ligand-receptor interaction	1.33e-02
4	04270	Vascular smooth muscle contraction	1.78e-02
5	04062	Chemokine signaling pathway	4.79e-02

Table B.15: List of Top 5 significant enriched KEGG terms of community $\mathcal{C}_2$. KEGG terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

B.2.5 Top 5 enriched GO terms and pathways for community $\mathcal{C}_3$

	GO ID	GO:BP TERM	P-value
1	GO:0015712	Hexose phosphate transport	3.47e-05
2	GO:0015748	Organophosphate ester transport	3.47e-05
3	GO:0015760	Glucose-6-phosphate transport	3.47e-05
4	GO:1901264	Carbohydrate derivative transport	2.42e-03
5	GO:0010649	Regulation of cell communication by electrical coupling	2.45e-03

Table B.16: List of Top 5 significant enriched GO terms of BP for community $\mathcal{C}_3$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0048471	Perinuclear region of cytoplasm	2.26e-03
2	GO:0044444	Cytoplasmic part	7.93e-03
3	GO:0031088	Platelet dense granule membrane	1.19e-02
4	GO:0005923	Tight junction	1.41e-02
5	GO:0070160	Occluding junction	1.41e-02

Table B.17: List of Top 5 significant enriched GO terms of CC for community $\mathcal{C}_3$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0005515	Protein binding	2.27e-03
2	GO:0004925	Prolactin receptor activity	2.27e-03
3	GO:0008127	Quercetin 2,3-dioxygenase activity	2.27e-03
4	GO:0015119	Hexose phosphate transmembrane transporter activity	2.27e-03
5	GO:0015152	Glucose-6-phosphate transmembrane transporter activity	2.27e-03

Table B.18: List of Top 5 significant enriched GO terms of MF for community $\mathcal{C}_3$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	KEGG ID	KEGG pathway	P-value
1	04662	B cell receptor signaling pathway	2.08e-02
2	00062	Fatty acid elongation	2.15e-02
3	04142	Lysosome	4.81e-02

Table B.19: List of 3 significant enriched KEGG terms of community $\mathcal{C}_3$. KEGG terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

B.2.6 Top 5 enriched GO terms and pathways for community $\mathcal{C}_4$

	GO ID	GO:BP TERM	P-value
1	GO:0009372	Quorum sensing	8.76e-04
2	GO:0032470	Elevation of endoplasmic reticulum calcium ion concentration	8.76e-04
3	GO:0048874	Homeostasis of number of cells in a free-living population	8.76e-04
4	GO:0052097	Interspecies quorum sensing	8.76e-04
5	GO:0052106	Quorum sensing involved in interaction with host	8.76e-04

Table B.20: List of Top 5 significant enriched GO terms of BP for community $\mathcal{C}_4$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0031673	H zone	7.17e-04
2	GO:0031095	Platelet dense tubular network membrane	5.72e-03
3	GO:0031094	Platelet dense tubular network	7.14e-03
4	GO:0031672	A band	1.21e-02
5	GO:0015629	Actin cytoskeleton	1.77e-02

Table B.21: List of Top 5 significant enriched GO terms of CC for community $\mathcal{C}_4$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0022804	Active transmembrane transporter activity	8.7e-04
2	GO:0004909	Interleukin-1, Type I, activating receptor activity	1.39e-03
3	GO:0015382	Sodium:sulfate symporter activity	1.39e-03
4	GO:0022890	Inorganic cation transmembrane transporter activity	1.85e-03
5	GO:0015226	Carnitine transporter activity	2.09e-03

Table B.22: List of Top 5 significant enriched GO terms of MF for community $\mathcal{C}_4$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	KEGG ID	KEGG pathway	P-value
1	004640	Hematopoietic cell lineage	4.94e-02

Table B.23: Significant enriched KEGG term of community $\mathcal{C}_4$.

B.2.7 Top 5 enriched GO terms and pathways for community C_5

	GO ID	GO:BP TERM	P-value
1	GO:0032672	Regulation of interleukin-3 production	1.02e-04
2	GO:0042223	Interleukin-3 biosynthetic process	1.02e-04
3	GO:0045399	Regulation of interleukin-3 biosynthetic process	1.02e-04
4	GO:0045401	Positive regulation of interleukin-3 biosynthetic process	1.02e-04
5	GO:0032632	Interleukin-3 production	3.05e-04

Table B.24: List of Top 5 significant enriched GO terms of BP for community C_5 . GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0072357	PTW/PP1 phosphatase complex	1.32e-03
2	GO:0043233	Organelle lumen	3.34e-03
3	GO:0031974	Membrane-enclosed lumen	4.45e-03
4	GO:0005887	Integral to plasma membrane	7.26e-03
5	GO:0044459	Plasma membrane part	9.23e-03

Table B.25: List of Top 5 significant enriched GO terms of CC for community C_5 . GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0030331	Estrogen receptor binding	1.21e-04
2	GO:0035258	Steroid hormone receptor binding	3.01e-04
3	GO:0005345	Purine nucleobase transmembrane transporter activity	9.76e-04
4	GO:0015205	Nucleobase transmembrane transporter activity	2.02e-03
5	GO:0016504	Peptidase activator activity	2.96e-03

Table B.26: List of Top 5 significant enriched GO terms of MF for community $\mathcal{C}_5$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	KEGG ID	KEGG pathway	P-value
1	04710	Calcium signaling pathway	2.26e-02
2	00785	Salivary secretion	2.34e-02
3	04950	Neuroactive ligand-receptor interaction	3.19e-02
4	05016	Vascular smooth muscle contraction	3.57e-02

Table B.27: List of 4 significant enriched KEGG terms of community $\mathcal{C}_5$. KEGG terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

B.2.8 Top 5 enriched GO terms and pathways for community $\mathcal{C}_6$

	GO ID	GO:BP TERM	P-value
1	GO:0006955	Immune response	3.30e-04
2	GO:0006477	Protein sulfation	3.50e-04
3	GO:0007187	G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	5.40e-04
4	GO:0006040	Amino sugar metabolic process	7.72e-04
5	GO:0043252	Sodium-independent organic anion transport	8.32e-04

Table B.28: List of Top 5 significant enriched GO terms of BP for community $\mathcal{C}_6$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:CC TERM	P-value
1	GO:0031228	Intrinsic to Golgi membrane	1.03e-03
2	GO:0005730	Nucleolus	7.17e-03
3	GO:0031300	Intrinsic to organelle membrane	8.92e-03
4	GO:0005589	Collagen type VI	1.36e-02
5	GO:0030688	Preribosome, small subunit precursor	1.36e-02

Table B.29: List of Top 5 significant enriched GO terms of CC for community $\mathcal{C}_6$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	GO ID	GO:MF TERM	P-value
1	GO:0001517	N-acetylglucosamine 6-O-sulfotransferase activity	2.24e-04
2	GO:0015347	Sodium-independent organic anion transmembrane transporter activity	2.24e-04
3	GO:0008514	Organic anion transmembrane transporter activity	1.21e-03
4	GO:0005344	Oxygen transporter activity	1.44e-03
5	GO:0031593	Polyubiquitin binding	4.47e-03

Table B.30: List of Top 5 significant enriched GO terms of MF for community $\mathcal{C}_6$. GO-terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.

	KEGG ID	KEGG pathway	P-value
1	00533	Glycosaminoglycan biosynthesis	2.09e-03
2	05144	Malaria	2.29e-03
3	00130	Ubiquinone and other terpenoid-quinone biosynthesis	2.66e-02
4	00430	Metabolism; Metabolism of other amino acids	4.74e-02

Table B.31: List of 4 significant enriched KEGG terms of community $\mathcal{C}_6$. KEGG terms are ranked based on their corresponding p-values which evaluated by hypergeometric test.