Path-based sampling and community detection methods for biological and other applications

Andrew Elliott
St Peter's College
University of Oxford

A thesis submitted for the
Doctor of Philosophy
Michaelmas 2016

Acknowledgements

There are so many people to thank for making this possible. First and foremost I would like to thank my supervisors Felix Reed Tsochas, Gesine Reinert, Elizabeth Leicht and Alan Whitmore, without their guidance, ideas, huge amounts of time and their commitment to their students I doubt this would exist. I also have to especially thank Gesine Reinert and Felix Reed Tsochas for going through this document several times, using up a lot of Oxford's supply of red ink.

I would like to thank my family for their support and help during this process.

I would also like to thank all the members of the CABDyN research group either still here or off to pastures new for companionship, discussions and day to day help. I would like to thank Harriet A. Keane for very helpful conversations on Parkinson's Disease.

I would also like to particularly thank Griffith Rees for many helpful discussions and help on a wide range of things.

Finally, I would like to thank Marta Sarzynska for putting up with me while I write this, for proofreading this document and helping me in more ways than I can express.

Funding:

I acknowledge e-Therapeutics plc and the UK's Engineering and Physical Sciences Research Council (EPSRC) for funding, via a studentship at the Systems Approaches to Biomedical Science Industrial Doctorate Centre at the University of Oxford.

Contents

List of Figures

Glossary

Benchmark: An artificially constructed problem which is believed to be related to a real world problem. A benchmark is used to test the performance of algorithms to solve the given problem. It is especially useful when a real world example of said problem with known properties is not available.

BioGrid: A database of Protein Protein Interactions (PPIs).

Bounded: A function f is bounded if there is a finite b such that $\forall x \ |f(x)| < b$.

Clustering: An ambiguous term, usually referring to the clustering coefficient - a network statistic which is related to the number of triangles in the network compared with the number of possible triangles. The clustering coefficient can be measured network wide or individually for each node. The term can also refer to performing community detection on a network, although this thesis does not use it for this purpose.

Community: A group of nodes which are more densely connected than we would expect at random. See also: partition and community detection.

Clique: A set of nodes which are completely connected, so that there is an edge between every pair of nodes in the set.

Community detection: A blanket term for a family of methods which attempt to take a network and find groups of nodes that are believed to more densely connected than we would expect at random. There are many different approaches to this particular problem, see for example Ref. [57] for a comprehensive review.

Community detection method: An algorithm usually performed by a computer which finds groups of nodes (communities) that are more densely connected than expected at random.

Community structure: A division of a network into groups of nodes that are more strongly connected to each other than would be expected at random.

Configuration model: A null model for network construction in which the edges are randomised while preserving the degree of each node. This can also be thought of as breaking each edge, leaving a series of stubs attached to each node, and then randomly connecting the stubs.

Degree: The number of edges/interactions in which a given node is involved.

Degree distribution: The frequency of each degree in a network.

Density: The number of edges divided by the number of possible edges.

Edge: One of two fundamental parts of a network. A network consists of objects and the dyadic relationships between them, and an edge is the mathematical description of the relationship between two nodes. An edge can be binary i.e. either present or not, or it can be weighted, representing the strength of the association. We can represent a binary edge as a pair of nodes for example (v_1, v_2).

Eigenvalue: A value λ defined with respect to a given matrix M such that $Mv = \lambda v$ for some vector v; this vector is then an eigenvector of M.

Eigenvector: A vector v defined with respect to a given matrix M such that $Mv = \lambda v$ for some constant λ; then λ is an eigenvalue of M.

Ensemble: A possibly infinite set of objects which match a certain set of conditions. An example would be the set of all graphs on n nodes with m edges.

Erdős Rényi Graph (ER Graph): A random graph model with n nodes, where each edge is a Bernoulli random variable with probability p.

Gene: A section of DNA which can be used to construct a functional molecule e.g. a protein or a functional RNA.

Graph: See Network

Laplacian: In the context of this thesis, a blanket term for a number of different matrices which are related to the adjacency matrix. In each case, properties of each Laplacian relate to features of the underlying network, for example graph cuts.

Modularity: A measure of the quality of a community structure (see Community detection). A common method of community detection involves maximising this quantity.

Monte Carlo: A computational approach which uses randomness to approximate or estimate the results of problems which are usually analytically or computationally intractable.

Network: A mathematical object consisting of a collection of objects, known as nodes and relationships between these objects, known as edges.

Newman Girvan: The most commonly used null model for community detection. The underlying model can either be interpreted as a configuration model (see Configuration model) or can be seen as the Chung Lu model (a network model

where each edge is a Bernoulli random variable with different probabilities such that the expected degree is preserved). In either case the functional form is given by $\frac{k_i k_j}{2m}$.

Node: One of two fundamental parts of a network. A network consists of objects and the dyadic relationships between them, and the objects are known as nodes.

Null model: Typically a mathematical construction describing what we would expect the system to look like if the effect we are interested in does not exist. For example in community detection, a standard way of assessing if a pair of nodes are more strongly connected than would be expected by chance is to compare it with the number of edges that would be expected under an appropriate null model.

Number of edges: A network statistic, the count of unique edges in a network/graph.

Number of nodes: A network statistic, the count of unique nodes in a network/graph.

OMIM: An online database (Online Mendelian Inheritance in Man), which stores genes that are believed to be involved in various diseases.

PPI: A Protein Protein Interaction, two (or more) proteins which physically interact.

PPI networks: A network consisting of nodes which represent proteins and edges that represent Protein Protein Interactions (PPIs).

Partition: The division of a set of objects into one or more groups. Usually used in conjunction with community detection, to represent the resultant non-overlapping community structure as a partition of the nodes.

Path: An ordered set of adjacent edges, such that if the edges in question are $(v_{1,1}, v_{1,2}), (v_{2,1}, v_{2,2}), ..., (v_{n,1}, v_{n,2})$ then $v_{i,2} = v_{(i+1),1}$ for $i = 1, .., n-1$.

Path Sampling: A procedure to construct a sampled network from a wider network using paths between seed nodes. In this thesis, refers to Path$\leq k$ sampling, where sampled network is constructed using a set of seeds and a parameter k. The new network is formed from the edges involved in paths between seed nodes that are of length less than or equal to k.

Protein: Proteins are the building blocks of cells. They are chains of amino acids which form complex 3-D structures to help fulfil many (but not all) of a cell's functions. The ordering of the amino acids for each protein is represented by a gene. While every protein comes from one gene, it is possible that one gene can produce several proteins.

Random variable: A variable which takes different values with a probability given by an underlying probability distribution. For example, $X \sim B(1, 0.5)$ is 0 or 1 with probability 0.5.

Redundant seed nodes: A node on a seed list for a given sampling technique is a redundant seed node if its removal does not change the resultant sampled network.

Resolution limit: A concept in community detection referring to the fact that modularity maximising community detection methods may fail to detect communities that are below a certain size. This can also affect other community detection methods.

Resolution parameter: A parameter in modularity based community detection which modifies the size of communities detected (the resolution).

Sampling: Extracting a subset of a collection of objects. Usually in this document sampling refers to network sampling, where we select a subset of the nodes and a subset of the edges and construct a new graph using these subsets.

Sampled network: A network which is by construction is a subnetwork (i.e. a subset of the nodes and edges) of a larger network. Examples include networks constructed from incomplete data and networks constructed from a known network by a sampling technique.

Sampling technique: A method of taking a network and (optionally) additional information and producing a subnetwork. See Snowball sampling, Path sampling and Shortest Path sampling.

Seed/Seed list/Seed nodes: A node (or set thereof in the case of a seed list) which is used as a parameter in many sampling methods. They are often believed to be related to the process of interest, for example individuals who are drug users, or proteins believed to be involved in a disease.

Snowball sampling: A network sampling method which takes a set of nodes and a number of hops, and constructs a new network including all nodes which are at a distance less than or equal to the number of hops away from at least one node in the seed list. There is an implementation choice over what we term in this document cross-edges. These are edges which connect nodes which meet the distance criteria but are not used as part of the sampling. In our implementation we include all edges between the nodes, however other implementations only include edges which are part of a path which is of distance less than or equal to the number of hops.

Stochastic block model: A generalisation of an ER graph in which each of the n nodes is assigned to a group from $1, ..., g$ (either randomly or with some other scheme). Each edge is a Bernoulli random variable with probability a function of the group of each node.

Shortest Path Sampling: A procedure to construct a sampled network from a wider network and a set of seeds. A new network is formed from the edges

involved in all shortest paths between nodes in the seed list in the wider network.

Subnetwork: A network that is formed by taking a subset of the nodes and edges of a given network. In order for the network to be well defined the nodes present in each edge must be present in the set of nodes in the subnetwork.

Triangle: A collection of 3 nodes which are all connected. (Also known as a 3-clique).

Uniform null model: A null model for community detection, in which the underlying null model is a ER graph with either an estimated or parameterised probability. This model is called the constant Potts model.

VOI: (Variation of Information). A measure of the distance between two classifications. It is defined as the sum of the entropy of each of the classifications minus twice the mutual information between the sequences. A high VOI indicates a large difference between the classifications and a low VOI indicates similar classifications.

Chapter 1

Introduction

The roots of this thesis lie in a biological application, namely the use of biological network data to help target drugs. Over the last 15 years the amount of widely accessible biological network data has increased substantially, encouraging the development of new methods that can take advantage of the biological network data now available. Much of this new data is network related, from *nanometre scale* (for example interactions between proteins [161]), through *millimetre scale* (for example interactions between brain regions [16, 28]), to the *metre scale* (for example contact networks for the study of disease [162]).

A universal set of methods is unlikely to provide the most effective tools for network analysis, and indeed there are many different methods that pertain to different domains and are very data-specific. For example, the experimental methods used to collect data often do not translate between fields, and statistical tests in one area may not be sufficient or appropriate in another.

Despite the need to be cautious about the universal applicability of network tools developed with a particular use case in mind, many methods from the field of network science have in fact successfully been applied to multiple domains. Network science is concerned with the development of tools to analyse network data often with some

expectation that network tools applied in specific domains will also prove informative in other contexts. For example, community detection methods (which divide a network into comparatively dense groups) have been applied to protein structure [149]; protein-protein interaction networks [108]; metabolic networks [75], the airline network [76, 143]; social networks [170] and financial networks [17, 115] and networks related to disease patterns [147].

In fact, from the standpoint of network science biological network data is not very different from network data obtained in other disciplines. The power of network science is that network methods that are first developed to develop important nodes in one field, for instance sociology, can then be applied other fields, such as biology, because fundamentally the language is the same.

No thesis could sensibly aim to cover the full breadth of methods covered by network science. Hence it is necessary to narrow the scope of this thesis to a few areas under the more general heading. One key issue which recurs through much of the thesis is the need to develop a better understanding of the effect of sampling on network analysis. Questions about sampling are also related to mesoscopic structure of in networks and the detection of communities. Community detection methods again feature significantly across multiple chapters of this thesis. Hence, this thesis focuses on methods and approaches related to three related areas: (1) null models for samples from networks; (2) community detection methods and sampled networks, (3) the application of network models to biological problems.

Indeed, while the methods constructed and analysed in this thesis seek to make more general methodological contributions, the starting point for this thesis is a particular area of biology, namely network pharmacology. Network pharmacology seeks to use biological data, for example protein-protein interaction networks, to select multiple proteins that can be targeted for intervention, rather than the traditional approach of identifying a single protein suitable for intervention [13, 84, 193, 194].

Therefore, we explore and develop methods for use by the general network science community but with the needs of network pharmacology in mind. In fact, the common theme across most methods considered here are, sampling and looking at subgroups of nodes and edges. For example, Chapter 3 explores selecting between methods for constructing sampled networks related to a disease, and in Chapter 4 we look at performing community detection on those sampled networks.

Here is a brief introduction to the different ingredients that are relevant to this thesis.

1.1 A network science perspective

A network (or a graph) is a set of objects (nodes) and a list of connections between them (edges), where edges can either be binary (i.e. exist or not exist) or weighted (with a comparative strength assigned to each edge). The edges can also be directed, so that a connection from node i to node j does not imply a connection from node j to node i. In this thesis, networks will mostly be undirected, unweighted networks.

Network science has been used to study many areas involving social, biological and engineered systems, for example friendship networks [158, 170], food webs [51] and the organisation of the Internet [54].

While the study of networks and network ideas can be traced back many years, a key moment in their development is the work on graph theory by Paul Erdős and collaborators developing and exploring properties of the Erdős Rényi network used in this and many other works on network science especially on random graphs. Many fields and disciplines have contributed to Network science, from sociologists to computer scientists, physicists, mathematicians and statisticians.

The following categories illustrate some of the different approaches that network methods have taken:

1. Network models - General or domain specific models of networks. These models can be analysed for their own sake [34, 35] or as a null model for real world data [126]. They can broadly be divided into two main types: general probablisitic ensembles of networks, e.g. Erdős Rényi networks, and more domain-specific models. There is considerable overlap between these areas.

2. Network global statistics - Summary statistics that give a sense of the overall structure of the network, for example density (the number of edges divided the number of possible edges) or average degree (the average number of edges of a node).

3. Network local level statistics - Statistics designed to measure node and edges properties, for example node and edge betweenness (the number of shortest paths going through a node or edge).

4. Prediction methods - Methods used to predict features based on observed data, for example node attributes or missing edges given the network structure.

5. Subgraph methods - Methods that are based on identifying, counting and potentially allocating a function to small, well-defined groups of nodes, for example a triangle (a set of three nodes where every node is connected to every other node).

6. Methods to look for larger structure in networks - Methods that look for structure at a larger level than subgraphs but at a smaller level than the whole network, the (so-called) mesoscale. The focus is to identify groups of nodes which have interesting properties, where 'interesting' depends on the research question.

In this thesis, many of the methods belong to the final class of methods, with a focus on sampling and community detection. Thus in the next section we will briefly

review these areas.

1.2 Sampling network methods and community detection

The methods we discuss in all of the substantive chapters of this thesis can be broadly divided into methods related to sampled networks and methods related to community detection.

Our application of these methods mainly focuses on methods based on paths between pairs of nodes.

Community detection

The number of methods to perform community detection has exploded in the last 15 years. While it is difficult to identify when the field started with highly related problems in different discipline (for example cut partitioning [153, 180] or the study of block models [83]), interest in the field vastly increased due to some influential papers in the early 2000s see for example Refs. [70, 126, 132].

There are now many community detection methods in the literature; for a good review see Ref. [57]. Two reasons for such a abundance of methods are as follows. There are many different methods to detect communities for two reasons. First, this is a computationally intensive problem and thus algorithms trade quality of partition for run time. Second, certain community detection methods prioritise different things.

The methods we will focus on in this thesis are the so-called modularity-based methods, which seek to maximise the number of connections within a group of nodes above that expected by chance. This framework is very flexible because we can use specifically tailored null models for different applications [53, 115, 147].

Sampled Networks

Network sampling is a set of procedures by which a subnetwork is extracted from a large network. Sampled networks exist in a wide variety of application domains for very different reasons, and thus with very different consequences.

One famous use of network sampling is sampling hard-to-reach or hidden populations in sociology [55, 61, 62]. A hard-to-reach population consists of individuals with an incentive to avoid detection (e.g. injecting drug users), where those individuals are nevertheless embedded in a social network. In this case, it is very difficult to observe the overall network, but we may be interested in understanding either the network structure or individual attributes of the people involved. Here, sampling schemes such as snowball sampling [55] are used, which traverse the network using referrals to gain a sample of the underlying network. This set of methods is both concerned with extracting a sample, and also correcting for the fact that the sample is not a simple random sample [61].

The sampling work done in this thesis is designed to tackle a related but different problem. In network pharmacology, we are interested in analysing a network related to a disease. Thus, we have a global (albeit incomplete) picture of the biological network, and we want to identify subnetworks related to the disease of interest. Thus, unlike in snowball sampling, we know the whole network, but we cannot be confident in the use of referrals as we cannot question biological molecules. The aim is to extract a small network which could then be studied in detail, a problem we will discuss in Chapter 3.

The underlying large network itself may also suffer from sampling bias. For example, in protein-protein interaction networks it is often not reported when there is experimentally evidence that an interaction does not exist (although there is some work to attempt to identify them [22]).

Finally, sampling in networks is also used for summarising networks (making a

small network that has the properties of a larger one), or for further study of the local neighbourhood of particular nodes (ego networks) [105], or to form an ensemble of subnetworks from one network in order to perform statistical analysis on the resultant distribution [3].

1.3 Network pharmacology and protein-protein interactions

Network pharmacology is based on the premise that biological network data can help to identify biological information that is relevant to healthcare. Fundamentally this is motivated by the idea that biological systems and processes are complex and may be resistant to single interventions [84, 193], where this robustness could be a consequence of evolution [118]. Thus, multiple interventions may have to be made in a biological system in order to change the functional behaviour of the system as desired [84, 193]. Network pharmacology is informed by network data and network models [84, 193, 194]. Indeed many current drugs do not only modify one specific process as is illustrated by the widespread existance of pharmacological side effects.

In this thesis, we will explore protein-protein interaction (PPI) data from a network pharmacology standpoint. A protein-protein interaction is given by a pair of proteins that undergo physical molecular docking [43]. PPIs are identified by multiple different experimental methods, each with their own biases. Combining many of these interactions creates a network with proteins as nodes and interactions as edges, known as a PPI network. Despite many false positives and negatives in the experimental data, PPI data has been used to successfully gain insight into biological systems, such as predicting protein function [2, 106] and predicting undiscovered PPIs [2, 77].

Complete PPI networks are typically too large for in-depth analysis of disease

mechanics. Thus several studies have attempted to sub-sample PPI networks, in order to generate smaller networks that better reflect the disease or the cellular processes of interest [65, 86, 94, 112]. However, this procedure is not straightforward and there is no technique that can be considered the gold standard for subsampling PPI networks. Many of the methods to create subsampled networks of relevance to a particular disease (referred to as 'disease networks' in this thesis) require a list of proteins known to be involved in the disease (referred to as a 'seed list'). It is important to note that there is not consensus on the constituent components of disease networks, the inclusion or exclusion of a protein or family of proteins in the seed list is subjective.

1.4 Overview of the thesis.

The dissertation is structured as follows: Chapter 3 is a stand-alone paper detailing a method we propose for selecting between different subnetworks that are candidates for further study. It has a focus on networks related to Parkinson's disease, and it includes a review of the relevant biological literature. Chapter 4 is a stand-alone preprint that considers the appropriateness of performing community detection on sampled networks, and proposes and tests a solution to this problem. Chapter 5 explores the community detection methods we constructed in Chapter 4. We generalise this approach to work on non-sampled networks, and show that for a sparse Erdős Rényi graph with a fixed community structure, a statistic related to our generalised community detection approach is asymptotically normally distributed.

A chapter by chapter summary of the thesis follows.

Chapter 3 - Selecting a Sampling Method

The first contribution of this thesis is a systematic study of disease network construction methods through sub-sampling and seed lists. We compare 12 disease networks

related to Parkinson's disease, each with a different set of construction rules and/or based on a different seed list. We then construct a method to prioritise these networks for further study. This method involves computing the significance of a basket of network statistics using a novel null model which we derive. The novel null model is required, as we demonstrate that the commonly used configuration model is not suitable for measuring the significance of certain network statistics in sampled networks. We derive some novel analytic results to compute the significance of simple statistics, and we use a Monte Carlo approach for more complex statistics. We verify this approach on a benchmark network set which consists of networks sampled from a stochastic block model. We then prioritise networks that seem to be significantly different from random, under the assumption that the significant networks might prove to be interesting, leading us to focus on two networks out of the original 12.

Chapter 4 - Community Detection in Sampled Networks

Once we have a subnetwork, we can use community detection to try to identify small functional modules. However, when the subnetwork comes from a subsampling scheme, then artefacts induced by this sampling scheme may affect community detection. The second contribution detailed in this work is related to community detection in such sampled networks. Community detection is a technique that allows us to partition the nodes into groups that we believe to be better connected than we would expect at 'random'. Deviation from randomness is traditionally measured using the configuration model, which may not be appropriate for sampled networks. We focus on networks sampled using shortest path sampling, as community detection in these networks performs very badly. We attempt to construct a new null model to adjust for the features of sampled networks, using a Monte Carlo approach to estimate the expected number of connections between pairs of nodes. Testing this approach in a set of model networks we discover that this approach does not recover the structure.

We believe this failure to recover the structure occurs because the sampling technique induces greater than pairwise structure on the network.

Using this observation we construct Path Modularity, an extension of modularity which incorporates information about simple paths. We fit a null model which accounts for features of the sampling scheme. Testing this approach, we discover that it performs favourably in comparison to standard modularity approaches, our modified approach that does not consider paths, and stability (another community detection method that uses greater than pairwise information).

Chapter 5 - Path Based Community detection

In this chapter, we generalise the community detection method we developed in Chapter 4 to non-sampled networks. Following the work of Franke and Wolfe [63], we then relate our new community detection method to a normal distribution with a view to measuring the statistical significance of our version of modularity with a fixed community structure.

We use Stein's method, an approach to bound the difference between two probability distributions. Stein's method can be especially powerful when we have variables that are not independent of each other which applies in this case due to the overlapping edges between paths in our statistic.

The statistic we focus on is the change in path modularity between a community that contains all nodes and a fixed partition of the nodes.

We show that this statistic is asymptotically normally distributed for $n \to \infty$ with an explicit bound (in Wasserstein distance) on the distance to the normal distribution. The statistic is a weighted cut of a fixed partition over the ensemble of sparse ($p = \frac{c}{n}$) Erdős-Rényi graphs for paths of length l, where $2 \leq l < \frac{n}{2}$.

We also give a specialised bound for the $l = 2$ case which performs well. The more general bound for $l > 2$ performs less well but demonstrates the same bound on the

rate of convergence which is of order $\frac{1}{\sqrt{n}}$.

Chapter 2

Background

Symbol List

Symbol	Description
B_n	The nth bell number
C	a set of nodes belonging to a community
D	Diagonal matrix with the degree of each node on the diagonal i.e. $D_{ii} = k_i$
D_{max}	The size of the largest dependency neighbourhood
$F(d)$	Expected number of connections at distance d
F_W	Set of f_h for $h \in H_W$
H, H_1, H_2	Sets of functions
H_K	Set of functions related to the Kolmogorov distance
H_W	Set of functions related to the Wasserstein distance
H_{bW}	Set of functions related to the bounded Wasserstein distance
J	The supremum of |f"(x)| for a given function f
K_{X_i}	The dependency neighbourhood of X_i if X_i is independent of $\{X_j : j \notin K_i\}$.
N_i	Population of region i
P_{ij}^{SPA}	The expected number of edges under a spatial null model
$P_M(C, t)$	Probability that a random walker is in a set of nodes C representing a community at time 0 and at time t with Markov process M
P_{art}	A set of sets of nodes in each community.

P_{ij}	The expected number of edges between node i and node j
$R_M(t)$	The sum of the $P_M(C,t) - P_M(C,\infty)$ over all of the communities in the network.
$R_n(t), R_c(t)$	R_M using the Markov process described in the Combinatorial and Normalised Laplacians
V_{ik}	The remaining terms in the W_i not included in W_{ik} such that $W_i = W_{ik} + V_{ik}$
W	Random variable which is the sum of not necessarily independent random variables. Often it is required to have mean 0 and variance 1
$W^{(M)}$	Random walker under markov process M
$W_t^{(M)}$	Random variable representing the position of the random walker at time t
W_i	A summation of X_i which is independent of X_i
W_{ik}	A summation that is independent of X_i and X_k
X, Y	Random variables
X_i	A random variable.
Z	Standard normal distribution
Z_i	A summation of random variables that are not included in W_i, so that $W = W_i + Z_i$
ϵ	A symbol used to represent the derived bound on $\lvert E(Wf(W) - f'(W))\rvert/C$
λ	The resolution parameter in community detection
$\mathbf{p}$	Vector of probabilities of a random walker being at each node
$\mathfrak{F}_C$	The set of all functions with bounded first and second derivatives
$\mathfrak{F}_W$	Superset of F_W, which maintains the restriction on the absolute value of the function and its first and second derivative.
$\mathfrak{H}$	Set of all $h'(x) - xh(x)$ for all absolutely continuous h which also obeys the following inequality $E\lvert h'(Z)\rvert < \infty$
c	Vector of community assignment
c_{max}	The vector of community assignment which maximises the modularity
$d_H(u,v)$	A distance function between probability distributions which is a function of the set of functions H
d_{ij}	The physical distance in space between point i and point j
f	A generic function
f_h	For given h the solution to f(x)-xf'(x)=f(x)-E[h(Z)]
h	Generic Function

i	Index for number of nodes
k_i	Degree of node i
m	Number of edges in the network
n	Number of nodes in the network
u, v	Probability distributions
x, y	Dummy variable used in multiple proofs

2.1 Introduction

This chapter contains the background and literature review for the results chapters (Chapters 3, 4 and 5). As the problems studied in the results chapters of this thesis are very different some sections of the literature review will apply to multiple chapters, for example, the review of community detection in section 2.6 whereas others will only apply to one chapter, for example, the review of Stein's method in section 2.7.

It should be noted that as chapter 3 has been adapted from a paper which did not have this literature review. To preserve the flow of the chapter, some of the topics in the literature will be briefly revisited in this chapter 3.

2.2 Background on sampling methods for selecting subnetworks of interest in biological systems

Some studies of PPI networks analyse the whole interactome to attempt to draw conclusions about their processes of interest, for example lethality studies [86]. Other studies sample the network producing a subnetwork which is supposed to be more informative about the process in question than the original network (examples of this approach are discussed later in this section). An advantage of sampling a network is that standard network summary statistics, such as clustering coefficient or betweenness, may become very informative on a disease network, as they are only summarising

the interactions related to the disease, while they may not be as informative about the whole interactome. Further, on a small network an in-depth analysis, such as verifying existing links, may be feasible.

A drawback of using sampling techniques is that since the PPI interactome has already in effect been sampled based on the experimental method used, and evidence on different interactions not homogeneous. Sampling a network which has already been sampled could introduce many problems, especially if the sampling technique is based on the topology of the network, since this makes it very difficult to assess the effect of the initial sampling induced by experimental methods. Further, most sampling techniques will contain subjective choices (e.g. determining seed lists and parameters) which can bias the results.

There is no gold standard method for constructing a network for a process in a cell. Studies have in the past used one of the approaches below:

1. Pre-selected sets of proteins and experimentally verified all paths of interactions [112]
2. Tested interactions between a small subset of proteins believed to be very important to a disease and a larger set of proteins that might be related to the disease process [71].
3. Experimentally located proteins present in the same cellular compartment as the process of interest and added edges using a PPI database e.g. BioGRID [65]
4. Taken a network directly from a pathway database e.g. KEGG [86].
5. Used *in silico* techniques to leverage data in the public domain to approximate disease networks e.g. network sampling in Ref. [94]

Some circumstances favour some approaches over others. For example, the first technique of using a pre-selected set of proteins and testing all interactions may produce networks that are biologically relevant, if it is well understood which proteins are important in the disease process. If, on the other hand the processes underlying

the disease state are not fully understood (as is the case in Parkinson's disease), then using this technique will limit the scope of the analysis. The second technique, takes this issue even further by limiting the set of nodes and the set of possible interactions that can exist in the network.

The third approach is mostly data-driven, so the coverage of this technique is likely to be very good, and the topology of the network has not been interfered with by the researcher sampling the network. However, it relies on a PPI database to add the edges, which potentially introduces bias to the network, as the validation of edges will not be uniform and it is possible that some edges may not have been . Further, in some studies the process of interest is not restricted to one compartment, making this technique hard to apply widely.

The fourth technique of taking a network directly from a pathway database can be advantageous in a very similar way to the first two techniques. It allows a network analysis of current knowledge, and may be able to highlight potentially unexplored important bio-molecules in the process. However, it also suffers from the same limitations as the first two techniques. Since the topology of the network is entirely determined by the database, it is very difficult to find new important processes. Further, the interactions included in the pathway are subjective, calling the results into question.

In practice, due to prohibitively expensive experimental methods otherwise required, disease networks are often constructed following the fifth technique of using *in silico* methods extracting information from online databases of PPIs (approach 5). For example, a plug-in to the popular software package Cytoscape expands outwards from seed proteins using a procedure from the social sciences called snowball sampling [116, 151]. Social scientists have developed such techniques to attempt to sample hidden populations such as individuals infected with HIV or drug users [55]. Snowball sampling is implemented according to the following steps. First, start with

a list of nodes (called a seed list) known to have the characteristic of interest, and form a network. Second, add all nodes that are linked to the nodes from the previous step to the network. [61] Repeat the previous step until a condition met for example a fixed number of times [61]. The inclusion probability in this scheme is proportional to the eigenvector centrality of each node in the hidden network [125]. This technique allows the construction of a large network of nodes that are believed to belong to a particular group. However, unlike applications in the social sciences, when this technique is applied to PPI networks, there is no guarantee that interactions that are being traversed belong to the same group as the nodes. This potentially leads to a large amount of non-relevant information being added to the network.

Steiner trees have also been used to construct disease networks [187]. A Steiner tree is defined as follows: given a weighted network and a set of nodes of interest, a Steiner tree is a subnetwork which connects all of the nodes of interest while minimising the sum of the weights of the edges [187]. One possible use of edge weights is to reduce the number of hubs that are part of the sub-sampled network by assigning the directed edge weights to be equal to the degree of the node at the destination of each edge, thus increasing the cost of visiting any hub [187]. The advantage of this technique is that it includes less redundant information than snowball sampling, as only edges that link important proteins are included. This technique could be useful for identifying links between proteins of interest - the original paper identified modules in a colorectal cancer network [187]. However, it also excludes non-optimal paths between nodes of interest, which could result in the loss of important pathways, limiting usefulness for creating disease networks.

Snowball sampling could include too many nodes for the research question of interest, and Steiner trees could include too few. Another approach (later called 'all paths less than or equal to k' or "Path$\leq$k") implemented in the Genes2Network tools [19] attempts to include more relevant pathways, while not including too many

interactions that are not related to the process [19]. This approach creates a sub-network by sampling all paths between seed proteins that are of length less than or equal to a parameter k [19]. This means that it only includes paths between seed proteins; these are not limited to paths that are defined as optimal. For example when k=1 this technique samples in a similar way to technique one where we tested all interactions between a given set of proteins; this is also known as an induced network.

Another technique that has been used is shortest path sampling [94, 110]. In this technique important nodes are identified by constructing all shortest paths between all pairs of nodes on a seed list. Note that if there are two shortest paths then both paths are included. In the implementation used in this thesis only edges that are part of the paths are then included in the sampled network, however in Ref. [94] all edges between these proteins were included.

Shortest path sampling in either form is good if we think that the short paths in the network are relevant to the disease process, for example if we imagine this as a communication network. However, it can be misled either by the inclusion of large degree nodes [110] or by a disease with two distinct processes.

Another method is DIAMOnD which has been designed to identify disease modules in networks [68] and has been used to do so in Refs. [69, 152]. DIAMOnD takes a set of proteins and a network and computes a rank-ordering of nodes based on the significance (p-value) of the number of connections they share with disease proteins controlling for the degree [68]. Adding the top ranked protein to the list of disease nodes the significance is then recalculated with the additional disease node and the procedure is repeated [68]. A parameterised version is also available which allows the non-seed nodes to have lower weight than the original seed list [68]. This procedure in theory will continue until all nodes are present in the sample. To alleviate this they suggest selecting the size of the network by thresholding either by (a) estimating the

recall at different sample sizes by removing seed nodes or (b) using addition data to test enrichment of the added proteins [68].

Seed List Construction Many of the above sampling techniques require a list of nodes (known as a seed list) that are believed to be involved in the process. Similarly to network construction, there is no gold standard method to construct a seed list, nor is it understood what makes a good seed list. Methods include:

- Literature curation, see for example [71],
- Localisation studies, see for example [65],
- Pathway data, see for example [86],
- Genes experimentally linked to the disease, see for example [112].

Each method has its advantages and disadvantages. Data driven approaches (e.g. localisation studies and disease genes) can suffer from a lack of coverage, as it is often only a subset of proteins that are tested, and may miss out important processes that do not occur at the level of data collection. Curated seed lists, such as literature curation and pathway data can be very subjective. There is no guarantee that two researchers would produce the same seed list even if they are given the same research papers. In chapter 3 we will assess the effect of such choices compared to choosing proteins at random.

2.3 Background on network pharmacology

The broader context that motivates this work has been described as network pharmacology. Between 1995 and 2005 the dominant drug development strategy was single target approaches, i.e. selecting a molecule in a pathway and interfering with it [146]. This period also saw a steady decline in new drugs [146]. It has been theorised that the large amount of resources required for target based drug discovery and the small

number of targets might be to blame [84, 146]. Target-driven approaches to drug discovery rely on simplified linear or branched pathway models to identify important elements (such as proteins) that may serve as drug targets [134]. Drugs are designed to interfere exclusively with one stage in the pathway to modify the disease process [84]. However, this approach ignores the true complexity of biological systems: disease states are often a combination of many small defects [131, 148, 194].

An alternative approach is provided by the emerging field of network pharmacology, building from some of the ideas in systems biology. Network pharmacology believes that in many cases multiple targeted interactions are required to have an impact on biological networks [84]. This is supported by evidence from yeast single gene knockout studies, where in ideal conditions only 15% of gene knockouts are lethal and 34% of gene knockouts have some impact on fitness [82, 84]. Thus in ideal conditions 66% of the yeast genome appears to be robust to deletion. However, in non-ideal conditions when presented with environmental changes or exposure to molecules in a large and diverse small molecule library (4000 molecules including approved drugs and chemical probes) only 3% of genes display no impact on fitness. Many genes may be redundant to one set of conditions but very few are redundant to all possible conditions [82, 84]. It has been proposed that biological systems have evolved to be robust to small changes in order to promote greater adaptability to new conditions [118]. It is therefore highly plausible that in many cases single targeted interactions will not produce the large effects that are required of drugs, and that it might be the case that multiple interactions are required.

One of the many current challenges for drug discovery motivated by network pharmacology is in the selection of multiple target sets such that they maximise efficacy while minimising risk. One possible solution is to screen large libraries of molecules against their effect on the 'disease network' and select molecules that appear to perturb the network in the correct direction [148]. Another potential solution is

to attempt to select targets *in silico*, leveraging the large amount of data that has emerged from the field of systems biology in the last 10 years [38, 193]. For example, Yang et al. used a differential equations on a directed inflammation network to determine the candidate set of multiple intervention points in the network [191]. Hormozdiari *et al.* used a min-cut approach to identify multiple target interactions in bacterial networks [85]. Hwang et al. used a modified betweenness measure 'bridging centrality' to assess potential targets in a cardiac network [86]. In a slightly different approach Vazquez used heuristics and exact enumeration to compute cancer drug combinations which would be resistant against different strains [177].

A recent breakthrough was achieved by Keane *et al.* [94], using ideas from network pharmacology and network science to predict and subsequently experimentally verify the effect of multiple interventions on a Parkinson's Disease network and cell model. This work started with a literature curated seed list; in this case the seed list was clearly well selected. In chapter 3 we provide a statistical methodology to assess such seed lists.

2.4 Background on Parkinson's disease

Partly motivated by the work Keane *et al.* [94], Parkinson's disease is also the starting point for the research on sampled networks presented in chapter 3.

Parkinson's disease is a neurodegenerative disorder that manifests with tremors, rigidity, postural instability and bradykinesia (slow movement) [67]. Age is a strong risk factor with studies suggesting that 1 in 20 people between the ages of 75-79 have the disease [45]. Studies also seem to indicate that the neuronal damage that causes PD start 4-7 years before the onset of the disease [163]. The disease is not easy to characterise, with a study showing that 25% of supposed PD patients have an alternative diagnosis in post-mortem [42].

A study in Rotterdam suggests that there may be a gender bias towards males [44], which is supported by data from the National Ambulatory Medical Care Survey [173] (analysis provided by WolframAlpha [189]), although the evidence is not robust [45].

PD is characterised by the degeneration of dopamine producing neurons (also known as dopaminergic neurons) in an area of the brain known as the substantia nigra pars compacta (SNPC) [107] and the formation of intracellular Lewy bodies. Lewy bodies (LB) are intracellular inclusions containing aggregated proteins, mostly α-synuclein [128], and they are found in several areas of the brain including the SNPC [178]. Patients that have similar symptoms but lack LB are known as having Parkinsonism.

There is no cure for PD. The current drug treatments slow down the progression of the disease by attempting to correct the dopamine deficit caused by the dying cells [32].

Why Dopaminergic Neurons? PD appears to affect dopaminergic neurons (DPN) disproportionately [10, 178]. It has been proposed that cytoplasmic dopamine may undergo oxidation to form reactive oxygen species (ROS), thus predisposing such cells to oxidative damage [178]. In support of this it has been shown that an increase of the proportion of cytoplasmic dopamine relative to VMAT2 (a protein responsible for dopamine transport into vesicles) increases the level of cell death [178]. Another important source of ROS in DPN is the electron transport chain (ETC). The high metabolic demands of the DPNs result in higher production of ROS from mitochondria [178], either through increased activity in the electron transport chain (ETC) [114] or through increased levels of mitochondrial DNA damage (exacerbated by clonal expansion) in the SNPC, which have been shown to cause problems in the ETC [98].

2.4.0.1 Possible disease mechanisms

The etiology of PD is unknown. Loss of dopaminergic neurons in the SNPC occurs by caspase-dependent apoptosis [24] but the precise mechanisms triggering programmed cell death remain poorly characterised. Various processes are postulated to contribute including protein accumulation, mitochondrial dysfunction, kinase signalling and oxidative stress [107].

The contribution of genetic factors to PD is complex, poorly understood, and remains the focus of intense study. Klein *et al.* state that there are 28 genome regions that have been "more or less" convincingly related to PD, however only 6 of these contain genes in which mutations are known to cause PD [96]. Specific gene mutations are believed to account for 10% of cases of PD - so called familial (heritable) forms [192]. In sporadic forms of the disease (i.e. with no familial relation) these 6 genes only account for 3-5% of the risk of contracting the disease [96]. Genetic risk factors including mutations in α-synuclein, Tau, LRRK2, Parkin, PINK1 and DJ-1 can increase the probability of developing PD substantially and contribute to the remaining 90% [10, 156].

Importance of α-synuclein The protein α-synuclein is a major constituent of LB and there is evidence linking PD with mutations [156] and triplication [157] of the α-synuclein gene. This protein is also linked to PD through studies in: transgenic nematodes and mice, overexpression in yeast and in neurotoxin based animal models [26]. Consistent with this, some studies have shown increased levels of α-synuclein mRNA in the brains of PD sufferers, but the evidence is not conclusive [178]. There appears to be some link between the post-translational modifications of α-synuclein and PD, with several different modifications being linked to PD either through mechanistic arguments or data from PD patients [26].

The function of α-synuclein is not fully understood [10, 50]. It appears to regulate

the production of dopamine by inhibiting the production and function of an enzyme in the synthetic pathway of dopamine [178]. Another possible function of α-synuclein is to regulate synaptic vesicle formation (synaptic vesicles are enclosed intracellular bodies used for rapid excretion [1]) [10, 50, 178]. This mechanism proposes that in normal conditions α-synuclein helps to regulate the release of vesicles and the availability of vesicles in the reserve, releasable and recycling pools [178]. Lowered α-synuclein levels are believed to reduce the size of the reserve pool of vesicles and increase the release of vesicles; whereas raised levels of α-synuclein are believed to reduce the number of vesicles in the recycling pool and possibly inhibit vesicle release [178].

Further, α-synuclein appears to be toxic in some cases. While the mechanism is not fully understood, it is believed that several α-synuclein monomers interact to form cytotoxic oligomers [50, 178]. The soluble oligomers can combine to form insoluble aggregates, which may be cytoprotective [178].

Importance of autophagy Autophagy is the process by which cells remove spent cellular components including cytoplasmic organelles, long lived proteins and protein aggregates [32]. Autophagy occurs in cells in three known ways: macroautophagy, microautophagy and chaperone mediated autophagy (CMA) [32, 192]. The link between autophagy and PD is supported by clinical evidence of dysregulation of autophagy [32], and through genetic studies that link autophagy related genes to PD [107].

It appears that autophagy is important in α-synuclein toxicity, with two of the PD linked mutations in α-synuclein resulting in a protein that cannot be destroyed by CMA [32]. Soluble α-synuclein is removed by the proteasome and CMA, whereas insoluble α-synuclein can be removed by macroautophagy [107]. Removing either autophagic pathway results in the accumulation of α-synuclein [32, 179].

2.4.0.2 Parkinson's disease models

A biological model in this context is a biological system which exhibits similar enough properties to the biological system of interest that it is useful to study said system to learn something about it. This is relevant to this chapter 3, as we use data from a disease model to construct one of our seed lists.

For example one may be interested in drug dosage in humans. Conducting this experiment with human beings directly would not be ethical, so companies that want to introduce drugs to the market perform the initial experiments on animals, and only after showing that it is safe perform drug trials on humans [175]. Thus the animals in this case are used as a biological model for a human being. In the case of PD scientists need to construct systems which have many of the same features of PD, but in a cell culture or animal. In the following section we will discuss some of the methods that are used.

MPP+ model The disease model we use to inform our proteins of interest, is based on a neurotoxin (MPP+) which produces PD-like symptoms in humans and animals. In 1982 four individuals who had taken synthetic narcotics with a large impurity of a substance known as MPTP, developed symptoms of persistent Parkinsonism [103]. The subsequent study determined that MPTP (unlike MPP+) was able to pass through the blood brain barrier where it was converted into MPP+ by glial monoamine oxidase B [150, 155]. MPP+ was identified as the toxic agent responsible for the observed symptoms and signs [155]. Due to MPTP's ability to pass through the blood brain barrier MPTP/MPP+ can be easily used to produce both *in vivo* and *in vitro* models of Parkinson's disease.

MPP+ is taken up by dopamine transporters and preferentially accumulates in the mitochondria of dopaminergic neurons [21, 155]. MPP+ then inhibits Complex I of the electron transport chain (ETC) [155]. The primary function of the ETC is

to produce ATP (the energy currency of the cell) [1], thus inhibiting the ETC results in impaired ATP production. Further, the ETC is known to be a major contributor of reactive oxygen species (ROS) which can cause oxidative stress [114, 182], and the presence of MPP+ in mitochondria has been shown to lead to increased ROS production [41]. Lowered ATP production may also increase oxidative stress by reducing the cells' ability to maintain dopamine homeostasis resulting in dopamine leakage into the cytosol [41]. In conclusion, MPP+ treatment to cells results in lowered ATP production and increased levels of ROS, and therefore damage to dopaminergic neurons.

MPP+ models show several similar features to PD. Neurons display increased levels of α-synuclein [41] and in animal models manifest cellular inclusions. There is evidence suggesting that α-synuclein may potentiate MPP+ toxicity [107]. Mice lacking α-synuclein show no response to MPP+ [60, 107], however neurons lacking α-synuclein do show response to Rotenone (another complex I inhibitor), indicating that α-synuclein appears to be related to the toxicity of MPP+ [107].

The model has some possible shortcomings. Bio-molecules can be promiscuous, thus MPP+ may affect other processes. In a human cell line MPP+ activates the ERK/MAPK pathway [73] which transduces surface signals involved in cell-cell communication into gene expression responses particularly in relation to proliferation, differentiation and survival [1]. Inhibiting this pathway in MPP+ treated SH-SY5Y human cells resulted in lower cell mortality in MPP+, suggesting that ERK/MAPK signalling may be involved in MPP+ related cell death [73].

Other Possible Models There are two main other *in vivo* and *in vitro* models of PD briefly described below, namely Rotenone and 6-OHDA.

Rotenone is a naturally occurring herbicide which binds to Complex I of the ETC [21, 155]. Unlike MPP+ it affects all of the neurons of the brain rather than

just the dopamine producing ones [21]. It has been argued that this may make it a more realistic PD model. It also serves to suggest, because of a differential effect on mortality of dopaminergic neurons, that dopaminergic cells are differentially sensitive to damage to the mitochondria [21]. Rotenone can be used in both *in vivo* and *in vitro* experiments. In the initial *in vitro* experiments only 50% of one strain of rat developed PD like symptoms with other strains of rat being unaffected [155].

6-OHDA has a very similar structure to dopamine (as the name suggests it has additional -OH). Thus it is preferentially taken up by dopamine transporters [150]. Although it preferentially accumulates in dopaminergic cells it has also been shown to increase non-dopaminergic neuron mortality rates *in vitro* [4]. Once in the cell it accumulates in the cytosol [150]. It is believed to inhibit Complexes 1 and 4 of the ETC, leading to mitochondrial dysfunction [4]. 6-OHDA can be used *in vivo* and *in vitro*, but since it cannot cross the blood brain barrier [155], it must be directly injected into the brain in the *in vivo* mouse and rat models [4].

While keeping the shortcomings and the alternative in mind, here we focus on the MPP+ model. The MPP+ model is plausible because it replicates many of the features of the disease. It may not replicate the disease mechanism completely but it provides a good proxy to potentially test *in silico* results.

Further, MPP+ with BE(2)-M17 (dopaminergic) was used in Keane *et al.* [94] which is a motivation for this chapter 3.

2.5 Background on protein protein interaction networks

2.5.1 Protein-protein interaction networks

A protein-protein interaction (PPI) is a pair of proteins that undergo physical binding via molecular docking [43], and a PPI network is defined as a network formed with proteins representing nodes and edges representing interactions. Due to the invention of high throughput experiments, large data sets have emerged with the number of proteins and interactions on the order of 10^3 and 10^5 respectively [81]. To date there is no good generating model for PPI networks [2, 113, 139, 165].

PPI networks are very rich data sources. Over the last 10 years they have been used to predict the function of proteins [2, 106], to predict interactions for biological validation [2, 65], to propose mechanisms for evolution of cellular complexity [2] and attempt to locate biological modules in the cell [2, 108]. They have also been used to attempt to gain insight into diseases e.g. ataxias [112] and cardiac arrest [86]. It has to be kept in mind that many diseases interfere with non-protein parts of the interactome (for example the dopamine pathway in PD), and hence the proteome will only ever be part of the story.

False Positives and Negatives and Bias in PPI data PPI networks are not clean data sources; several problems have been identified with them that must be taken into account. The experimental methods that are used to establish the existence of PPIs are error prone. They introduce significant rates of false positives and false negatives into the network, resulting in some estimates predicting that the human proteome is only 10% complete [79, 81]. Further, the known sample of the proteome is highly skewed towards proteins that are of interest to experimentalists, and it has been shown that varying the level of evidence required for an interaction (i.e. how

well it is studied) can vastly change global characteristics of the network [188].

Many methods have been developed to attempt to calculate the error in PPI datasets, for example:

1. *Literature Curation* High throughput screens are verified by testing the results against a list of known PPIs, see for example [39].
2. *Expression Profile Reliability* High throughout experiments are verified by comparing expression profiles (a measure of how often both proteins in a PPI are expressed together) of the test data set with the profiles of a large number of known PPIs, see for example [46].
3. *Paralog Verification Method* PPI are deemed probable if paralogs (same protein in different species) of the proteins in the PPI in question interact [46], although care must be taken to select proteins that are very similar [109].
4. *Network based approaches* Methods based on network structure (e.g. clustering coefficient) which are believed to be relevant in PPI networks to assess validity of interactions, see for example [2].
5. *Sequence and Functional Annotation data* The probability of a PPI is calculated using a Bayesian approach based on sequence data and functional annotation, see for example [129].

These methods have allowed experimental techniques to be assessed for their potential error rate. One such assessment conducted by Mering et al [181]. can be seen in Figure 2.1. While the method used to calculate error (literature curation) can induce bias to certain types of interactions as seen above, the level of accuracy and coverage of many of the experimental techniques gives a sense of the level of error in this data. For example *in silico* predictions have 10% coverage and accuracy.

Recent work seems to suggest another bias in the data. It is postulated that high degree 'hub' proteins in PPI networks could be combinations of multiple proteins, each with a smaller number of interactions [172]. This is a large concern as sampling and

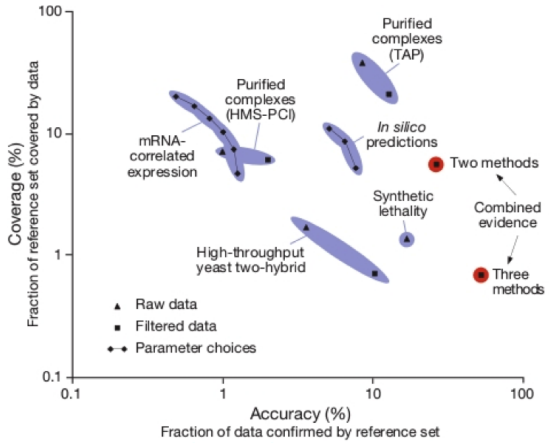

Figure 2.1: Estimates of the accuracy of different methods using a set of trusted interactions. Image taken from Von Mering et al. Comparative assessment of large-scale data sets of protein protein interactions Nature 2002 [181]

centrality measures (used in some attempts to select important genes) are correlated with degree.

2.5.2 Pathway, protein and PPI databases

Due to the explosion of protein and PPI data, several databases have been created to store the information for public use.

Pathway and Protein Databases PPI interaction networks cannot be interpreted without a detailed understanding of the biological context of each interaction. While this information for individual proteins is easy to obtain by reading several papers on the protein, it is very time consuming to do so for all proteins in the proteome. To address this, several protein and pathway databases have emerged to store contextual data in a format that is easy to extract (fixed nomenclature, web APIs etc). Some of the most important databases include Gene Ontology (GO) [9], KEGG [90], Uniprot [167] and OMIM [80].

Gene Ontology (GO) [9] is based on a structured vocabulary designed to describe biological properties. The GO contains 3 distinct root terms, namely cellular compo-

nent, biological process and molecular function, each relating to a distinct branch of biological knowledge. The structure is directed and hierarchical (acyclic), resulting in a vocabulary that becomes more specialised with greater distance from the root terms. The GO database is still being actively developed, which means that certain areas (for example those being studied extensively by the bioinformatics community) are likely to be over-represented and therefore contain a larger number of specialised terms. This is problematic for certain forms of analysis, which assume that all terms have equal probability of being assigned. Therefore the GO consortium has provided both a generic GO Slim database containing broad overview terms and the option of a user customisable GO database, which could for example be limited to terms with a certain level of evidence.

KEGG [90] is a pathway database containing high level diagrams detailing the directed iterations between bio-molecules in many biological processes. It includes information about diseases, genes, proteins, and standard biological modules. For example, KEGG includes a good summary of the processes believed to be involved in PD, as can be seen in Figure 2.2. KEGG provides an API (application programming interface) and a XML representation of KEGG diagrams to allow easy integration with new software and easy access to the data.

Uniprot [167] is a protein information resource providing manual and computer generated summaries of known information. It contains information about protein sequence, function, subcellular location, post-translational modifications and involvement in disease. Proteins are also cross referenced with other databases (including GO), resulting in one data source which summarises a large amount of useful protein information. Uniprot provides easily parsable formats which can be downloaded and extracted and a protein mapping service (http://www.uniprot.org/mapping/) available through a web interface and an established API.

The OMIM (Online Mendelian Inheritance in Man) [80] database collates known

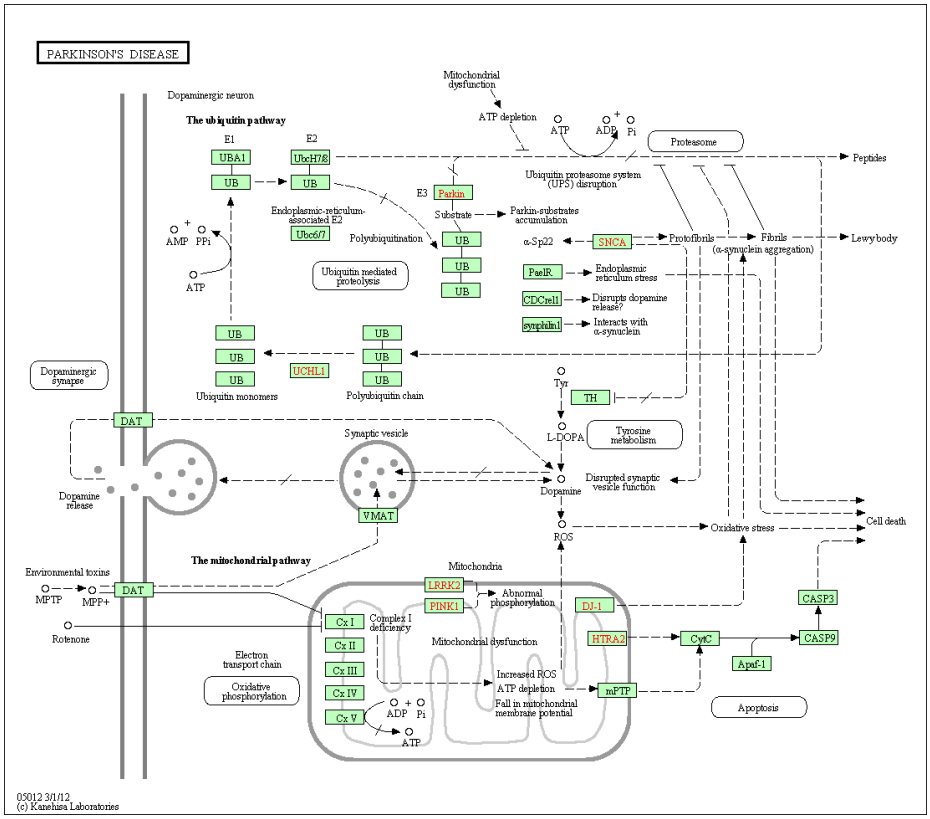

Figure 2.2: PD disease process diagram from the KEGG database. The diagram includes a summary of the processes that the authors believed to be important to the Parkinson's disease and the links between them [90].

relationships between genes and gene phenotypes. It also contains information about each gene, and the evidence underpinning the link to any one disease. OMIM provides data downloads of the whole database and an API which allows querying of the database.

PPI Databases There are many databases that store collections of PPIs. Different PPI databases have different curation rules and thus can contain very different sets of PPIs, often with only a small overlap [120]. A survey of PPI databases in 2005 revealed 78 different PPI databases [12]; the more common ones are: DIP [190], BIND [11], HPRD [95, 123, 130], MINT [29], IntAct [7] and BioGRID [31, 161].

Due to the lack of overlap between PPI databases, there has been a drive to combine databases to construct larger and potentially more complete databases. Examples include iRefIndex [136], MiMI [87], BIANA [66] and the Sysbiomics database [116]. While these databases give much better coverage it is potentially at the expense of curation, resulting in a large number of potentially low confidence interactions. In this document we will primarily use the BioGRID databases and a brief description can be found below.

2.6 Background on community detection

Community detection is a general term for selecting sets of nodes that are (in some sense) more strongly connected to each other than to the rest of the network. Some examples include teams in organisational networks, leagues in sports and friendship groups in social networks. Unlike some other clustering approaches, community detection is usually performed via an algorithm which also selects the number of communities.

There is a caveat to this: community detection by a technique which we will use in this chapter known as modularity maximisation can find it impossible to detect

communities below a certain size because of the so called resolution limit. [58] This was in part addressed by the introduction of the resolution parameter [137], however issues still exist - see Ref. [102]. For this chapter, we will assume that varying the resolution parameter will mostly address this issue. However, we are aware that there are partitions which modularity maximisation simply cannot find.

In this section we focus on some standard methods which will form the backdrop for our extension to modularity based community detection. The literature on community detection methods is too large to be adequately summarised here; for a good review see Ref. [57], or Ref. [59].

2.6.1 Newman-Girvan modularity

We will use a community detection approach known as modularity maximisation, where a quality function known as *modularity* is optimised over all possible partitions. The standard modularity quality function was defined by Newman and Girvan [126] as follows:

$$Q(c_i : i \in V) = \sum_{i,j \in V} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{2.1}$$

where V is the set of nodes, A is an $[0, 1]$ adjacency matrix where $A_{ij} = 1$ if there is an edge between i and j and $A_{ij} = 0$ otherwise, c_i is the community assignment of node i, $\delta(c_i, c_j)$ is a delta function which is 1 if $c_i = c_j$ and 0 otherwise, k_i the degree of node i and m is the total number of edges in the network.

The motivation behind modularity is as follows: modularity measures the difference between the observed number of edges and the expected number of edges within groups as specified by a null model which in Eq. (2.1) is the configuration model $(\frac{k_i k_j}{2m})$.

The optimal community is then defined as

$$c_{max} = \underset{c\in\{1,...,n\}^n}{\arg\max}\, Q(c_i : i \in V) = \underset{c\in\{1,...,n\}^n}{\arg\max} \sum_{ij\in V} (A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j), \tag{2.2}$$

where $n = |V|$. For a more general version of Eq. (2.1) replace $\frac{k_i k_j}{2m}$ by P_{ij} to represent the expected number of edges between i and j under a given null model.

We could compute c_{max} by checking all possible partitions of the nodes into groups. The number of possible partitions of length n is the nth Bell Number (B_n), which is defined using the recurrence relation [186]:

$$B_n = \sum_{i=0}^{n-1} \binom{n-1}{i} B_i, \tag{2.3}$$

where $B_1 = B_0 = 1$. This function is super-polynomial in n [77] and therefore for any large network checking all possible community vectors is not possible with current computational constraints. Hence approximation algorithms are employed. One popular approximation algorithm is given by the Louvain algorithm [23].

2.6.2 The Louvain method

The Louvain method to detect communities in networks [23] is a greedy algorithm that is general enough to allow the use of any null model (P_{ij} for $i, j = 1, ..., n$).

The algorithm places every node in its own community and makes single node community reassignments until there are no more changes that will increase modularity. Then it makes a meta-network where the nodes represent community assignments in the previous stage and the weight of the edges between the nodes is the sum of the weights between all pairs of nodes in respective communities, and then repeats the process on the new metanetwork until no more improvements are possible. This can be summarised into the following two phases which are applied iteratively:

Phase 1

1. Assign each node to its own community.
2. For each node i
   (a) Consider the increase in modularity if it is moved to the community of one of its neighbours.
   (b) If there is a move that increases modularity, make the move that increases modularity by the largest amount (in case of ties use a breaking rule).
3. If no moves have been made in step 2 then finish, else go back to step 2.

Phase 2

1. Make a network of meta communities. In this new aggregated network nodes represent the communities from the previous phase.
2. Add edges to this aggregated network with weight equal to the sum of the weights of the edges of the nodes in each community.

Phase 1 and 2 are performed iteratively until no more improvements in the modularity can be made. The key point of the method is that increasing modularity in the network of meta-communities also increases it on the wider network. The method is fast because the change in modularity in Phase 1 can be computed very quickly without computing the modularity for the whole network as moving one node only changes a small number of terms in the sum of Eq. (2.1).

In this document we make heavy use of the GenLouvain MATLAB code, and a generalised implementation of the Louvain algorithm which allows use of a generic null model [88].

2.6.3 Altering the null model

While the initial null model proposed by Newman and Girvan [126] is useful for certain networks, the assumptions of the null model can prove problematic in some cases.

One major limitation relates to 'resolution', i.e. the propensity for community detection by modularity maximisation to find communities of a given size. To deal with this issue Reichardt and Bornholdt [137] proposed an adjustment to modularity known as the resolution parameter (λ), resulting in the following adjusted modularity (fitness function):

$$Q = \sum_{ij}(A_{ij} - \lambda P_{ij})\delta(c_i, c_j). \tag{2.4}$$

The parameter λ can be thought as "zooming" in and out of the null models. With higher resolution parameter λ there is a much larger penalty for missing expected edges, and with lower resolution parameter λ the opposite is true. Hence the calculation of expected edges is crucial.

The standard Newman-Girvan null model calculates the expected number of edges between a pair of nodes through the configuration model: edges are randomly reshuffled but the degree of the each node is maintained. The configuration model does create multi-edges and self edges, which may not be appropriate. Clearly there are instances where this null model is not appropriate for community detection, for example in bipartite networks [57, 78] and correlation networks [72, 115].

This problem extends to networks which have an external driver (e.g. geography) which may influence the formation of edges, if we wish to observe communities that are in some sense 'surprising' after taking account of the external driver. For example in a network formed of all scientists where edge weight is the number of co-publications, it is likely (although not guaranteed) that standard community detection would find communities based on subject area, with a small amount of noise from interdisci-

plinary work. Therefore, by controlling for subject area through adjusting P_{ij}, we may observe additional structure, such as institutions.

An illustrative example is the analysis by Expert *et al.* [53] of a network constructed from mobile phone data in Belgium. The nodes in the network represent regions and the edge weight is the number of calls between each region in a 6 month period. The standard Newman-Girvan null model uncovers sets of nodes that are spatially close to each other. Expert *et al* [53]. constructed a null model based on population flows, which was based on gravity models. A gravity model is a model where the relationship between objects decays with distance, this is often modelled a negative power of the distance between the objects [15, 53]. They chose this null model due to earlier work which showed that communication in this dataset could be explained using a gravity model [100]. The null model in question takes the following form:

$$P_{ij}^{SPA} = N_i N_j F(d_{ij}) \tag{2.5}$$

where N_i is the population in region i, and d_{ij} is the distance between region i and j. The function F then fits the expected number of edges at a given distance resulting in the following formulation:

$$F(d) = \frac{\sum_{i,j|d_{ij}=d} A_{ij}}{\sum_{i,j|d_{ij}=d} N_i N_j}, \tag{2.6}$$

where A is the adjacency matrix. To address the fact that only a very small number of locations are precisely at the same distance, the distances are binned and thus we can think of the d_{ij} in Eq. (2.5) and Eq. (2.6) as the bin to which the actual distance belongs. The resultant null model is that the number of calls between two populations at the same approximate distance are independent and identically distributed.

Applying P^{SPA} to the network, the two largest communities accounting for 75% of the nodes correspond to a linguistic partition with respect to the Flemish speaking

community and the French speaking community [53].

However, further work has shown that if the attributes of the nodes are too strongly correlated with the geographic features, then the method of Expert *et al.* can fail [30]. In this case accounting for the correlation in the null model can recover performance [30].

2.6.4 Stability: An alternative approach

An alternative approach to community detection by modularity maximisation makes use of a quality function known as 'Stability' [101], that is constructed by considering ergodic random walkers on the network. Stability uses the dynamics of a random walker to define a similarity score between nodes, where the similarity score between two nodes is high when there is a higher likelihood than would be expected at random of a walker moving from one node to the other. The Louvain algorithm [23] is in then used on the resultant matrix of similarity scores to produce communities.

A random walker in this context can be thought of as an object that is moving from node to node with probability defined by the underlying process. The movement from node to node could either be in discrete time, i.e. walkers jump from node to node at a fixed intervals; or in continuous time where the waiting time at a particular node is random. Further, the random walk on a given network must be ergodic, i.e. have a stationary distribution for the system into which all initial states converge. This also imposes some limitations on the network the random walker is traversing, for example bipartite undirected networks are not always possible as a random walk would not be aperiodic for certain Markov processes.

Under the conditions noted above, Lambiotte *et. al.* defined the following quantity [101] for random walker $W^{(M)}$ under Markov process M, which is placed at time 0 at each node with probability given by the stationary distribution. Using slightly different notation to Ref. [101] we define the random variable $W_t^{(M)}$ as the position

of the random walker at time t. Then the probability that the random walker stops inside a community at time t is:

$$P_M(C,t) = Prob\left((W_0^{(M)} \in C) \cap (W_t^{(M)} \in C)\right) \tag{2.7}$$

where C is a set of nodes corresponding to a community. Further, Lambiotte defines the following:

$$P_M(t,\infty) = \lim_{t\to\infty} P_M(C,t) = \lim_{t\to\infty} Prob\left((W_0^{(M)} \in C) \cap (W_t^{(M)} \in C)\right) = Prob(W_0^{(M)} \in C)^2 \tag{2.8}$$

where the final equality comes from the fact that as $t \to \infty$ then the $W_t^{(M)}$ should tend to the stationary distribution. Putting this together Lambiotte then constructs the following quality function known as *Stability*:

$$R_M(t) = \sum_{C \in P_{art}} \left(P^{(M)}(C,t) - P^{(M)}(C,\infty)\right) \tag{2.9}$$

where M is the Markov process of interest and P_{art} is a set of communities, e.g. $P_{art} = \{\{1,2\},\{2,4\}\}$. This form is inspired by Delvenne *et al.* [49] (which shares three of the four co-authors with Ref. [101]) comparing partitions at different time periods using an autocovariance function of a Markov process. [49, 101]

The quantity $R_M(t)$ is then maximised at different values of t which can be thought of as similar to the resolution parameter in the formulation by Reichardt and Bornholdt [137] which is given in Eq. (2.4). Lambiotte showed that the ad-hoc resolution parameter is in fact equivalent to t in a linearised version of $R_M(t)$ for a particular Markov process. This approach is very different to the 'classic' approaches which only consider pairwise connections.

By using different dynamics for the random walk, Lambiotte *et al.* [101] derive two different forms for stability. The first form known as the normalised version

(also referred to as the normalised Laplacian due to the similarity [101]) is based on a random walk where walkers move from each node at a constant rate. [101] This results in dynamics (and therefore a Markov process) which is described by the Kolmogorov equation [101]:

$$\frac{d\mathbf{p}}{dt} = (AD^{-1} - I)\mathbf{p} \tag{2.10}$$

where A is the adjacency matrix, p is the probability of being at a given node, and D is a diagonal matrix with the degree of each node on the diagonal, i.e. $D_{ii} = \sum_j A_{ij} = k_i$. Note we have in the preceding formula slightly changed the notation and the presentation from Ref. [101]. Noting that the stationary solution of this equation is $\frac{k_i}{2m}$, therefore:

$$R_n(t) = \sum_{i,j} \left((e^{(AD^{-1}-I)t})_{i,j} \frac{k_j}{2m} - \frac{k_i}{2m}\frac{k_j}{2m} \right) \delta(c_i, c_j) \tag{2.11}$$

The second form is known as the combinatorial version (also referred to as the combinatorial Laplacian), where the walkers jump from each node with a rate proportional to the node's degree. [101] This results in the following dynamics:

$$\frac{d\mathbf{p}}{dt} = \frac{n}{2m}(A - D)\mathbf{p} \tag{2.12}$$

where n, m, and A are respectively the number of nodes, number of edges and adjacency matrix as before. The stationary distribution for this is given by $\frac{1}{n}$ [101], thus giving the following form:

$$R_c(t) = \sum_{i,j} \left((e^{t(A-D)\frac{n}{2m}})_{i,j} \frac{1}{n} - \frac{1}{n}\frac{1}{n} \right) \delta(c_i, c_j) \tag{2.13}$$

Uses and Extensions This method has been extended to directed networks [101], and to use higher order Markov Models. [145] It has also been used to analyse Twitter networks of riots [18], used to segment images [149], investigate protein structure. [149]

However, in this document we will use the original version for undirected graphs.

2.7 Background on Stein's method

In the following section we will give a brief background and overview of Stein's method, covering the basic approach. This review is inspired by the comprehensive introductory review in Ref. [142],

We will then explore more in depth the specific results which we will use in this chapter.

2.7.1 Distances between distributions

Stein's method is a tool to assess distributional distances. Before we describe Stein's method we discuss distances between two distributions. In fact there are several distances which are typically used for this task, depending on what we are interested in measuring. For example we could be interested in the difference between the cumulative distribution functions (CDFs) of the distribution of interest and the distribution we are approximating, or we could be interested in the difference between the probability density functions.

In our work on Stein's method we focus on three families of distance measures of the form:

$$d_H(u,v) = \sup_{h\in H} \left| \int h(x)du(x) - \int h(x)dv(x) \right|, \tag{2.14}$$

where H is a set of functions, and u and v are two probability distributions [142].

The three families of interest are, the Kolmogorov distance, and the Wasserstein distance and the bounded Wasserstein distance.

Kolmogorov Distance For this distance measure we define the set of functions H_K as follows [142]:

$$H_K = \{h_\alpha(x) : \alpha \in \mathbb{R}\}, \text{ where } h_\alpha(x) = \begin{cases} 1 & \text{if } x \leq \alpha; \\ 0 & \text{if } x > \alpha. \end{cases} \tag{2.15}$$

Taking the expectation of a function h_α in this set over a random variable measures the cumulative distribution function at a given point (α) for random variables defined on the real line (or a subset thereof).

Therefore, for u and v probability distributions defined on (a subset of) the real line, $d_{H_K}(u, v)$ is the largest difference in absolute value between the two CDFs.

Wasserstein Distance The set of functions for this distance measure is as follows [138, 142]:

$$H_W = \{h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y)| \leq |x - y|\}. \tag{2.16}$$

This is also known as the set of Lipschitz continuous functions with Lipschitz constant 1. We note that $\int h(x)du(x) = E[h(X)]$, where we let X be a random variable with distribution u [138]. Therefore:

$$d_{H_W}(u, v) = \sup_{h \in H_W} |E[h(X)] - E[h(Y)]| \tag{2.17}$$

where we let Y be a random variable with probability distribution v, so that X and Y are defined on the same probability space.

Bounded Wasserstein Distance The set of functions for this distance measure is as follows [6]:

$$H_{bW} = \{h : \mathbb{R} \to \mathbb{R} : \sup_{x \neq y} \frac{|h(x) - h(y)|}{|x - y|} + \sup_{x} |h(x)| \leq 1\} \tag{2.18}$$

Further, $H_{bW} \subseteq H_W$ as $\sup_{x \neq y} \frac{|h(x)-h(y)|}{|x-y|} \leq 1$ implies a Lipschitz continuous function with Lipschitz constant 1.

Relations between the distances There are cases where we may be interested in converting a bound in one distance to that of another [142].

Clearly, if $H_1 \subset H_2$ then $d_{H_1}(u, v) \leq d_{H_2}(u, v)$ therefore:

$$H_{bW} \subseteq H_W \implies d_{H_{bW}}(u, v) \leq d_{H_W}(u, v) \tag{2.19}$$

for arbitrary probability distributions u and v.

In this thesis we will be concerned with bounding the Kolmogorov distance as we are concerned with the CDF. It is possible to bound the Kolmogorov distance using the Wasserstein distance [142]. In Ref. [142] it is shown that: relation is as follows:

$$d_{H_K}(u, v) \leq \left(\frac{2}{\pi}\right)^{\frac{1}{4}} \sqrt{d_{H_W}(u, v)}. \tag{2.20}$$

for v standard normal.

Similarly, for the bounded Wasserstein distance, from Anastasiou and Reinert [6]:

$$d_{H_K}(u, v) \leq 2\sqrt{d_{H_{bW}}(u, v)}. \tag{2.21}$$

2.7.2 Stein equation and Stein's method

In this section, we will give an overview of Stein's method mainly based on Ref. [142].

First we define a set of functions as follows:

$$\mathfrak{H} = \{f : \mathbb{R} \to \mathbb{R},\ f(x) = h'(x) - xh(x) \text{ for a function } h : \mathbb{R} \to \mathbb{R} \qquad (2.22)$$
$$\text{with } h \text{ absolutely continuous and } E|h'(Z)| < \infty\}$$

where Z has the standard normal distribution.

The key statement of Stein's method for normal approximation, adapted from Ref. [142], then states that:

Theorem 1. ***Stein's Lemma*** *If $Z \sim N(0,1)$ then for all $f \in \mathfrak{H}$, $E[f(Z)] = 0$. Further for any random variable X, if $\forall f \in \mathfrak{H}$ $E[f(X)] = 0$ then $X \sim N(0,1)$.*

This result can be extended to general normal distribution $N(\mu, \sigma^2)$ by replacing $h'(x) - xh(x)$ with $\sigma^2 h(x) - (x - \mu)h(x)$ see Ref. [138].

Bounding the error In this section following Ref. [142], we abuse notation, where for X a random variable with distribution u and Y a random variable with distribution v we let $d_H(u, v) = d_H(X, Y)$.

We are interested in distributional distances that are of the form $|E[h(X)] - E[h(Z)]|$ (for $Z \sim N(0,1)$). Clearly, $d_H(X, Z) = 0$ if $X \sim N(0,1)$. We also know that $E[h'(X) - Xh(X)] = 0$, if $X \sim N(0,1)$ [142]. Taking these facts together we can see that it will be useful [142] to construct a function f such that $f'(x) - xf(x) = h(x) - E[h(Z)]$, so that after taking the expectation of both sides, we can attempt to bound the right hand side only using properties of X and $f(x)$ [142].

The unique bounded solution to this equation is given as follows [138, 142]:

$$f_h(x) = e^{\frac{x^2}{2}} \int_x^\infty e^{-\frac{y^2}{2}} (E[h(Z)] - h(y))dy. \qquad (2.23)$$

We omit the proof of boundedness but it can be found in Ref. [142]. A further related result that we shall state without proof which is referred to in Ref. [142] is as follows:

For h absolutely continuous and f_h as given in (2.23)

$$\sup_x |f_h(x)| \leq 2 \sup_x(|h'(x)|), \qquad \sup_x(|f_h'(x)|) \leq \sqrt{\frac{2}{\pi}} \sup_x(|h'(x)|), \tag{2.24}$$
$$\text{and } \sup_x(|f_h''(x)|) \leq 2 \sup_x(|h'(x)|).$$

Further, a function that is Lipschitz continuous is also absolutely continuous. Therefore for functions that are Lipschitz continuous with Lipschitz constant 1 (i.e. all functions in H_W), $sup_x(|h'(x)|) \leq 1$ where h' is the derivative of h which exists (almost surely) by Rademacher's Theorem.

We can now put all of these results together to obtain a key result which allows Stein's method to bound the distributional distance between a distribution and the standard normal. Let

$$F_W = \{f_h : h \in H_W\} \tag{2.25}$$

where f_h is the solution to the Stein equation for function h and given in (2.23). Further, let us construct the following set:

$$\mathfrak{F}_W = \left\{ f : \mathbb{R} \to \mathbb{R} \text{ such that } \sup_x(|f(x)|) \leq 2, \ \sup_x(|f'(x)|) \leq \frac{2}{\pi} \text{ and } \sup_x(|f''(x)|) \leq 2 \right\}. \tag{2.26}$$

Clearly as we know that all $h \in H_W$ are Lipschitz continuous with Lipschitz constant 1 it follows (2.24) by the result for absolutely continuous functions above that $F_W \subseteq \mathfrak{F}_W$.

Putting these results together for X a random variable with the distribution we want to approximate and $h \in H_W$ [142] we obtain:

$$h(X) - E[h(Z)] = f_h'(X) - X f_h(X). \tag{2.27}$$

Taking the expectation and the absolute value of both sides gives:

$$|E[h(X)] - E[h(Z)]| = |E[f_h'(X) - Xf_h(X)]|. \text{ Hence} \tag{2.28}$$

$$d_{H_W}(X, Z) = \sup_{h \in H_W} |E[h(X)] - E[h(Z)]| = \sup_{f_h \in F_W} |E[f_h'(X) - Xf_h(X)]| \tag{2.29}$$

$$\text{and } d_{H_W}(X, Z) \leq \sup_{f_h \in \mathfrak{F}_W} |E[f_h'(X) - Xf_h(X)]|. \tag{2.30}$$

where the final inequality comes from the fact that $F_W \subseteq \mathfrak{F}_W$. The functional form $|E[f_h'(X) - Xf_h(X)]|$, can then be bounded for random variables with specific properties. [142].

2.7.3 Bounding the sum of dependent random variables

Our statistic of interest Path Modularity is a sum of dependent random variables and we want to approximate its distribution by a normal distribution.

Before we do this let us define a few terms. Let $\{X_i : 1 \leq i \leq n\}$ be a set of possibly dependent random variables such that $E[X_i] = 0$ and $E[X_i^4] < \infty$.

If we now let $W = \sum_i X_i/\sigma$, where $\sigma = \text{Var}(\sum_i X_i)$, then $E[W] = 0$ and $\text{Var}(W) = 1$. The quantity we are interested in bounding is $d_{H_W}(W, Z)$ for a given n and dependencies between the X_is.

The set $\mathbb{K}_i$ is called a dependency neighbourhood of X_i if X_i is independent of $\{X_j : j \notin \mathbb{K}_i\}$. Often dependency neighbourhoods are constructed such that $X_j \in \mathbb{K}_i$ implies $X_i \in \mathbb{K}_j$.

A simple but crude bound on $d_{H_W}(W, Z)$ accounts for the size of the largest dependency neighbourhood $D_{max} = \max_{i=1,\ldots,n}(|\mathbb{K}_i|)$. The bound given in Ref. [142] Theorem 3.1, is as follows:

$$d_{H_W}(W, Z) \leq \frac{D_{max}^2}{\sigma^3} \sum_{i=1}^{n} E[|X_i|^3] + \frac{\sqrt{28} D_{max}^{\frac{3}{2}}}{\sqrt{\pi}\sigma^2} \sqrt{\sum_{i=1}^{n} E[X_i^4]}. \tag{2.31}$$

A more involved but better bound than (2.31) can be found in Barbour *et al.* [14]. This bound requires a slightly more complex set up. Let us now assume that $\mathrm{Var}(\sum_{i\in I} X_i) = 1$. Let X_i and $W = \sum_{i\in I} \frac{X_i}{\sigma} = \sum_{i\in I} X_i$ be as before and $E[W] = 0$ and $E[W^2] = 1$, but we now require W to be *decomposable*.

We aim to construct a decomposition of the summation into two components one of independent and one which is potentially dependent on an element of the summation under certain conditions, and then to repeat the procedure on the independent component with respect to another term in the summation. We will represent this composition as a series of sets. Formally, decomposable is defined as follows (taken from [14]):

For $i \in I$, $I = \{1, \ldots, n\}$ there exists a set $\mathbb{K}_i$, such that if $j \notin \mathbb{K}_i$ then X_i and X_j are independent.

Further, there exists sets of variables which are square integrable: $\{X_i\}$, $\{W_i\}$, $\{Z_i\}$, $\{Z_{ij}\}$ for $i \in I$ and $j \in \mathbb{K}_i$ such that:

$$W = W_i + Z_i, \tag{2.32}$$

where $Z_i = \sum_{j\in\mathbb{K}_i} Z_{ij}$, and W_i is independent of X_i. Further for $j \in \mathbb{K}_i$, there are sets $\{W_{ik}\}$ and $\{V_{ij}\}$ such that

$$W_i = W_{ij} + V_{ij}, \tag{2.33}$$

with W_{ij} independent of (X_i, X_j). Thus

$$W = Z_i + V_{ij} + W_{ij}, \tag{2.34}$$

Using this Barbour *et. al* [14] proves the following result:

Theorem 2. ***(Lemma 1 from Ref. [14])*** *Let W be decomposable. Then for every*

function $f : \mathbb{R} \to \mathbb{R}$ *which is bounded and has bounded first and second derivatives, we have*

$$|E(Wf(W) - f'(W))| \leq J\epsilon \tag{2.35}$$

where $J = \sup_x |f''(x)|$ *and*

$$\epsilon = \left(\frac{1}{2} \sum_{i \in I} E[|X_i| Z_i^2] + \sum_{i \in I} \sum_{k \in K_i} (E[|X_i Z_{ik} V_{ik}|] + E[|X_i Z_{ik}|] E[|Z_i + V_{ik}|]) \right). \tag{2.36}$$

As an illustration of Stein's method we will give an overview of the proof of this statement.

Proof Note, the full proof is in Ref. [14]; a shortened version is given here for transparency.

Barbour *et al.* [14] first relates $|E(Wf(W) - f'(W))|$ to the error (2.35) as follows.

Let $\mathfrak{F}_J$ be the set of all functions with bounded first and second order derivatives and let $f \in \mathfrak{F}_J$. Then:

$$\begin{aligned} E[Wf(W) - f'(W)] =& E[Wf(W)] - \sum_{i \in I} E[X_i Z_i f'(W_i)] + \sum_{i \in I} E[X_i Z_i f'(W_i)] \\ & - \sum_{i \in I} \sum_{k \in \mathbb{K}_i} E[X_i Z_{ik} f'(W_{ik})] + \sum_{i \in I} \sum_{k \in \mathbb{K}_i} E[X_i Z_{ik} f'(W_{ik})] \\ & - E[f'(W)] \end{aligned} \tag{2.37}$$

Ref. [14] notes that $\sum_i \sum_{k \in \mathbb{K}_i} E[X_i Z_{ik}] = \sum_i E[X_i Z_i] = \sum_i E[X_i W] = E[W^2] = 1$.

Therefore:

$$E[Wf(W) - f'(W)] = \Big(E[Wf(W)] - \sum_{i\in I} E[X_i Z_i f'(W_i)]\Big) \tag{2.38}$$
$$+ \Big(\sum_{i\in I} E[X_i Z_i f'(W_i)] - \sum_{i\in I}\sum_{k\in \mathbb{K}_i} E[X_i Z_{ik} f'(W_{ik})]\Big)$$
$$+ \sum_{i\in I}\sum_{k\in \mathbb{K}_i} E[X_i Z_{ik}](E[f'(W_{ik})] - E[f'(W)])$$

By Taylor expansion, noting that $|f''|$ is bounded and using the restrictions on the decomposition one can show (see Ref. [14] for full details) that for $J = \sup_x |f''(x)|$:

$$|\Big(E[Wf(W)] - \sum_{i\in I} E[X_i Z_i f'(W_i)]\Big)| \le \frac{1}{2} J \sum_{i\in I} E[|X_i| Z_i^2], \tag{2.39}$$

$$\Big(\sum_{i\in I} E[X_i Z_i f'(W_i)] - \sum_{i\in I}\sum_{k\in \mathbb{K}_i} E[X_i Z_{ik} f'(W_{ik})]\Big) \le J \sum_{i\in I}\sum_{k\in \mathbb{K}_i} E[|X_i Z_{ik} V_{ik}|], \tag{2.40}$$

$$\text{and } |E[f'(W_{ik})] - E[f'(W)]| \le J E[|Z_i + V_{ik}|]. \tag{2.41}$$

Putting this together we obtain:

$$E[Wf(W) - f'(W)] \le \frac{1}{2} J \sum_{i\in I} E[|X_i| Z_i^2] + J \sum_{i\in I}\sum_{k\in \mathbb{K}_i} E[|X_i Z_{ik} V_{ik}|] \tag{2.42}$$
$$+ \sum_{i\in I}\sum_{k\in K_i} E[X_i Z_{ik}] J E[|Z_i + V_{ik}|]$$
$$= \frac{1}{2} J \sum_{i\in I} E[|X_i| Z_i^2] \tag{2.43}$$
$$+ J \sum_{i\in I}\sum_{k\in \mathbb{K}_i} \Big(E[|X_i Z_{ik} V_{ik}|] + E[X_i Z_{ik}] E[|Z_i + V_{ik}|]\Big),$$

thus giving the form in the theorem. QED.

The proof can easily be adapted to Wasserstein distance. To be consistent with previous notation we slightly deviate from the approach in Barbour *et al.* [14]. Following previous notation we let f_h be as in (2.23) for $h \in H_W$. and we let $F_W = \{f_h : h \in H_W\}$. As $h \in H_W$ is Lipschitz continuous with Lipschitz constant at most 1, it

is absolutely continuous, by (2.24) f_h is bounded and is bounded in first and second derivatives, (in fact $\sup_x(|f_h''|) < 2\sup_x|h'|$). Therefore as it has bounded first and second derivatives we can apply Theorem 2 to f_h.

$$d_{H_W}(X,Z) = \sup_{h\in H_W} |\Big(E[h(X)] - E[h(Z)]\Big)| \tag{2.44}$$

$$= \sup_{f\in F_W} |E[f'(W) - Wf(W)]| \tag{2.45}$$

$$\leq \epsilon \sup_x |f_h''| \leq 2\epsilon \sup_x(|h'|) \leq 2\epsilon \tag{2.46}$$

where the final inequality uses the fact that h is Lipschitz 1 and therefore $\sup_x|h'| \leq 1$.

Simplifying the Bound Barbour *et al.* [14], uses an alternative form for ϵ which is easier to work with in a graph context.

This requires setting $Z_{ij} = X_j$ and $V_{ij} = \sum_{k\in\mathbb{K}_j\setminus\mathbb{K}_i} X_k$ and requiring that $i \in \mathbb{K}_j \longleftrightarrow j \in \mathbb{K}_i$. [14]. Barbour *et al.* [14] then obtains the following bound (which we will state without proof).

$$\epsilon = 2\sum_{i\in I}\sum_{j,k\in\mathbb{K}_i} E[|X_iX_jX_k|] + E[|X_iX_j|]E[|X_k|] \tag{2.47}$$

Combining this result with Theorem 2 and (2.44) we obtain:

$$d_{H_W}(X,Z) \leq 4\sum_{i\in I}\sum_{j,k\in\mathbb{K}_i} E[|X_iX_jX_k|] + E[|X_iX_j|]E[|X_k|] \tag{2.48}$$

This is the form we will use in this thesis.

Chapter 3

Selecting a Sampling Method

List of Symbols

Symbol	Description
A_{cc}	The statistic accuracy
$B_a(M)$	The set of nodes within a hops of the nodes in M.
C_1, C_2	Set of sampled and not sampled nodes respectively
$D(M, r)$	The number of elements in M of degree r
E	Set of edges
$F(t, u)$	Function which counts the number of elements of t which are equal to u
L	Dummy variable used for tested length of seed list
M	A set of nodes
P	P-value
$P_{arb}^{(V)}(x, y), P_{arb}^{(E)}(x, y)$	Set of node and edges sampled with a given path based technique using the path between x and y
P_{ur}	The statistic purity
R	Number of random realisations
S_1, S_2	Arbitrary sets
$T(X)$	Test statistic
T_{obs}	An observation of a statistic used the calculation of a p-value
$U(t)$	Function which returns the unique elements of t

V	Set of nodes
V_t, E_t	Set of nodes and edges after sampling with a given technique
X_i	An indicator variable for the presence of node i in the sample
Y	Sum of X_i over all nodes in the network
χ^2	A standard type of distribution of random variable
τ	Correlation statistic known as Kendall's τ
a	Number of hops
f_{arb}	Arbitrary sampling technique
$h(M, s)$	The probability that there does not exist a node within a hops of M on a randomly selected seed list of size s.
$h_d(M, s)$	The probability that there does not exist a node within a hops of M on a randomly selected seed list of with degree sequence t.
k	Parameter to one of the techniques, controls the maximum path length considered
l_1, l_2	Seed lists
p, q	Internal and external edge probabilities
s	Number of elements in a seed list
t	Degree sequence
u	Arbitrary node degree
y, z	Example Nodes

This chapter represents a collaboration between Felix Reed-Tsochas, Gesine Reinert, Elizabeth Leicht, Alan Whitmore and myself and I am the first author. The paper (Ref. [52]) is currently under-review by Bioinformatics.

3.1 Introduction

Network sampling is used in many different fields, such as biology [112] and sociology [20, 62]. Many studies sample a known network to produce a subnetwork that is believed to be more relevant to their research goals than the initial network. Fre-

quently protein-protein interaction (PPI) networks are sampled to form subnetworks that are associated with the disease or cellular process of interest e.g. [33, 65, 68, 71, 86, 112, 152]. An advantage of such sampling methods is that on a small network an in-depth analysis, such as verifying existing links, may be feasible. Network sampling can also reflect empirical limitations such as the availability of partial data for a given network [20, 62], or the exclusion of vertices that cannot be detected [144], with consequences for measured network statistics [97].

Subnetwork construction methods are not without their own problems, since they may induce artefacts in the subnetworks that they generate. Even the use of a PPI interactome as a starting point intrinsically reflects the effects of sampling, since different experimental methods vary in their ability to identify particular interactions. There are also inherent biases in the levels of evidence available for different interactions, and generally PPI networks are known to exhibit high rates of both false positives and false negatives [2]. Notably, there is no gold standard method for constructing a network representing a cellular process, although several techniques have been proposed to achieve this aim. Some studies test interactions experimentally between a subset of proteins believed to be important to a disease process [112]. Another approach is to locate proteins present in the same cellular compartment as the process of interest, and add edges between these proteins using a PPI database [65]. One can also form subnetworks from a larger PPI dataset using a seed list in conjunction with a construction method, e.g. snowball sampling [116], path based methods [19], Steiner trees [187], or the inclusion of nodes based on significance testing [68, 152]. Finally, one can also take a network directly from a pathway database e.g. KEGG [86].

The sampling techniques in this chapter start from a list of seed nodes and apply what we call a *construction method* to generate a subnetwork from these seed nodes. Seed nodes are typically believed to have certain attributes or associations,

e.g. proteins implicated in a disease. As the underlying PPI network is available, this approach is subtly different from the standard use of these network sampling techniques, namely sampling a large unknown network with the aim of estimating global properties [20, 61, 125]. The construction methods used in this chapter following prior work on biological systems [19, 110, 116, 154] are as follows: (i) snowball sampling; (ii) all paths up to a given length; (iii) all shortest paths between seed nodes. Snowball sampling has been used in biological systems through easy to use plug-ins to popular software applications; for instance the Cytoscape plug-in Bisogenet [116], and to find hidden populations (e.g. drug users) in sociology [20, 61, 144]. All paths up to a given length has been used in biology through the Genes2Networks web app [19]. We are not aware of a published software package that uses shortest-path sampling, although Ref. [110] have used shortest-path sampling in a study on colorectal cancer and Ref. [94] has used shortest-path sampling to study Parkinson's disease. It is important to note that in general the effect of network sampling on network statistics is non-trivial, and only well understood for very limited combinations of sampling methods and underlying networks. For instance, it has been shown that the degree distribution of a network uniformly sampled from a scale free network is not itself scale free [166]. To select good subnetworks, guidance about typical samples, or indeed atypical subnetworks, is required.

Here we provide a method to assess when a given subnetwork differs significantly from randomly generated subnetworks. A subnetwork which differs significantly from a random network could be viewed as containing relevant information, assuming that the comparison with the random network is meaningful. Hence a key question concerns the rules for constructing an appropriate null model, or a correctly randomised subnetwork.

As there is no generally accepted parametric model of PPI networks [139], we are unable to construct a general null model based on an ensemble of PPI networks.

Instead, our method compares a statistic of interest against that obtained for an ensemble of subnetworks generated from the same underlying network using a set of seed lists which are randomly chosen under certain constraints. First, we match the degree of the seeds with those of the original seed list. By contrast, the popular configuration model would match the degree of all nodes in the subnetwork. Second, there is a further feature in our null model, which relates to redundancy in the seed list. Many different seed lists may give rise to the same subnetwork. Hence, given the construction method, we must also control for the construction of the seed list. Some of these seed lists can be constructed by removing nodes from the original seed list so that the subnetwork that is generated from the modified seed list is identical to the original one. We refer to the seed nodes that can be removed without changing the subnetwork as 'redundant seed nodes'. On this basis, we can then construct a meaningful null model, using subnetworks generated at random with the same (approximate) degree distribution as the smallest subset of the original seed list that generates the same network (the minimum seed list). We use this null model in a nonparametric significance test for features of sampled networks. Our null model allows us to assess the significance of network features given a construction method, rather than given a construction method and a fixed seed list.

The test is first illustrated using simulated stochastic block model data for a network with two groups. A stochastic block model assigns each node to a group and then places edges between a pair of nodes with a fixed probability based on the group that the node is assigned to. We demonstrate that significance under our test is correlated in all but one case with two well known measures: accuracy (a measure of the completeness of sample) and purity (a measure of the ability of the sampling method to select nodes from the correct group). Although for completeness one of the correlations is very weak. We then compare subnetworks generated by two seed lists related to Parkinson's disease (PD), namely gene data from the OMIM

database [80], and a seed list derived from expression data of a PD cell model [36]. We find that the networks generated from the expression data seed list under the "all shortest paths between seed nodes" sampling scheme and under the "all paths up to 2" (including of paths of length 2) sampling scheme have significant results under our null model (although the latter is only partially robust to parameter choice), and therefore may have interesting properties for further analysis for our work on PD.

We demonstrate the effect of redundant seed nodes, first through simulations with randomly selected seed lists. Second, we investigate the effect in our two seed lists related to Parkinson's disease (PD), finding that redundant seed nodes can have a strong impact on the perceived significance of network statistics. We also demonstrate that our method compares favourably to the configuration model on this class of network sampling problems.

3.2 Methods

In this chapter we focus on a small set of sampling techniques and derived methods that are applicable to the wider class of sampling schemes involving a construction method and a seed list.

3.2.1 Network sampling and seed lists

The methods presented in this chapter focus on techniques to form subnetworks using a given seed list, where we use the following three sampling techniques:

1. **Snowball Sampling** includes all nodes that are less than a given graph distance from the nodes in a seed list; an example can be found in Fig. 3.1A. Depending on the implementation, the subnetwork can include only edges that were involved in the sampling process, or also include additional edges between

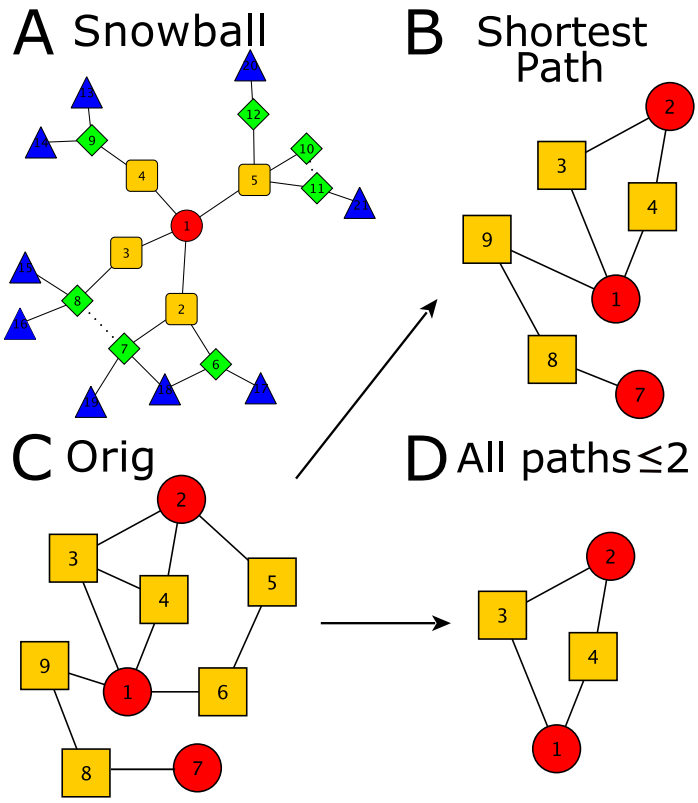

Figure 3.1: **A** 2-Hop Snowball Sampling Example: The seed list consists of node 1 (circle), node shape represents distance from seed protein square, representing nodes 1 hop from the seed, diamond 2 hops from a seed, triangle 3 hops from a seed. Dashed edges represent cross-edges in a 2-hop snowball sample. The network in (**C**) represents the unsampled network. **B and D** show the network in **C** sampled with the 'All Shortest Paths' (**B**) 'Paths≤2' (**D**) respectively. Seed nodes are represented by circles.

sampled nodes, which we call cross edges. In this chapter we write Snow1 for 1-hop snowball sampling, and Snow2 for 2-hop snowball sampling.

2. **All Paths $\leq$ k** (abbreviated Path2, Path3, Path4) includes all nodes and edges that are on a path between seed nodes that is less than or equal to k in length. An example can be found in Fig. 3.1D.

3. **Shortest Path Sampling** (abbreviated Shortest) includes all shortest paths between all pairs of seed proteins. An illustration can be found in Fig. 3.1B.

To illustrate the method, and following the approach of Ref. [135], we use a basket of commonly used network summary statistics, namely assortativity, average degree, clustering coefficient and number of nodes, using the following definitions. Other approaches are also available, see for example [168].

- **Assortativity:** The Pearson's correlation coefficient between the degree of nodes on either side of an edge;

- **Average Degree:** The mean number of edges per node;

- **Average local Clustering Coefficient:** The average of the local clustering coefficient of each node. The local clustering coefficient is defined as the number of triangles a node is involved in divided by the number of possible triangles (i.e. the number of pairs of edges that a node has).

- **Number of Nodes:** The number of nodes in the sampled network.

We choose these summary statistics, because they are commonly used and have low computational complexity. In the case that assortativity is not defined, for example because in the path $\leq k$ sampling method there are no paths $\leq k$ between seed nodes, or because all nodes in the sampled network have the same degree, the value

of assortativity is set to 0. Similarly, when there are no possible triangles (i.e. no nodes with degree greater than 1), the clustering coefficient is set to 0.

For Path≤k sampling, when a seed node is not connected to any of the other seed nodes with a path of length less or equal to k, this seed node is ignored for the calculation of the summary statistics. This choice is made to help interpret comparisons between subnetworks generated by different seed lists.

3.2.2 Network Data used in this document

To create our PPI network we use the yeast 2 hybrid (Y2H) experimental results in the BioGRID database version 3.4.127 [31, 161]. We remove all interactions that do not include a human protein. We then reduce the Y2H BioGRID network to its largest connected component, i.e. the graph formed from the largest group of nodes for which there is a path between any pair of nodes. The resulting network has 8,292 nodes, 25,062 edges, a density of 0.00073, and an average local clustering coefficient of 0.045.

3.2.2.1 Seed Lists

We compare two different seed lists for PD. For the first seed list, which we abbreviate *OMIM*, we assemble a list of genes known to be involved in the disease taken from the OMIM database [80]. We convert the genes to proteins in the BioGRID database using the relations provided in the BioGRID database [31, 161], resulting in a seed list with 16 proteins, of which 13 are present in our network, which form the OMIM seed list.

We construct the second seed list from differential expression data of 1185 genes in SH-SY5Y cells (a human cell line) before and after treatment with MPP+ (a toxin used as a model for PD) [36]. In [36] 313 genes were differentially expressed, 48 of which were deemed to be statistically significant. This list includes genes that are

up and down regulated by the cell when presented with MPP+. We convert the 48 significant genes to BioGRID gene identifiers using the BioGRID website. There are multiple mappings for some of these genes, resulting in 54 proteins, of which 46 are present in our network and these form the expression seed list.

It should be noted that two proteins in the expression list were not successfully identified at the time when the analysis for this chapter was performed. These two proteins were subsequently identified and are being included in follow on work.

Tables consisting of all of the proteins which we use in each of the seed lists can be found in the appendix in section A.1.

There is no overlap between the Expression seed list and the OMIM seed list.

3.2.3 Minimum seed lists and redundant seed nodes

We define a 'redundant seed node' as a node in a seed list that can be removed without changing the resulting subnetwork. For a given seed list we then define the (set of) minimum seed list(s) as the smallest non redundant subset (or set of subsets) of the original seed list which produces the same subnetwork.

As an example consider a triangle with nodes 1, 2 and 3. In 1-hop snowball sampling with a seed list consisting of all nodes, the set of minimum seed lists is $\{\{1\},\{2\},\{3\}\}$. In contrast, the set of minimum seed lists using the seed list $\{1,2\}$ would be $\{\{1\},\{2\}\}$. Note that $\{3\}$ is not present, as it is not a subset of the original seed list.

Computing the minimum seed list for a given subnetwork by considering all possible seed lists is computationally prohibitive. If we can guarantee that removing seed nodes does not add any previously unseen nodes or edges to the subnetwork (which all tested techniques in this chapter satisfy), we can use the procedure below:

1. Remove each seed in turn and check if the number of nodes and edges in the subnetwork do not change. If not, then add the node to the list of redundant

seeds.

2. Form a list of the remaining seeds.

3. Define a dummy variable L and set $L = 0$

4. For lists of redundant seeds of length L

   (a) Test if sampling with the list of the remaining seeds and the selected redundant seeds produce the same network.

   (b) Store all seed lists which pass the test.

5. If there are no seed lists which pass the test, set $L \to L + 1$ and go to step 4.

6. Return the smallest seed list(s) that produce the same network.

However, it should be noted that it is possible that there is a smaller seed list that generates the same network that is not a subset of the original seed list. As an advantage our technique is globally applicable and computationally tractable.

If the above procedure proves to be computationally prohibitive (which was not the case for results presented in this chapter), we can reduce the problem to a NP complete problem and use current best known algorithms to solve the problem. In the case of snowball sampling finding the global minimum seed list reduces to set cover. For further explanation and some other optimisations for this problem can be found in the appendix A.2.

3.2.4 The null model

For significance testing we would ideally want to use a parametric null model (in this case a parametrised random network ensemble), but currently no suitable parametric null model exists - for discussions on PPI networks see e.g. Ref. [2] or Ref. [139]. As an alternative we create a null model using an ensemble of subnetworks that have

been generated from a random seed list of the same size and (approximate) degree sequence as the original seed list. To adjust for redundancy, we use the smallest possible seed list that generates the same subnetwork. This model replicates the effect of the sampling procedure on the network.

It is difficult to construct analytic results, due to the dependence between seed nodes, while we can construct some results see Section 3.3, it is not computationally feasible to apply them to large networks. Thus we use Monte Carlo methods.

Monte Carlo Test We create a null model by estimating the underlying distribution using an ensemble of networks sampled using random seed lists of the same length and (approximate) degree distribution as the minimum seed list. We then calculate the p-values for the statistic of interest using a Monte Carlo test. Hence our procedure is as follows:

1. Construct the minimum seed list.
2. Generate many random seed lists with the same length and (approximate) degree distribution as the minimum seed list.
3. Generate a subnetwork for each seed list using the construction method under consideration (as described above).
4. Calculate the test statistic on each of the subnetworks.
5. Compute the p-value by counting the number of subnetworks with at least as extreme a test statistic as the subnetwork in question.

A p-value is defined as the probability, under the null model, of getting a value as least as extreme as the observed value. If $T(X)$ is a test statistic and we observe T_{obs}, then $p(T_{obs}) = P(T(X) \leq T_{obs})$.

Strictly enforcing the degree distribution may introduce problems in finite networks as there is a finite number of nodes of a given degree, possibly leading to a small number of random choices for some seed nodes. In order to alleviate this bias, the nodes are binned by degree from the left, such that each bin contains at least a predefined number of nodes, and the random selection of nodes is performed inside each bin. Stability testing is then performed over different bin sizes (5, 10, 20, 30 and 50) to guarantee that the result is robust to the bin size. Here we show results for bin size 20 only. Results for other bin sizes can be found in the appendix A.3; the conclusions drawn in this chapter are robust to the bin size unless otherwise stated. Thus, our method for constructing the null model specifies the (approximate) degree distribution and not the exact degree distribution.

3.2.5 Benchmarking the approach

To gauge where it is appropriate to use this method, we test when the method successfully selects subnetworks that better represent the network of interest on a simple benchmark network. We construct the benchmark network with known groups from a stochastic block model and then sample from it using a randomly selected seed list. We start with 4000 nodes, and assign the first 2000 nodes to Block 1 and the second 2000 nodes to Block 2. We place an edge between every pair of nodes in the same block with probability $p = 0.01$, and we place an edge between every pair of nodes in different blocks with probability $q = 0.001$. We randomly select 20 nodes from Block 1 to form the seed list. We sample a network using this seed list and the construction methods of interest; we record the p-value under the null model proposed in this chapter.

We then measure the success of the sampling by looking at the *accuracy* (a measure of the completeness of the sample) and the *purity* (also called precision, a measure of the ability of the sample to select nodes from the correct block) of the classification

which would classify all nodes in the sampled subnetwork as Block 1 nodes. We define C_1 as the set of nodes that are selected in the sample and C_2 is the set of nodes that have not been selected.

In this context we define accuracy A_{cc} as:

$$A_{cc} = \frac{|C_1 \cap \text{ Block 1}| + |C_2 \cap \text{ Block 2}|}{|C_1| + |C_2|} \tag{3.1}$$

and purity P_{ur} as:

$$P_{ur} = \frac{|C_1 \cap \text{ Block 1}|}{|C_1|}. \tag{3.2}$$

Due to computational demands of comparing these experiments over the ensemble, we restrict the minimum bin size to 20. Here we compare seed lists for a fixed method. For an exploratory analysis we can also compare different methods for a fixed seed list.

3.2.6 Assessing a null model fit

To evaluate the significance of any statistic with respect to the possibility of it being generated by random chance, the result must be compared against a credible null model.

One basic test of the applicability of a null model to a particular random process is to test if the distribution of p-values of randomly generated results is uniform provided that the null distribution is continuous. We can assess this hypothesis using the following procedure:

1. Create R random seed lists from a given network.
2. For each seed list create a subnetwork with the given technique.
3. Measure the statistic of interest on the subnetwork.
4. Use the null model of choice to calculate the p-value for the statistic of interest.

If T_{obs} is drawn uniformly at random from the distribution of T_{obs} and if this distribution is continuous, then under the null hypothesis the random variable $p(T_{obs})$ is uniformly distributed on $[0, 1]$. We can therefore assess the appropriateness of the null model by performing a χ^2 goodness of fit test on the distribution of p.

3.3 Analytic results

3.3.1 Number of nodes in a snowball sampled network

The interdependence between seed nodes severely limits the range of sampling techniques and statistics for which we can define tractable analytic expressions for the distribution of the statistic of interest over an ensemble. However, one case where we can derive an analytic solution is the number of nodes in snowball sampling with a hops with a seed list of size s.

The analytic results shown here have been extended from results in [61] and follow some of the notation. Note, Ref. [61] focuses on the classic problem of an unknown network with observed samples, in this case we have the related problem where we know that the network but we want to explore features of the samples. We define the function $h(M, s)$ as the probability that there does not exist a node within a hops of M on a randomly selected seed list of size s. Let $B_a(M)$ be the set of nodes within a hops of the nodes in M. If the selection is uniformly at random from all subsets of nodes of size s then we can use a hypergeometric distribution to derive an expression for $h(M, s)$. We can do so as follows, we randomly select s seeds from V, however if we select a seed from $B_a(M)$ then at least one node in M will be included. Thus we require the seeds to all be chosen from $V \setminus B_a(M)$. Thus through the hypergeometric

distribution this results in the following form for $h(M, s)$:

$$h(M,s) = \frac{\binom{|V|-|B_a(M)|}{s}\binom{|V|-(|V|-|B_a(M)|)}{s-s}}{\binom{|V|}{s}}, \tag{3.3}$$

which simplifies to:

$$h(M,s) = \frac{(|V|-s)!(|V|-|B_a(M)|)!}{(|V|-|B_a(M)|-s)!|V|!} = \prod_{i=0}^{s-1} \frac{|V|-|B_a(M)|-i}{|V|-i}, \tag{3.4}$$

If we wish to fix the degree sequence of the seeds (or with a small adjustment binned degree), we can use a similar approach to the uniform case, to derive an expression for a degree sequence version, $h_d(M, t)$ where t is the degree sequence. This results in the following form for $h(M, t)$:

$$h_d(M,t) = \prod_{u \in U(t)} \prod_{i=0}^{F(t,u)-1} \frac{D(V,u) - D(B_a(M),u) - i}{D(V,u) - i}, \tag{3.5}$$

where t is the degree sequence of the seed list. Here $F(t, u)$ is a counting function, it counts how many instances of u there are in t (e.g. $F([1, 2, 3, 4, 1], 1) = 2$), $U(t)$ returns the unique elements of l and $D(M, r)$ is the number of elements in M of degree r. The seeds of different degrees are selected independently so we can calculate the probability for each unique degree in the seed degree list and then multiply them to get the final probability.

Deriving the Mean and Variance If we let X_i be an indicator variable for the presence of node i in the sample, then the number of nodes in a sample is $Y = \sum_i X_i$. We can compute the mean number of nodes as follows:

$$E[Y] = E[\sum_i X_i] = \sum_i E[X_i] = \sum_i 1 - h(\{i\}, s) = |V| - \sum_i h(\{i\}, s). \tag{3.6}$$

To compute the variance we can use the correlated variables formula to get:

$$\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j). \tag{3.7}$$

We can imagine each X_i as being a Bernoulli random variable with $p = 1 - h(\{i\}, s)$, therefore,

$$\mathrm{Var}(X_i) = p(1-p) = h(\{i\}, s)(1 - h(\{i\}, s)). \tag{3.8}$$

The covariance between the two variables is defined as:

$$\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]; \tag{3.9}$$

thus:

$$E[X_i X_j] = E[(1-(1-X_i))(1-(1-X_j))] = 1-E[1-X_i]-E[1-X_j]+E[(1-X_i)(1-X_j)] \tag{3.10}$$

$$= 1 - h(\{i\}, s) - h(\{j\}, s) + h(\{i, j\}, s). \tag{3.11}$$

Combining and simplifying all of the terms we obtain:

$$\mathrm{Var}(\sum_i X_i) = (\sum_i h(\{i\}, s) - h(\{i\}, s)^2) + 2(\sum_{i<j} h(\{i, j\}, s) - h(\{i\}, s)h(\{j\}, s)). \tag{3.12}$$

We can then factor the expression to obtain:

$$\mathrm{Var}(\sum_i X_i) = \sum_{\alpha=1}^{2} \alpha \sum_{\substack{L \subseteq V \\ |L|=\alpha}} h(L, s) - (\sum_i h(\{i\}, s))^2. \tag{3.13}$$

where L is a dummy summation variable.

Figure 3.2 illustrates the effect of seed list size on the distribution of the number of nodes in a 1-hop snowball sample in the BioGRID PPI network [31, 161]. As we

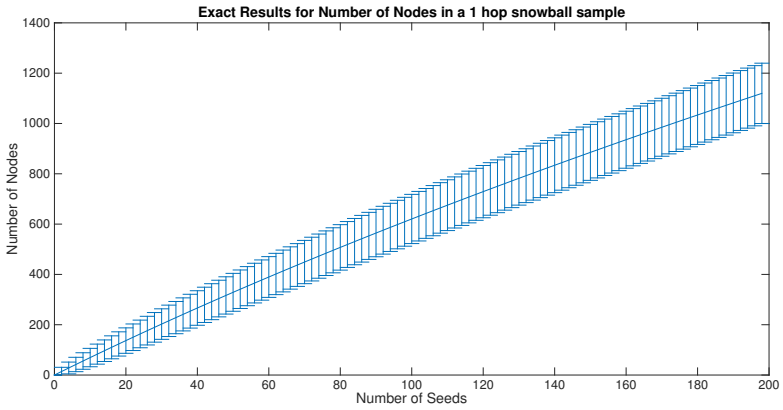

Figure 3.2: Exact Results: Points represent the mean number of nodes given a number of seed nodes in a 1-hop snowball sample from the BioGRID PPI network. The error bars represent one standard deviation of the same quantity.

do not have a seed list of interest in this case, we have assumed that there is no restriction on the degree sequence of the seed nodes.

As expected, the larger the number of seed nodes, the larger the average number of nodes in the resulting network. Further, a small change in the number of seed nodes can have a large impact on the expected size of the network. The number of nodes in a 2-hop snowball sample on, say 20 proteins from the PPI network may appear small when compared to subnetworks randomly generated from 30 seeds but large when compared to such subnetwork generated from 10 seeds.

3.3.2 Finding redundant seed nodes: testing for isomorphism between sampled graphs

To discover redundant seed nodes we need to be able to guarantee that a pair of labelled networks are identical. The trivial way to do this is to compare the node lists and the edge lists, and if they are equal then the networks are equal. As we will have to check equality a large number of times, this approach can can prove computationally constraining. Making use of the fact that we are adding or removing

nodes from the seed list, we can derive a simpler condition for this problem.

We take an arbitrary network with node set V and edge list E and a seed list l_1. Let $[V_{l_1}, E_{l_1}] = f_{arb}(V, E, l_1)$, where f_{arb} is a arbitrary network sampling function, l_1 is a seed list and V_{l_1} and E_{l_1} are the subset of nodes and edges that are in the subnetwork.

Let us assume that we have two seed lists l_1 and l_2 and $l_1 \subseteq l_2$. Further, let us assume that our sampling technique (f_{arb}) guarantees that $V_{l_1} \subseteq V_{l_2}$ and $E_{l_1} \subseteq E_{l_2}$ if $l_1 \subseteq l_2$.

Trivially for sets S_1 and S_2 if $S_1 \subseteq S_2$ and $|S_1| = |S_2|$, therefore $S_1 = S_2$. Thus, if $|E_{l_1}| = |E_{l_2}|$ and $|V_{l_1}| = |V_{l_2}|$ then $E_{l_1} = E_{l_2}$ and $V_{l_1} = V_{l_2}$. Therefore we can use the condition:

$$|E_{l_1}| = |E_{l_2}| \text{ and } |V_{l_1}| = |V_{l_2}|. \tag{3.14}$$

Note, we must condition both on the set of edges and the set of nodes as they are both required to fully define the subnetwork.

Therefore if we can guarantee using our sampling techniques that if $l_1 \subseteq l_2$ then $V_{l_1} \subseteq V_{l_2}$ and $E_{l_1} \subseteq E_{l_2}$ then we can simply test for equality in the number of nodes and edges.

Snowball Sampling In snowball sampling the contribution from each of the nodes on the seed list is independent, as it is simply the number of nodes within a certain radius of the each of the seeds. Therefore the expression for the V_{l_1} is as follows:

$$V_{l_1} = \bigcup_{x \in l_1} V_x. \tag{3.15}$$

If we take $l_1 \subseteq l_2$, then $V_{l_2} = V_{l_1} \cup V_{l_2 \setminus l_1}$ and therefore trivially $V_{l_1} \subseteq V_{l_2}$. We can use the same argument for E_{l_1}. We can therefore use the condition in (3.14) for snowball sampled networks given that the seed list is a subset or a superset of the original seed list.

Deterministic Path Based Sampling Techniques We define a deterministic path based sampling techniques in undirected networks as a procedure for which the sample as is function of every pair of nodes on the seed list and the underlying network. Note that Path≤k and shortest path sampling both fall into this category. Therefore,

$$V_{l_1} = \{x \in P_{arb}^{(V)}(y,z) : y,z \in l_1\}, \text{ and } E_{l_1} = \{x \in P_{arb}^{(E)}(y,z) : y,z \in l_1\}, \tag{3.16}$$

where $P_{arb}^{(V)}(y,z)$ and $P_{arb}^{(E)}(y,z)$ are the set of the sampled nodes and edges respectively on the path(s) between y and z defined by the sampling technique in question. Let l_1 and l_2 be seed lists and let $l_1 \subseteq l_2$, then

$$V_{l_2} = V_{l_1} \cup V_{l_2 \setminus l_1} \cup \{x \in P_{arb}^{(V)}(y,z) : y \in l_1,\ z \in l_2 \setminus l_1\} \cup \{x \in P_{arb}^{(V)}(y,z) : z \in l_2,\ y \in l_2 \setminus l_1\}. \tag{3.17}$$

where the right hand side is the union of the nodes included by the seed list l_1 (first term), the nodes that are included by the seed list $l_2 \setminus l_1$ (second term), and the third and fourth terms represent the contributions from paths which connect a node from l_1 to a node from $l_2 \setminus l_1$. Therefore $V_{l_1} \subseteq V_{l_2}$, and by similar argument $E_{l_1} \subseteq E_{l_2}$. Thus we can use the condition in (3.14) for all deterministic path sampling techniques.

3.4 Simulation results

3.4.1 Evaluation on the benchmark data

To test whether there is a negative relationship between the p-value of our test and accuracy or purity, we use Kendall's τ statistic which is a measure for association. The value is in $[-1, 1]$; the closer to ± 1; the stronger the association. We measure Kendall's τ with respect to the minimum of the p values of the two tails, as we do

not specify in which direction the statistics differ.

Method	Kendall's τ for Accuracy	Kendall's τ for Purity
Snow1	-0.231	-0.184
Snow2	-0.115	-0.113
Shortest	0.534	-0.222
Path2	-0.594	-0.022
Path3	-0.265	-0.173
Path4	-0.113	-0.112

Table 3.2: Kendall's τ for the relationship between test p-value and accuracy or purity in the benchmark data set. The benchmark is a stochastic block model, consisting of two blocks of size 2000 with an internal connection probability of 0.01 and an external edge probability of 0.001. Further, details of this benchmark can be found in section 3.2.5

The results in Table 1 show that for all of the single construction methods except for the Shortest construction method, there is the desired negative association for both methods (the smaller the p-value, the better the assignment to the block). Although, in the case of the Path2 method the correlation with purity is small, however the correlation with accuracy is much stronger.

Note, here we compare seed lists for a fixed method. For an exploratory analysis, we can also compare different methods for a fixed seed list. Further, we note that the results obtained here are not independent of the parameter choices.

For differentiating between subnetwork construction methods we also investigated the trade-off between accuracy and purity. The results in Figure 3.3 show that care must be taken in selecting the correct construction method for the problem at hand by considering the trade off between purity and accuracy of each of the methods. The Path4 method can achieve high accuracy but does not achieve high purity, while Path2 achieves the highest purity overall, but has low accuracy.

The behaviour of the parameters in these sampling techniques is in line with what we would expect. In the case of snowball sampling, we observe that increasing the number of hops from 1 to 2 increases the accuracy but decreases the purity. In

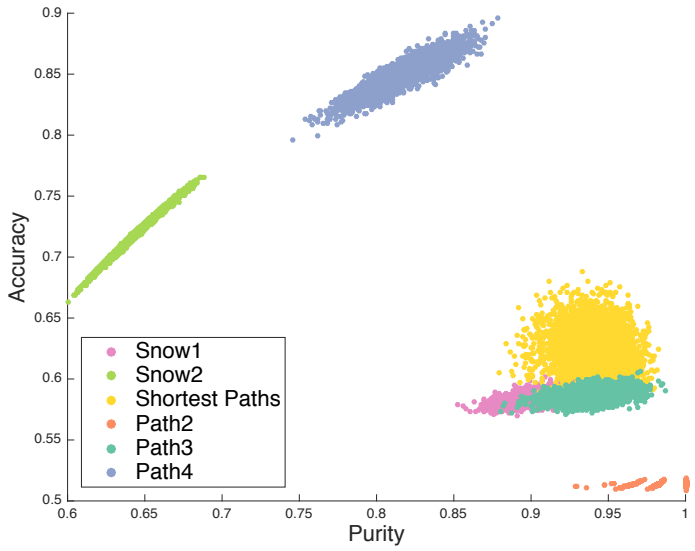

Figure 3.3: A scatter diagram of accuracy versus purity of benchmark networks in which the sample is significant under our test ($P < 0.05/8$ due to a 2-tailed adjustment and a Bonferroni correction) where colour represents the construction method used. An ideal method would have accuracy = 1 and purity = 1.

the stochastic block model the immediate neighbourhood of the seed nodes is small (and thus results in a subnetwork with a small number of nodes), but a much higher percentage of the nodes will be in the same group at the original node. Thus in 1-hop snowball sampling there is high purity as the immediate neighbourhood is mostly the same group, but low accuracy as this immediate neighbourhood is relatively small. By increasing the number of hops from 1 to 2 this lowers the purity as the further we are from the original seed list, the less likely the nodes are to be in the same group, but accuracy increases since as simply more of the Block 1 nodes are included.

The relative behaviour of Path2, Path3 and Path4 can be explained by the same mechanism. Increasing the path length increases the likelihood of including more of the Block 1 nodes, thus increasing accuracy, but this also increases the likelihood of including nodes that are not part of Block 1 thus lowering purity.

3.4.2 Comparing sampling methods and seed lists for PD

When trying to construct a subnetwork which reflects a disease process, one is faced with a plethora of choices. In order to address this problem in our work on Parkinson's disease (PD) we compare how far the subnetworks deviate from random according to the null model as described in Section 3.2.4. While we do not know if the subnetwork that deviates the most from random will contain more (or less) biological information than other subnetworks, it is possible that there are certain subsets of the sampling techniques described in Section 3.2.1 that identify interesting structural features which may also be biologically meaningful. As we cannot test all possible summary statistics, we use the statistics described in Section 3.2.1 as a comparison.

To illustrate our approach we compare our two different seed lists for PD, (OMIM and Expression see Section 3.2.2.1 for details), across all of our sampling techniques and a reasonable parameter range.

To contrast the different sampling techniques, we compute the significance of all of the statistics in our set and select the smallest p-value. We test in both tails, at significance level 0.025, and as we compare 4 statistics we apply a Bonferroni correction resulting in a significance test at level $0.025/4 = 0.00625$. The significance results for the OMIM seed list and the expression seed list can be found in Fig. 3.4 respectively. The p-values are plotted on a log-scale: the higher the box, the smaller the p-value. The 0.00625 threshold is marked by a red line.

Two networks show promising deviation from randomly sampled networks. In the expression seed list Path2 and shortest path are significant, the Path2 method is robust in all but one bin size (50) and the shortest path is robust in all bin sizes. In the OMIM seed list while Snow1 is not significant it is approaching significance with a p-value of 0.0063 and as such may deserve further consideration. While we cannot claim that the other networks do not have any information about the disease condition, the significance of these networks suggests that these may be a good networks on which

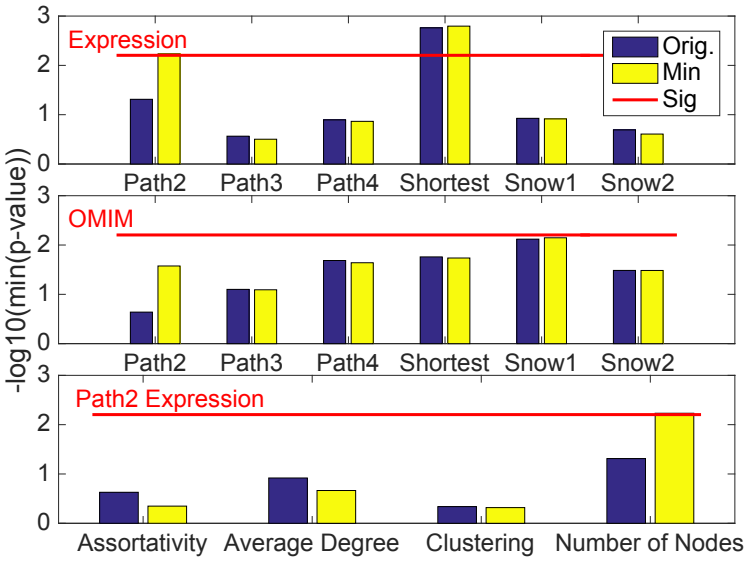

Figure 3.4: Test results for different seed lists: smallest p-value, on a negative log scale. Results are shown for the OMIM seed list (first panel); Expression seed list (middle panel); and a breakdown of the p-value for the 4 statistics evaluated for the Path2 Expression network (final panel). Blue: original seed list; yellow: minimum seed list; red: significance level (0.025/4). Note due to the negative log scale on the y axis, values above the red line are significant.

to focus in depth analysis.

We also explored how many networks in each sampling technique have assortativity values which are assigned a value of 0. Most construction methods very rarely experience this, however 11% of the OMIM Path2 Monte Carlo test null network ensemble generated and 27% of the minimum OMIM Path2 seed list Monte Carlo test null network ensemble have assortativity values that are set to 0. This is mostly due to the short seed list.

In view of Figure 3.3 which shows poor accuracy for Path2 sampling, our preferred subnetworks are the networks created from the Expression seed list via all shortest paths and the OMIM seed list via Snow1. Some summary statistics for these networks can be found in Table 3.3.

	Expression Shortest Path	OMIM Snow1	BioGRID
Number of Nodes	1383	134	8292
Number of Edges	4252	168	25062
Density	0.0044	0.0189	0.00073
Average local clustering coefficient	0.044	0.044	0.021

Table 3.3: Summary statistics for a selected set of sampled networks we have selected for further study.

3.4.3 Redundant seed nodes in PPI networks

As our null model starts with random seed lists of the same length and (approximate) degree distribution as the chosen minimum seed list, our test relies crucially on a minimum seed list. Without reducing the original seed list to a minimum seed list, the test results could be very different - we call these resulting p-values *perceived p-values*, the p-values which we would perceive if we were not to correct for redundant seeds.

To demonstrate the effect of redundant seed proteins on perceived p-value of network features, first we add redundant seed nodes to randomly selected seed lists in

the BioGRID PPI network, and second we compare the perceived p-value on the networks based on PD seed lists. We illustrate our results for assortativity, average degree, clustering and number of nodes.

We investigate the importance of accounting for redundant seed proteins by comparing the significance of two seed lists that generate the same network. We construct an ensemble of random seed lists of length 25 sampled uniformly at random from all possible seed lists. For each seed list, we construct the longest seed list that generates the same network. The algorithm used to construct the longest possible seed list can be found in Section A.2. For simplicity in cases where there is more than 1 longest seed list we select one randomly. We compute the difference between the perceived left p-value in the original seed list and the left p-value of the longest seed list. If there is little difference we would expect the results to be close to 0.

On the BioGRID PPI network, (Fig. 3.5) we observed a large difference in p value in some techniques. For example in two hop snowball sampling we observe a large change in all statistics. Further, while there are some techniques that are more robust to these changes it is clear that the size of effect in many of the techniques is large. Fig. 3.4 shows that while adjusting for minimum seeds often does not make a large difference to perceived p-value, in the case where it does (Fig. 3.4 Expression Seed list Path2), the change can be large.

3.4.4 Comparison with the configuration model

A popular null model in network science is the configuration model, which has been widely used as a null model across application domains. In the configuration model, the network is rewired randomly while preserving the degree distribution of the network [127].

The configuration model may not preserve the structural features of the original network. For example in the 2-hop snowball sample in Fig. 3.1A, there is a very

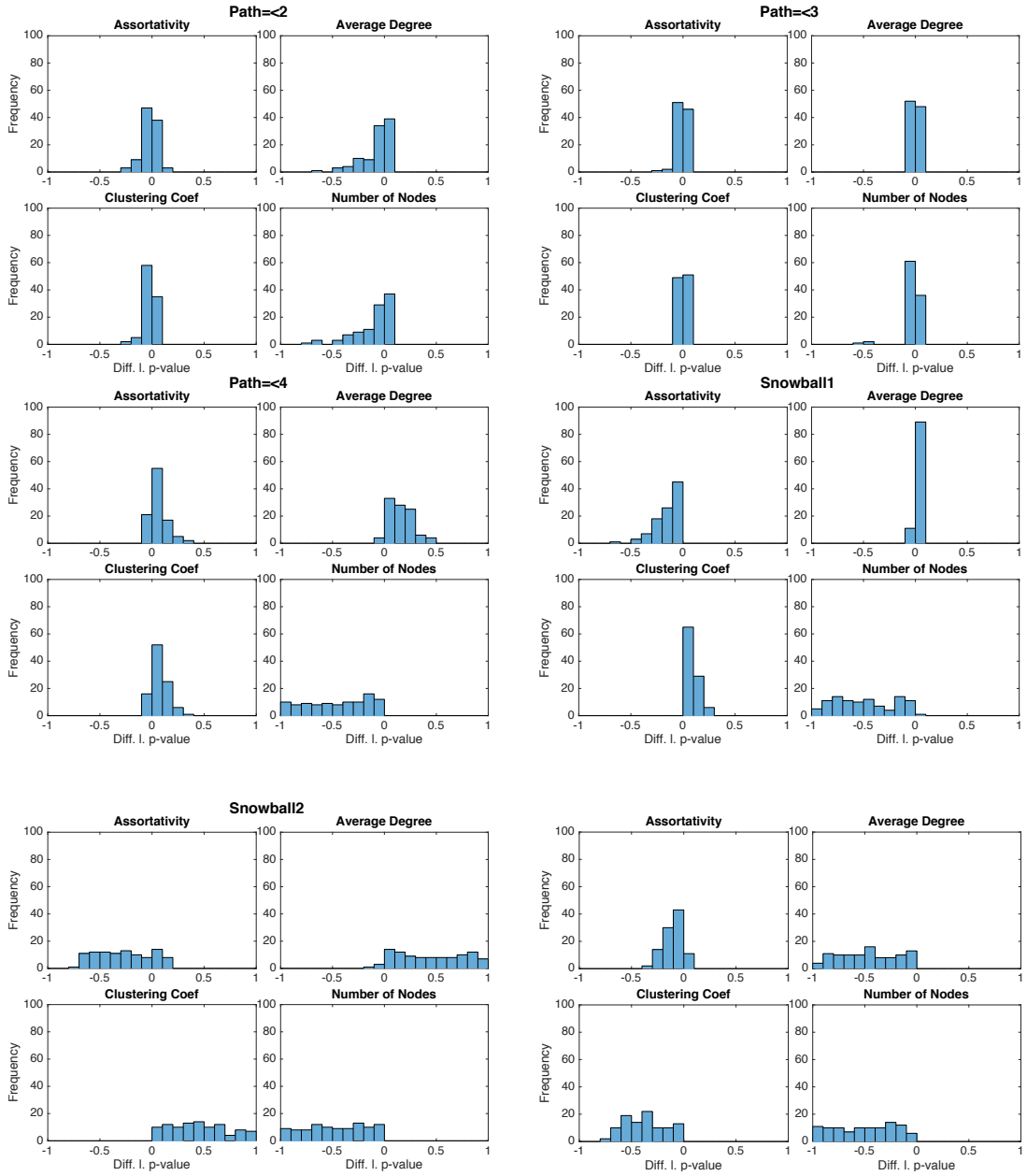

Figure 3.5: Histogram of differences in left p-values between networks sampled from BioGRID with 25 initial random seed nodes and the longest possible seed list that will generate the same network. The histogram has 100 observations and the panels represent the different techniques. For most sampling methods and summary statistics the difference of p-values can be close to 1 or -1, and hence redundant seeds can have a large influence.

clear structure, with a maximum path length of 4 between any pair of nodes. This structure will not be preserved in the configuration model. Fig. 3.6 shows a simple comparison between the distribution of shortest path length in the snowball sampled network compared against an ensemble of configuration models of the same network. We also compared the configuration model with all of the other sampling techniques using the method described in Section 3.2.6. The results can be seen in Figure 3.7. We use a χ^2 test to compare the distributions with the uniform distribution taking as observations the p-values of the statistic of interest from 1000 networks generated by selecting 25 random seeds (Table 3.4). We see that in all sampling techniques we reject the null hypothesis for the configuration model and for Snow1, Snow2 and all shortest paths we do not reject the null hypothesis for our null model. For Path3 in clustering we reject the null hypothesis at the 5% level but not at the 1% and further visual inspection of the distribution also does not draw any concern.

In the case of the clustering in Path2, triangles can occur only when two seed nodes are less than or equal to 2 hops apart and one of them is part of a triangle with $8,292$ nodes and an average shortest path length of 4.35 this is unlikely to happen. For non-continuous distributions, rather than the uniform distribution, we would expect to see the generalised inverse of the cumulative density function as distribution of the p-values. Fig. 3.8 shows that the distribution of average local clustering coefficients for Path2 networks sampled with 25 random seeds is indeed discontinuous and therefore we should not be surprised to see a non-uniform null distribution. In contrast with the same distribution from Snow1 shown in Fig. 3.9 is approximately continuously distributed and therefore the uniform distribution appears as the null distribution as expected.

While we cannot generalise from these results to all possible networks ensembles, and it is highly likely that there are network models and parameters ranges where the configuration model performs well in subnetworks, the configuration model does

	Our Null		Configuration model	
	Assortativity	Clustering	Assortativity	Clustering
Path3	0.3115	0.0255	0	0
Path4	0.9659	0.0734	0	0
Shortest	0.8343	0.4788	0	0
Snow1	0.4654	0.4788	0	0
Snow2	0.2380	0.9522	0	0

Table 3.4: Goodness of fit χ^2 p-value results for different sampling techniques under our null model and the configuration model. 1000 networks are generated by selecting 25 random seeds and sampling with the given technique. Assortativity and average local clustering coefficient are calculated on the resultant network. We measure the p-value of these statistics under our null model and the configuration model. If the null hypothesis is correct and the distribution under the null hypothesis is continuous then these p-values, viewed as random observations, should be approximately uniformly distributed on the interval $[0, 1]$. This table gives the results from a χ^2 test for goodness to the uniform distribution. The results for the configuration model indicate that the distribution of p-values is far from uniform indicating that the configuration model is not a good null model , whereas in our model we do not reject the null hypothesis, lending support to its use as a null model.

not perform well in general when comparing subnetworks based on seed lists. This demonstrates the need for an alternative to the configuration model for this task.

3.5 Discussion

There is a need for a robust and reliable nonparametric test when testing the significance of summary statistics for sampling techniques based on seed lists. Depending on the research question the configuration model does not fulfil this role. Here we propose an alternative null model that is based on an ensemble of seed lists generated from the minimum seed list.

We focus on the significance of network features, given a construction method, rather than given a construction method and a fixed seed list, as different seed lists may result in the same subnetwork.

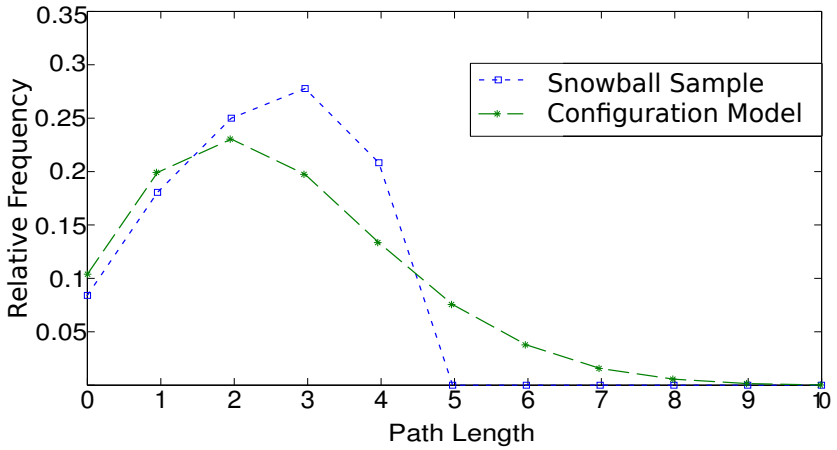

Figure 3.6: Comparison of distribution of shortest path lengths in the original network and over an ensemble of networks generated by a configuration model from the same network.

.

We have demonstrated that accounting for seed list construction is important, by artificially increasing the significance of a randomly chosen seed lists in a biological network, and through observing the effect of this increase on the biologically motivated seed lists.

We have also shown through our benchmark that p-values from our test are negatively correlated in all but one case with measures of purity and accuracy of the sample (i.e. on average small p-values result in more accurate/pure networks).

Our null model is not without issues. Notably, it is rare but possible for there to be more than one minimum seed list which then requires a comparison with multiple seed lists. A further problem is that the seed list does not have to be a global minimum; it is possible that there is a seed list that is smaller than the supposed 'minimum seed list'. Finding this minimum seed list for an arbitrary technique is computationally prohibitive. We believe that the very tractable null model presented in this chapter is superior to the model based on a globally minimum seed list, due to its applicability to many different problems.

For PPI networks, our nonparametric test allows us to choose a subnetwork which

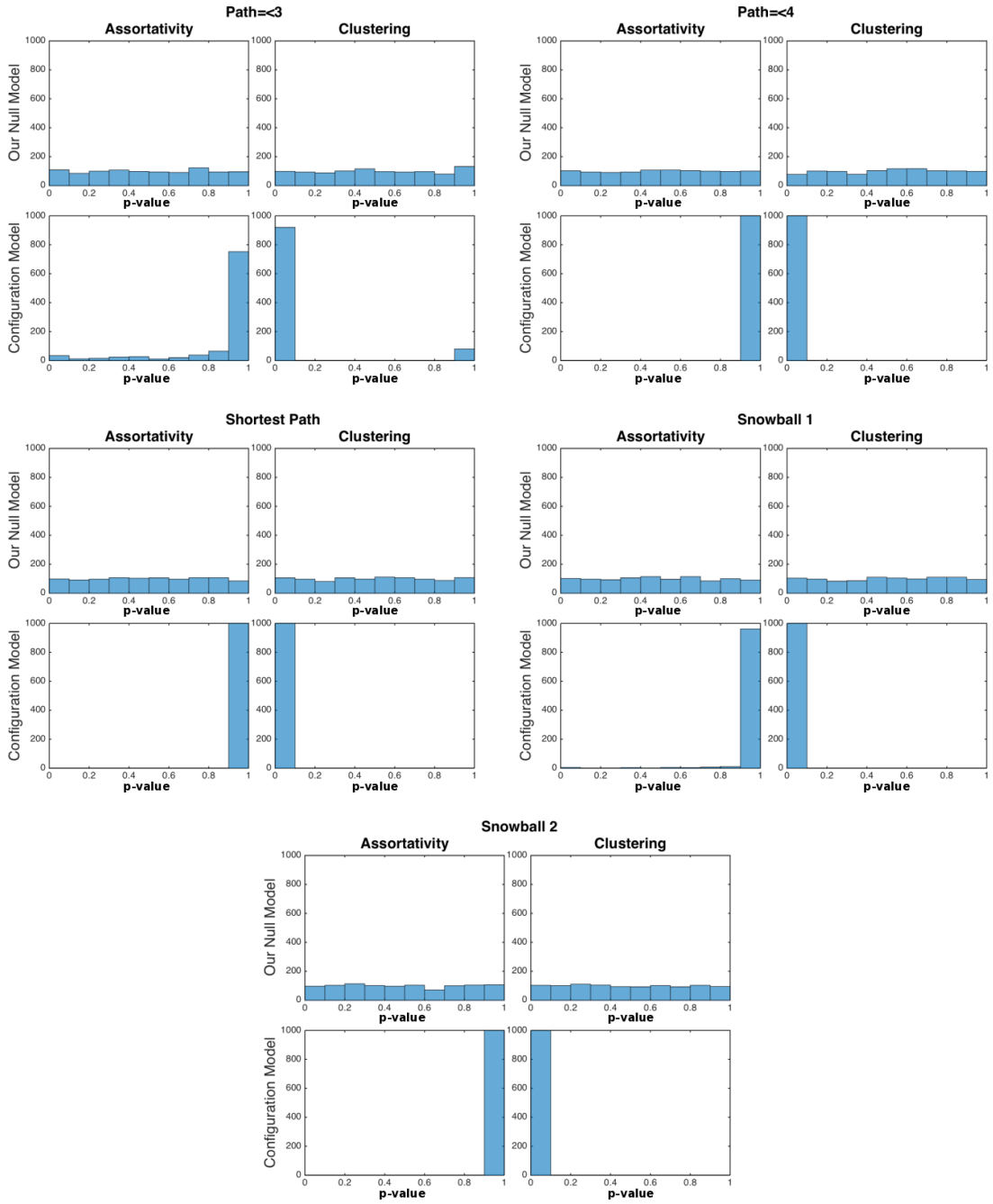

Figure 3.7: Distribution of p-value results for different sampling techniques under our null model and the Configuration Model. 1000 networks are generated by selecting 25 random seeds, assortativity and average local clustering coefficient are calculated. For each network we observe the p-value for deviation from a random network (our null model and the configuration model). For our null model the p-values are approximately uniformly distributed, indicating a reasonable fit of the null model. For the configuration model the distribution is far from uniform, indicating a poor fit.

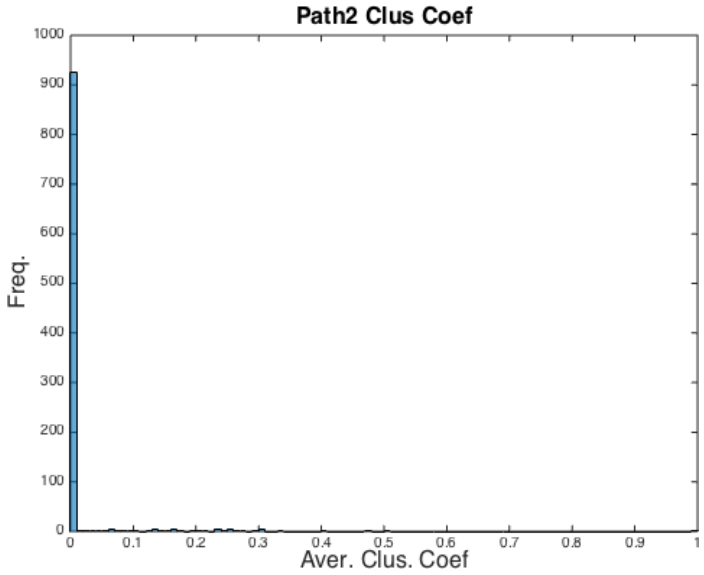

Figure 3.8: Distribution of Average Clustering Coefficient Path2 sampled from BioGRID network with 25 randomly chosen seeds. There is a strong concentration at 0 and a few discrete values which are not equal to 0.

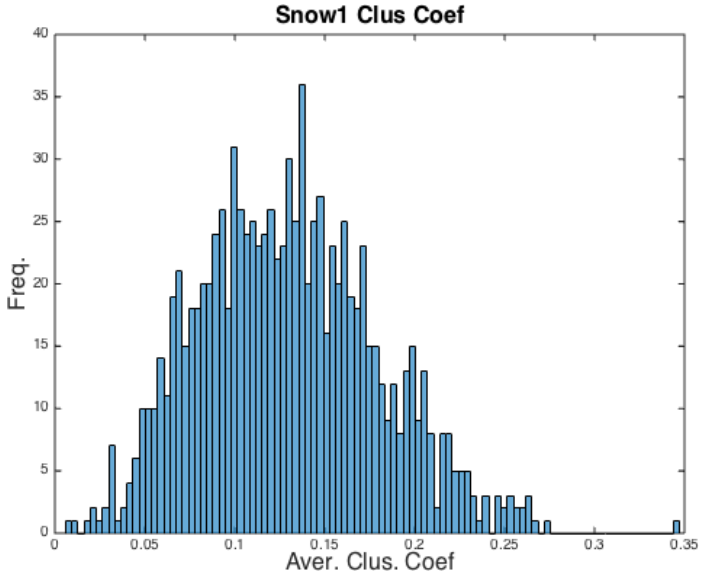

Figure 3.9: Distribution of Average Clustering Coefficient Snow1 sampled from BioGRID network with 25 randomly chosen seeds.

may have interesting properties for further analysis for our work on PD. On the statistics tested many of the generated networks do not appear to deviate significantly from random, unlike the results from the expression seed list using Path2 and shortest paths. Our work also highlights the need to focus more attention on generative models of biological networks in order to generate parametric models of these systems.

Another possible application of these approaches is to learn about seed lists directly rather than attempting to select sampled networks. We could do this in a number of ways. First, we could use the number of redundant seed nodes in a seed list as a test statistic and compute the significance using the random seed list approach that we used in this chapter. If we used a biologically motivated sampling scheme, this could allow us to select between seed lists. Second, inspired by Ref. [68] we could also attempt to find additional nodes that should be present on a seed list by looking for nodes that are redundant with respect to a sampling scheme. However, care would have to be taken in the selection of the sampling technique as the results would be greatly affected by this choice. Finally, we could attempt to classify seed nodes with respect to the nodes and edges they add to the sampled network to construct (possibly overlapping) groups of seed nodes that are similar with respect to the chosen sampling technique.

Chapter 4

Path-based techniques for community detection on sampled networks

Symbol List

Symbol	Description
A	Adjacency matrix
$A_{ij\Upsilon}$	The number of non-self intersecting paths between nodes i and node j of length Υ
B_n	The nth bell number
C	a set of nodes belonging to a community
D	Diagonal matrix with the degree of each node on the diagonal i.e. $D_{ii} = k_i$
E	Set of edges in the graph
$F(d)$	Expected number of connections at distance d
$H(X)$	The entropy - a measure of the uncertainity associated with a random variable.
I	Identity matrix

$MI(x, y)$	The mutual information- a measure of the amount of information about x and y that cannot be gain by considering the other
M_{Katz}	A matrix related to Katz similarity and centrality.
N_i	Population of region i
N_{et}	An example network
$P^{(1)}, P^{(2)}$	Fixed partitions also known as a vector of community assignment
$\mathfrak{P}_{ab}^{N_{et}}$	Average number of edges between a node of degree a and a node of degree b in the network N_{et}
P_{ij}^{SPA}	The expected number of edges under a spatial null model
P_{ij2}^{inde}	The expected number of paths of length 2 if we assume that all edges are independent
$P_M(C, t)$	Probability that a random walker is in a set of nodes C representing a community at time 0 and at time t with Markov process M
$\mathfrak{P}_{ab}^{(f_{samp})}(s), \mathfrak{P}_{ab}^{(f_{samp})}$ $\hat{\mathfrak{P}}_{ab}^{(f_{samp})}$	The expected number of connections between a node of degree a and b in a network sampled with f_{samp} and the seed list s. $\hat{\mathfrak{P}}_{ab}^{(f_{samp})}$ is the estimate of the same quanity over a number of samples.
P_{art}	A set of sets of nodes in each community.
$P_{ij\Upsilon}$	The expected number of paths of length l between i and j
P_{ij}	The expected number of edges between node i and node j
$Prob(X = 1)$	Probability that (for example) the random variable X is 1
Q	Newman Girvan Modularity
Q_1^{Path}, Q_2^{Path}, Q_3^{Path}, Q_l^{Path}	Path modularity using paths of length 1, 2, 3 and l respectively
Q^{var}	Modularity where we let $P_{ij} = P_{ij}^{(f_{samp-rescale})} + \sigma_{ij}^{f_{samp-rescale}}$
Q^{dist}	Modularity which scales based on the distribution of connection probabilities over the ensemble.
Q^{dist1}	Modularity which scales based on a unrescaled variance, see Q^{var} for final version of the model.
$R_M(t)$	The sum of the $P_M(C, t) - P_M(C, \infty)$ over all of the communities in the network.

$R_n(t), R_c(t)$	R_M using the Markov process described in the Combinatorial and Normalised Laplacians
$S_a^{N_{et}}$	Set of all nodes in the network N_{et} with degree a
V	Set of all nodes in the graph.
$VOI(x, y)$	The variation of information - a measure of the amount of additional information in and x and y that cannot be gain by considering the other
$W^{(M)}$	Random walker under markov process M
$W_t^{(M)}$	Random variable representing the position of the random walker at time t
X	A discrete uniform random variable from 1 to n
Z	Number of seeds selected uniformly at random in constructing random seed lists.
γ	Resolution parameter associated with the variance, note that $\gamma \in \mathbb{R}$ unlike $\lambda \in \mathbb{R}^+$
$\mathfrak{B}_{in}$	A collections of degrees, used to combine observation of different degrees to improve the accuracy of the estimation.
$\hat{\mathfrak{P}}_{ab}^{(N_{et})}$	The probability of observing a connection between a node of degree a and a node of degree b in the network N_{et}
$\hat{\mathfrak{P}}_{ab}^{(f_{samp}-rescale)}$	The estimated expected number of paths rescales to that the expected number of edges is equal to the observed number of edges in the graph
$\hat{\mathfrak{P}}_{abl}^{(f_{samp}-rescale)}$	The estimated expected number of paths of length l between a node of degree a and b rescaled to the number of observed paths in the network.
$\hat{\sigma}_{k_i k_j}^{(f_{samp}-rescale)}$	The rescaled version of the $\hat{sigma}$
λ	The resolution parameter in community detection
λ_1, λ_2	Resolution parameters for path modularity
$\mathbf{p}$	Vector of probabilities of a random walker being at each node
$\mathfrak{N}(s)$	The set of all sampled networks with seed list of size $\|s\|$.
$\mathfrak{N}_a(s)$	The set of all sampled networks with seed list of size $\|s\|$ which has at least one node of degree a

$\mathfrak{N}_{ab}(s)$	The set of all sampled networks with seed list of size $\|s\|$ such that if $a \neq b$ there is at least one node of degree a and one node of degree b, or if $a = b$ then at least two nodes of degree a.
$\mathfrak{N}_a^{(sample)}(s)$	A sample of the element of $\mathfrak{N}_a(s)$ constructed by randomly sampling from $\mathfrak{N}(s)$ and selecting the networks which match the conditions detailed in $\mathfrak{N}_a$
$\mathfrak{N}_{ab}^{(sample)}(s)$	A sample of the element of $\mathfrak{N}_{ab}(s)$ constructed by randomly sampling from $\mathfrak{N}(s)$ and selecting the networks which match the conditions detailed in $\mathfrak{N}_{ab}$
a, b	Arbitrary degrees used as indices for certain matrices
c	A community vector, such that c_i is the community id which is assigned to node i
c_{in}, c_{out}	Probability of connections between node in the same group and in different groups respectively
c_{max}	The vector of community assignment which maximises the modularity
d_{ij}	The physical distance in space between point i and point j
f_{samp}	An arbitrary sampling technique
k_i	Degree of node i
m	Number of edges in the network
n	Number of nodes in the network
$q_{ab}(\lambda)$	The value required such that $\frac{\lambda}{2}$ percent of networks in $\mathfrak{N}_{ab}(s)$ will have $P_{ab}^{(N_{et})} < q_{ab}(s)$
r	The amount that someone is more likely to make friends inside a class
s	The original seed list
w	The arbitrary degree of a node
$\sigma_{ij}^{(f_{samp})}, \hat{\sigma}_{ij}^{(f_{samp})}$	Standard deviation and estimated standard deviation of the expected number of edges between node i and node j

4.1 Introduction

A central motivation of this dissertation is to construct a biologically meaningful way of interpreting a subnetwork related to Parkinson's disease, which can then be used to inform network pharmacology.

In this context, community detection can be of help. Community detection methods partition the network into groups which are denser internally and sparser between groups than would be expected at random. These groups could then form the input for further analysis in network pharmacology, for example selecting proteins that optimally interrupt the edges between communities.

Community detection algorithms are usually based on topological features of the network, yet the nodes within the detected groups (communities) often also have similar properties or attributes. Community detection can also be used as a basis for identifying the roles of nodes with respect to different communities [75].

Some examples for successful community detection include: identifying protein complexes (proteins that combine together for long periods of time to perform a specific function) in PPI networks [160]; identifying metabolites that are more conserved across species in metabolic networks [75]; identifying missing interactions in PPI networks [77]; showing that communities in PPI networks are related to biological function [108]; identifying spatial and linguistic divides in countries [53]; showing that communities in Facebook networks of different universities can be organised by different underlying factors [170, 171]. For more applications of community detection see the survey in Ref. [57].

The underlying assumption behind community detection is that sets of nodes that have more edges between them than would be expected at 'random', given a particular definition of random, are interesting. However, the appropriate choice of 'random' reference state often referred to as a null model is very important. This basic principle can be illustrated by considering the hypothetical example of a friendship network

in a school, where we try to assess if the friendship structure is predominately a function of attributes such as gender, religion, race or economic background. One null model might be that people want to have a fixed number of friends, and they select these friends at random from all of the people in the network. A small group of individuals who have a large number of friendships between them would be surprising and therefore could be considered to be a community under this null model. However, the null model in this example could be badly chosen. It is conceivable that students would be more likely to make friendships within their own classes than between classes simply due to the length of time they spend together. Thus to investigate whether the friendship structure is a function of the attributes above, we should include class structure in the null model, for example by assuming that students are r times more likely to make friends within their own class than outside of it; r would have to be estimated from the data.

The key point is to select a null model that is appropriate so that community detection yields communities that are not solely a function of methodological constraints on the topology of the network or of some properties of the nodes which for purposes of community detection are considered to be not of interest. Some examples of networks where the appropriate null model is not trivially obvious include bipartite networks [57, 78], geographical networks [53], correlation networks [72, 115] and geographic correlation networks [147]. For further discussion of the construction of null models for community detection see Section 2.6.3.

The work in the previous chapter illustrates the potential problems with the use of the configuration model as a null model in sampled networks, as it does not replicate the particular structure that the sampling technique induces into the network. In this chapter we will focus on shortest path sampling. Here we demonstrate that the same problem that we have already identified for significance testing in sampled networks, namely that the configuration model is not a good null model, also applies for com-

munity detection in sampled model networks. We derive a solution to this problem using a novel community detection quality function which we call Path Modularity, and demonstrate that it performs substantially better than traditional community detection methods on such networks.

This method extends modularity (a common quality function for community detection) to incorporate information about simple paths of length greater than 1. Similar approaches have been used by other authors, see for example Refs. [64, 99, 121, 143, 145, 184].

4.2 Measuring performance and exploring the limitations of the configuration model

In this section we will develop a method to test the performance of community detection methods in sampled networks and then use this method to test the performance of the configuration model demonstrating its limitations.

We are motivated by the results of the previous chapter (Chapter 3), namely that the configuration model is an inappropriate null model for certain statistics in sampled networks as it does not preserve the features of the sampling procedure.

As the null model in the original formulation of modularity (Eq. (2.1)) is based on the configuration model, using this approach in community detection could result in communities that are simply a function of the sampling procedure rather than some feature of the underlying data.

4.2.1 Specifying the problem

To summarise the wider motivation, we want to sample a large network using a construction method and a seed list. We want to use the technique we developed in Chapter 3 to help select a subnetwork. We would then like to be able to detect

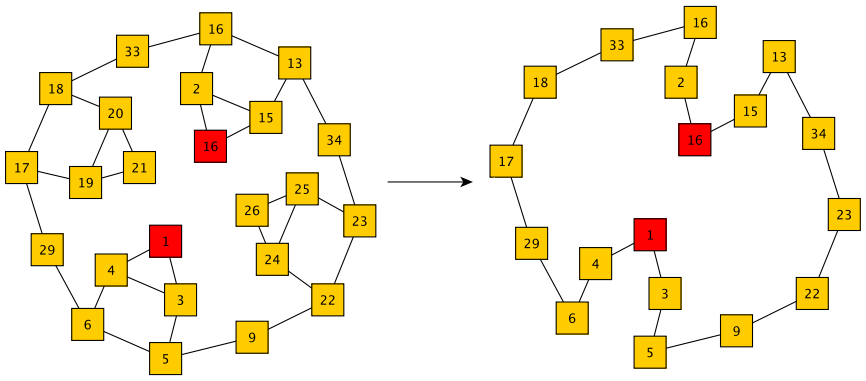

Figure 4.1: An example of shortest path sampling from a graph.

communities in the sampled network. However some sampling features can affect community detection null models. We want communities that are more densely connected than we would expect at random given the construction method and thus we need to incorporate the sampling technique into the null model.

Further, computationally testing the different sampling schemes used in Chapter 3 indicated that the communities obtained in a shortest path sampled network could be strongly affected by the sampling scheme. As we saw in Chapter 3, shortest path sampling is a promising method, hence we focus on this method.

An example of how damaging shortest path sampling can be to the community structure can be found in Figure 4.1. The red nodes are seed nodes and there are clearly (by design) mini-communities around each of the triangles (e.g. 1,3,4,5,6). In contrast, shortest path sampling reduces this network to a ring with no apparent community structure, as all of the 'dense' regions which are used for community detection have been removed by the sampling.

In this chapter we explore possibilities to adapt the null model to adjust for the structure induced by shortest path sampling in order to use community detection.

4.2.2 Testing suitability: Model networks

In order to test the suitability of a new null model, we construct a model for a sampled network which has a known partition. This model will then be used to check if our method can recover the structure. Unfortunately, constructing a network with both the structure of the sampling scheme and embedded communities is non-trivial. Thus the following method is used.

We make the assumption that if the overall network has a suitably strong community structure then this should also be the community structure of the sampled network at some resolution, and thus this community structure should be detectable at some resolution in the sampled network with the correct null model. Thus we construct networks with strong community structure and then sample them. As is the case for any benchmark, it is possible that in real-world use cases the sampled community structure would not be a function of the wider community structure, as it is possible that methods that fail on this benchmark could work on networks which are sampled from networks with weaker communities.

The model we have chosen to use is the Stochastic Block Model, originally developed in sociology [83] and previously used to test detection and classification methods [75, 77, 99, 145, 159], and as a tool to improve community detection [77]. It can be seen as a generalisation of an ER graph, where each node is assigned a category and each edge is an independent Bernoulli trial with a probability defined by the category of each end node. In our model we simplify the probability structure to have only two different edge probabilities, c_{in} for the edge probability for edges between nodes of the same type, and c_{out} for edge probability between nodes of different types. To produce a community structure, c_{in} is set to be substantially larger than c_{out}.

To fix the range of the probabilities c_{in} and c_{out} in our test networks we consider the graph density. We would like the overall graph density to be similar to that found in graphs in our domain of interest. PPI networks typically have densities that are

less than 0.01 [139], so our graphs are chosen to have an expected density of 0.005. We then vary the ratio $\frac{c_{in}}{c_{out}}$ while keeping the expected density of the graph fixed.

We fix the remaining parameters as follows: we set the number of communities in our stochastic block model at 6 with 1000 nodes in each community, we fix the number of seeds at 10 and we vary the ratio between internal and external connections ($\frac{c_{in}}{c_{out}}$) from 1 (no community structure in the stochastic block model) to 100 (a very strong community structure in the stochastic block model). To obtain a sense of the average behaviour we perform 50 repeats for each ratio.

Putting this all together, we use the following procedure to construct each member of our ensemble. For a given ratio ($\frac{c_{in}}{c_{out}}$,) the procedure to generate model networks is as follows:

1. Create a graph from a stochastic block model with a given number of nodes of each category and probabilities c_{in} and c_{out} which are constructed to obtain the desired expected density and ratio ($\frac{c_{in}}{c_{out}}$).

2. Select 'Z' seed nodes uniformly at random without replacement.

3. Sample using the given technique on these seed nodes.

To keep the results comparable between different methods we use the same ensemble sampled networks for each method.

To investigate the ability of a given community detection quality function to recover the embedded community structure we then compare the embedded structure to the detected structure. However, we must also account for the resolution, as the assumption that we made earlier in this section is that the embedded structure should exist at some resolution but not necessarily at all resolutions. Thus, rather than simply comparing at one resolution we construct communities at multiple values of the resolution parameter for each method and then select the community structure which is the most similar. In order to keep the methods comparable we test approximately

the same number of resolution parameters for each method unless otherwise stated. We also report the resolution parameter which performs best on each of the statistics, in the event of a tie between two or more parameters, a parameter is randomly selected from those that performed best.

In all of the communities detection work in this chapter with the exception of the work with stability we use a generalised version of the Louvain algorithm [23], known as GenLouvain [88]. The work with stability uses code which is designed to work with stability [48] which also uses the Louvain algorithm [23].

4.2.3 Assessing the similarity of partitions

To assess the similarity between the induced partition and the observed partition, we use four different commonly used statistics. The statistics are as follows, Variation of Information, Arithmetically Normalised Mutual Information, Geometrically Normalised Mutual Information, and a standardisation of the Rand coefficient known as the z-score of the Rand or simply Zrand. In the following section we will describe each of these methods, and we will briefly discuss the relative merits.

4.2.3.1 Variation of Information

Variation of information is a metric first proposed by Ref. [122], in which it is described as the sum of the information lost from swapping from one partition to another and the information gained performing the same action. Alternatively, for partitions $P^{(1)}$ and $P^{(2)}$ it can be seen as the amount of information required to fully determine $P^{(1)}$ given we know $P^{(2)}$ plus the amount of information required to fully determine $P^{(2)}$ given we know $P^{(1)}$ [122].

More formally, following the definition of Ref. [122], Variation of Information (VOI) is defined as follows. Let us impose an arbitrary ordering on the nodes of the network. Let $P^{(1)}$ and $P^{(2)}$ be fixed partitions such that $P_i^{(j)}$ is the community of the

ith node in the jth partition. Then let $X \sim \mathcal{U}\{1, n\}$, a discrete uniform distribution over the integers 1 to n. Then VOI is defined as follows:

$$V(P^{(1)}, P^{(2)}) = H(P_X^{(1)}) + H(P_X^{(2)}) - 2I(P_X^{(1)}, P_X^{(2)}), \tag{4.1}$$

where $H(Y)$ is the entropy of a random variable Y, and $I(X, Y)$ is the mutual information of random variables X and Y. For discrete random variables $P_X^{(1)}$ and $P_X^{(2)}$ the entropy is defined as

$$H(P_X^{(1)}) = -\sum_{i=1}^{\max(P^{(1)})} Prob(P_X^{(1)} = i) \log_2(Prob(P_X^{(1)} = i)) \tag{4.2}$$

where $Prob(X)$ is the probability of X, and $\max(P^{(1)})$ is the maximum community assignment in partition 1. The entropy can be interpreted as the number of bits that are required on average to represent each assignment. For example, a partition where all nodes are in the same community would have an entropy equal to 0, whereas a sequence of coin flips of an unbiased coin would have an entropy of 1.

Further, mutual information is defined as follows:

$$\begin{aligned} I(X, Y) = \sum_{i=1}^{\max(P^{(1)})} \sum_{j=1}^{\max(P^{(2)})} & Prob\left((P_X^{(1)} = i) \cap (P_X^{(2)} = j)\right) \\ & \times \log_2 \left(\frac{Prob\left((P_X^{(1)} = i) \cap (P_X^{(2)} = j)\right)}{Prob(P_X^{(1)} = i) Prob(P_X^{(2)} = j)} \right). \end{aligned} \tag{4.3}$$

The mutual information measures the amount of information that is gained in the state of $P_X^{(1)}$ if we know the value of $P_X^{(2)}$ (or of $P_X^{(2)}$ given $P_X^{(1)}$ as the measure is symmetric). Thus variation of information is 0 if $P_X^{(2)} = P_X^{(1)}$ which implies $P^{(1)} = P^{(2)}$ as $H(P^{(1)}) = H(P^{(2)}) = I(P^{(1)}, P^{(2)})$. Further, it is equal to $H(P_X^{(1)}) + H(P_X^{(2)})$ if the partitions share no information. Thus this measure can be thought of as measuring the difference in information content of the partitions.

The variation of information is bounded from below by 0 and bounded from above by $log(n)$ [57, 122]. Thus, as observed by Refs. [57, 59, 91, 122], variation of information is not strictly comparable between graphs of different sizes. The sampled networks which we consider have a range of sizes as they are sampled using a random seed list. Further, the distribution of the number of nodes in each of the network is function of the ratio $(\frac{c_{in}}{c_{out}})$, thus we cannot compare the distribution of variation of information between different ratios as they do not have the same distribution of graph sizes. To alleviate this bias, following the suggestion of Refs. [57, 59, 91, 122] we divide the variation of information by $log(n)$ to make the results more comparable between different graph sizes, which also has the added advantage of standardising the measure to be between 0 and 1.

4.2.3.2 Normalised mutual information

Another similarity measure based on information theory is normalised mutual information (NMI). As the name suggests this measure uses mutual information from (4.3) and normalises this quantity to make the result comparable between partitions. It is often used to compare partitions in the literature [59].

There are two forms of normalised mutual information, which in this document we will refer to as arithmetically normalised and geometrically normalised. In either case the statistic is bounded between 0 and 1 which can be advantageous in some contexts.

The arithmetically normalised mutual information was defined by Ref [5] as follows:

$$\frac{2I(P^{(1)}, P^{(2)})}{H(P^{(1)}) + H(P^{(2)})} \tag{4.4}$$

We can see that as $I(P^{(1)}, P^{(2)}) \leq \min(H(P^{(1)}), H(P^{(2)}))$ this expression is bounded between 0 and 1. [183].

Ref. [164] defines the geometric version define as follows:

$$\frac{I(P^{(1)}, P^{(2)})}{\sqrt{H(P^{(1)}) + H(P^{(2)})}}. \tag{4.5}$$

Notice again that as $I(P^{(1)}, P^{(2)}) \leq \min(H(P^{(1)}), H(P^{(2)}))$ then the geometric version will be bounded between 0 and 1. [183].

Ref. [183] also notes that in the case where the entropy of both of partitions is 0, i.e. when there is a single partition, then the measures are not defined. Due to the construction of our networks it is very unlikely that we will observe a case where the embedded partition only has a single community (in fact it never happens across the ensemble), however it is possible for the detected community to consist only one community. In the case of the geometrically normalised version this results in a $\frac{0}{0}$ [183]. In the circumstance of a $\frac{0}{0}$ we replace the undefined quantity with 0 as this measure clearly has 0 mutual information.

4.2.3.3 Zrand

The final measure that we use in this chapter is a standardised version of the Rand coefficient. The Rand Coefficient is defined for a pair of partitions $P^{(1)}$ and $P^{(2)}$ as the sum of the number of pairs of nodes that are in the same community in $P^{(1)}$ and in $P^{(2)}$ and the number of pairs of nodes that are in different communities in $P^{(1)}$ and $P^{(2)}$ divided by the total number of pairs [133].

Following the notation we used earlier this is equivalent to the following expression:

$$\frac{1}{\binom{|P^{(1)}|}{2}} \left(\sum_{i=1}^{|P^{(1)}|} \sum_{j=1}^{|P^{(2)}|} \delta(P_i^{(1)}, P_j^{(1)}) \delta(P_i^{(2)}, P_j^{(2)}) + (1 - \delta(P_i^{(1)}, P_j^{(1)}))(1 - \delta(P_i^{(2)}, P_j^{(2)})) \right) \tag{4.6}$$

where δ is the Kronecker delta.

Ref. [170] considers a z-score of the Rand Coef. A z-score is a standard approach

which is defined for a random variable X as follows:

$$\frac{X - E[X]}{\sqrt{Var(X)}} \tag{4.7}$$

where $E[X]$ is the expectation of X and $Var(X)$ is the variance. Ref. [170] observes that if the model of randomness preserves the number of nodes in each size of each of the community (and by extension the number of communities) then several pair counting scores are linear functions of each other. Thus, under a model of randomness that preserves this feature the z-score of each of this pair counting scores is equivalent. Thus, using analytic results for the mean and variance of a different but related pair counting score from Ref. [25] which uses a hypergeometric model, Ref. [170] constructs the z-score of a quantity which is a linear function of the Rand coefficient and thus constructs the Zrand score. The form of the variance is complex and does not add to the discussion and thus we do not repeat it here, however the expression for both the mean and the variance can be found in Ref. [170].

The code we use for the implementation comes directly from the website of the authors of Ref. [170]. We only make one adjustment, in certain cases the variance in the denominator can become zero thus giving an undefined quantity, if the numerator is a also 0 we replace the undefined quantity with a 0.

4.2.3.4 Advantages and disadvantages of each approach

There are many different ways of measuring the similarity between partitions, each with their own advantages and disadvantages. For general reviews of the approaches see Refs. [57, 122, 183] and Ref. [170] which discusses so called pair counting methods methods. In this section we will briefly visit the advantages of each approach.

Variation of information does have in general some mathematically advantageous properties. First, it is a metric i.e. is symmetric, non negative, and obeys the

triangle inequality [59, 122]. Thus, a score of 0 implies that the partitions are identical. As discussed in the Section 4.2.3, following the approach suggested by Refs. [91, 122] to make the VOI for graphs of different sizes comparable we divide by $log(n)$. However, Ref. [122] also notes that VOI is also bounded from above by $2log(\max(\max(P^{(1)}), \max(P^{(2)})))$ [122]. Thus, null models that favour small numbers of communities could have a much smaller possible range of VOI/log(n) than other methods.

In both forms normalised mutual information is also bounded between 0 and 1. However, Ref [195] demonstrates that arithmetic normalised mutual information can be misleading in small networks favouring partitions with more communities due to finite size effects, a flaw which the geometric case should also share. Further, it is possible variation of information may share this issue. Ref. [195] also proposes an adjusted index in which the expectation of normalised mutual information is subtracted from the value. While in test cases this appears to overcome the issue, the statistic is no longer in $[0, 1]$.

The Zrand score is unbounded, and unlike the other measures (not counting the adjusted measure proposed by Ref. [195]) it measures a deviation from random, thus a high score in this measure means that there is a significant association between the true communities and the detected communities. However, unlike the other measures it does not have a simple value which indicates exact agreement between embedded and detected communities. Further, it should be noted that Ref. [170] advocates the use of Zrand to deal with the case of two partitions where the mutual information is small in comparison to the entropy, to detect non trivial association between the classifications.

In some sense the choice is a little arbitrary, Refs. [59, 170] recommends using variation of information (as it is a metric) for partitions that are quite similar, however NMI is a standard approach for these problems [59], further each method has their

own drawbacks as discussed above. Thus, rather than selecting a statistic we will use all four statistics and investigate the methods that perform best with respect to all of them.

4.2.3.5 Accounting for the limitations of these statistics

We discussed in the previous section the limitations of some of these statistics at measuring similarity. By considering these limitations we can construct a minimal comparison for these statistics.

For a method to perform well with respect to a given statistic we would expect it to at least outperform the following trivial partitions. We consider two trivial partitions, a single community consisting of all of the nodes and a community structure where each node is assigned to a different community. Each of these approaches is inspired by limitations we discussed in the previous section. In the case of the community consisting of a single partition, it is the the upper bound on the VOI of $2\log(\max(\max(P^{(1)}), \max(P^{(2)})))$ from Ref. [122], as a single community minimises this upper bound while containing no additional information about the underlying communities. The partition of singletons is inspired by the observation of Ref. [195] that partitions with more communities can do well even when there is no structure in the network. A partition of singletons is the structure which maximises this while containing no additional information.

Thus, for each of the four statistics in Section 4.2.3 we measured the similarity between each of the trivial partitions and the community structure from each member of the ensemble of sampled networks we constructed in Section 4.2.2. The results can be seen in Figure 4.2.

If we consider VOI, in which lower values indicate a more similar partition we observe that the single community markedly outperforms the partition of singletons. Thus for the analysis in this chapter we will compare our results to the single com-

munity rather than the partition of singletons.

For the two forms of NMI it is the opposite situation, larger values indicate a more similar partition and in both cases the partition of singletons outperforms the single community which agrees with the observation of Ref. [195]. In fact for the Geometric NMI the statistic is not defined for the case of a single community [183] as the entropy of a single community is zero and thus the geometric average is zero. Thus the correct comparison is with the partition of singletons.

Finally, the results related to the Zrand score indicate that this method is not affected by either of these cases and thus we cannot use these trivial partitions to benchmark the behaviour of this method.

4.2.4 A problem with using the configuration model.

To demonstrate the limitations of the configuration model we test the performance of this form of modularity using the procedure and the ensemble of networks detailed in Section 4.2.2 and Section 4.2.3.

We perform community detection with the Reichardt and Bornholdt formulation of modularity with a resolution parameter using the configuration model as a null model as described in Section 2.6.1. This results in the following quality function from (2.1):

$$Q^{conf}(c_i : i \in V) = \sum_{i,j \in V} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{4.8}$$

We obtain the community structure via the Louvain algorithm which is described in Section 2.6.2. Following the procedure discussed in section 4.2.2 we compute the similarity at several different resolution values. For the configuration model we use $\lambda = 0.1,0.3,0.5,...,1.9$ (note, that due to the increments λ never equals 1), and then select the best partition with respect to each of the statistics across all of the resolutions.

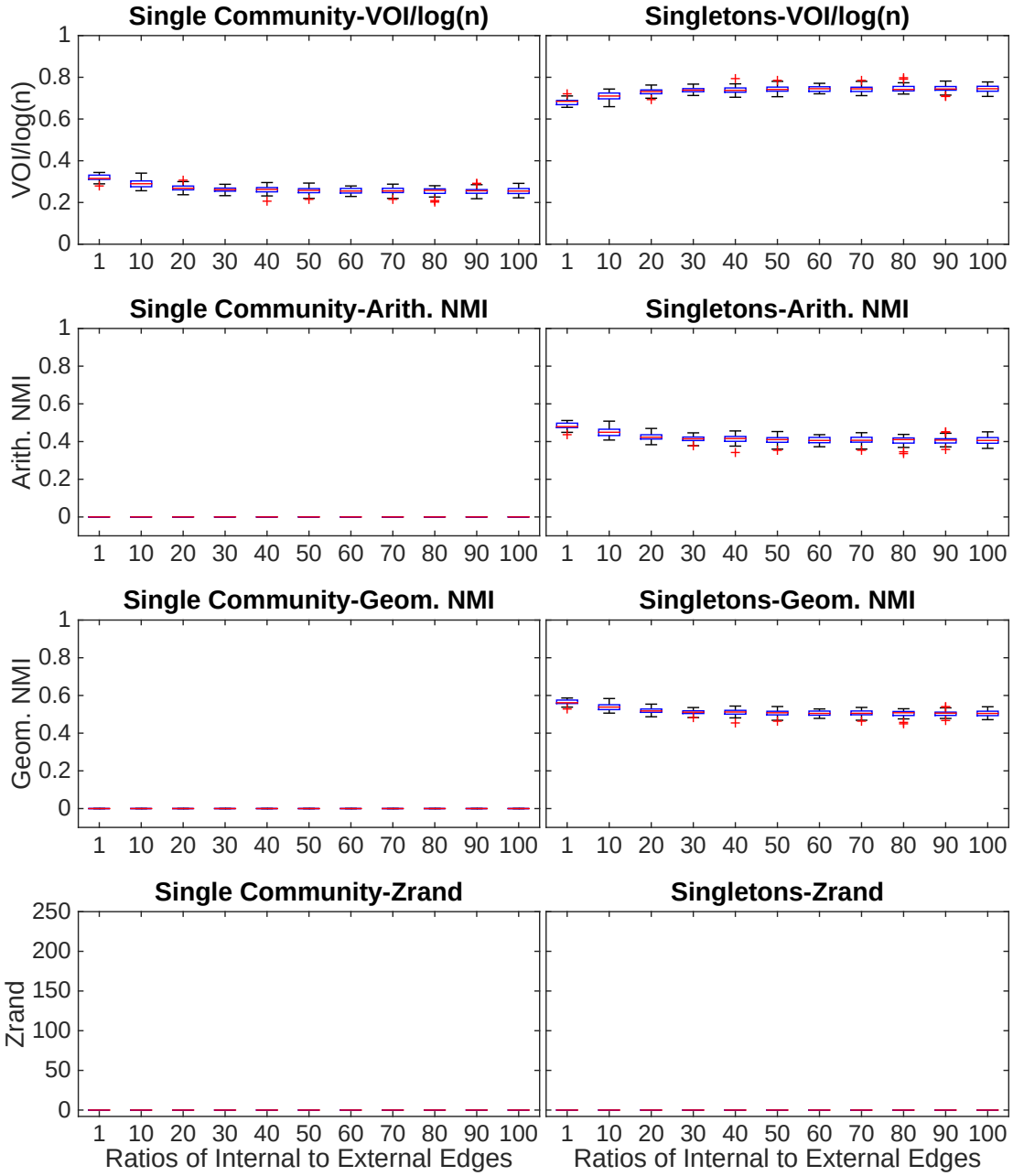

Figure 4.2: The performance of two trivial partitions at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). The first trivial partition is a community consisting of all nodes (Single Partition) and the second is a community structure where each node belongs to a different community (Singletons). Performance is measured using 4 statistics: VOI, Arith. NMI, Geom. NMI and Zrand (for details see 4.2.3).

To interpret the similarity of these partitions we need to understand typical values for each of the statistics in this system. To this end we construct two comparisons. The first is to compare the statistics on an ensemble of subnetworks sampled from a network with no community structure (i.e. that has $\frac{c_{in}}{c_{out}} = 1$). The distribution of the statistics in this ensemble should represent random variation with no structure present. The second comparison is to the trivial partitions statistics we defined in section 4.2.3.5

The results can be found in Figure 4.3. We observe that the values of VOI at each of the ratios ($\frac{c_{in}}{c_{out}}$) greater than 1 tend to be lower than the values when the ratio is 1, potentially indicating that we discover more structure. However, when we compare to the single partition results (green triangles in VOI plot) we observe that the result under the configuration model and the single group partition are very similar. The lower panel of the figure shows the value of λ at which the best performing community structure was detected. Looking at the results for VOI, we observe that many of the values for λ are close to or at the minimum possible value. Further, as the most likely community structure for small values of the resolution parameter is the single community and thus it is unsurprising that the average VOI we observe is close to the average VOI for a single community.

The performance in NMI is mixed, the arithmetic NMI shows that for a ratio ($\frac{c_{in}}{c_{out}}$) greater than 60 the configuration model outperforms the average NMI of the singletons. Although, with the exception of the result for a ratio of 100, for these high ratios the average of the trivial partition is still within the first and third quartiles. In the case of the geometric NMI the opposite behaviour is observed. The range of λs used by these methods is broadly similar, however interestingly it is markedly different to the range used by the VOI.

Finally, the performance in Zrand indicates a strong association between the embedded communities and the detected communities when the ratio is above 1. Inter-

estingly the range of λs used for large ratios appear to be similar to those used by the NMI measures but very different from the values used for the variation of information.

In conclusion for two of the measures the configuration model does not perform better than a trivial partition, namely VOI and Geometric NMI. For Arithmetic NMI it outperforms the trivial partition for large values of the ratio on average finding a non-trivial solution. The Zrand score indicates that the partitions detected have a much higher Rand Coef. than would be expected at random,

Thus finding a suitable quality function and null model for shortest path sampled networks which would perform well on all of the statistics is a worthwhile problem that we may approach in a similar fashion to the approach by Expert et. al. [53] which is described in section 2.6.3 but based on sampling arguments.

4.3 Constructing an appropriate null model and quality function

In the previous section we showed that the configuration model had difficulty detecting the underlying structure in the benchmark networks. One possible reason for this failure is that changing the null model in modularity is a pairwise adjustment, whereas using a path sampling technique influences the structure across paths of length greater than one. Therefore accounting for the structure at this level may not be sufficient.

To gain intuition about networks constructed from paths let us consider an example (Figure 4.4(a)). We observe that the network is constructed from 5 paths in different colours which each consist of two or more edges. In traditional community detection we think of edges as being somewhat independent observations, and thus it is natural to think of an adjacency matrix A_{ij}, where the ij^{th} entry indicates existence of an edge between a particular pair of nodes i and j. For modularity based

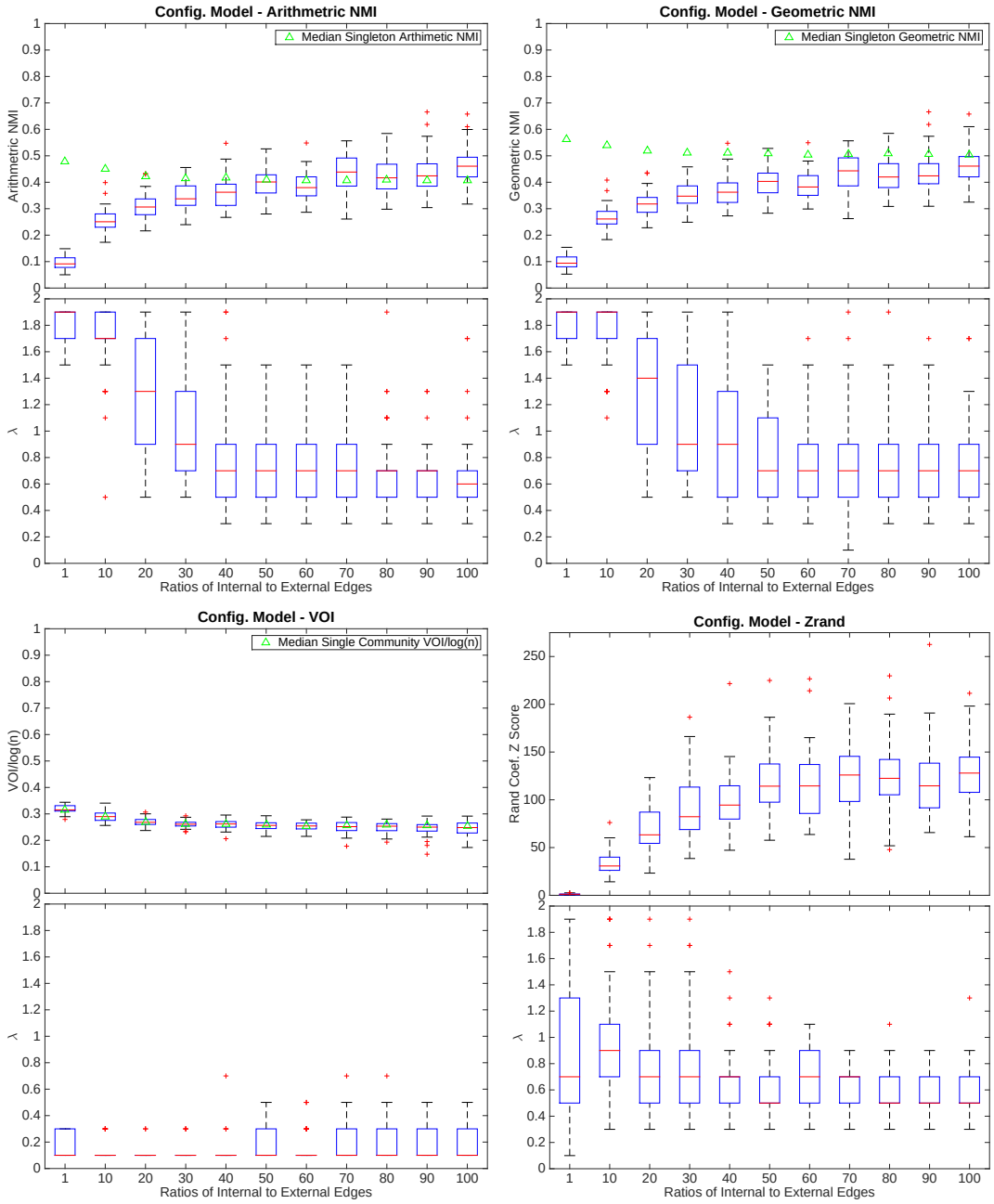

Figure 4.3: Performance of community detection using Q^{conf} modularity at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $\lambda = 0.1, 0.3, ..., 1.9$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

community detection we then define the expected number of edges between i and j (P_{ij}) by accounting for any constraints that we know about the system. We then use A_{ij} and P_{ij} to construct and then maximise modularity (2.4).

In a network made of paths the natural equivalent of the adjacency matrix is a collections of tensors where each tensor stores the existence of ordered paths between tuple of nodes of a certain length. Thus in the example in Figure 4.4(a), the information about the paths of length 2 (green) would be stored in one tensor and in the existence of path of length 3 (orange, violet, purple and blue) would be stored in another. We could then equally define the expected number of ordered paths between nodes attempting to account for any constraints on the paths. For example, in Figure 4.4(a), yellow nodes are never at the beginning or the end of paths.

However, as we would like to perform community detection using a standard algorithm (namely Louvain [23]), we need a similarity measure between nodes ($A_{ij} - P_{ij}$ in modularity) rather than a similarity measure between paths which obtain from the tensors. Thus, we construct a variable $A^*_{ij\Upsilon}$ which computes the number of times we observe node i and node j, Υ nodes apart in a path, i.e. the number of subpaths of length Υ which connects nodes i and j. Similarly we can construct $P^*_{ij\Upsilon}$, the expected number of such paths. To detect communities in this system we would then sum the (possibly weighted) similarities at different values of Υ to get a similarity measure between node i and node j.

Let us now imagine we are in the case shown in Figure 4.4(b), where the paths present in Figure 4.4(a) have been transformed to binary edges. We know that underlying structure of the network is dictated by the underlying paths but we can no longer observe them and thus we cannot construct $A^*_{ij\Upsilon}$ and $P^*_{ij\Upsilon}$. However, we can estimate them by considering all paths of a length Υ.

For example, if we consider a sampled network that is a line graph (a graph consisting of a line of nodes each connected only to its immediate neighbours), it is

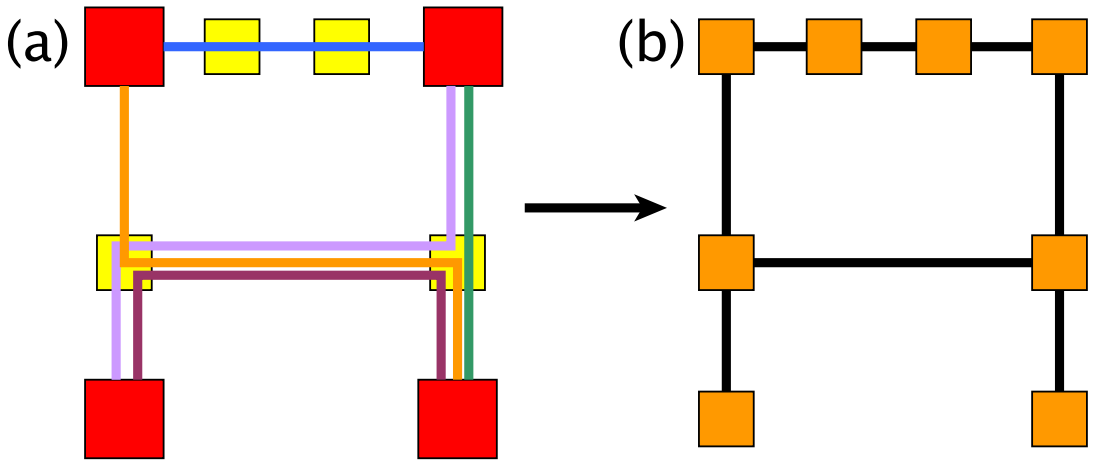

Figure 4.4: An example of a network constructed of paths rather than individual edges.

not inconceivable to suppose that pairs of nodes that are 2 apart (say the first and third node) are likely to be in the same community.

Note, this is different but related to the approach used in stability which measures the probability of a random walker moving from a node to another node (for a review of this method see Section 2.6.4, and we discuss the differences in more depth in Section 4.3.5).

A key consideration when constructing a similarity measure from paths is cycles. When considering the number of paths of a given length between a pair of nodes, cycles in the graph become a confounding factor. If we consider Figure 4.5, we observe that the number of paths of length three is between node 1 and node 8 is seven, whereas the number of paths of length three from node 11 to node 1 is only one. However, the large number of paths between nodes 1 and 8 are simply due to the high degree of node 1. While controlling for the degree in the null model will help to mitigate this issue, binning for rare degrees will reduce the effectiveness as nodes with different degree will be affected differently by this bias. Further, for larger path lengths we will have to control for the number of cycles of different lengths that each

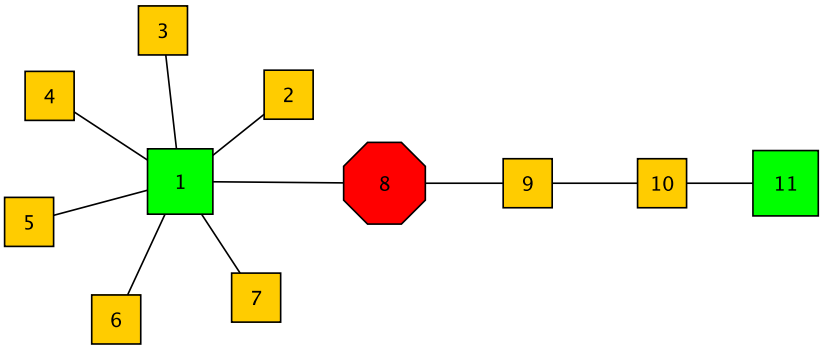

Figure 4.5: An example of the importance of considering only simple paths rather than all paths. We note the that number of paths of length 3 from node 1 to node 8 is 7, whereas the number of paths of length 3 between node 11 and node 1 is 1, despite the fact that many of these paths are simply due to the degree of the node.

node has (for example number of triangles for paths of length 4). A simpler solution is to consider only simple paths, i.e. paths that only visit each node once. This is also in line with what we would expect from shortest path sampled networks. For the example in Figure. 4.5, this reduces the number of paths of length three from node 1 to node 8 to zero from seven while preserving the one path from node 11 to node 8.

This leads to the following modified form of modularity which we call Path Modularity:

$$Q_l^{Path} = \sum_{\Upsilon=1}^{l} \lambda_1^{\Upsilon} \sum_{i,j} (A_{ij\Upsilon} - \lambda_2 P_{ij\Upsilon}) \delta(c_i, c_j) \tag{4.9}$$

where $A_{ij\Upsilon}$ is the number of non-self intersecting paths between nodes i and node j of length Υ, and $P_{ij\Upsilon}$ is the expected number of paths of length l between i and j, and l is the length of paths considered. The choice of l heavily influences the computational complexity.

Note, incorporating a null model to account for the sampling technique into stability would require redefining the dynamics of the random walker to include the features of the null model of interest, i.e. the sampling structure, which would be prohibitive.

Similar adjustments have also been made in other approaches for community detection, although to our knowledge the form used in the equation above is novel. For example, the spectrum of the so called non-backtracking matrix has been used very successfully for community detection. [99] The non-backtracking matrix is constructed from a network such that a random walker cannot return to a node they have visited in one step [99]. It is similar to the path idea, in that random walks on this matrix of length 2 are simple and walks of length 3 with distinct end points are simple.

Other community detection approaches have been constructed around similar ideas. Ref. [121] constructs a path-based modularity based on a given power of the adjacency matrix (and therefore counting all paths between pairs of nodes of a given length). A similar idea is Katz similarity: a measure that compares the similarity of a pair of nodes by computing the number of paths between them downweighting longer paths based on the Katz centrality [93], see for example Ref. [111]. Formally, the Katz centrality is defined as follows. Consider the following matrix [93]:

$$M_{Katz} = \sum_{\Upsilon=1}^{\infty} \sum_{j} \tau^{\Upsilon}(A^{\Upsilon}) = (I - \tau A)^{-1} - I \tag{4.10}$$

where I is the identity matrix and τ is a parameter. The Katz centrality is the column sums of the matrix M_{Katz} [93] and the Katz similarity is the i, jth element of this matrix [111]. We note the similarity to our measure with the exponential downweighting of path lengths and the combination of information from different path lengths.

Independent of this work other authors have developed similar measures. Ref. [184] constructs a community detection method which uses the matrix $\sum_{\Upsilon=1}^{a}(2)^{-\Upsilon}A_{ij\Upsilon}$, clearly related to the matrix we use. However, the methods differ as Ref. [184] partitions the network by then applying a variant of k means clustering to the resultant

n dimensional space constructed from the columns of the matrix. Also, Ref. [64] constructed a modified modularity function that shares many features of the function that we developed. The key differences are that we fit the null model by simulation (for further details see section 4.3.1) and they do not have any explicit resolution parameters and they normalise their matrices differently.

Further, a similar approach (using the spectra of $A_{ij\Upsilon}$ for fixed Υ) was also used to prove part of the detection limit [119]. The detection limit illustrates that the underlying community structure in sparse stochastic block models (where probability of an edge scales like $\frac{1}{n}$) cannot be detected if the internal and external edge probabilities fulfil a certain condition (the detection limit) [47, 124]. The approach in Ref. [119] showed that using $A_{ij\Upsilon}$ they could produce a non-trivial reconstruction of the underlying community structure up to the detection limit.

4.3.1 Constructing a null model for paths of length 1

In order to construct communities using our new version of modularity we need to define the null model $P_{ij\Upsilon}$. To construct this null model we use the following strategy: we first create a null model for paths of length 1 in a sampled network, and we then generalise this approach to arbitrary path lengths.

We require that our null model satisfies the following conditions:

1. Replicates or approximates the sampling scheme. It is clear from the example in Fig. 4.1 that shortest path sampling can induce artificial structural features in a network which should be taken into account.

2. Easy to obtain the expectation of an individual edge as we want to use modularity maximising community detection.

3. Parameterised by degree. Following Newman and Girvan [126] we believe that nodes of high degree may behave differently to nodes of low degree and thus

should be addressed in the null model.

Of these conditions the first is the most difficult to satisfy. The induced artificial structural features are hard to characterise for shortest paths as they involve global network information, for example assessing whether a path via node x is shorter than a path via node y. Further, as in Chapter 3, obtaining analytics for certain graph types might be possible but would be very difficult in general.

Thus, following Chapter 3 we approximate the sampling scheme by defining an ensemble of networks with the desired property name the structure induced by shortest path sampling. Conditions 2 and 3, can then be fulfilled by computing expectations over the ensemble.

Again following Chapter 3 we construct an ensemble of sampled graphs by sampling the observed graph with an ensemble of seed lists that are selected uniformly at random from the set of seed lists of the same length as the original seed list. Note, this is different to the null model in the Chapter 3 where we fixed the (approximate) degree distribution of the seed list, and we do not adjust for redundant seed nodes for computational reasons. Thus we formally define our ensemble as follows. For a given seed list s we define the set of subnetworks in our ensemble as:

$$\mathfrak{N}(s) = \{f_{samp}(V, E, s') : s' \subset V,\ |s'| = |s|\} \tag{4.11}$$

where f_{samp} is an arbitrary sampling technique (here shortest path sampling) which outputs a sampled network, V and E are the sets of nodes and edges in the network and s is the original seed list. Further, we define $\mathfrak{N}_{ab}$ as the set of subnetworks which have at least one potential connection between a node of degree a and a node

of degree b, i.e.

$$\mathfrak{N}_{ab}(s) = \begin{cases} \{f_{samp}(V,E,s') : s' \subset V, |s'| = |s|, |S_a^{f_{samp}(V,E,s')}||S_b^{(f_{samp}(V,E,s'))}| > 0\} & \text{if } a \neq b \\ \{f_{samp}(V,E,s') : s' \subset V, |s'| = |s|, |S_a^{f_{samp}(V,E,s')}| > 1\} & \text{if } a = b \end{cases} \tag{4.12}$$

where $S_a^{N_{et}}$ is the set of all nodes in N_{et} with degree a.

Having constructed the ensemble we now try to estimate the expected number of paths. In order to satisfy the third condition we estimate the connection probability between nodes of given degrees across the ensemble. There are two different types of randomisation that we could use to estimate the expected number of connections between a node of degree a and a node of degree b, which are as follows:

1. Select a pair of nodes uniformly at random from all pairs of nodes with the correct degrees with both nodes in the same graph in the ensemble.

2. Select a graph uniformly at random from $\mathfrak{N}_{ab}$. Then select a pair of nodes with the correct degrees uniformly at random from this graph.

We select the second type of randomisation for three reasons. First, in some sense we want the average sampled subnetwork (and not the average potential edge). Second, this type of randomisation means that disproportionate weight is not given in our estimate to networks which have a large number of nodes of a given degree. Finally this gives us a distribution of edge weights from which we can compute the variance over the set of all sampled networks which we use in section B.2.

This leads to the following null model formulation; let $\mathfrak{P}^{(N_{et})}$ be an $n \times n$ matrix where $\mathfrak{P}_{ab}^{(N_{et})}$ is the average number of edges between a node of degree a and a node of degree b for the network N_{et}, i.e.:

$$\mathfrak{P}_{ab}^{(N_{et})} = \begin{cases} \frac{1}{|S_a^{N_{et}}||S_b^{N_{et}}|} \sum_{(i,j) \in S_a^{N_{et}} \times S_b^{N_{et}}} A_{ij} & \text{if } a \neq b \\ \frac{1}{|S_a^{N_{et}}|(|S_b^{N_{et}}|-1)} \sum_{i \in S_a^{N_{et}}} \sum_{j \in S_b^{N_{et}} \setminus \{i\}} A_{ij} & \text{if } a = b \end{cases} \tag{4.13}$$

To obtain the null model following the second type of randomisation proposed above, we now perform the average over the ensemble. We define the $n \times n$ matrix $\mathfrak{P}^{(f_{samp})}$, where $\mathfrak{P}_{ab}^{(f_{samp})}$ is the average expected number of edges over the ensemble following the second randomisation, i.e.

$$\mathfrak{P}_{ab}^{(f_{samp})}(s) = \frac{\sum_{N_{et} \in \mathfrak{N}_{ab}(s)} \mathfrak{P}_{ab}^{(N_{et})}}{|\mathfrak{N}_{ab}(s)|.} \tag{4.14}$$

For notational convenience, we drop the "(s)" and we write $\mathfrak{P}_{ab}^{(f_{samp})}$ instead of $\mathfrak{P}_{ab}^{(f_{samp})}(s)$.

Computing $\mathfrak{P}_{ab}^{(f_{samp})}$ is non-trivial, as the number of terms required to compute $\mathfrak{P}_{ab}^{(f_{samp})}$ for all a and b is given by $|\mathfrak{N}(s)| \approx \frac{|V|!}{(|V|-|s|)!}$. The approximation is due to the redundant seed nodes causing some of the networks in the ensemble to be identical and thus we could reduce the number of calculated terms. Therefore, for a seed list of size 10 this scales approximately as $|V|^{10}$. Thus $\mathfrak{P}_{ab}^{(f_{samp})}$ is estimated from an ensemble of sampled networks, to obtain $\hat{\mathfrak{P}}_{ab}^{(f_{samp})}$. Note, as we cannot construct $\mathfrak{N}_{ab}(s)$ without sampling all of the networks, we sample from $\mathfrak{N}(s)$ and then construct a sample of $\mathfrak{N}_{ab}(s)$ which we denote $\mathfrak{N}_{ab}^{(sample)}(s)$.

4.3.1.1 Estimating probabilities for rare events

As we are only able to take a relatively small number of samples (in this case 10,000 networks), the estimates for the probabilities of rare events can be very inaccurate. To alleviate this problem, as in Chapter 3 the nodes are binned by degree. We make the assumption that finite differences in $\mathfrak{P}_{ab}^{(f_{samp})}$ (e.g. $\mathfrak{P}_{(a+1)b}^{(f_{samp})} - P_{ab}^{(f_{samp})}$ or $\mathfrak{P}_{a(b+1)}^{(f_{samp})} - \mathfrak{P}_{ab}^{(f_{samp})}$) are small such that they can be well represented by the average of the terms. There are clearly sampling techniques where this would not be the case, for example if we 'sampled' by removing all edges where both end nodes have a degree less than w, then there would be a clear disconnect in connection probability between nodes of degree greater than w and nodes of degree less than w. We do not

believe that this is not an issue in these networks, but it should be considered before applying this approach to other sampling schemes.

We also cannot bin the degrees using the degree distribution of the sample in question, as the frequency of a degree in our sampled network may not correlate to the frequency of the degree across the ensemble.

To address this we use the following approach. To compute the expected number of paths we construct $10,000$ random sampled networks which we combine with our test network to give 10001 sampled networks, and average over them following the second model of randomness as described in section 4.3.1. To obtain an estimate of the number of paths of a given length between a pair of nodes of given degrees from a graph in the ensemble we require at least one of pair of nodes of the correct degrees. Further to obtain a good average over the ensemble we require the number of graphs in the ensemble that fulfil this condition to be large, i.e. that for a pair of arbitrary degrees a and b that $|\mathfrak{N}_{ab}^{(sample)}|$ is large. While we could optimise $\min_{a,b}(|\mathfrak{N}_{ab}^{(sample)}|)$ directly, instead we use the following much simpler procedure. We let:

$$\mathfrak{N}_a = \{f_{samp}(V, E, s') : s' \subset V,\ |s'| = |s|,\ |S_a^{(f_{samp}(V,E,s'))}| > 0\}, \tag{4.15}$$

and following the notation for $\mathfrak{N}_{ab}$ we let $\mathfrak{N}_a^{(sample)}$ be a sample of $\mathfrak{N}_a$ constructed from the $10,001$ randomly sampled networks. Thus, the probability that a randomly selected network has a node of degree a is given by $\frac{|\mathfrak{N}_a^{(sample)}|}{10001}$. Therefore, assuming independence the probability of a network having a node of degree a and a node of degree b is given by:

$$\frac{|\mathfrak{N}_a^{(sample)}|}{10001}\frac{|\mathfrak{N}_b^{(sample)}|}{10001}. \tag{4.16}$$

Thus, using this observation we control the average number of networks which contain a pair of nodes of given degrees by binning degrees. Thus, we construct bins consisting of one or more of the observed degrees such that for a bin $\mathfrak{B}_{in}$, $\frac{|\bigcup_{a\in\mathfrak{B}_{in}}\mathfrak{N}_a^{(sample)}|}{10001}$

is greater than a constant. Then by assuming independence we have an estimate of the of the number of graphs in the ensemble that fulfil the criteria.

The existence of a node of degree a and the existence of a node of degree b are of course not independent, they will both be positively correlated with network size and therefore the probability of observing a network with a node of degree a and a node of degree b ($\frac{|\mathfrak{N}_{ab}^{(sample)}|}{10001}$) could be a lot lower than the expression above. This approach also does not let us easily estimate the number of graphs which contain two nodes of the same degree.

Thus, as this likely to be an over-estimate we fix the constant at 0.5, i.e. 50% of the graphs, which assuming independence would give on average ≈ 2500 graphs with the binned degrees, and we then we check that the number of networks goes below 1000. We bin the degrees from the right giving the following procedure

1. Compute an ordered list of all degrees that appear in the ensemble.
2. Start a new bin from the smallest degree which does not have a bin assigned.
3. If the probability of selecting a graph which contains an element of current bin is greater than or equal to 0.5, go to **2**, otherwise add smallest degree which does not have a bin assigned to the current bin and go to **3**, if there are no additional unassigned degrees go to **4**.
4. If the probability of selecting a graph which contains an element of the final bin is less than 0.5 then combine it with the previous bin.
5. Finally check that $\min_{a,b} |\mathfrak{N}_{ab}^{(sample)}| > 1000$

4.3.1.2 Fixing the expected number of edges in the network

In the method proposed above we estimate $\mathfrak{P}_{ab}^{(f_{samp})}$, the probability that an edge exists between two randomly selected nodes of degree a and b in a network selected

uniformly at random from $\mathfrak{N}_{ab}(s)$ (Eq. (4.14) and surrounding text). Fitting the null model to the ensemble which can have very different numbers of edges means that unlike when using the configuration model $\sum_{ij}\left(A_{ij}-\hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp})}\right)$ is not necessarily zero.

Clearly, as Equation. (2.4) has an arbitrary parameter (λ) multiplied by the null model, the null model can be rescaled without affecting the outcome. Therefore, in order to make the resolution parameter easier to scan over and in order to make the resolutions comparable between the configuration model and the sampled null model is rescaled as follows, we let:

$$\hat{\mathfrak{P}}_{ab}^{(f_{samp}-rescale)}(s,\mathbf{k})=\hat{\mathfrak{P}}_{ab}^{(f_{samp})}(s)\frac{\sum_{ij}A_{ij}}{\sum_{ij}\hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp})}(s)} \tag{4.17}$$

The dependence on $\mathbf{k}$ reflects the fact that the rescaling is a function of the sampled network considered. As before for notational convenience we drop the "$(s,\mathbf{k})$" and we write $\mathfrak{P}_{ab}^{(f_{samp}-rescale)}$ instead of $\mathfrak{P}_{ab}^{(f_{samp}-rescale)}(s,\mathbf{k})$. Under this rescaling the Reichardt and Bornholdt version of modularity with the configuration model (2.4) and Q_1^{Path} from (4.9) are comparable for $\lambda=1$.

4.3.1.3 Constructing Q_1^{Path}

We can then use our estimated expected number of paths of length 1 ($\mathfrak{P}_{ab}^{(f_{samp}-rescale)}$) from (4.17) as a null model in (4.9) as follows:

$$P_{ij1}=\mathfrak{P}_{k_ik_j}^{(f_{samp}-rescale)} \tag{4.18}$$

Leading to the following form of modularity:

$$Q_1^{Path}=\lambda_1\sum_{i,j}(A_{ij1}-\lambda_2P_{ij1})\delta(c_i,c_j) \tag{4.19}$$

$$=\lambda_1\sum_{i,j}(A_{ij1}-\lambda_2\mathfrak{P}_{k_ik_j}^{(f_{samp}-rescale)})\delta(c_i,c_j) \tag{4.20}$$

The performance of $Q_1^{(Path)}$ on the benchmark used in this chapter which described in section 4.2.2 can be found in the appendix in section B.2. In summary the approach does not seem to outperform the results from the configuration model.

We also constructed an alternative modularity model (Q^{var}) with a null model based on the summation of the estimated mean and estimated variance of the number of edges over the ensemble of networks. The motivation behind this model was the concern about the multiplicatively scaling for paths with a large estimated expected number of paths. Detailed results on the motivation and construction of this model can be found in the appendix in section B.1. The performance in the sampled networks community detection benchmark can also be seen in section B.2, this model also does not outperform the configuration model.

4.3.2 Generalising the approach to arbitrary path lengths

One option that starts from ideas similar to Leicht et al. [104] is to predict the number of simple paths of length l between any two nodes using the expected number of connections between all pairs of nodes (P_{ij1}) using the model we just derived. For example, the number of paths of length 2 between node i and node j can then be modelled as the nodes in $V \setminus \{i, j\}$ that connect to both i and j. Thus the number of paths is distributed as the sum of non-independent and non-identically distributed Bernoulli trials (i.e. distinct probabilities for a length 2 path via different nodes), known as a Poisson Binomial distribution. The expected number of paths of length 2 between a pair of nodes with independent edges is:

$$P_{ij2}^{inde} = \sum_{k=1}^{n} P_{ik1} P_{kj1} \tag{4.21}$$

where $P_{ii11} = 0$ by definition.

This approach makes the same assumption that modularity makes, namely that

the behaviour of simple paths of length greater than 1 can be accurately estimated using the pairwise connections. It is very likely that path sampling induces structural correlation between nodes that are not neighbours, and hence the expected value of any path-based on pairwise connections in these networks may be inaccurate.

Instead, we use the non parametric bootstrap that we used to approximate the expected number of pairwise connections, namely taking an ensemble average over the estimated expected number of paths of length Υ between a node of degree a and a node of degree b in each network,

$$\mathfrak{P}_{ab\Upsilon} = \frac{\sum_{N_{et} \in \mathfrak{N}_{ab}(s)} \mathfrak{P}_{ab\Upsilon}^{(N_{et})}}{|\mathfrak{N}_{ab}(s)|}. \tag{4.22}$$

We again estimate this by sampling using randomly constructed seed lists of the same length, and essentially approximating the summation using Monte Carlo methods. Note, the same ensemble is used to estimate each of the quantities, and thus the errors are correlated. Following the notation used earlier we label the estimate $\hat{\mathfrak{P}}_{ab\Upsilon}^{(f_{samp})}$.

To account for the fact that $\sum_{i,j}(A_{ij\Upsilon} - \hat{\mathfrak{P}}_{k_i k_j \Upsilon}^{(f_{samp})}) \neq 0$, a similar approach to the pairwise case is used to rescale each of the $\hat{\mathfrak{P}}_{ij\Upsilon}^{f_{samp}}$ so that the $\sum_{i,j}(A_{ij\Upsilon} - \hat{\mathfrak{P}}_{ij\Upsilon}^{(f_{samp}-rescale)}) = 0$ for all l, resulting in the following formula:

$$\hat{\mathfrak{P}}_{ab\Upsilon}^{(f_{samp}-rescale)} = \hat{\mathfrak{P}}_{ab\Upsilon}^{(f_{samp})} \frac{\sum_{ij} A_{ij\Upsilon}}{\sum_{ij} \hat{\mathfrak{P}}_{k_i k_j \Upsilon}^{(f_{samp})}} \tag{4.23}$$

4.3.2.1 Constructing Q_l^{Path}

Following the approach we used to construct Q_1^{Path} we use the rescaled estimates of $\mathfrak{P}_{ab}^{f_{samp}}$ from (4.23) to construct the null model, resulting in the following expression,

$$P_{ijl} = \mathfrak{P}_{k_i k_j l}^{(f_{samp}-rescale)} \tag{4.24}$$

Substituting this null model into (4.9) we obtain the general path modularity quality function for sampled networks:

$$Q_l^{Path} = \sum_{\Upsilon} \lambda_1^{\Upsilon} \sum_{i,j} (A_{ij\Upsilon} - \lambda_2 P_{ij\Upsilon}) \delta(c_i, c_j) \tag{4.25}$$

$$= \sum_{\Upsilon} \lambda_1^{\Upsilon} \sum_{i,j} (A_{ij\Upsilon} - \lambda_2 \mathfrak{P}_{k_i k_j \Upsilon}^{(f_{samp} - rescale)}) \delta(c_i, c_j) \tag{4.26}$$

The performance of this model is explored in the section 4.3.4.

4.3.3 Gaining intuition about the resolution parameters

The resolution parameter in standard modularity "zooms" in and out by multiplying the null model by larger and smaller values. This behaviour is replicated in Eq. (4.9) by placing a parameter λ_2 in front of the expected number of paths for each length thus "zooming" in and out the community structure in each case.

The second resolution parameter (λ_1) in Eq. (4.9) adjusts for the fact that for small l the number of non-self-intersecting paths of length l is likely to be larger than the number of length non-self-intersecting paths of length $l-1$. The parameter λ_1 weights the evidence from different path lengths and can be thought of as being similar in spirit to the parameter in Katz centrality which down-weights paths of longer length - in the case of Katz centrality this is due to the probability of edge effectiveness, but the idea is very similar. [93] This parameter is also similar to the time parameter in stability and to a free parameter in Ref. [104]. Thus the larger the parameter λ_1 the more weight is given to longer paths. This can be interpreted as a resolution parameter in the sense that if high weight is assigned to paths of low length we will get small communities, but if low weight is assigned to paths of low length we will get large communities.

4.3.4 Results from test networks

We tested versions of Path modularity using paths of length ≤ 2 (Q_2^{Path}), ≤ 3 (Q_3^{Path}) and ≤ 4 (Q_4^{Path}) using the same procedure as the other community detection null models. Following the approach detailed in section 4.2.2 in order keep the number of tests constant we fix $\lambda_2 = 1$ and we then let $\lambda_1 = 0.1, 0.3, ..., 1.9$ giving us the same number of tests as all of the previous tests on the detectability of the induced community by each null model.

The results can be seen in Figure 4.6 (Q_2^{Path}), Figure 4.7 (Q_3^{Path}) and Figure 4.8 (Q_4^{Path}). The results show a considerable improvement over previously tested methods, with average VOI's distinctly below the trivial partition for a ratios greater than 60 in Q_2^{Path}, greater than 40 in Q_3^{Path} and greater than 40 in Q_4^{Path}. The detected communities also show promising results with respect to arithmetic and geometric NMI. In arithmetic NMI it distinctly outperforms the trivial partition for a ratio greater than 20 in all paths lengths. Although, in certain statistics for some of the smaller ratios the average score of the trivial partitions are close to or within the quartiles. We also observe an improvement in this statistic by increasing the length of the path considered with a large difference between Q_2^{Path} and Q_3^{Path} and a smaller improvement observed between Q_3^{Path} and Q_4^{Path}. The results of the geometric are similar except that they only outperform when the ratio is greater than 40. Further, for higher ratios it markedly outperforms the trivial partition which may imply that it is discovering the underlying structure.

Similar to the performance in the configuration model Zrand indicates a strong association for a ratio greater than 1. Further, like in the other statistics the Zrand improves between Q_2^{Path} and Q_3^{Path}.

In contrast, for VOI the results for $\frac{c_{in}}{c_{out}} = 1$ and $\frac{c_{in}}{c_{out}} = 10$ are substantially worse than the previous methods, although the previous methods do not outperform the trivial partition. A possible explanation is that the induced partitions are not strong

enough and the other methods are simply finding a single partition, whereas due to fixing $\lambda_2 = 1$, we can only take off the expected number of edges which makes finding the single partition much harder.

The parameters in which these results were discovered are also interesting. For the larger ratios the distribution of parameters used by each of the statistics are very similar and only include a small number of parameter values. This is unlike the results from the configuration model, where the parameters used for VOI are very different. This is interesting as it implies that the same or similar communities structures are outperforms in each of the statistics.

4.3.5 Comparison with stability

Our novel adjustment to modularity clearly out-performs both the standard modularity and modularity with the other null models we tested. However, to do this we have had to generalise the concept of modularity to consider paths of length 2, 3 and 4. Thus it is interesting to compare this to stability, a community detection method which also accounts for longer paths.

We recall Equation (2.11) and Equation (2.13) from Section 2.6.4 which state the quality functions of stability.

$$R_n(t) = \sum_{i,j} \left((e^{(AD^{-1}-I)t})_{i,j} \frac{k_j}{2m} - \frac{k_i}{2m}\frac{k_j}{2m} \right) \delta(c_i, c_j) \tag{4.27}$$

$$R_c(t) = \sum_{i,j} \left((e^{t(A-D)\frac{n}{2m}})_{i,j} \frac{1}{n} - \frac{1}{n}\frac{1}{n} \right) \delta(c_i, c_j) \tag{4.28}$$

where $D_{ii} = k_i$ and A, m, k_i, n are the Adjacency matrix, number of edges, degree of node i and number of nodes respectively. For further discussion of stability see Section 2.6.4.

Stability is different to our novel measure in two ways. Firstly, the underlying statistic each measure uses is different: our novel measure considers all non-

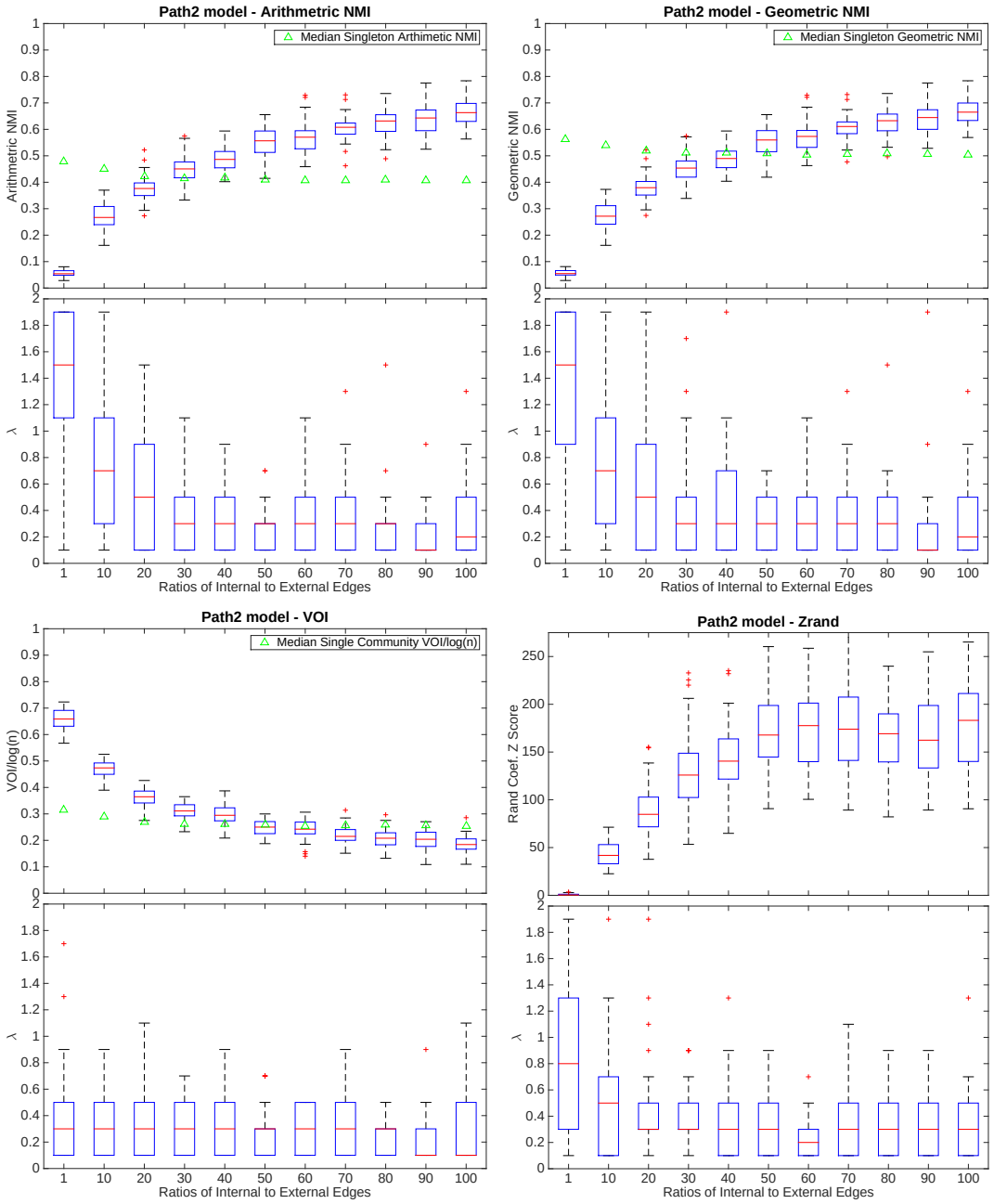

Figure 4.6: Performance of community detection using Q_2^{Path} modularity at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $\lambda_1 = 0.1, 0.3, ..., 1.9$ and $\lambda_2 = 1$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

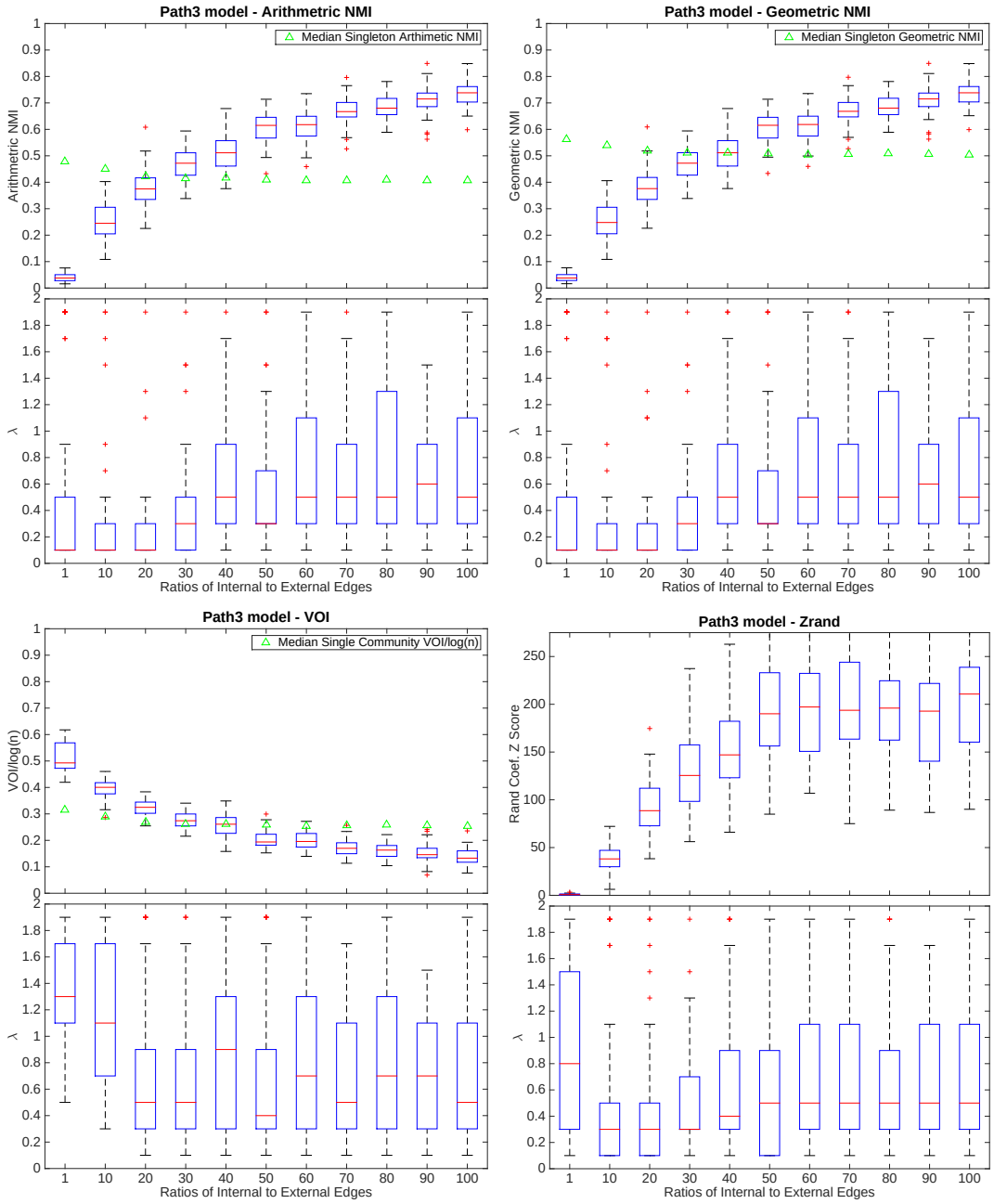

Figure 4.7: Performance of community detection using Q_3^{Path} modularity at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $\lambda_1 = 0.1, 0.3, ..., 1.9$ and $\lambda_2 = 1$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

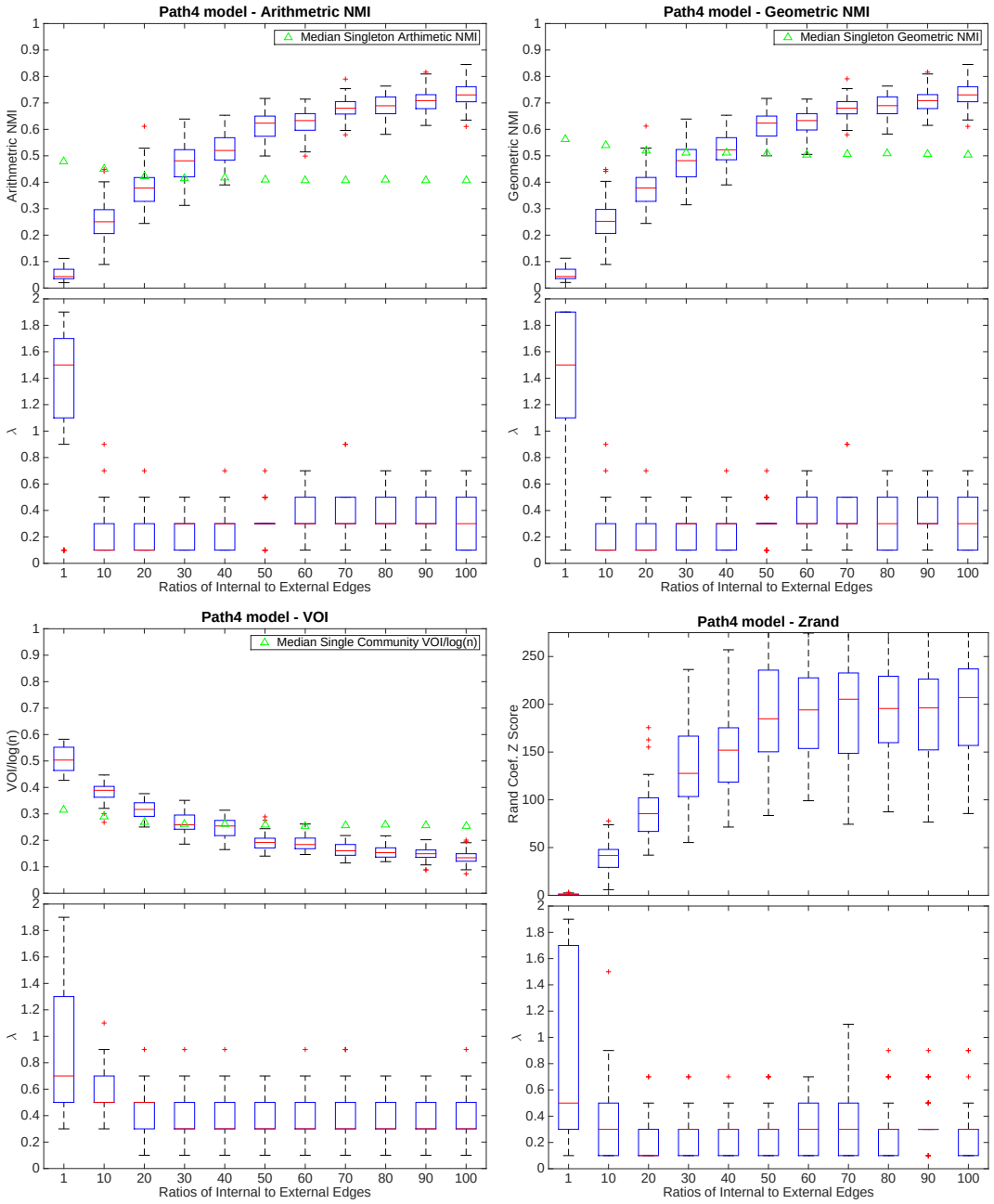

Figure 4.8: Performance of community detection using Q_4^{Path} modularity at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $\lambda_1 = 0.1, 0.3, ..., 1.9$ and $\lambda_2 = 1$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

intersecting paths between node i and j equally, whereas stability weights a random walk by the probability of randomly walking along that path. Secondly the null model in both forms of stability (R_n and R_c) is a function of either the number of nodes or the degree distribution of the network. Importantly it does not and cannot take into account any additional structural features of the network. In contrast the path-based modularity method allows us to account for some non-pairwise interactions that we believe not to be a relevant feature of the network by fitting a separate null model for each of the path lengths. In this case the uninteresting features are features of the sampling procedure.

We run the comparison we performed on the other null models, on stability [1]; To adjust for the fact that we do not have a good understanding of appropriate resolution parameter value (Markov Time) t for our test, we use the range 10^{-2}-10^{2} and we carry out 4 times as many measurements $(10^{-2}, 10^{-1.9}, ..., 10^{2})$ as in the other null models. However, to make the results comparable to the other approaches we only use one optimisation of the Louvain algorithm at each resolution.

The results from the combinatorial Laplacian can be found in Figure. 4.9 and the results from the normalised Laplacian can be found in Figure 4.10.

We observe that stability performs much better than the non path-based methods at discovering the structure of the network. With respect to VOI the normalised Laplacian outperforms the trivial partition for ratios greater than 40 and the combinatorial Laplacian distinctly outperforms the trivial partition for ratios greater than 60. For arithmetic NMI the trivial partition is outperformed for ratios above 30 and 40 for the normalised and combinatorial respectively and in the case of the geometric NMI for ratios greater than 40 and 60. Echoing previous results the Zrand score shows that in both versions of stability the partitions uncovered are associated with the underlying partition. We also observe that similar to Q_l^{Path} for large ratios the

[1]Ran using software found at `https://github.com/michaelschaub/PartitionStabilit://github.com/michaelschaub/PartitionStability`

distribution of successful λs have very similar distributions in all of the statistics in the normalised case. Interestingly, in all four of the similar measures the normalised Laplacian appears to perform better than the combinatorial Laplacian.

These observations clearly indicate that stability is able to discover structure the embedded partition in these networks in a way that the configuration model is not able to do. In the next section we will compare the performance of stability and Q_l^{Path}.

4.3.5.1 Comparing the different approaches

To compare the relative performance of each of the approaches we have discussed in this chapter we compare the mean performance with respect each similarity measure across the ensemble for each ratio $\frac{c_{in}}{c_{out}}$. The result can be found in in Figure 4.11.

We observe that with respect to VOI and the two NMI statistics with a ratio ($\frac{c_{in}}{c_{out}}$) of 10 or 20 none of the greater than pairwise measures produce communities that in any way approximate the supposedly induced communities in that they do not outperform the trivial partitions. If the induced community structure still exists in the sampled network then it is too subtle for any of the current methods to detect with respect to these statistics. The result from Zrand indicates that the communities being uncovered are related to the underlying structure indicating that some of the structure remains but it is unclear if a method could be constructed to uncover enough structure such that the information theory based statistics can detect the similarity.

For ratios of 30 and 40 the picture is mixed with the greater than pairwise methods performing well in Arithmetic NMI, however this performance is not replicated in the other information theory based statistics.

The results for ratios $\frac{c_{in}}{c_{out}} > 40$ are more conclusive. In this range we observe that Path Modularity based on paths of length 3 and the version paths of length 4 performs favourably compared to all other methods. The null model based on paths of length 2

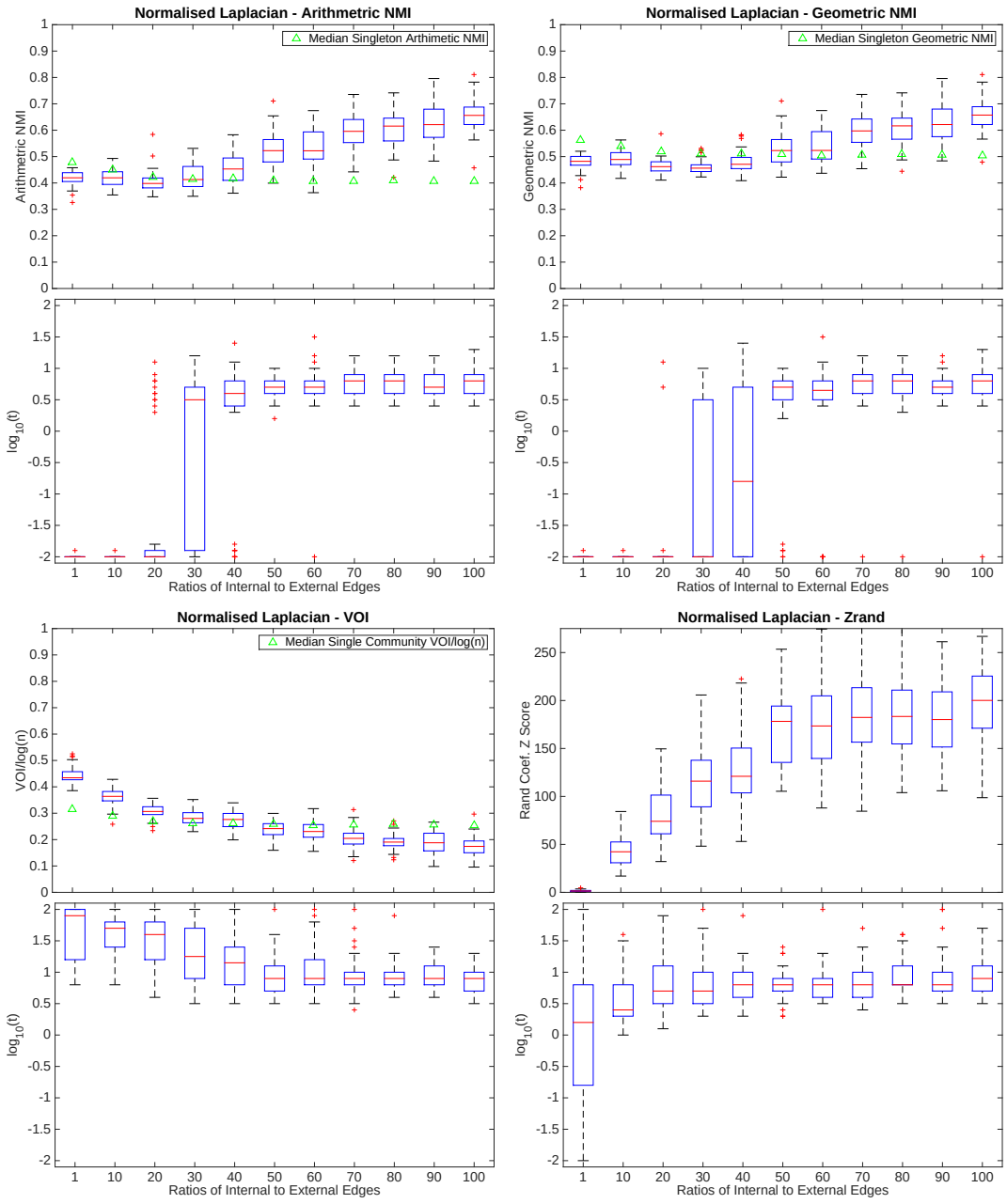

Figure 4.9: Performance of community detection using Norm. Laplacian ($R_n(t)$) at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $t = 10^{-2}, 10^{-1.9}, ..., 10^{2}$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

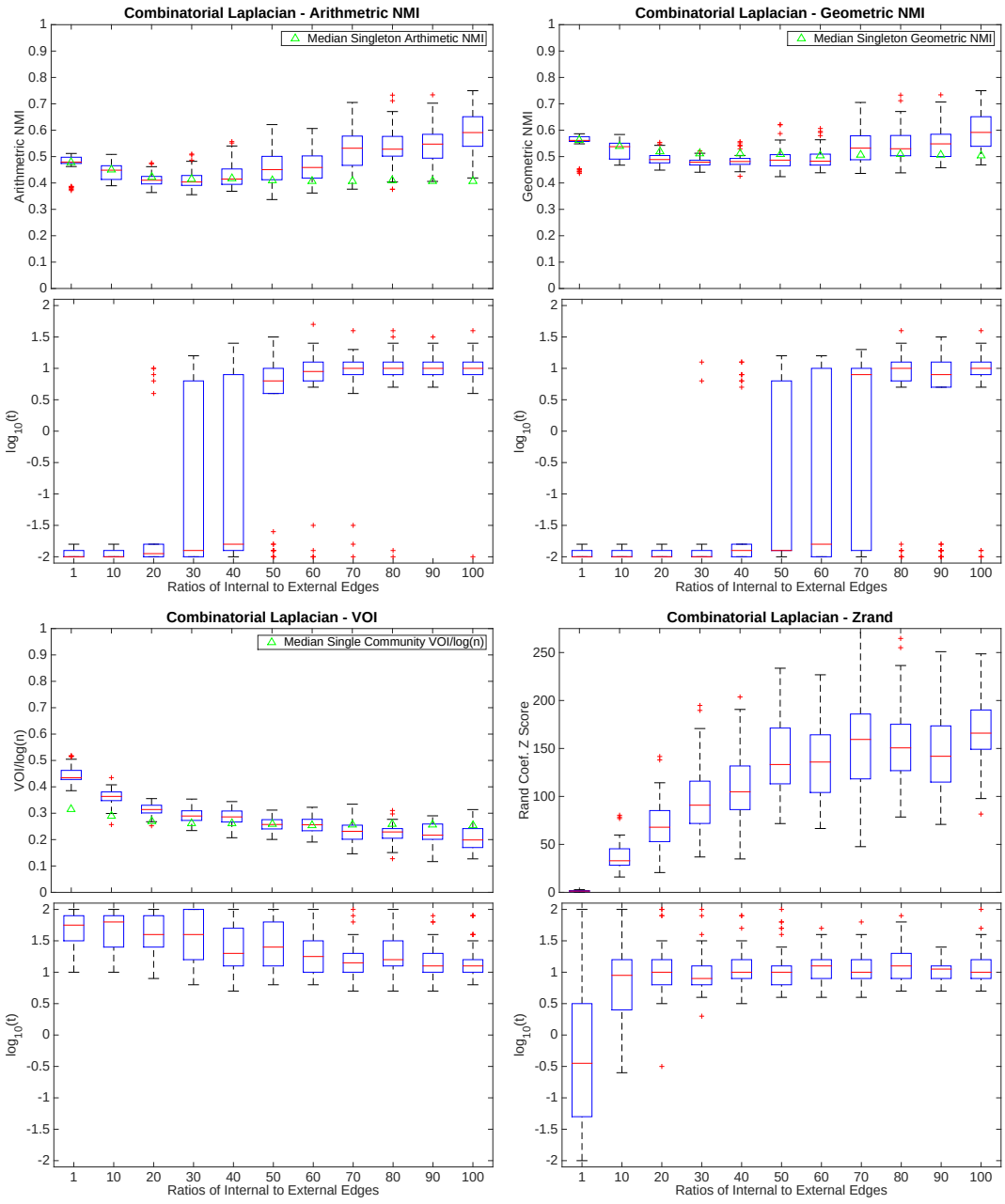

Figure 4.10: Performance of community detection using Comb. Laplacian ($R_c(t)$) at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $t = 10^{-2}, 10^{-1.9}, ..., 10^{2}$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

performs better or comparably to the Stability measures in certain statistics, namely the NMI based measures however in the other statistics this method is outperformed.

Thus for path lengths greater than 2 the Path Modularity compares favourably to Stability for ratios greater than 40 and performs at least as well for the other ratios $\frac{c_{in}}{c_{out}}$ tested when we consider the effect of trivial partitions.

A possible explanation why our path-based extension to modularity outperforms Stability is simply that constructing the null models for paths of length three and four account better for some features of the network. A second explanation could be that a shortest path sampled network is better represented by the non-intersecting path counts rather than random walks. Further study is needed to fully understand why this method outperforms Stability.

4.4 Conclusion

In this chapter, we constructed a novel community detection quality function which incorporates both longer simple paths and Monte Carlo sampled estimates for the properties of the sampling scheme. To test the performance of our community detection quality function we constructed a benchmark based on shortest path samples from a stochastic block model. We performed community detection and measured performance using four similarity measures.

Our novel formulation of modularity (Path Modularity) which considers paths of length greater than one appears to to able to detect communities on these simple networks. This suggests that to detect communities in networks sampled using shortest path sampling we need to take into account greater than pairwise structure in the network.

For ratios $\frac{c_{in}}{c_{out}} > 40$ we are able to detect partitions that have that produce a stronger similarity score in all four statistics than the trivial partitions. Further,

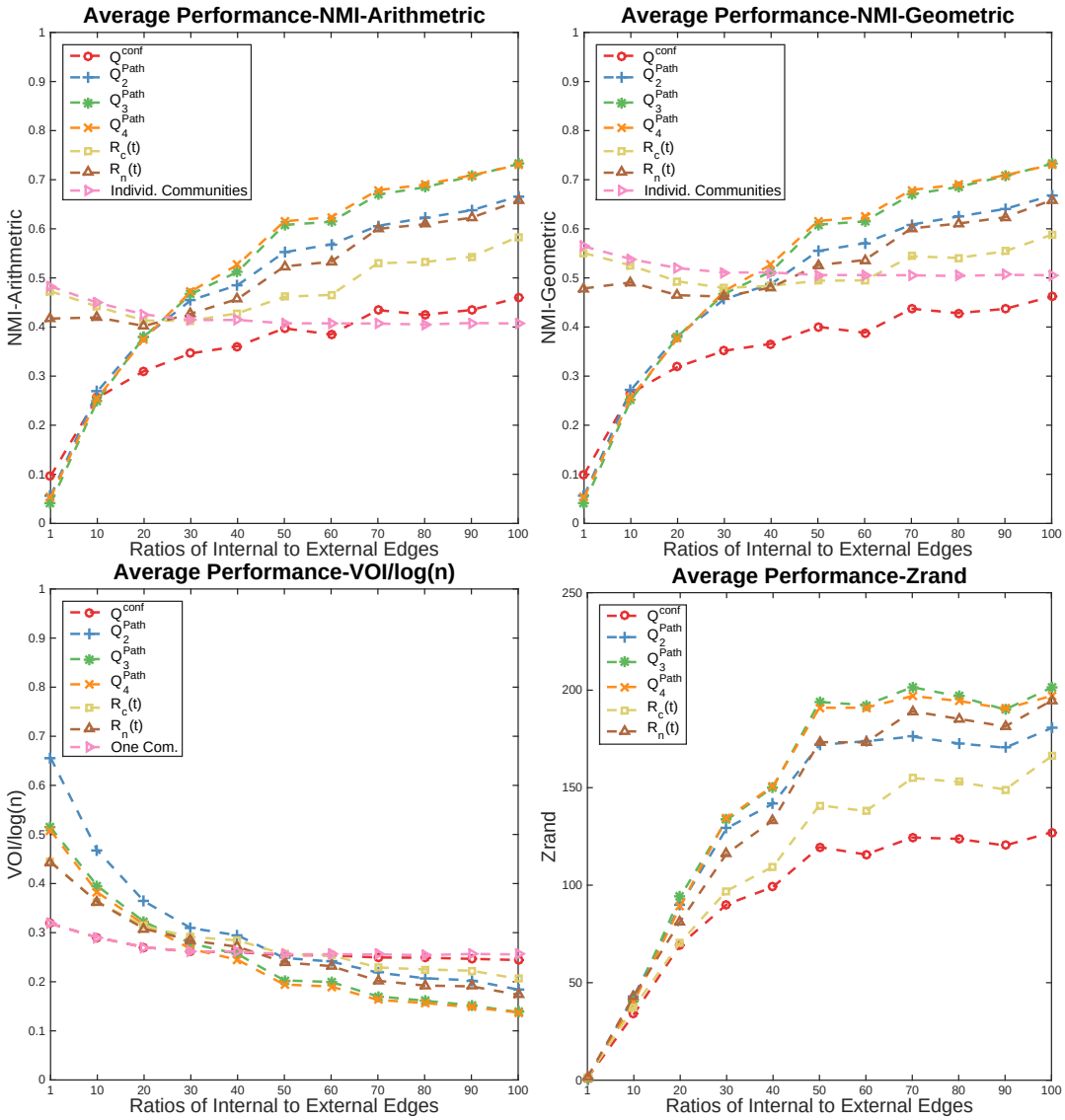

Figure 4.11: Comparison of the performance of different community detection approaches at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$). (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). At each resolution we plot the mean performance with 50 replicates for each ratio. Community detection is performed at multiple values of the resolution parameter and the best performance is used for the average.

again for ratios greater than 40, with path lengths of 3 or 4 the path method also performs consistently as well as Stability for lower ratios $\frac{c_{in}}{c_{out}}$ and substantially better

for higher ratios, even through for Stability we test four times more parameters than for our null model.

Further, on inspection of the parameters that produce the best performing statistics in Q_4^{Path} we discover that it is dominated by small values of λ_1 indicating that there perhaps are better results that could be extracted using this method if we concentrated the search on this area.

Further, when interpreting these conclusions we will have to bear in mind that this is conditioned on the assumption we made to test these approaches. First, that using networks with strong community structure induces a community structure in the sample. Second, we fix both the number of seeds and the density of the network, preliminary computational evidence suggests that these parameters could have a substantial impact on performance, and thus care must be taken to generalise these results to different parameter regimes.

Regardless, the method appears to outperform the other tested approaches on this benchmark, which indicates that it can pick up global structure better than the other approaches, and if the assumption holds, also sampled structure.

However, further work is needed to develop a full understanding of these performance differences.

Is our path-based approach simply capturing more information and fitting a more complex null model or it there more to it. One way to shed light on this would be to test our path based approach against the recently published higher order Markov version of Stability (Ref. [145]), which with the correct selection of parameters would allow a non-backtracking random walk. The interest in community detection approaches related to the approach discussed in this chapter (using simple paths [119] or non backtracking walks [99]) may suggest that such a comparison may be interesting to explore. In Chapter 5 we will make some initial steps in this direction in the case of a general unweighted network.

Another avenue worth pursing is the detection of so called non-clique like communities. An example of this is the ring of rings in Ref. [149]. This paper reformulates modularity as a random walker process with one step and demonstrates Stability's ability to utilise more than one step. Testing our method directly on these benchmark graphs would be difficult as for now it is only defined on unweighted networks, and a key element of the benchmark is the weight of the connections. However, our method and other similar methods (see section 4.3) incorporate combining information from different path lengths, and therefore may do well on this benchmark with the suitable generalisation to weighted graphs.

Further extensions of this work could be to test this approach on other path sampling techniques and to take this back to biological networks and test if the communities detected using the path-based modularity have any particular biological meaning. Another potentially interesting application of this work would be on other networks that are fundamentally constructed from paths rather than from individual nodes with given connections. Some examples include metabolic networks, biological pathway networks and traceroute networks (a survey technique used to construct the internet graph by sending packets to other machines around the world).

Chapter 5

Path based community detection

Symbol List

Symbol	Description
$A_{ij\Upsilon}$	The number of non-self intersecting paths between nodes i and node j of length Υ
$B^*(m, r, \nu)$	Part bound on function that does not dependent on n
$C(x)$	The number of contiguous subsequences that are generated by the edge set x
D_{max}	The size of the largest dependency neighbourhood
$D(r, k, m, \nu)$	The part of the upper bound for a function that does not dependent on n
$F(x)$	Cumulative density function.
F_W	Set of f_h for $h \in H_W$
$G(n, \frac{c}{n})$	A Erdős Rényi graph in the sparse case
$G(n, p)$	An Erdős Rényi graph, with n nodes and each edge exists with probability p
H, H_1, H_2	Sets of functions
H_K	Set of functions related to the Kolmogorov distance
H_W	Set of functions related to the Wasserstein distance
H_{bW}	Set of functions related to the bounded Wasserstein distance

I	An index set, for the $l = 2$ case this refers to the union between I_1 and I_2
$I^{(l)}$	Union of all index sets upto l
I_x	Index sets containing tuples which consist of nodes in order of appearance of a path. Paths of length x are in I_x.
J	The supremum of |f"(x)| for a given function f
$K_b^{\max}$	Upper bound on the number of contiguous paths formed from the edges present in both α and β
$K_g^{\max}$	Upper bound on the number of contiguous paths formed from the edges present in γ which are also present in α and β
$\mathbb{K}_{X_i}$	The dependency neighbourhood of X_i, X_i is independent of $\{X_j : j \notin K_i\}$.
M_i	Number of occurances of an ith largest unique element in the sequence of random variables
N_{samp}	Number of random samples drawn to construct an empirical CDF
O_i	Summation indices which restricts the number of edges in a path or the number of overlapping edges between paths
$P_{ij\Upsilon}$	The expected number of paths of length Υ between i and j
$Q_l^{Path-Diff}(g)$	The difference between the path-modularity and the path modularity with a single community mathematically this is equivalent to $Q_l^{Path}(\mathbf{1}) - Q_l^{Path}(g)$,
$Q_2^{Path}, Q_3^{Path}, Q_l^{Path}$	Path modularity using paths of length 2, 3 and l respectively
V_1, V_2	Two sets of nodes such that $V_1 \cup V_2 = V$, $V_1 \cap V_2 = \emptyset$ and $\|V_1\| = \|V_2\| = \|V\|/2$
V_{ik}	The remaining terms in the W_i not included in W_{ik} such that $W_i = W_{ik} + V_{ik}$
W	Random variable which is the sum of potentially dependent random variables. Often it is required to have mean 0 and variance 1
W_i	A summation of X_i which is independent of X_i
W_{ik}	A summation that is independent of X_i and X_k
S_1, S_2	Arbitrary sets

$T, \hat{T}$	The cumulative density function of $\tilde{\zeta}$/The estimated cumulative density function of $\tilde{\zeta}$.
X, Y	Random variables
X_α	The random variable associated with path α. The product of an indicator variable for the existence of the path associated with α and the weight associated with the path.
X_i	A random variable.
Y_α	Standardised version of X_α, $Y_\alpha = \frac{1}{(\mathrm{Var}(\zeta))^{0.5}}(X_\alpha - E[X_\alpha])$
Z	Standard normal distribution
Z_i	A summation of random variables that are not included in W_i, so that $W = W_i + Z_i$
Φ_1, Φ_2	In the $l = 2$ represents the contribution to the bound by terms for which the index in the first summation is a member of I_1 for Φ_1 and I_2 for Φ_2
$\Phi_1^{(i,j)}(\alpha)$	The contribution to the $l = 2$ bound by terms for which the first index in the summation is from I_1, the second index in the summation is from I_i and the third is from I_j
$\Phi_2^{(i,j)}(\alpha)$	A further term representing the contribution to the bound by terms that for which the first index in the summation is from I_1 the second index in the summation is from $\mathfrak{K}_\alpha^{(i)}$ and the third is from $\mathfrak{K}_\alpha^{(j)}$
$\Psi_O(\alpha, \beta, \gamma)$	Subset of $(I^{(l)})^3$ such that for $(\psi_1, \psi_2, \psi_3) \in \Psi_O(\alpha, \beta, \gamma)$ $\mathfrak{O}(\alpha, \beta, \gamma) = \mathfrak{O}(\psi_1, \psi_2, \psi_3)$
$\Theta(n)$	A bound on the growth of a function, a function is $\Theta(n)$ if it is bounded from above and below by functions that are linear in n
Υ	Summation index over path lengths
α, α^*	An arbitrary tuple representing a path
τ	Number of blocks in a stochastic block model
β, β^*	An arbitrary tuple representing a path
δ_{ab}	Kronecker delta, which is 1 if $i = j$ and 0 otherwise
ϵ	A symbol used to represent the derived bound on $\lvert E(Wf(W) - f'(W))\rvert/C$

$\eta(\alpha)$	Set of each edge involved in path α, where each edge is represented by a set of the nodes involved in the edge
γ, γ^*	An arbitrary tuple representing a path
$\hat{\zeta}$	Sum of Y_α over all paths in I.
κ	The overlapping edge between two paths
κ_1, κ_2	Constants used in the derivation of simple bounds on f_1 and f_2
λ	For notational convenience we let $\lambda = \lambda_1$
λ_1, λ_2	Resolution parameters for path modularity
$\mathbb{N}$	Natural numbers
$\mathbf{1}$	A vector consisting of only of 1s used to represent a single community
$\mathbf{d}$	Vector of nodes which form the contiguous overlapping between two paths
$\mathbb{K}_\alpha(I)/\mathbb{K}_\alpha$	The set of $\beta \in I$ such that X_β is dependent on X_α. For notational convenience we drop the (I)
$\mathbb{K}_\alpha^{(i)}/\mathbb{K}_\alpha^{(i)}$	The set of $\beta \in I$ such that the path associated with α and β intersect on i edges
$\mathfrak{A}_{ij}$	In contrast to A_{ij} which is an observation of network, $\mathfrak{A}_{ij}$ is a random variable which is 1 with probability p and 0 otherwise
$\mathfrak{B}_1$	The number of ways of selecting the overlapping edges from one sequence to place in another
$\mathfrak{B}_2$	The number of ways of placing overlapping edges from one subsequence into another
$\mathfrak{B}_3$	The number of ways of selecting the remaining nodes in a subsequence after the overlapping edges have been placed
$\mathfrak{C}(a, b, g, k_b)$	A constant used to form a more concise expression, it is related to the number of ways of selecting and placing the overlapping edges from α to β
$\mathfrak{D}(a, b, g, k_b)$	A constant used to form a more concise expression, it is related to the number of ways of selecting and placing the overlapping edges from α and β to γ

$\mathfrak{E}$	A constant which collects many terms which do not rely on n to produce a more concise expression for the bound in the case $l > 2$
$\mathfrak{F}_C$	The set of all functions with bounded first and second derivatives
$\mathbb{F}_V, \mathbb{F}_W$	Expressions related to the variance in the general l case. They are used to form a more concise expression.
$\mathfrak{F}_W$	Superset of F_W, which maintains the restriction on the absolute value of the function and its first and second derivative.
$\mathfrak{H}$	Set of all $h'(x) - xh(x)$ for all absolutely continuous h which also obeys the following inequality $E\|h'(Z)\| < \infty$
$\mathfrak{I}_X$	A symbol representing the contributions to the bound when the condition X is true
$\mathfrak{K}_\alpha^{(1)}$	A set consisting of α and all elements of I_1 that have a dependence on α
$\mathfrak{K}_\alpha^{(2)}$	A set consisting of all $\beta \in I_2$ which shares one edge with the path associated with α
$\mathfrak{O}(\alpha, \beta, \gamma)$	A function which constructs a vector describing the overlap between each of the paths
$\mathfrak{S}(v)$	The set of tuples of all permutations of v
ν	A constant such that $0 < \nu < 1$ and $\nu n > 2l$.
ψ_1, ψ_2, ψ_3	An arbitrary tuple representing a path
ρ, ρ_1, ρ_2	Variables used to represent arbitrary quantities in derivations of upper bound
θ	An arbitrary node
$\omega(r, m)$	Number of ways to place β a tuple of length $m + 1$ into α a tuple of length $r + 1$
$\tilde{\zeta}^{(l)}$	A normalised version of ζ^l, with the mean removed and standardised to be variance 1, this is also equivalent to summing over Y_α rather than X_α
ζ	Sum of X_α over all paths in I.
$\zeta^{(l)}$	Sum of ζ_r for r from 1 to l
ζ_r	Sum of X_α over $\alpha \in I_r$
a	Path length
a_1, a_2	Arbitrary constants used for upper and lower bounds

a_{in}, a_{out}	In the sparse case of the stochastic block model these constants describe the probability of an edge at different network sizes, i.e. $p_{in} = \frac{a_{in}}{n}$ and $p_{out} = \frac{a_{out}}{n}$
$b(\alpha, m)$	A function counts the number of $\beta \in I_m$ which contains all edges from α
$b^*(m, r, \nu)$	The part of the lower bound for a function that does not dependent on n
c	Constant such that $p = \frac{c}{n}$
d	Arbitrary integer
$d(r, k, m, \nu)$	The part of the lower bound for a function that does not dependent on n
$d_H(u, v)$	A distance function between probability distributions which is a function of the set of functions H
f	A generic function. Often it is assigned to a polynomial in p and used in the bounding the polynomial.
$f_1(p), f_2(p)$	Functions used as examples of simple techniques to bound expressions from above
f_h	For given h the solution to f(x)-xf'(x)=f(x)-E[h(Z)]
g	A vector of community assignments or block assignments in a stochastic block model, such that g_i is the block which is assigned to node i.
g^*	Vector of community assignment with two communities of equal size
g'	A sparse Erdős Rényi graph
h	Generic Function
i	Arbitrary integer
k	The number of edges shared between paths
k_g, k_b	Summation indices representing the number of contiguous paths formed by overlapping edges
l	Largest path length considered
n	Number of nodes in the network
n_b	Number of nodes in each block of the stochastic block model
p	Probability of connection in a Erdős Rényi graph

p_{in}, p_{out}	Probability of connections between node in the same group and in different groups respectively in a stochastic block model
q	Arbitrary integer
r	An arbitrary path length
m	An arbitrary path length
s^*	Desired significance level
s	Number of edges in a path
$t(s), t, t^*(s, n, \lambda)$	Lower bound on the value of $Q_2^{Path-Diff}(g)$ to obtain a at least a given significance
u, v	Probability distributions
x, y	Dummy variable used in multiple proofs
x_i	Arbitrary numbers
x_i^*	The ith largest unique element in a random sample.
x_j, y_j, d_j	Arbitrary nodes
y_i	Arbitrary integer

5.1 Introduction

In this chapter we explore the community detection method Path Modularity which we developed in chapter 4 to detect communities a path sampled network in a different context. Rather than using an empirical null model which is specified by a Monte Carlo search of the resultant space, now we use an analytic null model.

We will explore this analytic null model in two distinct ways. First we show that in a stochastic block model, for certain parameters there are considerably more paths of length 2 between pairs of nodes in the same community than in different communities, indicating that in general networks there is additional information that can be exploited in length 2 paths and therefore incorporating these approaches may prove profitable.

Second, we explore the statistic that underlies the approach more directly. Focus-

ing on network bisection (splitting the network into two equally sized pieces), we show that path modularity given in Eq. (4.9) constructed in Section 4.3 is approximately normally distributed and that it converges to a normal distribution with a bound on the rate of convergence of leading order $n^{-0.5}$ in n, in the sparse Erdős Rényi model where $G(n,p) = G(n, \frac{c}{n})$ where c is a constant.

In essence we are interested in exploring how Path Modularity performs in a wider context. We take inspiration from the work of Franke and Wolfe [63], which shows that, for a fixed group assignment, Newman Girvan modularity for large network size and with suitable normalisation converges in distribution to a normal distribution as the network size goes to infinity. In Ref. [63] results are used to compute p-values for the modularity measured in real world networks with a fixed known division. Thus it can be tested if the modularity of this group assignment is significant.

In this chapter we explore Path Modularity and test make a similar but weaker claim about Path Modularity (Eq. (4.9)), namely that the weighted sum of paths of length upto l between two equally sized groups is asymptotically normally distributed. This result can then be (trivally) extended to apply to a generalised version of Eq. (4.9) with a uniform null model. One key challenge in performing this analysis on Path Modularity is that the paths are not independent and hence the Path Modularity is not a sum of independent random variables. To obtain a normal approximation in the presence of this dependence we use a method called Stein's method, which we is described in the detail in Section 2.7.

5.2 Path based null model

5.2.1 Paths in non path sampled networks

A key feature of Path Modularity (4.9) in Chapter 4 is to consider simple paths of length greater than 1 in community detection, and to fit separate null models for

each of these paths. It is clear that this approach makes sense in a system sampled from paths, but does it also make sense in a standard network, where all edges are independent?

To answer this question we estimate the expected number of paths of given length between a pair of nodes, both for nodes in the same block and nodes from different blocks, where the network is generated by a stochastic block model with n nodes made up of τ blocks of size n_b, probability p_{in} of an edge within a community and p_{out} of an edge between communities. For node i let g_i be the block to which node i is assigned.

First, we calculate the expected number of paths of length 2 between a pair of nodes in a stochastic block model as follows: if i and j are in the same block, then the expectation is

$$E[A_{ij2}|g_i = g_j] = \sum_{k \notin \{i,j\}} \left(\delta_{g_i g_k} p_{in}^2 + (1 - \delta_{g_i g_k}) p_{out}^2\right) = (n_b - 2)p_{in}^2 + p_{out}^2(n - n_b) \quad (5.1)$$

where δ_{xy} is 1 if $x = y$ and 0 otherwise. The first term in (5.1) concerns paths with three nodes in the same group, and the second term represents a path with a node of an alternative block in the path.

Similarly the number of paths of length 2 between pairs of nodes that are not in the same block is given by:

$$E[A_{ij2}|g_i \neq g_j] = (n - 2n_b)p_{out}^2 + p_{in}p_{out}(2n_b - 2). \quad (5.2)$$

The first term in (5.2) represents a path using a node which is not part of either of the blocks of node i and node j and the final term in (5.2) represents the possibility that the middle node is part of one of the groupings of node i and node j.

The difference between these expectations is (simplifying using $n_b - 2 \approx n$ and

$2n - 2 \approx 2n$):

$$E[A_{ij2}|g_i = g_j] - E[A_{ij2}|g_i \neq g_j] \approx n_b p_{in}^2 + (n - n_b - (n - 2n_b))p_{out}^2 - p_{in}p_{out}(2n_b). \tag{5.3}$$

Simplifying and collecting the terms:

$$E[A_{ij2}|g_i{=}g_j] - E[A_{ij2}|g_i \neq g_j] \approx n_b(p_{in} - p_{out})^2, \tag{5.4}$$

independent of n, the total number of nodes.

Therefore on a network with $p_{in} = 0.1$ and $p_{out} = 0.01$ and a group size of 1000 the average difference of the number of paths of length 2 between a randomly selected pair of nodes in the same block and the number of length 2 paths between nodes in different groups should be ≈ 8. If $p_{in} = 0.01$ and $p_{out} = 0.001$ we need a group size of 100,000 before we see the same difference.

Further, if we consider the sparse network case, that $p_{in} = \frac{a_{in}}{n_b}$ and $p_{out} = \frac{a_{out}}{n_b}$, then considering paths of length 1 and of length 2,

$$E[A_{ij1}|g_i = g_j] - E[A_{ij1}|g_i \neq g_j] = \frac{a_{in} - a_{out}}{n_b}, \tag{5.5}$$

$$E[A_{ij2}|g_i = g_j] - E[A_{ij2}|g_i \neq g_j] = \frac{(a_{in} - a_{out})^2}{n_b}; \tag{5.6}$$

they decay at the same rate with network size.

5.2.2 The Path Modularity under a uniform null model

The approach we explore is the Path Modularity developed in Chapter 4. While it was initially developed for path sampled networks, its ability to find communities was based on it exploiting the information that is present in simple (non self intersecting) paths, and thus maybe this method could work more generally. This heuristic is

further supported by the use of a very similar method to prove the upper bound of the detection limit. [119]

We note that as we observed in Chapter 4 there are related approaches to this problem. The Path Modularity statistic is related to the Katz similarity score [184]. Further, many approaches use paths to detect communities, for example Ref. [64, 99, 101] Independent of this work, Ref. [184] uses the weighted sum of simple paths as a similarity measure for a k-means based community detection. Finally, Ref. [121] has constructed a version of modularity based on non-simple paths. There is clearly considerable interest in this area, but to the author's knowledge we are the first to analyse this community detection statistic by relating it to a normal distribution with Stein's method.

Note, as mentioned in Chapter 4 independently of this work Ref. [64] developed a similar quality function. Our approach in this chapter is different as we have explicit resolution parameters and we fit a different model. However, it is possible with some extension that the mathematics we develop in this chapter could apply to this null model also.

The expression for Path Modularity as defined in Equation 4.9 in Chapter 4 is:

$$Q_l^{Path} = \sum_{\Upsilon=1}^{l} \lambda_1^{\Upsilon} \sum_{i,j=1}^{N} (A_{ij\Upsilon} - \lambda_2 P_{ij\Upsilon}) \delta(g_i, g_j) \tag{5.7}$$

where $A_{ij\Upsilon}$ is the number of simple paths between node i and node j of length Υ and $P_{ij\Upsilon}$ is the expected number of simple paths of length Υ between node i and node j.

The empirical null model for P_{ijl} used in the Chapter 4 is based on the particulars of shortest path sampling. Now we want to construct an analytic null model which is based on the use of simple paths. (If we were to allow non-simple paths, then any additional independent signal that we gain from paths could be diluted by the presence of repeated lower length paths - see Section 4.3 for further discussion).

Taking inspiration from the work of stability [101] (see section 2.6.4) and of Refs. [137, 141, 169], we construct a null model based on the uniform null model. Here the uniform null model is an ER graph. The edge probability can be estimated from the network or specified by parameter.

As edges are independent and identically distributed, a normal approximation for Q_1^{Path} is straightforward using the central limit theorem, when the assignment to groups is given. Note, Q_1^{Path} is the same (up to multiplication with λ_1) as the formulation of modularity in [137] and is strongly related to that in Ref. [169].

For example, with an ER graph with N nodes that is partitioned into α blocks of size n each:

$$Q_1^{Path}(g) = 2\lambda_1 \sum_{i>j} (A_{ij} - \lambda_2 p)\delta(g_i, g_j). \tag{5.8}$$

where g is a vector which represents the community assignment. For the case we are considering, i.e. α blocks of size n we let $g_i = i - 1 \mod \alpha$ for $i = 1, ..., \alpha n$.

Note, for simplicity we do not include the self loops. Therefore, we can compute the expectation of Q_1^{Path} by adding the expectation of each edge within a block, noting that there are $N\frac{(n-1)}{2}$ within block edges. We can compute the variance by using the fact that Q_1^{Path} is a sum of Bernoulli random variables, and hence once we remove the mean, the contribution of the null model in Q_1^{Path} is removed. Therefore we can use the the variance of a binomial distribution. This results in the following expressions:

$$E[Q_1^{Path}(g)] = 2\lambda_1(1 - \lambda_2)pN\frac{(n-1)}{2}, \tag{5.9}$$

$$\text{and } Var(Q_1^{Path}(g)) = 4\lambda_1^2(1 - p)pN\frac{(n-1)}{2}. \tag{5.10}$$

To compute results for higher path lengths we have two options. First we could use the same empirical approach we used in Chapter 4 and estimate the expected number of paths by simply averaging over the computed matrix of path lengths. This has the advantage of adjusting for some of the structure in the network, and

the approximation being used for each path length has similar levels of accuracy. However, this approach is not without problems. Notably, in small networks we run the risk of overfitting. As an alternative we mathematically construct the expected number of paths of length l between two given nodes i and j, given an underlying ER graph $G(n,p)$ with assumed known p. This is given by:

$$E[A_{ijl}] = E[\sum_{|P_a|=l} p^l] = \frac{(n-2)!}{(n-1-l)!}p^l \tag{5.11}$$

where P_a is a potential path between i and j. We then substitute the estimated value of p into the function to get an estimate of the number of paths we would expect.

Using the analytic form, we obtain a final form for Path Modularity under the ER null model as follows:

$$Q_l^{Path}(g) = \sum_{\Upsilon=1}^{l} \lambda_1^{\Upsilon} \sum_{i,j} \left(A_{ij\Upsilon} - \lambda_2 \frac{(n-2)!}{(n-1-\Upsilon)!}p^{\Upsilon} \right) \delta(g_i, g_j) \tag{5.12}$$

5.3 Constructing an appropriate test statistic

In order to better understand Path Modularity we use an approach similar to Franke and Wolfe [63], and relate the distribution of $Q_l^{(path)}(g)$ to a normal distribution. Here we are interested in the distribution of the path-modularity statistic with a random chosen community structure (with certain constraints) over an ensemble of networks.

For our ensemble we focus on a sparse Erdős Rényi graph with n nodes and $\frac{c}{n}$ edges. This is a very natural setting as our null model is also based on an Erdős Rényi graph.

However, rather than the Path Modularity we will focus on a slightly different quantity, the sum of the weight of the edges minus expectation between groups rather than within them which we will call $Q^{Path-diff}$.

. This is equivalent to the cut size of the weighted graph formed with edges

between node i and j with weight $\sum_{\Upsilon=1}^{l}(A_{ij\Upsilon} - P_{ij\Upsilon})$ where $P_{ij\Upsilon}$ is the null model we defined in the previous chapter. In fact as we will focus on the case of two equal groups this is related to a graph bisection, however crucially with a fixed partition, rather than the minimum bisection. Investigating graph cuts is well studied problem, see for example [27, 153, 180].

Mathematically we consider

$$Q_l^{Path-Diff}(g) = \sum_{\Upsilon=1}^{l} \lambda_1^{\Upsilon} \sum_{i,j} \left(A_{ij\Upsilon} - \lambda_2 \frac{(n-2)!}{(n-1-\Upsilon)!} p^{\Upsilon} \right) \delta(g_i \neq g_j) \tag{5.13}$$

for a given partition g. We note that this is equivalent to $Q_l^{Path-Diff}(g) = Q_l^{Path}(\mathbf{1}) - Q_l^{Path}(g)$, where $\mathbf{1}$ is a partition where all nodes are in the same group and $Q_l^{Path}(g)$ is defined in (5.12).

We focus on fixed partition with two equally sized communities or blocks, which we shall name V_1 and V_2. Thus for this partition our quantity of interest becomes:

$$\begin{aligned} Q_l^{Path-Diff} &= \sum_{\Upsilon=1}^{l} \lambda_1^{\Upsilon} \sum_{i \in V_1} \sum_{j \in V_2} \left(A_{ij\Upsilon} - \lambda_2 \frac{(n-2)!}{(n-1-\Upsilon)!} p^{\Upsilon} \right) && (5.14) \\ &= \left(\sum_{\Upsilon=1}^{l} \lambda_1^{\Upsilon} \sum_{i \in V_1} \sum_{j \in V_2} A_{ij\Upsilon} \right) - \lambda_2 \left(\sum_{\Upsilon=1}^{l} \lambda_1^{\Upsilon} \sum_{i \in V_1} \sum_{j \in V_2} \frac{(n-2)!}{(n-1-\Upsilon)!} p^{\Upsilon} \right). && (5.15) \end{aligned}$$

We observe that the second term is a non random function of the parameters of the null model, with the caveat that we use the probability of connection from our Erdős Rényi ensemble rather than estimating it from the resultant graph. Thus our strategy will be as follows. We will focus on bounding the difference in Wasserstein distance between a normal distribution and the first term using Stein's method. Then we will reintroduce the second term, which will shift the mean of the corresponding normal distribution.

Our interest in considering $Q_l^{Path-Diff}$ instead of Q_l^{Path} is inspired by the use of cut statistics in the analysis of the resolution limit [102], in which it is used to

decide whether a community should be split under modularity. Thus, this statistic was chosen with a view to explore the resolution limit of our modified community detection approach. However, this chapter only represents the first stage in this work.

Analysis of this statistic has other benefits. It can be used in a very similar way as modularity is used in Franke and Wolfe [63], in which significance of the modularity of a fixed partition is measured with respect to a graph ensemble. In this case, we can test the significance of the cut for a fixed partition over the graph ensemble. Note, the graph ensemble studied in this chapter is different and simpler than the ensemble considered in Ref. [63]. If the cut size is significantly lower than we would expect at random given the graph ensemble, then then this is evidence that this is a good partition in the sense that it increases path modularity.

In a different context, this statistic can be used as a test statistic for the existence of a block structure in a simple block model, with equal groups and internal probability of connection equal between the groups, under the null that this is a ER graph. This type of graph has been widely studied as part of work on the detectability threshold [47, 99, 124].

5.4 Analytic results on ER graphs

5.4.1 Analytic results for the case $l = 2$

We first focus on the case of $l = 2$, in which we bound the deviation between the weighted sum of paths of length 1 and 2 and a suitable normal distribution. In the following section, we generalise this to the case $l > 2$, both showing that the order of the approximation is the same as for the case $l = 2$ and deriving a bound on the Wasserstein distance to normal.

This section is organised as follows. We first compute the mean and variance for the weighted sum of the path lengths of length 1 and 2 between two predefined com-

munities. We then use an inequality from Ref. [14] to bound the difference between a standard normal distribution and the suitably standardised weighted sum of paths.

Rather than using the community vector g^* with two even sized communities, it is notationally convenient to replace this with two blocks of equal size as follows. We name the two equally sized blocks of nodes as V_1 and V_2, where $|V| \pmod 2 = 0$ and $|V_1| = |V_2| = \frac{n}{2}$. Assuming that these blocks are given, we can then define the weighted path bisection size ($\sum_{\Upsilon=1}^{l} \sum_{i \in V_1} \sum_{j \in V_2} A_{ij\Upsilon}$) in terms of existence of edges which are modelled as independent identically distributed Bernoulli random variables with probability p i.e. a $G(n,p)$ graph model. We are in the sparse case where $p = \frac{c}{n}$ for a constant c.

5.4.1.1 Computing mean and variance for $l = 2$

We now compute the mean and variance of the weighted sum of the path lengths in the $G(n, \frac{c}{n})$ model, but first we define the following index sets:

$$I_1 = \{(i,j) : i \in V_1 \; j \in V_2\}; \qquad \text{(for paths of length 1)} \quad (5.16)$$

$$I_2 = \{(i,k,j) : i \in V_1; \; j \in V_2; \; k \in V \setminus \{i,j\}\}; \qquad \text{(for paths of length 2)} \quad (5.17)$$

$$I = I_1 \cup I_2. \qquad (5.18)$$

The set I_1 contains a tuple corresponding to each path of length 1 which enumerates the vertices on said path. Similarly, I_2 performs the same role for paths of length 2. These sets allow us to easily sum over the different possible paths.

Note that $|I_1| = |V_1||V_2| = \frac{n^2}{4}$ and $|I_2| = |V_1||V_2|(|V| - 2) = \frac{n^2}{4}(n-2)$. We also define the following function, which maps an element of an index set to the set of all edges that are involved in the corresponding path. For $\alpha = (\alpha_1, ..., \alpha_{|\alpha|})$ a member of

the index set I, where $\alpha_i \in V$ we define:

$$\eta(\alpha) = \{\{\alpha_i, \alpha_{i+1}\} : 0 \leq i \leq |\alpha| - 1\}. \tag{5.19}$$

The following lemma is included as it will be useful to refer to later.

Lemma 3. For simple paths of length up to and including l, η is an injective function from the set of simple paths from V_1 to V_2, to the set of sets of unordered edges, and a bijective function from the set of simple paths from V_1 to V_2, to the set of sets of unordered edges of simple paths of length l or less.

Proof The second statement (bijection) is true if the first statement is true (injective) by a redefinition of the codomain as found in the lemma. Thus we simply need to prove injection.

We shall prove injection by showing that we can uniquely obtain α from $\eta(\alpha)$ for arbitrary α representing a simple path from V_1 to V_2. As the path is simple, each node will appear in at most 2 edges. We reconstruct the sequence of edges by placing edges which share nodes next to each other, which results in a reversible sequence of edges. The orientation of the edges is then fixed by enforcing that the path goes from V_1 to V_2. Therefore for given $\eta(\alpha)$ we can uniquely construct α. QED.

We define the following set of random variables. First, we let $\mathfrak{A}_{ij}$ be an indicator variable for the existence of an edge between node i and j. Note: we use a separate symbol to the standard adjacency matrix to emphasise the fact that this is a random variable. Note as the graph is symmetric $\mathfrak{A}_{ij} = \mathfrak{A}_{ji}$ and therefore we have a set of $\frac{n}{2}(n-1)$ indicator variables.

To express the weighted summation of paths, we define some random variables to sum over to compute the quantities of interest. We define X_α for tuple α as the indicator variable of the existence of the path associated with α multiplied by the

weight attached to this path, as follows:

$$X_\alpha = \begin{cases} \lambda \mathfrak{A}_{\alpha_1\alpha_2}, & \text{for } \alpha = (\alpha_1, \alpha_2) \in I_1 \\ \lambda^2 \mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3}, & \text{for } \alpha = (\alpha_1, \alpha_2, \alpha_3) \in I_2. \end{cases} \tag{5.20}$$

Note that X_α and X_β are independent if they do not share an edge, that is, if $\eta(\alpha) \cap \eta(\beta) = \emptyset$. We can compute the weighted summation of paths by summing over all α, giving the following expression

$$\zeta = \sum_{\alpha \in I} X_\alpha. \tag{5.21}$$

The quantity we are interested in for this work is $Q_l^{Path-Diff}$ from (5.13). From (5.15),

$$Q_l^{Path-Diff} = \left(\sum_{\Upsilon=1}^{l} \lambda_1^\Upsilon \sum_{i \in V_1} \sum_{j \in V_2} A_{ij\Upsilon} \right) - \lambda_2 \left(\sum_{\Upsilon=1}^{l} \lambda_1^\Upsilon \sum_{i \in V_1} \sum_{j \in V_2} \frac{(n-2)!}{(n-1-\Upsilon)!} p^\Upsilon \right). \tag{5.22}$$

Therefore, for $l = 2$ and letting setting $\lambda = \lambda_1$,

$$Q_2^{Path-Diff} = \zeta - \lambda_2 \left(\sum_{\Upsilon=1}^{2} \lambda_1^\Upsilon \sum_{i \in V_1} \sum_{j \in V_2} \frac{(n-2)!}{(n-1-\Upsilon)!} p^\Upsilon \right). \tag{5.23}$$

We will relate the first term ζ to a normal distribution as the second term is entirely specified by the parameters of the model.

For a normal approximation we first derive the mean and variance of ζ.

Lemma 4. If $p = \frac{c}{n}$, then:

$$E[\zeta] = \frac{\lambda c}{4}(n + \lambda c(n-2)). \tag{5.24}$$

Proof The calculation of the mean by linearity of expectation is as follows:

$$E[\zeta] = E[\sum_{\alpha \in I} X_\alpha] = \sum_{\alpha \in I_1} E[X_\alpha] + \sum_{\alpha \in I_2} E[X_\alpha] \tag{5.25}$$

$$= |I_1|\lambda p + |I_2|(\lambda p)^2 = \frac{\lambda p n^2}{4} + (\lambda p)^2 \frac{n^2}{4}(n-2). \tag{5.26}$$

Using $p = \frac{c}{n}$ we obtain the form we require. QED.

To compute the variance we require a few small lemmas which will be useful later. We choose as a dependency set $\mathbb{K}_\alpha$ of $\alpha \in I$ the set

$$\mathbb{K}_\alpha(I) = \{x \in I : |\eta(\alpha) \cap \eta(x)| > 0\}; \tag{5.27}$$

for notational convenience and brevity we drop (I) from the expression. Then X_α is independent of the set $\{X_\beta : \beta \notin \mathbb{K}_\alpha\}$. Further, we define the restricted dependency set as follows:

$$\mathbb{K}_\alpha^{(y)}(I) = \{x \in I \ : \ |\eta(\alpha) \cap \eta(x)| = y\}. \tag{5.28}$$

where again for brevity and notational convenience we drop the (I). Then:

$$\mathbb{K}_\alpha = \bigcup_y \mathbb{K}_\alpha^{(y)} \qquad \text{and} \qquad \mathbb{K}_\alpha^{(x)} \cap \mathbb{K}_\alpha^{(y)} = \emptyset \ \text{ if } x \neq y. \tag{5.29}$$

Lemma 5. For $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in I_2$, it holds that $\mathbb{K}_\alpha^{(2)} \cap I_2 = \{\alpha\}$. Further, if $\alpha_2 \in V_1$, then:

$$\mathbb{K}_\alpha^{(1)} \cap I_2 = \{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \cup \{(\alpha_2, \alpha_1, \theta) : \theta \in V_2\} \tag{R1}$$

$$\cup \{(\alpha_2, \alpha_3, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \cup \{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1, \alpha_2\}\},$$

and if $\alpha_2 \in V_2$, then:

$$\begin{aligned}\mathbb{K}_\alpha^{(1)} \cap I_2 =& \{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_2, \alpha_3\}\} \cup \{(\theta, \alpha_1, \alpha_2) : \theta \in V_2 \setminus \{\alpha_1\}\} \\ & \cup \{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1\}\} \cup \{(\theta, \alpha_3, \alpha_2) : \theta \in V_1\}.\end{aligned} \tag{R2}$$

Proof Let $\alpha \in I_2$. We first show that $\mathbb{K}_\alpha^{(2)} \cap I_2 = \{\alpha\}$. We note that as $|\eta(\alpha)| = 2 \implies \alpha \in \mathbb{K}_\alpha^{(2)} \cap I_2$ for $\alpha \in I_2$. Now let $\beta \in \mathbb{K}_\alpha^{(2)} \cap I_2$. Then $|\eta(\alpha) \cap \eta(\beta)| = 2 = |\eta(\alpha)| = |\eta(\beta)| \implies \eta(\alpha) = \eta(\beta)$. By Lemma 3 η is injective so $\eta(\alpha) = \eta(\beta) \implies \alpha = \beta$. Hence $\mathbb{K}_\alpha^{(2)} \cap I_2 = \alpha$ as required.

We now prove (R1) and (R2). Let us first assume that $\alpha_2 \in V_1$; the argument for $\alpha_2 \in V_2$ will be almost identical. Consider the tuples which overlap on only the first edge $(\alpha_1, \alpha_2) \in V_1 \times V_1$. Our choices are restricted as $I_2 \subset (V_1 \times V_1 \times V_2) \cup (V_1 \times V_2 \times V_2)$, giving a possible set:

$$\{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \cup \{(\alpha_2, \alpha_1, \theta) : \theta \in V_2\} \subseteq \mathbb{K}_\alpha. \tag{5.30}$$

Similarly we construct a set of tuples which overlap on only the second edge, $(\alpha_2, \alpha_3) \in V_1 \times V_2$:

$$\{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1, \alpha_2\}\} \cup \{(\alpha_2, \alpha_3, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \subseteq \mathbb{K}_\alpha, \tag{5.31}$$

where the set difference operations are to preserve the simplicity of the resultant paths. To compute $\mathbb{K}_\alpha^{(1)} \cap I_2$, we simply combine these terms noting that the sets are distinct, which we can prove as follows. Any element which belongs to both sets must contain both edges, but by condition on $\mathbb{K}_\alpha^{(1)}$, the overlap is only one edge, thus the sets are distinct. QED

For much of the remainder of this section, we will be concerned with the size of the intersections of a restricted dependency set of a given $\alpha \in I$, with I_1 and I_2,

i.e. the number of elements of I_x which share a given number of edges with α. The following lemma combines the results.

Lemma 6. For $l = 2$ and $\mathbb{K}_\alpha^{(x)}$ the restricted dependency set for α as defined above with n nodes,

	$\lvert\mathbb{K}_\alpha^{(1)} \cap I_1\rvert$	$\lvert\mathbb{K}_\alpha^{(1)} \cap I_2\rvert$	$\lvert\mathbb{K}_\alpha^{(2)} \cap I_1\rvert$	$\lvert\mathbb{K}_\alpha^{(2)} \cap I_2\rvert$	$\lvert\mathbb{K}_\alpha^{(1)}\rvert$	$\lvert\mathbb{K}_\alpha^{(2)}\rvert$
For $\alpha \in I_1$	1	$n-2$	0	0	$n-1$	0
For $\alpha \in I_2$	1	$2n-4$	0	1	$2n-3$	1

where the table is to be read as follows: the element in the top left states that for $\alpha \in I_1$, $\lvert\mathbb{K}_\alpha^{(1)} \cap I_1\rvert = 1$.

Proof As $l = 2$, we know that $I_1 \cap I_2 = \emptyset$ and $I_1 \cup I_2 = I$. As $\lvert\mathbb{K}_\alpha^{(1)} \cap I_1\rvert + \lvert\mathbb{K}_\alpha^{(1)} \cap I_2\rvert = \lvert\mathbb{K}_\alpha^{(1)}\rvert$, we simply need to demonstrate the results for each of the intersections and we obtain the general results.

We first prove the results for $\alpha \in I_1$, i.e. the first row of the table. The results for $\lvert\mathbb{K}_\alpha^{(2)} \cap I_x\rvert$ for $x = 1, 2$ are proved as follows:

$$\lvert\mathbb{K}_\alpha^{(2)} \cap I_x\rvert \leq \lvert\mathbb{K}_\alpha^{(2)}\rvert = \lvert\{x \in I \ : \ 2 = \lvert\eta(\alpha) \cap \eta(x)\rvert \leq \lvert\eta(\alpha)\rvert = 1\}\rvert = 0. \tag{5.32}$$

For the remaining results we note that for $\alpha \in I_1$, $\lvert\eta(\alpha) \cap \eta(\alpha)\rvert = \lvert\eta(\alpha)\rvert = 1$, therefore $\alpha \in \mathbb{K}_\alpha^{(1)} \cap I_1$. Further, for all $x \in I_1$, $\lvert\eta(x)\rvert = 1$, and if $\lvert\eta(\alpha) \cap \eta(x)\rvert = \lvert\eta(\alpha)\rvert = \lvert\eta(x)\rvert = 1$ then $\eta(x) = \eta(\alpha)$ and hence $\alpha = x$, as by Lemma 3 η is injective. Therefore $\lvert\mathbb{K}_\alpha^{(1)} \cap I_1\rvert = 1$. Next:

$$\lvert\mathbb{K}_\alpha^{(1)} \cap I_2\rvert = \lvert\{(\alpha_1, \alpha_2, \beta_3) : \beta_3 \in V_2 \setminus \{\alpha_2\}\}\rvert + \lvert\{(\beta_1, \alpha_1, \alpha_2) : \beta_1 \in V_1 \setminus \{\alpha_1\}\}\rvert = n - 2, \tag{5.33}$$

because as $\alpha_1 \neq \alpha_2$ the two sets on the right-hand side are disjoint. This finishes the

proof of the first row. For $\alpha \in I_2$, clearly:

$$|\mathbb{K}_\alpha^{(2)} \cap I_1| = |\{x \in I_1 \ : \ 2 = |\eta(\alpha) \cap \eta(x)| \leq |\eta(x)| = 1\}| = 0. \tag{5.34}$$

Further, we demonstrate that $|\mathbb{K}_\alpha^{(1)} \cap I_1| = 1$. In the proof we assume that $\alpha_2 \in V_1$. We omit the proof of the case $\alpha_2 \in V_1$ as it is almost identical to the case shown. We have

$$|\mathbb{K}_\alpha^{(1)} \cap I_1| = \sum_{\beta \in I_1} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| > 0) \tag{5.35}$$

$$= \sum_{\beta \in I_1} \mathbb{1}(|\{\{\alpha_1, \alpha_2\}, \{\alpha_2, \alpha_3\}\} \cap \{\{\beta_1, \beta_2\}\}| > 0) \tag{5.36}$$

$$= \sum_{\beta \in I_1} \mathbb{1}(|\{\{\beta_1, \beta_2\}\} \cap \{\{\alpha_1, \alpha_2\}\}| + |\{\{\beta_1, \beta_2\}\} \cap \{\{\alpha_2, \alpha_3\}\}| > 0). \tag{5.37}$$

Now, $(\alpha_1, \alpha_2) \in V_1 \times V_1$, $(\alpha_2, \alpha_3) \in V_1 \times V_2$ and $(\beta_1, \beta_2) \in V_1 \times V_2$. Hence, $|\{\{\beta_1, \beta_2\}\} \cap \{\{\alpha_1, \alpha_2\}\}| = 0$ and so

$$|\mathbb{K}_\alpha^{(1)} \cap I_1| = \sum_{\beta \in I_1} \mathbb{1}(\{\beta_1, \beta_2\} = \{\alpha_2, \alpha_3\}) = 1. \tag{5.38}$$

The function η is injective (Lemma 3) for $\alpha \in I_2$, therefore:

$$x \in \mathbb{K}_\alpha^{(2)} \cap I_2 \implies |\eta(x) \cap \eta(\alpha)| = |\eta(x)| = |\eta(\alpha)| \implies \eta(\alpha) = \eta(x) \implies \alpha = x \tag{5.39}$$

where the final equality comes from the injection. Therefore, $|\mathbb{K}_\alpha^{(2)} \cap I_2| = |\{\alpha\}| = 1$. Finally, let us now consider, $|\mathbb{K}_\alpha^{(1)} \cap I_2|$, if $\alpha_2 \in V_1$ (the case $\alpha_2 \in V_2$ is almost identical).

From Lemma 5, we know that:

$$\begin{aligned}\mathbb{K}_{\alpha}^{(1)} \cap I_2 =&\{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \cup \{(\alpha_2, \alpha_1, \theta) : \theta \in V_2\} \\ &\cup \{(\alpha_2, \alpha_3, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \cup \{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1, \alpha_2\}\}.\end{aligned} \tag{5.40}$$

We note that as $\alpha_1, \alpha_2, \alpha_3$ are distinct, the sets in the union are pairwise disjoint. Therefore:

$$\begin{aligned}|\mathbb{K}_{\alpha}^{(1)} \cap I_2| =&|\{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\}| + |\{(\alpha_2, \alpha_1, \theta) : \theta \in V_2\}| \\ &+ |\{(\alpha_2, \alpha_3, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\}| + |\{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1, \alpha_2\}\}|\end{aligned} \tag{5.41}$$

$$= \left(\frac{n}{2} - 1\right) + \left(\frac{n}{2}\right) + \left(\frac{n}{2} - 1\right) + \left(\frac{n}{2} - 2\right) = 2n - 4, \tag{5.42}$$

thus completing all of the results, QED.

We can now compute the variance as follows:

Theorem 7. *The variance of ζ from* (5.21) *is given by:*

$$Var(\sum_{\alpha \in I} X_\alpha) = \frac{\lambda^2 c(1-p)}{4}\Big(n + \lambda c(n-2)(2 + \lambda((2n-3)p + 1)\Big). \tag{5.43}$$

Proof We split the variance into three parts as follows:

$$\begin{aligned}\mathrm{Var}(\zeta) = E[(\zeta - E[\zeta])^2] &= E[(\sum_{\alpha \in I}(X_\alpha - E[X_\alpha]))^2] \\ &= \sum_{\alpha \in I_1}\sum_{\beta \in I_1} E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])] \qquad (5.44a) \\ &+2\sum_{\alpha \in I_1}\sum_{\beta \in I_2} E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])] \qquad (5.44b) \\ &+\sum_{\alpha \in I_2}\sum_{\beta \in I_2} E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta]))]. \qquad (5.44c)\end{aligned}$$

Clearly the summand in each of the sums is equal to zero if the pairs of elements from which it is constructed are independent, i.e. $E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])] = E[(X_\alpha - E[X_\alpha])]E[(X_\beta - E[X_\beta])] = 0$. For the sum over $\alpha \in I_1$ and $\beta \in I_1$ the terms are only dependent (i.e. non zero) when $\alpha = \beta$. Using the variance of a binomial random variable term (5.44a) equals

$$\sum_{\alpha \in I_1} \sum_{\beta \in I_1} E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])] = \lambda^2 |I_1| p(1-p) = \lambda^2 \frac{nc}{4}\left(1 - \frac{c}{n}\right). \quad (5.45)$$

For term (5.44b) ($\alpha \in I_1$ and $\beta \in I_2$), we first compute the summand, and then we will multiply by the number of terms in the summation. This is permitted as it will turn out that the summand does not depend on the α and β. First, we let $\alpha = (\alpha_1, \alpha_2)$ and $\beta = (\beta_1, \alpha_1, \alpha_2)$, and calculate the expectation as follows:

$$E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])] = \lambda^3 E[(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\beta_1\alpha_1}\mathfrak{A}_{\alpha_1\alpha_2} - p^2)] \quad (5.46)$$

$$= \lambda^3 (P(\mathfrak{A}_{\alpha_1\alpha_2} = 0)p^3 + P(\mathfrak{A}_{\alpha_1\alpha_2} = 1)(1-p)E[\mathfrak{A}_{\beta_1\alpha_1} - p^2]) \quad (5.47)$$

$$= \lambda^3 ((1-p)p^3 + p(1-p)(p - p^2) \quad (5.48)$$

$$= \lambda^3 p^2 (1-p)(p + (1-p)) = \lambda^3 p^2 (1-p). \quad (5.49)$$

We note that if instead we let $\alpha = (\alpha_1, \alpha_2)$ and $\beta = (\alpha_1, \alpha_2, \beta_3)$ then we would obtain $E[(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\beta_3} - p^2)]$ which is clearly equal to the form in the above expression $E[(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\beta_1\alpha_1}\mathfrak{A}_{\alpha_1\alpha_2} - p^2)]$ as all edges are equally likely. Therefore,

using $|\mathbb{K}_\beta^{(1)} \cap I_1| = 1$ from Lemma 6 we obtain the value of term (5.44b):

$$2\sum_{\alpha\in I_1}\sum_{\beta\in I_2} E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])] = 2\lambda^3 p^2(1-p)\sum_{\beta\in I_2}\sum_{\alpha\in I_1}\mathbb{1}(\eta(\alpha)\cap\eta(\beta)=1) \tag{5.50}$$

$$= 2\lambda^3 p^2(1-p)\sum_{\beta\in I_2}|\mathbb{K}_\beta^{(1)}\cap I_1| = 2\lambda^3 p^2(1-p)|I_2| = \lambda^3\frac{c^2}{2}(n-2)(1-p). \tag{5.51}$$

We now compute term (5.44c). By independence if $|\eta(\alpha)\cap\eta(\beta)| = 0$ then X_α and X_β are independent. Therefore the term (5.44c) is equal to:

$$\sum_{\alpha\in I_2}\sum_{\beta\in I_2}\mathbb{1}(\eta(\alpha)\cap\eta(\beta)\neq\emptyset)E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta]))] \tag{5.52}$$

$$= \sum_{a=1}^{2}\sum_{\alpha\in I_2}\sum_{\beta\in I_2}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = a)E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta]))]. \tag{5.53}$$

Let us now consider the value of $E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta]))]$ with $|\eta(\alpha)\cap\eta(\beta)| \in \{1,2\}$. For $|\eta(\alpha)\cap\eta(\beta)| = 1$ and $\alpha = (\alpha_1,\alpha_2,\alpha_3)$ and $\beta = (\alpha_1,\alpha_2,\beta_3)$

$$E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta]))] = \lambda^4 E[(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\beta_3} - p^2)] \tag{5.54}$$

$$= \lambda^4\Big((1-p)p^4 + pE[\mathfrak{A}_{\alpha_2\alpha_3} - p^2]^2\Big) \tag{5.55}$$

$$= \lambda^4\Big((1-p)p^4 + p(p(1-p^2) - p^2(1-p))^2\Big) \tag{5.56}$$

$$= \lambda^4 p^3(1-p). \tag{5.57}$$

The result for $\beta = (\beta_1,\alpha_2,\alpha_3)$ is identical. For $|\eta(\alpha)\cap\eta(\beta)| = 2$, which implies that $\alpha = \beta = (\alpha_1,\alpha_2,\alpha_3)$:

$$\lambda^4 E[(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)^2] = \lambda^4\Big(p^2(1-p^2)^2 + (1-p^2)p^4\Big) \tag{5.58}$$

$$= \lambda^4 p^2(1-p^2)\Big((1-p^2) + p^2\Big) = \lambda^4 p^2(1-p^2). \tag{5.59}$$

The above expressions simply rely on the fact that only one or two edges overlap, not on which edges overlap. Thus we can write term (5.44c) as:

$$\lambda^4 p^3(1-p)\Big(\sum_{\alpha\in I_2}\sum_{\beta\in I_2}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)|=1)\Big)+\lambda^4p^2(1-p^2) \tag{5.60}$$

$$\times\Big(\sum_{\alpha\in I_2}\sum_{\beta\in I_2}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)|=2)\Big)$$

$$=\lambda^4p^3(1-p)\left(\sum_{\alpha\in I_2}|\mathbb{K}_\alpha^{(1)}\cap I_2|\right)+\lambda^4p^2(1-p^2)\left(\sum_{\alpha\in I_2}|\mathbb{K}_\alpha^{(2)}\cap I_2|\right) \tag{5.61}$$

$$=\lambda^4p^3(1-p)|I_2|(2n-4)+\lambda^4p^2(1-p^2)|I_2| \tag{5.62}$$

$$=\lambda^4(1-p)\frac{c^2}{4}(n-2)(2p(n-2)+(1+p)), \tag{5.63}$$

where the values of $|\mathbb{K}_\alpha^{(1)}\cap I_2|$ and $|\mathbb{K}_\alpha^{(2)}\cap I_2|$, are taken from Lemma 6. We now combine terms (5.44a) (5.44b) and (5.44c) to compute the variance for $\sum_{\alpha\in I}X_\alpha$.

$$\begin{aligned}\mathrm{Var}(\sum_{\alpha\in I}X_\alpha)=&\lambda^2\frac{nc}{4}(1-\frac{c}{n})+\lambda^3\frac{c^2}{2}(n-2)(1-p) &(5.64)\\ &+\lambda^4(1-p)\frac{c^2}{4}(n-2)(2p(n-2)+(1+p))\\ =&\lambda^2\frac{c}{4}(1-\frac{c}{n})\Big(n+2c\lambda(n-2)+\lambda^2c(n-2)(2p(n-2)+(1+p))\Big) &(5.65)\\ =&\frac{\lambda^2c(1-\frac{c}{n})}{4}\Big(n+\lambda c(n-2)(2+\lambda((2n-3)p+1)\Big), &(5.66)\end{aligned}$$

QED.

5.4.1.2 Bounding the deviation from a normal distribution

In order to justify the normal approximation for $\zeta=\sum_{\Upsilon=1}^{2}\lambda_1^{\Upsilon}\sum_{i\in V_1}\sum_{j\in V_2}A_{ij\Upsilon}=\sum_{\alpha\in I}X_\alpha$ we show that the Wasserstein difference between the distribution we observe and the normal distribution goes to zero as n goes to infinity. We use a result from Ref. [14] detailed in Theorem 2 (from page 55). To use this result we require a variable which is a sum of mean zero variables, such that the sum has variance 1.

Finally the sum must obey a dissociated decomposition as in Section 2.7.3. Lemma 4 and Lemma 7 indicate that ζ is the sum of random variables but non-zero mean and not necessarily variance 1.

Thus, to apply Theorem 2 we standardise X_α to have zero mean and finite variance. To do so we let

$$Y_\alpha = \begin{cases} \frac{\lambda}{(\mathrm{Var}(\zeta))^{0.5}}(\mathfrak{A}_{\alpha_1\alpha_2} - p), & \text{for } \alpha = (\alpha_1, \alpha_2) \in I_1 \\ \frac{\lambda^2}{(\mathrm{Var}(\zeta))^{0.5}}(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2), & \text{for } \alpha = (\alpha_1, \alpha_2, \alpha_3) \in I_2 \end{cases} \tag{5.67}$$

We then construct $\tilde{\zeta}$ which sums over Y_α rather than X_α resulting in the following quantity:

$$\tilde{\zeta} = \frac{\zeta - E[\zeta]}{(\mathrm{Var}(\zeta))^{0.5}} = \sum_{\alpha \in I} Y_\alpha. \tag{5.68}$$

Clearly for all $\alpha \in I_1 \cup I_2$, we have $E[Y_\alpha] = E[\tilde{\zeta}] = 0$ and $\mathrm{Var}(\tilde{\zeta}) = 1$. Moreover $\tilde{\zeta}$ has a dissociated decomposition with a dependency neighbourhood given by $\mathbb{K}_\alpha$ given in (5.27) on page 161.

Thus, we can apply Theorem 2 from page 55 to $\tilde{\zeta}$ and obtain that:

$$d_{H_W}(\tilde{\zeta}, Z) \leq 4 \sum_{\alpha \in I} \sum_{\beta,\gamma \in K_x} \left(E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|] E[|Y_\gamma|] \right). \tag{5.69}$$

Applying the less precise bound. In the background section on Stein's method we discussed a less precise bound which similarly bounds the difference between a normal distribution and $\tilde{\zeta}$. If we apply the bound from (2.31) we obtain:

$$d_{H_W}(\tilde{\zeta}, Z) \leq \frac{D_{max}^2}{\sigma^3} \sum_{\alpha \in I} E[|Y_\alpha|^3] + \frac{\sqrt{28} D_{max}^{\frac{3}{2}}}{\sqrt{\pi}\sigma^2} \sqrt{\sum_{\alpha \in I} E[Y_\alpha^4]}, \tag{5.70}$$

where $D_{max} = \max_\alpha |\mathbb{K}_\alpha|$. Let us consider the first term of this bound. From Lemma 6 we know that $D_{max} = 2n - 2$. Further, from Theorem (5.43) we know that $\mathrm{Var}(\zeta)$ is of order n. By definition of $\tilde{\zeta}$ the $\mathrm{Var}(\tilde{\zeta}) = 1$ Therefore $\frac{D_{max}^2}{\sigma^3}$ is of order n^2.

Consider:

$$\sum_{\alpha\in I} E[|Y_\alpha|^3] = \sum_{\alpha\in I_1} E[|Y_\alpha|^3] + \sum_{\alpha\in I_2} E[|Y_\alpha|^3] \tag{5.71}$$

Clearly both terms in this expression are non-negative, we will therefore focus on the first term:

$$\sum_{\alpha\in I_1} E[|Y_\alpha|^3] = \sum_{\alpha\in I_1} \left(p\frac{\lambda^3(1-p)^3}{(\mathrm{Var}(\zeta))^{1.5}} + (1-p)\frac{\lambda^3(p)^3}{(\mathrm{Var}(\zeta))^{1.5}} \right) \tag{5.72}$$

$$= |I_1|\frac{\lambda^3 p(1-p)}{(\mathrm{Var}(\zeta))^{1.5}}(p^2 + (1-p)^2) \tag{5.73}$$

$$= \frac{\lambda^3 n^2}{4(\mathrm{Var}(\zeta))^{1.5}}(p - 3p^2 + 4p^3 - 2p^4) \tag{5.74}$$

$$= \frac{\lambda^3}{4(\mathrm{Var}(\zeta))^{1.5}}(cn - 3c^2 + 4\frac{c^3}{n} - 2\frac{c^4}{n^2}) \tag{5.75}$$

We know that $\mathrm{Var}(\zeta)$ is of order n and therefore, $\sum_{\alpha\in I_1} E[|Y_\alpha|^3]$ is of order $n^{-0.5}$.

As both terms in (5.71) are non negative this implies that $\sum_{\alpha\in I} E[|Y_\alpha|^3]$ is at least $n^{-0.5}$. We know that $\frac{D_{max}}{\sigma^3}$ is of order n^2. Thus the first term in the expression (5.70) is at least order $n^{1.5}$. The second term is positive and thus the expression is at least $n^{1.5}$ to leading order. Thus the bound (5.70) is not appropriate for this problem.

Deriving Basic Inequalities Note that we will not be attempting to find the best possible bounds in this section. We are simply looking for a simple way to achieve reasonable bounds.

In order to bound the RHS of (5.69) first we compute some bounds on some common expressions and demonstrate a simple approach which will make the computation simpler and less repetitive. This approach uses the following three observations.

1. Let $f_1(p) = 1 + \rho_1 p - \rho_1 p^2$ with $\rho_1 > 0$ and $0 \le p \le 1$. Then taking the derivative

   $\frac{d}{dp} f_1 = \rho_1 - 2\rho_1 p$ shows that f_1 is increasing for $p < \frac{1}{2}$ and decreasing for $p > \frac{1}{2}$.

Hence,

$$\max_{0\leq p\leq 1}(f_1(p)) = f_1\left(\frac{1}{2}\right) = 1 + \frac{\rho_1}{2} - \frac{\rho_1}{4} = 1 + \frac{\rho_1}{4}. \tag{5.76}$$

2. Let $f_2(p) = -\rho_1 p^{\kappa_1} + \rho_1 p^{\kappa_1+\kappa_2}$, with $\rho_1 > 0$ and $0 \leq p \leq 1$ then

$$f_2(p) = \rho_1 p^{\kappa_1}(p^{\kappa_2} - 1) \leq 0, \tag{5.77}$$

for $\kappa_1, \kappa_2 \in \mathbb{R}$ and $\kappa_1, \kappa_2 \geq 0$.

3. We use a simple differentiation to bound the function f under the condition that f is differentiable on the interval $(0, 1)$ and continuous on $[0, 1]$. We further assume that we can solve the equation $f'(p) = 0$ where f' is the derivative of f. The procedure for function $f(p)$ is as follows:

Let $p_1, \ldots p_q$ be all solutions to $f'(p) = 0$.

Then: $\text{Max}_{a\leq p\leq b}(f(p)) = \text{Max}(f(a), f(b), f(p_1), \ldots, f(p_q))$. (5.78)

Computing the Bound We will now apply the bound (5.69) to our problem. To do this we will split the sum into a sum over I_1 and a sum over I_2, i.e.

$$d_{H_W}(\tilde{\zeta}, Z) \leq 4(\Phi_1 + \Phi_2), \quad \text{where} \tag{5.79}$$

$$\Phi_1 = \sum_{\alpha\in I_1} \sum_{\beta.\gamma\in K_\alpha} \Big(E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]\Big) \tag{5.80}$$

$$\text{and } \Phi_2 = \sum_{\alpha\in I_2} \sum_{\beta,\gamma\in K_\alpha} \Big(E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]\Big). \tag{5.81}$$

We will then bound Φ_1 and Φ_2 separately.

We first focus on Φ_1, which is a sum over I_1. We compute the contribution of one term to the sum, and then we multiply over the number of terms in the sum to get the total contribution. Note this is possible as all elements of I_1 are interchangeable,

because they all consist of one element from V_1 and one from V_2. Recall that $|I_1| = \frac{n^2}{4}$. Let $\alpha \in I_1$, and note that $\mathbb{K}_\alpha = \mathbb{K}_\alpha^{(1)}$ because $|\eta(\alpha)| = 1$.

We split Φ_1 into different parts as follows:

$$\Phi_1^{(i,j)}(\alpha) = \sum_{\beta \in \mathbb{K}_\alpha \cap I_i} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_j} (E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) \quad \text{for } i,j \in \{1,2\}, \tag{5.82}$$

$$\text{so that: } \Phi_1 = |I_1| \sum_{i=1}^{2} \sum_{j=1}^{2} \Phi_1^{(i,j)}. \tag{5.83}$$

In order to make this calculation easy to read, we bound each of the $\Phi_1^{(i,j)}$ separately in the following set of lemmas, which we will then combine.

Lemma 8. We have

$$\Phi_1^{(1,1)} = \frac{\lambda^3}{\mathrm{Var}(\zeta)^{1.5}} p(1-p). \tag{5.84}$$

Proof By Lemma 6 the number of terms in this sum is $|\mathbb{K}_\alpha \cap I_1|^2 = 1^2 = 1$ i.e. $\alpha = \beta = \gamma$. This results in the following expression for $\Phi_1^{(1,1)}$:

$$\begin{aligned}
\Phi_1^{(1,1)} =& E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|] && (5.85)\\
=& \frac{\lambda^3}{\mathrm{Var}(\zeta)^{1.5}} (E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)^3|] + E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)^2|]E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)|]) && (5.86)\\
=& \frac{\lambda^3}{\mathrm{Var}(\zeta)^{1.5}} (p(1-p)(1-2p+2p^2) + 2(1-p)^2p^2) && (5.87)\\
=& \frac{\lambda^3}{\mathrm{Var}(\zeta)^{1.5}} p(1-p), && (5.88)
\end{aligned}$$

which is as stated in the Lemma. QED.

Second we consider $\Phi_1^{(1,2)}$. (Due to the lack of symmetry in the summand of (2.48), $\Phi_1^{(1,2)} \neq \Phi_1^{(2,1)}$.)

Lemma 9. We have

$$\Phi_1^{(1,2)} \leq \frac{3}{2} \frac{\lambda^4 (n-2)}{\mathrm{Var}(\zeta)^{1.5}} p^2(1-p). \tag{5.89}$$

Proof As $\mathbb{K}_\alpha \cap I_1 = \{\alpha\}$, we obtain:

$$\Phi_1^{(1,2)} = \sum_{\beta \in \mathbb{K}_\alpha \cap I_1} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} (E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) \tag{5.90}$$

$$= \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \Big(E[|Y_\alpha^2 Y_\gamma|] + E[|Y_\alpha^2|]E[|Y_\gamma|] \Big) \tag{5.91}$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \Big(E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)^2 (\mathfrak{A}_{\gamma_1\gamma_2}\mathfrak{A}_{\gamma_2\gamma_3} - p^2)|] \tag{5.92}$$

$$+ E[(\mathfrak{A}_{\alpha_1\alpha_2} - p)^2]E[|(\mathfrak{A}_{\gamma_1\gamma_2}\mathfrak{A}_{\gamma_2\gamma_3} - p^2)|] \Big)$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \Big(E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)^2 (\mathfrak{A}_{\gamma_1\gamma_2}\mathfrak{A}_{\gamma_2\gamma_3} - p^2)|] + 2p^3(1-p)^2(1+p) \Big). \tag{5.93}$$

As $\gamma \in K_\alpha$, either $(\gamma_1, \gamma_2) = (\alpha_1, \alpha_2)$ or $(\gamma_2, \gamma_3) = (\alpha_1, \alpha_2)$. Let us assume that $(\gamma_1, \gamma_2) = (\alpha_1, \alpha_2)$, and observe that the choice will not change the value of the summand.

$$\Phi_1^{(1,2)} = \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \Big(E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)^2 (\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|] \tag{5.94}$$

$$+ 2p^3(1-p)^2(1+p) \Big)$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \sum_{x=0}^{1} \Big(\Big(P(\mathfrak{A}_{\alpha_1\alpha_2} = x) E[|(x-p)^2 (x\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|] \Big) \tag{5.95}$$

$$+ 2p^3(1-p)^2(1+p) \Big).$$

By summing over x we obtain:

$$\Phi_1^{(1,2)} = \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \Big((1-p)p^4 + p^2(1-p)^3(2p+1) + 2p^3(1-p)^2(1+p) \Big) \quad (5.96)$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} (1-p)p^2(1+2p-2p^2) \quad (5.97)$$

$$= \frac{\lambda^4 |\mathbb{K}_\alpha \cap I_2|}{\mathrm{Var}(\zeta)^{1.5}} (1-p)p^2(1+2p-2p^2) \quad (5.98)$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} (n-2)(1-p)p^2(1+2p-2p^2) \quad (5.99)$$

$$\leq \frac{3}{2} \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} (n-2)p^2(1-p), \quad (5.100)$$

where we have used the fact that $|\mathbb{K}_\alpha \cap I_2| = n-2$ from Lemma 6, and the final inequality comes from the approach described in (5.76). The final form is as stated in the Lemma. QED.

We will now consider $\Phi_1^{(2,1)}$.

Lemma 10. We have

$$\Phi_1^{(2,1)} \leq \frac{8}{5} \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} (n-2)p^2(1-p). \quad (5.101)$$

Proof Note that $E[|Y_\alpha Y_\beta Y_\gamma|]$ does not change between $\Phi_1^{(1,2)}$ and $\Phi_1^{(2,1)}$, therefore we use $E[|Y_\alpha Y_\beta Y_\gamma|]$ from $\Phi_1^{(1,2)}$ corresponding to the first two terms from (5.96). We

then replace the second term with the appropriate term to obtain:

$$\Phi_1^{(2,1)} = \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \Big(p^2(1-p)(1-2p^2+2p^3)) \tag{5.102}$$

$$+ E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|]E[|\mathfrak{A}_{\alpha_1\alpha_2} - p|]\Big)$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} \Big((1-p)p^4 + p^2(1-p)^2(1+p(1-2p)) + 2p^3(1-p)^2(1+2p(1-p))) \Big) \tag{5.103}$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} |\mathbb{K}_\alpha \cap I_2| p^2(1-p)(1+2p-6p^3+4p^4) \tag{5.104}$$

$$= \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} (n-2) p^2(1-p)(1+2p-6p^3+4p^4) \tag{5.105}$$

where we have again used $|\mathbb{K}_\alpha \cap I_2| = n-2$ from Lemma 6. We bound as follows:

$$\text{Let } f(p) = 1 + 2p - 6p^3 + 4p^4, \quad f(0) = 1, \quad f(1) = 1. \tag{5.106}$$

To find the maxima, we compute

$$\frac{d}{dp} f(p) = f'(p) = 2(1 - 9p^2 + 8p^3) = 2(1-p)(1+p-8p^2).$$

Now $f'(p) = 0$ when $p = 1, \frac{1}{16} \pm \frac{\sqrt{33}}{16}$. As $\frac{1}{16} - \frac{\sqrt{33}}{16} < 0$ we ignore this solution.

Hence by (5.78):

$$f(p) \leq \mathrm{Max}\left(f(0), f(1), f\left(\frac{1}{16} + \frac{\sqrt{33}}{16} \right) \right) = f\left(\frac{1}{16} + \frac{\sqrt{33}}{16} \right) = \frac{5}{2048}(33\sqrt{33} + 433) < \frac{8}{5}$$

and so $\Phi_1^{(2,1)} < \frac{8}{5} \frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}} (n-2) p^2(1-p)$,

where the final inequality is the same as the Lemma. QED.

Finally for Φ_1 we have to compute $\Phi_1^{(2,2)}$.

Lemma 11. We have

$$\Phi_1^{(2,2)} \leq \frac{\lambda^5 (n-2)}{\mathrm{Var}(\zeta)^{1.5}} p^2(1-p) \left(1 + \frac{3}{2}(n-3)p + \frac{8}{3}(n-2)p^2 \right). \tag{5.107}$$

Proof We can express $\Phi_1^{(2,2)}$ as a sum as follows:

$$\Phi_1^{(2,2)} = \sum_{\beta \in \mathbb{K}_\alpha \cap I_2} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} (E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) \tag{5.108}$$

$$= \sum_{\beta \in \mathbb{K}_\alpha \cap I_2} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2 \setminus \{\beta\}} E[|Y_\alpha Y_\beta Y_\gamma|] + \sum_{\beta \in \mathbb{K}_\alpha \cap I_2} E[|Y_\alpha Y_\beta^2|] \tag{5.109}$$

$$+ \sum_{\beta \in \mathbb{K}_\alpha \cap I_2} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} E[|Y_\alpha Y_\beta|]E[|Y_\gamma|].$$

Let us start with the term $\sum_{\beta \in \mathbb{K}_\alpha \cap I_2} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]$, where we again use the fact that the choice of edge shared between α and β does not affect the outcome.

Let $\alpha = (\alpha_1, \alpha_2)$, $\beta = (\alpha_1, \alpha_2, \beta_3)$, and $\gamma = (\alpha_1, \alpha_2, \gamma_3)$, then

$$E[|Y_\alpha Y_\beta|]E[|Y_\gamma|] = E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\beta_3} - p^2)|]E[|(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|] \tag{5.110}$$

$$= \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}}(p^2(1-p))(1+2p-2p^2)(2p^2(1-p^2)) \tag{5.111}$$

$$= \frac{2\lambda^5(n-2)^2}{\mathrm{Var}(\zeta)^{1.5}}p^4(1-p)(1-p^2)(1+2p-2p^2)). \tag{5.112}$$

Now summing over all $\beta \in \mathbb{K}_\alpha \cap I_2$ and $\gamma \in \mathbb{K}_\alpha \cap I_2$ and exploiting the interchangability of β and γ in the previous calculation and the fact the $|\mathbb{K}_\alpha \cap I_2| = n-2$ from Lemma 6, we obtain:

$$\sum_{\beta \in \mathbb{K}_\alpha \cap I_2} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2} E[|Y_\alpha Y_\beta|]E[|Y_\gamma|] = \frac{2\lambda^5(n-2)^2}{\mathrm{Var}(\zeta)^{1.5}}p^4(1-p)(1+2p-3p^2-2p^3+2p^4) \tag{5.113}$$

$$< \frac{2\lambda^5(n-2)^2}{\mathrm{Var}(\zeta)^{1.5}}p^4(1-p)(1+2p-3p^2) \tag{5.114}$$

$$= \frac{2\lambda^5(n-2)^2}{\mathrm{Var}(\zeta)^{1.5}}p^4(1-p)\left(\frac{4}{3} - 3(p-\frac{1}{3})^2\right) \tag{5.115}$$

$$\leq \frac{8}{3}\frac{\lambda^5(n-2)^2}{\mathrm{Var}(\zeta)^{1.5}}p^4(1-p). \tag{5.116}$$

Next, we compute the second term from (5.109). Let $\alpha = (\alpha_1, \alpha_2)$ and $\beta = (\alpha_1, \alpha_2, \beta_3)$, then:

$$E[|Y_\alpha Y_\beta^2|] = E[|(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\beta_3} - p^2)^2|] \tag{5.117}$$

$$= \sum_{x=0}^{1} \frac{\lambda^5 P(\mathfrak{A}_{\alpha_1\alpha_2} = x)}{\mathrm{Var}(\zeta)^{1.5}} E[|(x-p)(x\mathfrak{A}_{\alpha_2\beta_3} - p^2)^2|] \tag{5.118}$$

$$= \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} \left((1-p)E[|(0-p)(0-p^2)^2|] + pE[|(1-p)(\mathfrak{A}_{\alpha_2\beta_3} - p^2)^2|]\right) \tag{5.119}$$

$$= \frac{\lambda^5 p^2(1-p)}{\mathrm{Var}(\zeta)^{1.5}} \left(1 - 2p^2 + 2p^3\right). \tag{5.120}$$

Now summing over all $\beta \in \mathbb{K}_\alpha \cap I_2$ (using Lemma 6) and using the fact that $1 - 2p^2 + 2p^3 \leq 1$,

$$\sum_{\beta \in \mathbb{K}_\alpha \cap I_2} E[|Y_\alpha Y_\beta^2|] \leq \frac{\lambda^5 p^2(1-p)}{\mathrm{Var}(\zeta)^{1.5}}(n-2), \tag{5.121}$$

where the final inequality comes from (5.77). For the final term of (5.109) let $\alpha = (\alpha_1, \alpha_2)$, $\beta = (\alpha_1, \alpha_2, \beta_3)$, $\gamma = (\alpha_1, \alpha_2, \gamma_3)$:

$$E[|Y_\alpha Y_\beta Y_\gamma|] = \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} \sum_{x=0}^{1} P(\mathfrak{A}_{\alpha_1\alpha_2} = x) E[|(x-p)(x\mathfrak{A}_{\alpha_2\beta_3})(x\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|] \tag{5.122}$$

$$= \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} \left((1-p)p^5 + p(1-p)E[|(\mathfrak{A}_{\alpha_2\beta_3} - p^2)(\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|]\right) \tag{5.123}$$

$$= \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} \left((1-p)p^5 + p(1-p)E[|(\mathfrak{A}_{\alpha_2\beta_3} - p^2)|]^2\right) \tag{5.124}$$

$$= \frac{\lambda^5 p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}} \left(1 + 2p - 2p^2 - 4p^3 + 4p^4\right) \tag{5.125}$$

Now summing over all $\beta \in \mathbb{K}_\alpha \cap I_2$ and $\gamma \in \mathbb{K}_\alpha \cap I_2 \setminus \{\beta\}$ and using the fact the $|\mathbb{K}_\alpha \cap I_2| = n-2$ from Lemma 6, we obtain:

$$\sum_{\beta \in \mathbb{K}_\alpha \cap I_2} \sum_{\gamma \in \mathbb{K}_\alpha \cap I_2 \setminus \{\beta\}} E[|Y_\alpha Y_\beta Y_\gamma|] = \frac{\lambda^5 (n-2)(n-3)p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}} \left(1 + 2p - 2p^2 - 4p^3 + 4p^4\right) \tag{5.126}$$

$$\leq \frac{\lambda^5 (n-2)(n-3)p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}} \left(1 + 2p - 2p^2\right) \tag{5.127}$$

$$\leq \frac{3}{2} \frac{\lambda^5 (n-2)(n-3)p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}}, \tag{5.128}$$

where the penultimate inequality comes from (5.77), and the final inequality comes from (5.76). Putting these terms together we obtain:

$$\Phi_1^{(2,2)} \leq \frac{8}{3} \frac{\lambda^5 (n-2)^2}{\mathrm{Var}(\zeta)^{1.5}} p^4(1-p) + \frac{\lambda^5 p^2 (1-p)}{\mathrm{Var}(\zeta)^{1.5}} (n-2) + \frac{3}{2} \frac{\lambda^5 (n-2)(n-3)p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}} \tag{5.129}$$

$$= \frac{\lambda^5 (n-2)}{\mathrm{Var}(\zeta)^{1.5}} p^2(1-p) \left(1 + \frac{3}{2}(n-3)p + \frac{8}{3}(n-2)p^2\right) \tag{5.130}$$

which is as stated in the Lemma, QED.

We can now bound Φ_1 from above.

Lemma 12. For Φ_1, n, λ and p as previously defined:

$$\Phi_1 \leq \frac{1}{4} \frac{\lambda^3 c(n-c)}{\mathrm{Var}(\zeta)^{1.5}} \left(1 + \frac{16}{5}\lambda c + \frac{8}{3}\lambda^2 \frac{c^3}{n} + \lambda^2 c + \frac{3}{2}\lambda^2 c^2\right). \tag{5.131}$$

Proof Combining the results from Lemma 8, Lemma 9, Lemma 10 and Lemma 11 we obtain:

$$\frac{\Phi_1}{|I_1|} = \sum_{i=1}^{2}\sum_{j=1}^{2}\Phi_1^{(i,j)} \tag{5.132}$$

$$= \frac{\lambda^3}{\mathrm{Var}(\zeta)^{1.5}}p(1-p) + \frac{3}{2}\frac{\lambda^4(n-2)}{\mathrm{Var}(\zeta)^{1.5}}p^2(1-p) + \frac{8}{5}\frac{\lambda^4}{\mathrm{Var}(\zeta)^{1.5}}(n-2)p^2(1-p) \tag{5.133}$$

$$+ \frac{\lambda^5(n-2)}{\mathrm{Var}(\zeta)^{1.5}}p^2(1-p)\Big(1 + \frac{3}{2}(n-3)p + \frac{8}{3}(n-2)p^2\Big)$$

$$= \frac{\lambda^3}{\mathrm{Var}(\zeta)^{1.5}}p(1-p)\Bigg(1 + \frac{31}{10}(n-2)\lambda p + \lambda^2(n-2)p\Big(1 + \frac{3}{2}(n-3)p + \frac{8}{3}(n-2)p^2\Big)\Bigg). \tag{5.134}$$

Clearly $(n-2) < n$ and $(n-3) < n$, thus:

$$\frac{\Phi_1}{|I_1|} = \frac{\lambda^3 p(1-p)}{\mathrm{Var}(\zeta)^{1.5}}\Bigg(1 + \frac{31}{10}n\lambda p + \lambda^2 np\Big(1 + \frac{3}{2}np + \frac{8}{3}np^2\Big)\Bigg). \tag{5.135}$$

Rearranging and using the fact that $p = \frac{c}{n}$ and $|I_1| = \frac{n^2}{4}$, we obtain:

$$\Phi_1 \leq \frac{1}{4}\frac{\lambda^3 c(n-c)}{\mathrm{Var}(\zeta)^{1.5}}\Bigg(1 + \frac{16}{5}\lambda c + \frac{8}{3}\lambda^2\frac{c^3}{n} + \lambda^2 c + \frac{3}{2}\lambda^2 c^2\Bigg), \tag{5.136}$$

which is as stated in the Lemma. QED.

We now consider Φ_2, where $\alpha \in I_2$. For notational convenience, $\mathfrak{K}_\alpha^{(i)}$ for $i = 1, 2$ is defined as follows:

$$\mathfrak{K}_\alpha^{(1)} = \mathbb{K}_\alpha^{(2)} \cup (\mathbb{K}_\alpha^{(1)} \cap I_1) = \{\alpha\} \cup \{\beta \in I_1 : |\eta(\alpha) \cap \eta(\beta)| = 1\}; \tag{5.137}$$

$$\mathfrak{K}_\alpha^{(2)} = \mathbb{K}_\alpha^{(1)} \cap I_2 = \{\beta \in I_2 : |\eta(\alpha) \cap \eta(\beta)| = 1\}. \tag{5.138}$$

This decomposition of $\mathbb{K}_\alpha$ is exhaustive as $\Big(\mathbb{K}_\alpha^{(2)} \cup (\mathbb{K}_\alpha^{(1)} \cap I_1)\Big) \cup \Big(\mathbb{K}_\alpha^{(1)} \cap I_2\Big) = \mathbb{K}_\alpha^{(2)} \cup (\mathbb{K}_\alpha^{(1)} \cap (I_1 \cup I_2)) = \mathbb{K}_\alpha$. Further, using Lemma 6, we can compute $|\mathfrak{K}_\alpha^{(1)}| = |\mathbb{K}_\alpha^{(2)}| +$

$|(\mathbb{K}_\alpha^{(1)} \cap I_1)| = 2$, and $|\mathfrak{K}_\alpha^{(2)}| = |\mathbb{K}_\alpha^{(1)} \cap I_2| = 2n - 4$. We can now split Φ_2 as we split Φ_1. We denote it $\Phi_2^{(i,j)}$, for $1 \leq i, j \leq 2$, where:

$$\Phi_2^{(i,j)}(\alpha) = \sum_{y \in \mathfrak{K}_\alpha^{(i)}} \sum_{z \in \mathfrak{K}_\alpha^{(j)}} (E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]), \tag{5.139}$$

$$\Phi_2 = \sum_{\alpha \in I_2} \sum_{i=1}^{2} \sum_{j=1}^{2} \Phi_1^{(i,j)}(\alpha) = |I_2| \sum_{i=1}^{2} \sum_{j=1}^{2} \Phi_1^{(i,j)}(\alpha), \tag{5.140}$$

in which there is the implicit underlying assumption that the value of $\Phi_2^{(i,j)}$ does not depend on the choice of α, which we will see in the derivations below.

We will again split the calculation into several lemmas. We first consider $\Phi_2^{(1,1)}$.

Lemma 13. For $\Phi_2^{(1,1)}$, n and c as previously defined:

$$\Phi_2^{(1,1)} \leq 2\frac{\lambda^4 p^2 (1-p)}{\mathrm{Var}(\zeta)^{1.5}} \left(\lambda^2 + 2\lambda + \frac{4}{5} \right). \tag{5.141}$$

Proof From Lemma 6, we know that $|I_1 \cap \mathbb{K}_\alpha| = 1$. Let $\{\kappa\} = I_1 \cap \mathbb{K}_\alpha$. Therefore, either $\kappa = (\alpha_1, \alpha_2)$ or $\kappa = (\alpha_2, \alpha_3)$ but crucially not both. The choice again does not affect the outcome, thus we fix $\kappa = (\alpha_1, \alpha_2)$. Again it can be seen that there is no difference by simply repeating the following arguments with the alternative choice. Therefore as $\mathfrak{K}_\alpha^{(1)} = \{\kappa, \alpha\}$ thus:

$$\Phi_2^{(1,1)} = \sum_{\beta \in \mathfrak{K}_\alpha^{(1)}} \sum_{\gamma \in \mathfrak{K}_\alpha^{(1)}} (E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) \tag{5.142}$$

$$= E[|Y_\alpha^3|] + 2E[|Y_\alpha^2 Y_\kappa|] + E[|Y_\alpha Y_\kappa^2|] + (E[|Y_\alpha^2|] + E[|Y_\alpha Y_\kappa|])(E[|Y_\alpha|] + E[|Y_\kappa|]) \tag{5.143}$$

$$= \frac{\lambda^4 p^2 (1-p)}{\mathrm{Var}(\zeta)^{1.5}} \Big(\lambda^2 (p+1) + 2\lambda (p(1-p)(1+3p^2-2p^4)+1) + 2p(1-p)^2(2p+1) + 1 \Big) \tag{5.144}$$

To bound (5.144) we consider each power of λ separately. We bound $\lambda^2(1+p) \leq 2\lambda^2$. The bound for the coefficient of λ is as follows:

$$2(p(1-p)(1+3p^2-2p^4)+1) = 2(1+p-p^2+3p^3-3p^4-2p^5+2p^6) \tag{5.145}$$

$$< 2(1+p-p^2+3p^3-3p^4) = 2(1+p(1-p+3p^2-3p^3) < 2(2-p+3p^2-3p^3) \tag{5.146}$$

Let $f(p) = -p+3p^2-3p^3$. By (5.78), $\text{Max}_{0\leq p\leq 1} f(p) \in \{f(0), f(1), f(x) : f'(x) = 0\}$ To find the maxima, note that $f(0) = 1,\ f(1) = 0,$ and $f'(p) = -9\left(p-\frac{1}{3}\right)^2$. So $f'(p) = 0$ if $p = \frac{1}{3}$. As $f\left(\frac{1}{3}\right) = -\frac{1}{9},$ it follows that $\text{Max}_{0\leq p\leq 1} f(p) = \text{Max}\left(0, -1, -\frac{1}{9}\right) = 0$ and hence $2(p(1-p)(1+3p^2-2p^4)+1) \leq 4$.

To bound the coefficient of λ^0 we can use the bound we derived in (5.106), using which we obtain:

$$2p(1-p)^2(2p+1)+1 = 4p^4-6p^3+2p+1 < \frac{8}{5}. \tag{5.147}$$

Substituting these bounds into the equation we derived for $\Phi_\alpha^{(1,1)}$ gives the bound in the Lemma. QED.

To compute $\Phi_2^{(1,2)}$ and $\Phi_2^{(2,1)}$, we require a more involved calculation of a pair of summations. Due to the length of the calculation, we placed it in the following Lemma:

Lemma 14. For $\alpha \in I_2$ and $\kappa \in I_1$ such that $|\eta(\alpha) \cap \eta(\kappa)| = 1$ then

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|] \leq \frac{8}{5}\frac{\lambda^5(n-3)}{\text{Var}(\zeta)^{1.5}}p^3(1-p), \tag{5.148}$$

$$\text{and} \sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|] \leq \frac{7}{5}\frac{\lambda^5(n-1)p^3(1-p)}{\text{Var}(\zeta)^{1.5}}. \tag{5.149}$$

Proof In previous proofs we have simply stated that the arbitrary choice that $\alpha_2 \in V_2$ did not affect the outcome. In this case, while true it is less obvious why this is the case. Thus, we will work through the choices and demonstrate that they do not have an effect on the result. Let $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in I_2$. We know by Lemma 5 that $\mathfrak{K}_\alpha^{(2)} = \mathbb{K}_\alpha^{(1)} \cap I_2$ is equal to:

$$\mathfrak{K}_\alpha^{(2)} = \mathbb{K}_\alpha^{(1)} \cap I_2 = \begin{cases} \{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \cup \{(\alpha_2, \alpha_1, \theta) : \theta \in V_2\} \\ \cup \{(\alpha_2, \alpha_3, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\} \\ \cup \{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1, \alpha_2\}\} & \text{if } \alpha_2 \in V_1, \\ \\ \{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_2, \alpha_3\}\} \\ \cup \{(\theta, \alpha_1, \alpha_2) : \theta \in V_1 \setminus \{\alpha_1\}\} \\ \cup \{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1\}\} \cup \{(\theta, \alpha_3, \alpha_2) : \theta \in V_1\} & \text{if } \alpha_2 \in V_2. \end{cases} \tag{5.150}$$

If $\kappa \in I_1$ such that $|\eta(\alpha) \cap \eta(\kappa)| = 1$, then $\kappa = (\alpha_1, \alpha_2)$ if $\alpha_2 \in V_2$ and $\kappa = (\alpha_2, \alpha_3)$ otherwise. Therefore:

$$\mathbb{K}_\kappa = \begin{cases} \{(\alpha_2, \alpha_3), \} \cup \{(\alpha_2, \alpha_3, \theta)\ \theta \in V_2 \setminus \{\alpha_3\}\} \\ \cup \{(\theta, \alpha_2, \alpha_3)\ \theta \in V_1 \setminus \{\alpha_2\}\} & \text{if } \alpha_2 \in V_1, \\ \{(\alpha_1, \alpha_2), \} \cup \{(\alpha_1, \alpha_2, \theta)\ \theta \in V_2 \setminus \{\alpha_2\}\} \\ \cup \{(\theta, \alpha_1, \alpha_2)\ \theta \in V_1 \setminus \{\alpha_1\}\} & \text{if } \alpha_2 \in V_2. \end{cases} \tag{5.151}$$

Combining these equations we obtain:

$$|\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa| = |\mathbb{K}_\alpha^{(1)} \cap I_2 \cap \mathbb{K}_\kappa| = \begin{cases} |\{(\alpha_2, \alpha_3, \theta) : \theta \in V_2 \setminus \{\alpha_3\}\}| \\ \quad +|\{(\theta, \alpha_2, \alpha_3) : \theta \in V_1 \setminus \{\alpha_1, \alpha_2\}\}| \\ = n-3 & \text{if } \alpha_2 \in V_1, \\ |\{(\alpha_1, \alpha_2, \theta) : \theta \in V_2 \setminus \{\alpha_2, \alpha_3\}\}| \\ \quad +|\{(\theta, \alpha_1, \alpha_2) : \theta \in V_1 \setminus \{\alpha_1\}\}| \\ = n-3 & \text{if } \alpha_2 \in V_2. \end{cases} \tag{5.152}$$

Thus, $|\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa| = n-3$, which implies that $|\mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa| = |\mathfrak{K}_\alpha^{(2)}| - |\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa| = (2n-4) - (n-3) = n-1$, where we have used $|\mathbb{K}_\alpha^{(1)} \cap I_2| = 2n-4$ from Lemma 6.

We can now bound the summand of the first of the sum of the Lemma.

$$E[|Y_\alpha Y_\kappa Y_\beta|] = \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} E[|(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\beta_3} - p^2)|] \tag{5.153}$$

$$= \frac{\lambda^5 p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}}(1 + 2(1-p)(p-2p^3)). \tag{5.154}$$

To bound this let $f(p) = p - 2p^3$ then using the standard approach, f takes its maximum at $p = \frac{\sqrt{6}}{6}$. Therefore, $(1 + 2(1-p)(p-2p^3)) \leq (1 + 2f(p)) \leq \frac{8}{5}$ and hence $E[|Y_\alpha Y_\kappa Y_\beta|] \leq \frac{8}{5}\frac{\lambda^5 p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}}$. (5.155)

Now summing over all $\beta \in \mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa$ we obtain:

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|] \leq \frac{8}{5}\frac{\lambda^5 (n-3) p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}}, \tag{5.156}$$

which is as stated in the Lemma. For the second remaining sum in the Lemma, let

$\beta = (\beta_1, \alpha_2, \alpha_3)$:

$$E[|Y_\alpha Y_\kappa Y_\beta|] = \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} E[|(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)(\mathfrak{A}_{\alpha_1\alpha_2} - p)(\mathfrak{A}_{\beta_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)|]$$
$$= \frac{\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} p^3(1-p)^2(1+2p(1+p)) = \frac{\lambda^5 p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}}(1+p-2p^3).$$

To bound this expression let: $f(p) = (1+p-2p^3)$ using the standard approach, f takes its maximum at $\frac{\sqrt{6}}{6}$ and $f(\frac{\sqrt{6}}{6}) < \frac{7}{5}$. Hence, $E[|Y_\alpha Y_\kappa Y_\beta|] \leq \frac{7}{5}\frac{\lambda^5 p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}}$.

Summing over $\beta \in \mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa$ we obtain:

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|] \leq \frac{7}{5}\frac{\lambda^5 (n-1) p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}}, \tag{5.157}$$

which is as stated in the Lemma. QED.

Using Lemma 14 we can now compute $\Phi_2^{(1,2)}$ and $\Phi_2^{(2,1)}$.

Lemma 15. With n, p, ζ, $\Phi_2^{(1,2)}$ and $\Phi_2^{(2,1)}$ as previously defined,

$$\Phi_2^{(1,2)} + \Phi_2^{(2,1)} \leq \frac{np^3(1-p)\lambda^5}{\mathrm{Var}(\zeta)^{1.5}}\Big(12(\lambda+p)+6\Big). \tag{5.158}$$

Proof We now combine $\Phi_2^{(1,2)}$ and $\Phi_2^{(2,1)}$ and we let $\kappa = (\alpha_1, \alpha_2)$ if $\alpha_2 \in V_2$ or $\kappa = (\alpha_2, \alpha_3)$ if $\alpha_2 \in V_1$ as above. We recall from (5.137) that $\mathfrak{K}_\alpha^{(1)} = \{\alpha, \} \cup \{\psi \in I_1 : |\eta(\alpha) \cap \eta(\psi)| = 1\}$. Therefore, for $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\mathfrak{K}_\alpha^{(1)} = \{\alpha, \kappa\}$. Thus,

$$\Phi_2^{(1,2)} + \Phi_2^{(2,1)} = \sum_{\beta \in \mathfrak{K}_\alpha^{(1)} = \{\alpha,\kappa\}} \sum_{\gamma \in \mathfrak{K}_\alpha^{(2)}} E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) \tag{5.159}$$
$$+ \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in \mathfrak{K}_\alpha^{(1)} = \{\alpha,\kappa\}} (E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|])$$
$$= \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} 2E[|Y_\alpha^2 Y_\beta|] + 2E[|Y_\alpha Y_\kappa Y_\beta|] \tag{5.160}$$
$$+ (E[|Y_\alpha^2|] + E[|Y_\alpha Y_\kappa|])E[|Y_\beta|]) + E[|Y_\alpha Y_\beta|](E[|Y_\alpha|] + E[|Y_\kappa|]),$$

where the final line is obtained by performing the summation over $\mathfrak{K}_\alpha^{(1)}$. Unfortunately this sum is not easily computable in its current form as Y_β and Y_κ may not be independent. To address this we split out the terms as follows:

$$\begin{aligned}
\Phi_2^{(1,2)} + \Phi_2^{(2,1)} = &\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \Big(2E[|Y_\alpha^2 Y_\beta|] + \left(E[|Y_\alpha^2|] + E[|Y_\alpha Y_\kappa|] \right) E[|Y_\beta|]) \qquad (5.161) \\
&+ E[|Y_\alpha Y_\beta|] \left(E[|Y_\alpha|] + E[|Y_\kappa|] \right) \Big) \\
&+ 2 \sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|] + 2 \sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|] \\
= &2 \frac{\lambda^5 p^3 (1-p)}{\mathrm{Var}(\zeta)^{1.5}} (2n-4) \lambda \left(1 + p(1-p^2)(3-2p^2)(2p^2+1) \right) \qquad (5.162) \\
&+ 2 \frac{\lambda^5 p^3 (1-p)}{\mathrm{Var}(\zeta)^{1.5}} (2n-4) p(1-p)(2+7p-6p^3) \qquad (5.163) \\
&+ 2 \sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|] + 2 \sum_{\beta \in \mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa} E[|Y_\alpha Y_\kappa Y_\beta|]. \qquad (5.164)
\end{aligned}$$

Before proceeding to the remaining terms, we bound the terms (5.162) and (5.163) we just computed. We start with (5.163): $(1-p)(2+7p-6p^3) = 2+5p-7p^2-6p^3+6p^4 \leq 2+5p-7p^2 = f(p)$.

Clearly, $f'(p) = 5-14p,\ f''(p) = -14$, which implies $p = \frac{5}{14}$ is the unique maximum:

$$(1-p)(2+7p-6p^3) \leq 2+5p-7p^2 \leq f\left(\frac{5}{14}\right) < 3. \qquad (5.165)$$

Next we bound (5.162) by differentiation:

$(1+p(1-p^2)(3-2p^2)(2p^2+1)) = 1+3p+p^3-8p^5+4p^7 \leq 1+3p+p^3-4p^5 = f(p)$;

so $f'(p) = 3+3p^2-20p^4$. The equation $f'(p) = 0$ is solved by $p = \pm\left(\frac{3 \pm \sqrt{249}}{40}\right)^{0.5}$.

The positive real solution is $p = \left(\frac{3+\sqrt{249}}{40}\right)^{0.5}$.

Hence, $f(p) \leq \mathrm{Max}\left(f(0), f\left(\left(\frac{3+\sqrt{249}}{40}\right)^{0.5} \right), f(1) \right) < 3. \qquad (5.166)$

We then substitute the bound from (5.165) and (5.166) into (5.163) and (5.162) respectively. We further note that the remaining sum is bounded by Lemma 13. Combining these results we obtain:

$$\Phi_2^{(1,2)} + \Phi_2^{(2,1)} \leq \frac{\lambda^5 p^3(1-p)}{\mathrm{Var}(\zeta)^{1.5}} \left(6(2n-4)\lambda + 6(2n-4)p + \frac{16}{5}(n-3) + \frac{14}{5}(n-1) \right) \tag{5.167}$$

$$\leq \frac{np^3(1-p)\lambda^5}{\mathrm{Var}(\zeta)^{1.5}} \left(12(\lambda+p) + 6 \right), \tag{5.168}$$

which is as stated in the Lemma, QED.

We now bound $\Phi_2^{(2,2)}$. Again it is more convenient to split this computation into a lemma which deals with the complex terms and a lemma which then combines this result into the final calculation.

Lemma 16. For $\mathfrak{K}$ as defined in (5.137) and (5.138) and for $\alpha \in I_2$:

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|]) \leq \frac{5}{2} \frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}} 2(n-1)(n-3)p^4(1-p), \tag{5.169}$$

and,

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|]) = 3 \frac{\lambda^6}{\mathrm{Var}(\eta)^{1.5}} (2n^2 - 10n + 14)p^4(1-p). \tag{5.170}$$

Proof Let us consider the first summation: $\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|])$. We first show that up to trivial rearrangement the form of $Y_\alpha Y_\beta Y_\gamma$ is equal for all

terms in the summation.

As $\mathfrak{K}_{\alpha}^{(2)} = \mathbb{K}_{\alpha}^{(1)} \cap I_2,$ therefore we have that $\forall x \in \mathfrak{K}_{\alpha}^{(2)}\ |\eta(\alpha) \cap \eta(x)| = 1.$

Therefore if $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\beta \in \mathfrak{K}_{\alpha}^{(2)}$, and $\gamma \in (\mathfrak{K}_{\alpha}^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_{\beta}$.

Then, $\eta(\alpha) \cap \eta(\beta)$ is either $\{\{\alpha_1, \alpha_2\}\},$ or $\{\{\alpha_2, \alpha_3\}\}$.

Further, $\eta(\alpha) \cap \eta(\gamma)$ is either $\{\{\alpha_1, \alpha_2\}\},$ or $\{\{\alpha_2, \alpha_3\}\}$.

Trivially $\mathbb{K}_{\beta} \cap ((\mathfrak{K}_{\alpha}^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_{\beta}) = \emptyset$ hence $\eta(\beta) \cap \eta(\gamma) = \emptyset = \eta(\beta) \cap \eta(\gamma) \cap \eta(\alpha)$

therefore $\eta(\alpha) \cap \eta(\beta) \neq \eta(\alpha) \cap \eta(\gamma)$ and we conclude that $(\eta(\alpha) \cap \eta(\beta), \eta(\alpha) \cap \eta(\gamma))$ equals $(\{\alpha_1, \alpha_2\}, \{\alpha_2, \alpha_3\})$ or $(\{\alpha_2, \alpha_3\}, \{\alpha_1, \alpha_2\})$.

We observe that the final choice of $(\eta(\alpha) \cap \eta(\beta), \eta(\alpha) \cap \eta(\gamma))$ does not affect the expectation in the summand. Therefore for $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\beta = (\alpha_1, \alpha_2, \beta_3)$ and $\gamma = (\gamma_1, \alpha_2, \alpha_3)$ we obtain:

$$E[|Y_{\alpha}Y_{\beta}Y_{\gamma}|] = \frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}} E[|(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\beta_3} - p^2)(\mathfrak{A}_{\gamma_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)|]. \tag{5.171}$$

We can then directly compute the expectation as follows:

$$\begin{aligned} E[|Y_{\alpha}Y_{\beta}Y_{\gamma}|] &= \frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}} \Big((1-p)p^4 E[|(\mathfrak{A}_{\gamma_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)|] \qquad (5.172) \\ &\qquad + pE[|\mathfrak{A}_{\alpha_2\beta_3} - p^2|]E[|(\mathfrak{A}_{\alpha_2\alpha_3} - p^2)(\mathfrak{A}_{\gamma_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)|] \Big) \\ &= \frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}} p^4(1-p)^2(1 + 4p + 6p^2 - 4p^4). \qquad (5.173) \end{aligned}$$

We sum over $\beta \in \mathfrak{K}_\alpha^{(2)}$ and $\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta$ to obtain:

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta} E[|Y_\alpha Y_\beta Y_\gamma|] = \frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta} p^4(1-p)^2(1+4p+6p^2-4p^4). \tag{5.174}$$

We bound this term as follows:

$$(1-p)(1+4p+6p^2-4p^4) = 1+3p+2p^2-6p^3-4p^4+4p^5 \tag{5.175}$$

$$\leq 1+3p+2p^2-6p^3 = f(p) \tag{5.176}$$

By (5.78) as $f'(p) = 0$ at $p = \dfrac{2+\sqrt{58}}{18}$ therefore:

$f(p) \leq \mathrm{Max}\left(f(0), \left(\dfrac{2+\sqrt{58}}{18}\right), f(1)\right) < \dfrac{5}{2}$ and hence

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|]) \leq \frac{5}{2} \frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}} \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta} p^4(1-p). \tag{5.177}$$

We observe that the summand does not depend on the particular choice of α, β or γ. We now compute the number of terms in the sum. This is more involved than it was in $\Phi_2^{(1,2)}$ as we must consider the identity of the overlapping edge between α and β. We could simply perform the sum

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} |(\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta|. \tag{5.178}$$

However, there is a simpler way to obtain these quantities. We know that only one of the edges of α is in I_1, and we also know from the derivation of the expression for

$\Phi_1^{(1,2)}$ that:

$$|\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa| = n-3, \tag{5.179}$$

$$|\mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa| = n-1, \tag{5.180}$$

where κ is the edge from V_1 to V_2 in $\eta(\alpha)$. Clearly as $|\eta(\alpha)\cap\eta(\beta)| = |\eta(\alpha)\cap\eta(\gamma)| = 1$, and by condition on the sum index, $|\eta(\beta)\cap\eta(\gamma)| = 0$, therefore if $\beta \in \mathbb{K}_\kappa$ then $\gamma \notin \mathbb{K}_\kappa$ and vice versa. Therefore we can compute the sum as follows:

$$\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}} |(\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta| = |\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa||\mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa| + |\mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa||\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa| \tag{5.181}$$

$$= 2|\mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa||\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa| = 2(n-1)(n-3). \tag{5.182}$$

We can use the same approach to compute the other quantity:

$$\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}} |(\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta| = |\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa|(|\mathfrak{K}_\alpha^{(2)} \cap \mathbb{K}_\kappa| - 1) + |\mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa|(|\mathfrak{K}_\alpha^{(2)} \setminus \mathbb{K}_\kappa| - 1) \tag{5.183}$$

$$= (n-1)(n-2) + (n-3)(n-4) = 2n^2 - 10n + 14. \tag{5.184}$$

We note that both the total number of terms in the pair of summations $(2n-4)(2n-5)$ and the sum of the number of terms in each summation are equal to $4n^2 - 18n + 20$. Therefore:

$$\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}} \sum_{\gamma\in(\mathfrak{K}_\alpha^{(2)}\setminus\{\beta\})\setminus\mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|]) \leq \frac{5}{2}\frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}} 2(n-1)(n-3)p^4(1-p). \tag{5.185}$$

We now bound the second sum of Lemma 16, $\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}} \sum_{\gamma\in(\mathfrak{K}_\alpha^{(2)}\setminus\{\beta\})\cap\mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|])$. First we explore the possible functional forms that $E[|Y_\alpha Y_\beta Y_\gamma|]$ can take for $\alpha \in$

$I_2,\ \beta \in \mathfrak{K}_\alpha^{(2)},\ \gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta$. One option is for α, β and γ to share the same edge. For example $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\beta = (\alpha_1, \alpha_2, \beta_3)$ and $\gamma = (\alpha_1, \alpha_2, \gamma_3)$, which gives:

$$E[|Y_\alpha Y_\beta Y_\gamma|] = \frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}}(E[|(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\beta_3} - p^2)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|]) \tag{5.186}$$

We now show that α, β and γ sharing a single edge is the only way that we can construct these paths under the given conditions. To show this, it is helpful to state what is implied by the conditions on the summation. We enforce that Y_β and Y_γ are dependent on Y_α; Y_β and Y_γ are dependent on each other. Note, by definition of the path $\alpha_1, \beta_1, \gamma_1 \in V_1$ and $\alpha_3, \beta_3, \gamma_3 \in V_2$. Further, α, β, γ are all distinct, $\alpha \in I_2$, $\beta \in \mathfrak{K}_\alpha^{(2)}$, and $\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta$. This implies that:

$$|\eta(\alpha) \cap \eta(\beta)| = |\eta(\alpha) \cap \eta(\gamma)| = |\eta(\beta) \cap \eta(\gamma)| = 1. \tag{5.187}$$

Note that $|\eta(\beta) \cap \eta(\gamma)| = 1$, as $\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta$ implies that $\gamma \neq \beta$ and therefore $|\eta(\beta) \cap \eta(\gamma)| < 2$. But, as $\gamma \in \mathbb{K}_\beta$, therefore $|\eta(\beta) \cap \eta(\gamma)| > 0$. Combining these we obtain $|\eta(\beta) \cap \eta(\gamma)| = 1$.

Without loss of generality let us assume that α and β share the edge (α_1, α_2), i.e. $\{\alpha_1, \alpha_2\} \in \eta(\alpha)$ and $\{\alpha_1, \alpha_2\} \in \eta(\beta)$. Clearly $\{\alpha_1, \alpha_2\} \in \eta(\gamma)$ would satisfy both conditions and would imply that they share the same edge. Let us assume that

$\{\alpha_1, \alpha_2\} \notin \eta(\gamma)$[1].

Then for $\{\alpha_1, \alpha_2\} \notin \eta(\gamma), \quad \{\alpha_2, \alpha_3\} \in \eta(\gamma)$.

Hence, $\eta(\gamma) = \{\{\alpha_2, \alpha_3\}, \{\gamma_b, \alpha_2\}\}$ or $\{\{\alpha_2, \alpha_3\}, \{\gamma_b, \alpha_3\}\}$.

Similarly, $\eta(\beta) = \{\{\alpha_1, \alpha_2\}, \{\beta_a, \alpha_2\}\}$ or $\{\{\alpha_1, \alpha_2\}, \{\beta_a, \alpha_3\}\}$

Further, $|\eta(\beta) \cap \eta(\gamma)| = 1,$ therefore let us consider the edge on which they overlap:

Case 1: If $\{\alpha_1, \alpha_2\} = \{\alpha_2, \alpha_3\},$ then α is non-simple !

Case 2: If $\{\alpha_1, \alpha_2\} = \{\gamma_b, \alpha_{2 \text{ or } 3}\},$

Then only simple path solution is $\gamma = (\alpha_1, \alpha_2, \alpha_3) = \alpha$!

Case 3: If $\{\alpha_2, \alpha_3\} = \{\beta_b, \alpha_{1 \text{ or } 2}\},$

Then only simple path solution is $\beta = (\alpha_1, \alpha_2, \alpha_3) = \alpha$!

Case 4: If $\{\gamma_b, \alpha_{2 \text{ or } 3}\} = \{\beta_b, \alpha_{1 \text{ or } 2}\} = \{\alpha_2, \gamma_b\}$ or $\{\alpha_1, \alpha_3\},$ then

Case 4.1: If it is: $\{\alpha_1, \alpha_3\}$ then $\gamma = (\alpha_1, \alpha_3, \alpha_2)$ which implies that $\alpha_2 \in V_2$

Further, $\beta = (\alpha_2, \alpha_1, \alpha_3)$ which implies that $\alpha_2 \in V_1$!

Case 4.2: If it is:$\{\alpha_2, \gamma_b\}$ then $\beta = (\alpha_1, \alpha_2, \gamma_b)$ which implies that $\gamma_b \in V_2$.

Further, $\gamma = (\gamma_b, \alpha_2, \alpha_3)$ which implies that $\gamma_b \in V_1$!

We therefore that α, β and γ share the same edge and thus confirm that Eq.(5.186) is the functional form of $E[|Y_\alpha Y_\beta Y_\gamma|]$ for $\alpha \in I_2, \ \beta \in \mathfrak{K}_\alpha^{(2)}, \ \gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta$. Hence with $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\beta = (\alpha_1, \alpha_2, \beta_3)$ and $\gamma = (\alpha_1, \alpha_2, \gamma_3)$ we obtain:

$$E[|Y_\alpha Y_\beta Y_\gamma|] = \frac{\lambda^6}{\text{Var}(\eta)^{1.5}} E[|(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\alpha_3} - p^2)(\mathfrak{A}_{\alpha_2\alpha_3}\mathfrak{A}_{\alpha_2\beta_3} - p^2)(\mathfrak{A}_{\alpha_1\alpha_2}\mathfrak{A}_{\alpha_2\gamma_3} - p^2)|] \tag{5.188}$$

$$= \frac{\lambda^6}{\text{Var}(\eta)^{1.5}} \left((1-p)p^6 + pE[\mathfrak{A}_{\alpha_2\alpha_3} - p^2]^3\right). \tag{5.189}$$

[1]We represent contradiction by the symbol !

Summing over all $\beta \in \mathfrak{K}_\alpha^{(2)}$ and $\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta$ we obtain

$$\sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta} E[|Y_\alpha Y_\beta Y_\gamma|] = \frac{\lambda^6}{\mathrm{Var}(\eta)^{1.5}}(2n^2-10n+14)p^4(1-p)(p^2+(1-p)^2(2p+1)^3). \tag{5.190}$$

Here, we use (5.183) to compute the number of terms in the summation.

We now bound the expression above for conciseness let:

$$p^2 + (1-p)^2(2p+1)^3 = 1 + 4p + 2p^2 - 10p^3 - 4p^4 + 8p^5 \leq 1 + 4p + 2p^2 - 6p^3 = f(p). \tag{5.191}$$

Using 5.78, we note that $f'(p) = 0$ when $p = \dfrac{1+\sqrt{19}}{9}$ (ignoring the negative solution), (5.192)

Therefore: $f(p) \leq \mathrm{Max}\left(f(0), f(1), f\left(\dfrac{1+\sqrt{19}}{9}\right)\right) < 3.$ (5.193)

Combining these results we obtain the result stated in the Lemma. QED.

Lemma 17. For $n > \frac{7}{5}$ we have

$$\Phi_2 \leq \frac{\lambda^4 c^2 (n-c)}{4\mathrm{Var}(\zeta)^{1.5}}\left(2\Big(\lambda^2 + 2\lambda + \frac{4}{5}\Big) + 11(c\lambda)^2 + 15c\lambda^2 + 6c\lambda\frac{c^2\lambda}{n}(\frac{88}{5}\lambda c + 12)\right). \tag{5.194}$$

Proof Following the approach we used in computing $\Phi_2^{(1,2)}$, we split the sum into four components as follows:

$$\Phi_2^{(2,2)} = \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in \mathfrak{K}_\alpha^{(2)}} (E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) \tag{5.195}$$

$$= \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in \mathfrak{K}_\alpha^{(2)}} (E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) + \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} (E[|Y_\alpha Y_\beta^2|]) \tag{5.196}$$

$$+ \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \setminus \mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|]) + \sum_{\beta \in \mathfrak{K}_\alpha^{(2)}} \sum_{\gamma \in (\mathfrak{K}_\alpha^{(2)} \setminus \{\beta\}) \cap \mathbb{K}_\beta} (E[|Y_\alpha Y_\beta Y_\gamma|]).$$

In the same fashion as we for the previous terms we can compute the first term as follows:

$$\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}}\sum_{\gamma\in\mathfrak{K}_\alpha^{(2)}}(E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) =\frac{\lambda^6}{\mathrm{Var}(\zeta)^{1.5}}\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}}\sum_{\gamma\in\mathfrak{K}_\alpha^{(2)}}p^3(1-p)(1+4p-4p^3)(2p^2(1-p^2)) \tag{5.197}$$

$$=\frac{2\lambda^6(2n-4)^2}{\mathrm{Var}(\zeta)^{1.5}}p^5(1-p)(1-p^2)(1+4p-4p^3). \tag{5.198}$$

We bound this term as follows let:

$$(1-p^2)(1+4p-4p^3) = 1+4p-p^2-8p^3+4p^5 \leq 1+4p-8p^3+4p^5 = f(p) \tag{5.199}$$

so that, using (5.78), we note that $f'(p)=0$ when $p=\frac{\sqrt{5}}{5}$ or $p=1$ (ignoring the negative solutions), therefore: $f(p) \leq \mathrm{Max}\left(f(0), f\left(\frac{\sqrt{5}}{5}\right), f(1)\right) < \frac{11}{5}$.

Hence, $\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}}\sum_{\gamma\in\mathfrak{K}_\alpha^{(2)}}(E[|Y_\alpha Y_\beta|]E[|Y_\gamma|]) < \frac{11}{5}\frac{2\lambda^6(2n-4)^2}{\mathrm{Var}(\zeta)^{1.5}}p^5(1-p).$ (5.200)

We compute the second term as follows:

$$\sum_{\beta\in\mathfrak{K}_\alpha^{(2)}}(E[|Y_\alpha Y_\beta^2|]) = \sum_{\beta\in\mathfrak{K}_\alpha^{(2)}}p^3(1-p)(1+2p-2p^2-2p^3+2p^4) \tag{5.201}$$

$$=\frac{\lambda^6(2n-4)}{\mathrm{Var}(\zeta)^{1.5}}p^3(1-p)(1+2p-2p^2-2p^3+2p^4). \tag{5.202}$$

We will use a similar approach to bound this term using (5.76). It is easy to see that:

$$1+2p-2p^2-2p^3+2p^4 \leq 1+2p-2p^2 \leq \frac{3}{2}. \tag{5.203}$$

$$\text{Hence, } \sum_{\beta\in\mathfrak{K}_\alpha^{(2)}}(E[|Y_\alpha Y_\beta^2|]) \leq \frac{3}{2}\frac{\lambda^6(2n-4)}{\mathrm{Var}(\zeta)^{1.5}}p^3(1-p). \tag{5.204}$$

Using Lemma 16 for the remaining sum, we can now bound Φ_2 from its constituent parts into one term,

$$\Phi_2 = |I_2| \sum_i \sum_j \Phi_2^{(i,j)} \tag{5.205}$$

$$\leq |I_2| \frac{\lambda^4 p^2 (1-p)}{\mathrm{Var}(\zeta)^{1.5}} \Bigg(2\Big(\lambda^2 + 2\lambda + \frac{4}{5}\Big) + np\lambda\Big(12(p+\lambda)+6\Big) + \frac{11}{5} 2\lambda^2 (2n-4)^2 p^3 \tag{5.206}$$

$$+ 3\lambda^2 (n-2)p + 5\lambda^2 (n-1)(n-3)p^2 + 3\lambda^2 (2n^2 - 10n + 14)p^2 \Bigg).$$

Note that $n - |\rho| < n$ and $2n^2 > 2n^2 - 10n + 14$ if $n > \frac{7}{5}$. Thus $|I_2| = \frac{n^2}{4}(n-2) < \frac{n^3}{4}$. Further we recall that $p = \frac{c}{n}$ therefore:

$$\Phi_2 \leq \frac{\lambda^4 c^2 (n-c)}{4\mathrm{Var}(\zeta)^{1.5}} \Bigg(2\Big(\lambda^2 + 2\lambda + \frac{4}{5}\Big) + 11(c\lambda)^2 + 15c\lambda^2 + 6c\lambda \frac{c^2\lambda}{n} \Big(\frac{88}{5}\lambda c + 12\Big) \Bigg). \tag{5.207}$$

which is as stated in the Lemma. QED.

We can now give the final result.

Theorem 18. *For $n > 4$ and $c \geq 4$ and $\lambda \geq 0$:*

$$d_{H_W}(\tilde{\zeta}, Z) \leq 8 \frac{1 + 5c\lambda + 30c^2\lambda^2 \max(1,\lambda) + 11(c\lambda)^3 + \frac{44c^3\lambda^2}{n}\Big(\frac{2}{5}\lambda c + \frac{1}{3}\Big)}{c^{0.5}(n-c)^{0.5}\Big(1 + c\lambda^2(2c-7)\Big)^{1.5}}. \tag{5.208}$$

Proof From (5.79) we know that:

$$d_{H_W}(\tilde{\zeta}, Z) \leq 4(\Phi_1 + \Phi_2). \tag{5.209}$$

Further from Lemma 12 and Lemma 17 we have bounds on Φ_1 and Φ_2 as follows:

$$\Phi_1 \leq \frac{1}{4}\frac{\lambda^3 c(n-c)}{\mathrm{Var}(\zeta)^{1.5}}\left(1+\frac{16}{5}\lambda c+\frac{8}{3}\lambda^2\frac{c^3}{n}+\lambda^2 c+\frac{3}{2}\lambda^2 c^2\right), \tag{5.210}$$

$$\Phi_2 \leq \frac{\lambda^4 c^2(n-c)}{4\mathrm{Var}(\zeta)^{1.5}}\left(2\Big(\lambda^2+2\lambda+\frac{4}{5}\Big)+11(c\lambda)^2+15c\lambda^2+6c\lambda\frac{c^2\lambda}{n}(\frac{88}{5}\lambda c+12)\right). \tag{5.211}$$

Combining Φ_1 and Φ_2 we obtain:

$$\begin{aligned}\Phi_1+\Phi_2 \leq & \frac{1}{4}\frac{\lambda^3 c(n-c)}{\mathrm{Var}(\zeta)^{1.5}}\left(1+\frac{16}{5}\lambda c+\frac{8}{3}\lambda^2 c^2 p+\lambda^2 c+\frac{3}{2}\lambda^2 c^2\right) \qquad (5.212)\\ & +\frac{\lambda^4 c^2(n-c)}{4\mathrm{Var}(\zeta)^{1.5}}\left(2\Big(\lambda^2+2\lambda+\frac{4}{5}\Big)+11(c\lambda)^2+15c\lambda^2+6c\lambda\frac{c^2\lambda}{n}(\frac{88}{5}\lambda c+12)\right)\Bigg)\\ \leq & \frac{\lambda^3 c(n-c)}{4\mathrm{Var}(\zeta)^{1.5}}\left(1+c\lambda\left(2\lambda^2+5\lambda+\frac{24}{5}+c\lambda\Big(11c\lambda+15\lambda+\frac{15}{2}\Big)\right)\right. \qquad (5.213)\\ & \left.+\frac{44c^3\lambda^2}{n}\Big(\frac{2}{5}\lambda c+\frac{1}{3}\Big)\right).\end{aligned}$$

We now consider the variance. From Theorem 7 we know that:

$$\mathrm{Var}(\zeta) = \frac{\lambda^2 c(1-p)}{4}\Big(n+\lambda c(n-2)(2+\lambda((2n-3)\frac{c}{n}+1)\Big). \tag{5.214}$$

As this is the denominator of the bound, if we want to approximate this expression we use a lower bound. Clearly:

$$\mathrm{Var}(\zeta) \geq \frac{\lambda^2 c(n-c)}{4}\Big(1+\frac{1}{n^2}(\lambda c)^2(n-2)(2n-3)\Big). \tag{5.215}$$

Consider $\frac{(2n-3)(n-2)}{n^2}$. We know that $4 \leq n$ and n even, if n was smaller than this

paths of length 2 would not be possible. Further, by assumption $c \geq 4$. Therefore:

$$f(n) = \frac{(2n-3)(n-2)}{n^2} = 2 - \frac{7}{n} + \frac{6}{n^2} \tag{5.216}$$

$$f'(n) = \frac{7}{n^2} - \frac{12}{n^3} \tag{5.217}$$

Note: $f'(n) > 0$ for $n > \frac{12}{7}$.

Therefore for $x > 4 > \frac{12}{7}$, if $n \geq x$ then $f(n) \geq f(x)$.

By assumption, $c \geq 4$. Set $x = c$, gives that $f(n) \geq 2 - \frac{7}{c} + \frac{6}{c^2} > 2 - \frac{7}{c}$.

Substituting this into the variance bound we obtain:

$$\text{Var}(\zeta) \geq \frac{\lambda^2 c(n-c)}{4}\Big(1 + c\lambda^2(2c-7)\Big) \text{ and} \tag{5.218}$$

$$\text{Var}(\zeta)^{1.5} \geq \frac{\lambda^3 c^{1.5}(n-c)^{1.5}}{8}\Big(1 + c\lambda^2(2c-7)\Big)^{1.5}. \tag{5.219}$$

Performing the division we obtain:

$$\Phi_1 + \Phi_2 \leq 2\frac{1 + c\lambda\Big(2\lambda^2 + 5\lambda + \frac{24}{5} + c\lambda\Big(11c\lambda + 15\lambda + \frac{15}{2}\Big)\Big) + \frac{44c^3\lambda^2}{n}\Big(\frac{2}{5}\lambda c + \frac{1}{3}\Big)}{c^{0.5}(n-c)^{0.5}\Big(1 + c\lambda^2(2c-7)\Big)^{1.5}}. \tag{5.220}$$

Using that $c \geq 4$ let us simplify the following bracket,

$$c\lambda\Big(2\lambda^2 + 5\lambda + \frac{24}{5} + c\lambda\Big(11c\lambda + 15\lambda + \frac{15}{2}\Big)\Big) \tag{5.221}$$

$$\leq c\lambda\Big(\frac{24}{5} + \max(1,\lambda)\lambda c(2 + 5 + 15 + \frac{15}{2}) + 11(c\lambda)^2\Big) \tag{5.222}$$

$$\leq c\lambda\Big(5 + 30\max(1,\lambda)\lambda c + 11(c\lambda)^2\Big). \tag{5.223}$$

Recalling (5.79), we obtain the final bound as follows:

$$d_{H_W}(\tilde{\zeta}, Z) \leq 8\frac{1 + 5c\lambda + 30c^2\lambda^2\max(1,\lambda) + 11(c\lambda)^3 + \frac{44c^3\lambda^2}{n}\Big(\frac{2}{5}\lambda c + \frac{1}{3}\Big)}{c^{0.5}(n-c)^{0.5}\Big(1 + c\lambda^2(2c-7)\Big)^{1.5}}, \tag{5.224}$$

which is as stated in the theorem QED.

The bounds are of order $n^{-0.5}$ for fixed c, l and λ as $n \to \infty$. This is the expected order for sums of n weakly dependent variables. As ζ sums over order of n^3 variables, the bound might not be optimal.

We now construct some motivating examples to see how this bound performs as a function of n.

Example 1 Our first motivating example is $\lambda = \frac{1}{c}$. This makes the relative contribution of each path length approximately equal. If we do so and make a few other bounding approximations we obtain:

$$d_{H_W}(\tilde{\zeta}, Z) \leq 8 \frac{47 + 33cn^{-1}}{c^{0.5}(n-c)^{0.5}\Big(3 - \frac{7}{c}\Big)^{1.5}}. \tag{5.225}$$

Example 2 Another possible option is to fix $\lambda = 1$ i.e. weight each of the paths evenly. This results in the following expression:

$$d_{H_W}(\tilde{\zeta}, Z) \leq 8 \frac{1 + 5c + 30c^2 + 11c^3 + \frac{44c^3}{n}\Big(\frac{2}{5}c + \frac{1}{3}\Big)}{c^{0.5}(n-c)^{0.5}\Big(1 + c(2c-7)\Big)^{1.5}}. \tag{5.226}$$

We plot $4(\Phi_1 + \Phi_2)$ (i.e. a bound on the Wasserstein distance) against n with $c = 100$ with the two schemes, as can be seen in Figure. 5.1. The bound tends to 0 as $n \to \infty$, further, the simplified bound for $\lambda = c^{-1}$ is out performed by $\lambda = 1$ even when $n = 1000$.

5.4.1.3 Comparing the computed bounds to simulated data

To evaluate the effectiveness of the resultant bound to the Wasserstein distance between a normal distribution and our distribution of interest we estimate the Wasserstein distance computationally. The formulation of the Wasserstein distance in 2.7.1, i.e. the supremum over all Lipschitz continuous functions with Lipschitz constant 1,

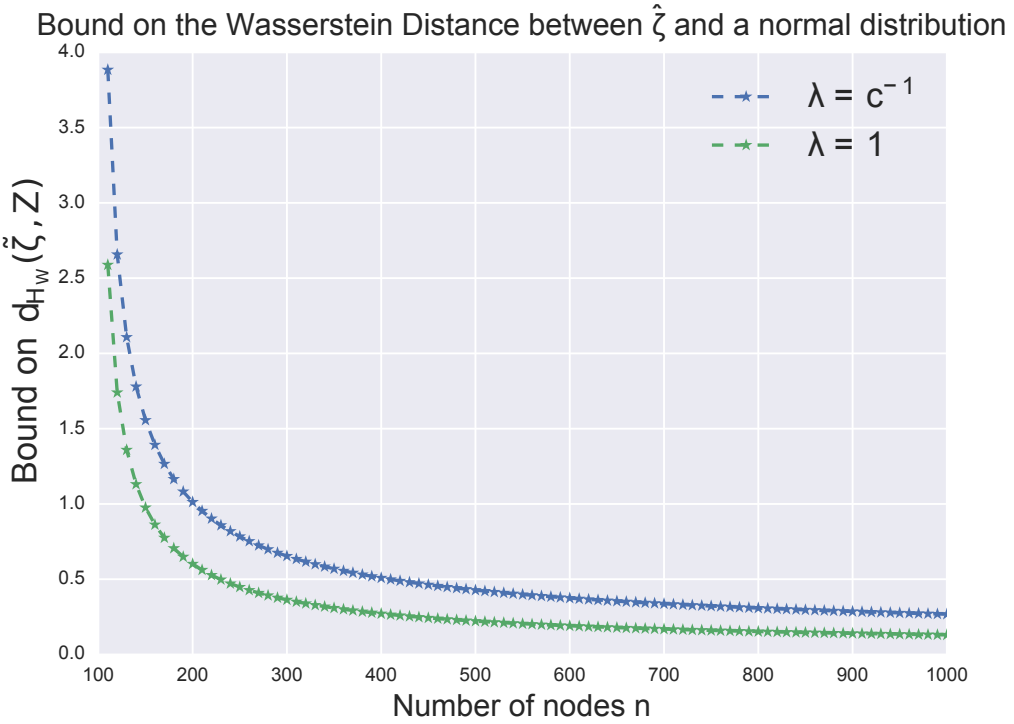

Figure 5.1: The bound from Theorem 18 for $l = 2$, n is plotted against $4(\Phi_1 + \Phi_2)$. Note that for small n the bound is restrictively large. Here the bound for $n = 100$ is not plotted as this results in a division by zero, the first result is the bound for $n = 110$.

is not a particularly useful formulation in this context. However, we can use a result from [174], which states that:

$$d_{H_W}(F, G) = \int_{-\infty}^{\infty} |F(x) - T(x)| dx \tag{5.227}$$

where $F(x)$ and $T(x)$ are the cumulative density functions of the two random variables in question. Using this formulation and the fact that one of the distributions is a normal random variable we can then use the following procedure:

1. Generate many random samples from the non-normal distribution.
2. Use these samples to construct an empirical CDF of the non-normal distribution.
3. Use (5.227) to estimate the difference.

Mathematically this is equivalent to the following. We estimate the CDF by numerically constructing N_{samp} samples from the distribution, which we name $x_1, x_2, ..., x_{N_{samp}}$, and then we can estimate $T(x)$ as follows:

$$\hat{T}(x) = \frac{1}{N_{samp}} \sum_{i=1}^{N_{samp}} \mathbb{1}(x \leq x_i) \tag{5.228}$$

As the weighted sum of paths is a discrete random variable we must adjust for potential multiplicity. Let x_i^* be a sequence of the unique ordered elements in x_i, i.e. $x_1^* < x_2^* < ... < x_{N_{samp}^*}^*$, where N_{samp}^* is the number of unique x_i. Further, let M_i be the multiplicity of each of the x_i^* in x_i, which is equivalent to the following expression:

$$M_i = \sum_j \mathbb{1}(x_i^* = x_j). \tag{5.229}$$

Using the x_i^* and M_i we can construct an equivalent formulation of $\hat{G}$ as follows:

$$\hat{T}(x) = \frac{1}{N_{samp}} \sum_{i=1}^{N_{samp}^*} M_i \mathbb{1}(x \leq x_i^*) \tag{5.230}$$

Putting this all together we obtain the following estimate for d_{H_W}:

$$
\begin{aligned}
d_{H_W} &= \int_{-\infty}^{\infty} |F(x) - T(x)| dx && (5.231)\\
&\approx \int_{-\infty}^{\infty} |F(x) - \hat{T}(x)| dx && (5.232)\\
&= \int_{-\infty}^{x_1^*} |F(x) - \hat{T}(x)| dx + \left(\sum_{i=1}^{N_{samp}^*} \int_{x_i^*}^{x_{i+1}^*} |F(x) - \hat{T}(x)| dx \right) && (5.233)\\
&\quad + \int_{x_{N_{samp}^*}^*}^{\infty} |F(x) - \hat{T}(x)| dx
\end{aligned}
$$

The final issue we must address is the accuracy of the approximation of the CDF. Trivially, if this approximation is not accurate then this could lead to an poor estimation.

We set the numbers of random samples taken at $1,000,000$ and explored the convergence of each of the statistics with different number of samples. Even at 10^6 samples there is some volatility in the estimation and thus this should not be considered a accurate estimate. However, testing with different numbers of samples appears to indicate that the order of magnitude is reasonable in all cases. For some of the more problematic cases, namely for $n = 500, 600, 700$ (both for $\lambda = 1$ and $\lambda = \frac{1}{c}$) and $n = 400$ (for $\lambda = \frac{1}{c}$), we increased the number of samples to $2,000,000$ in order to improve our estimation. Even with the large number of samples what we obtain should be considered an order of magnitude estimation.

For computational reasons further adjustments had to be made to the expression in (5.233). The small difference between the upper and lower bound of the integral caused by the large number of samples resulted in errors from the numerical integrator (the quad routine from SciPy [89]). Thus, rather than performing a separate integral for each unique element, we combine adjacent integrals, which resolves the problem for the most cases. For the case of $n = 500$ with $\lambda = \frac{1}{c}$ we increase the number of adjacent intervals combined to 10.

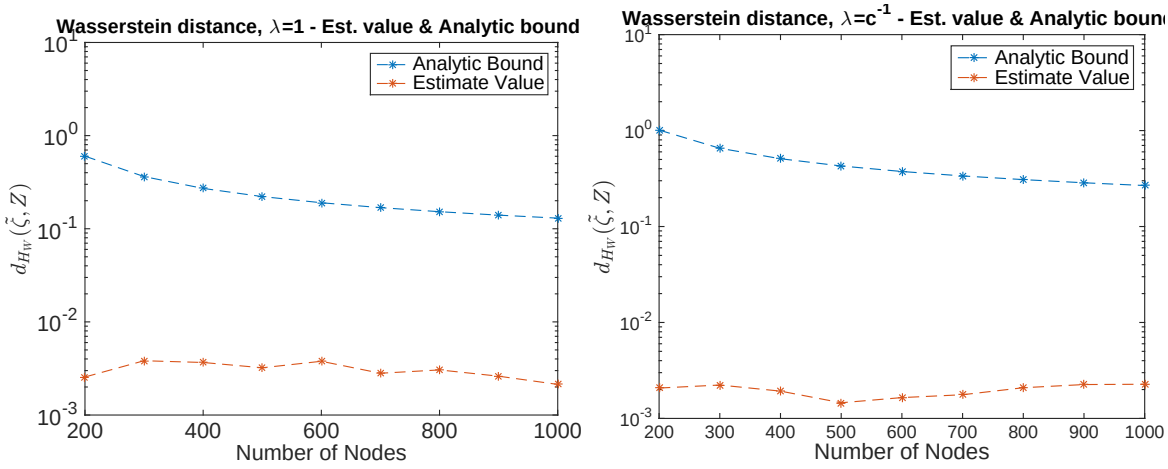

Figure 5.2: Plot of the estimated value of $d_{H_W}(\tilde{\zeta}, Z)$ for $\lambda = 1$ and $\lambda = \frac{1}{c}$ against the simplified form of the bound on $d_{H_W}(\tilde{\zeta}, Z)$ from Theorem 18 in (5.225) and (5.226).

In Figure 5.2 we plot the simplified form of the bound from (5.225) and (5.226) that we obtained in Theorem 18 against the estimated distance we derived in this section. We note that the estimated actual value of the Wasserstein distance is substantially lower than the bound for small n.

5.4.2 The general case $l > 2$

We derived explicit bounds for the case $l = 2$. For the general case $l > 2$ this is more difficult as the increased length of path gives a large number of ways in which two paths can overlap.

Here we consider $l < \frac{n}{2}$ so that paths of length l could stay in the same group in all but one node, and for asymptotic results l will be considered as fixed while $n \to \infty$. There is a relationship between this work and Theorem 2 in Ref. [14] where the convergence of the normalised number of copies of a given graph in $G(n, p)$ is considered. To our knowledge the statistic in question (the weighted sum of different path lengths) has not been studied and Theorem 2 in Ref. [14] does not apply to it.

5.4.2.1 Computing the mean and variance to leading order

We first define some terms, many of which will simply be higher path length versions of previously defined quantities. We are still working in the same framework, where we have a set of nodes V containing two blocks of nodes V_1 and V_2 such that $V = V_1 \cup V_2$. Further, $|V| \bmod 2 = 0$ and $|V_1| = |V_2|$.

For a set v let the function $\mathfrak{S}(v)$ to be the set of tuples of all permutations of v e.g.:

$$\mathfrak{S}(\{1,2,3\}) = \{(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)\} \tag{5.234}$$

Using this function we can then define the index set for general l as follows:

$$I_1 = \{(i,j) : i \in V_1,\ j \in V_2\} \tag{5.235}$$

$$I_a = \{(i, w_1, \ldots, w_{a-1}, j) : i \in V_1,\ j \in V_2,\ v \subset V \setminus \{i,j\},\ w \in \mathfrak{S}(v)\} \text{ for } a \geq 2 \tag{5.236}$$

$$I^{(l)} = \bigcup_{i=1}^{l} I_a \tag{5.237}$$

For $\alpha = (\alpha_1, ..., \alpha_{|\alpha|}) \in I$ we also define a random variable for a weighted path as follows:

$$X_\alpha = \lambda^{|\alpha|-1} \prod_{i=1}^{|\alpha|-1} \mathfrak{A}_{\alpha_i \alpha_{i+1}} \tag{5.238}$$

where $\mathfrak{A}_{ij}$ is an indicator variable for the existence of an edge between i and j. Notice that the definition of the index sets and the random variable X_α are consistent between the $l = 2$ case and the general case, see equations (5.16) (5.17) (5.18) and (5.20).

In order to compute the variance and the bound between the weighted sum, we perform comparisons with paths of different lengths. To make this notationally convenient we introduce the following variables:

We let ζ_r be:

$$\zeta_r = \sum_{i \in I_r} X_i \tag{5.239}$$

summing over all paths between V_1 and V_2 of length r.

Further, we let $\zeta^{(l)}$ be:

$$\zeta^{(l)} = \sum_{r=1}^{l} \zeta_r = \sum_{i \in I^{(l)}} X_i, \tag{5.240}$$

summing over all paths between V_1 and V_2 of length less than or equal to l.

There is an analogous relation to $Q_l^{Path-Diff}$ and $\zeta^{(l)}$, as between $Q_2^{Path-Diff}$ and ζ. In that $\zeta^{(l)}$ is equivalent to the first term in (5.15) for general l and thus approximating the distribution of $\zeta^{(l)}$ can help us to understand $Q_l^{Path-Diff}$.

We now find some basic properties of ζ_r and $\zeta^{(l)}$:

Lemma 19. The expectation of ζ_r is:

$$E[\zeta_r] = (\lambda p)^r \frac{n^2}{4} \frac{(n-2)!}{(n-1-r)!} \tag{5.241}$$

and the expectation of $\zeta^{(l)}$ is given by:

$$E[\zeta^{(l)}] = \sum_{r=1}^{l} (\lambda p)^r \frac{n^2}{4} \frac{(n-2)!}{(n-1-r)!}. \tag{5.242}$$

Proof The expectation of ζ_l can be computed by using the independence of $\mathfrak{A}_{ij}$, and the linearity of expectations as follows:

$$E[\zeta_r] = \lambda^r \sum_{i \in V_1} \sum_{j \in V_2} \sum_{\substack{v \subseteq V \setminus \{i,j\} \\ |v| = r-1}} \sum_{w = (w_1, \ldots, w_{r-2}) \in \mathfrak{S}(v)} E[\mathfrak{A}_{iw_1} \mathfrak{A}_{jw_{r-1}} \prod_{t}^{r-2} \mathfrak{A}_{w_t, w_{t+1}}]. \tag{5.243}$$

As the graph is a $G(n,p)$ we can rewrite this as

$$E[\zeta_r] = \lambda^r |V_1||V_2| \sum_{\substack{v \subseteq V \setminus \{i,j\} \\ |v|=r-1}} \sum_{\tilde{v} \in \mathfrak{S}(v)} p^r = \frac{n^2}{4} \frac{(n-2)!}{(n-1-r)!} (\lambda p)^r. \tag{5.244}$$

Moreover:

$$E[\zeta^{(l)}] = \sum_{r=1}^{l} E[\zeta_r] = \sum_{r=1}^{l} \frac{n^2}{4} \frac{(n-2)!}{(n-1-r)!} (\lambda p)^r. \tag{5.245}$$

Before we state the major result of this section we must discuss one additional concept - asymptotic notation. For brevity we will only discuss the notation that we use, namely Θ. A guide to the remaining notation can be found in Ref. [37].

Following the notation from Ref. [37], Θ is formally defined as follows: for a function $f(n)$:

$$\begin{aligned} \Theta(f(n)) = \{g(n) : &\forall g(n) \text{ where } \exists c_1, c_2 > 0 \text{ such that for} \\ &\forall n \geq n_0 \quad c_1 f(n) \leq g(n) \leq c_2 f(n)\} \end{aligned} \tag{5.246}$$

where n_0 is a finite number. In words, $\Theta(g(n))$ is a set of functions which are bounded from above and below by a constant multiple of $g(n)$ for all n greater than a threshold. Therefore a function that is in $\Theta(1)$ must have a finite upper bound and a finite lower bound for inputs larger than a given constant. More interestingly a function that is in $\Theta(n)$ must by definition be bounded above and below by a linear function and therefore cannot go to infinity faster than linearly. Formally, we would write $f(n) \in \Theta(g(n))$, however in this document following the notation of Ref. [37] we use the expression $f(n) = \Theta(g(n))$.

We can now state the major result of this section; to prove it we need a few intermediate results.

Theorem 20. *Let $l \in \mathbb{N}$ and $2 < l < \frac{n}{2}$. Let $g' = (V, E)$ be a sparse Erdős Renyi*

graph, where $p = \frac{c}{n} < 1$ *for* c *fixed with* $n = |V|$ *and* n *even. Further, let* $0 < \nu < 1$ *be such that* $\nu n > 2l$. *Let* $V = V_1 \cup V_2$ *where the union is disjoint and* $|V_1| = |V_2| = \frac{n}{2}$. *Let* ζ *be as given in* (5.240). *Then for* $\lambda > 0$ $Var(\zeta^{(l)}) = \Theta(n)$ *as* $n \to \infty$, *for every fixed* l. *Further,*

$$Var(\zeta^{(l)}) \geq \frac{n}{8} l(l-1)(1-p)\frac{(c\lambda)^{2l}}{c}(1-\nu)^{2l-2}. \tag{5.247}$$

Recall Eq. (5.19), where for $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{|\alpha|}) \in I_{|\alpha|-1}$ we let $\eta(\alpha) = \{\{\alpha_i, \alpha_{i+1}\}, 0 \leq i \leq |\alpha| - 1\}$. Then X_α is independent of $\{X_\beta : \beta \in I^{(l)}, \eta(\alpha) \cap \eta(\beta) = \emptyset\}$.

To prove Theorem 20, first we assess the number of paths which share a given number of edges (Lemma 21), and then we derive an expression for the covariance between the weighted sum of paths of the same length i.e. ζ_r (Lemma 22). Proposition 23 assesses the lower bound of this covariance. The proof of Theorem 20 then combines these results.

The following Lemma is related to the counting of edge-disjoint paths. Edge-disjoint paths occur when finding non-intersecting paths through graphs and they are also of interest in computer science.

Lemma 21. Fix $k, m, r \in \mathbb{N}$ so that $k \leq m \leq r \leq l$ and $l < \frac{n}{2}$. Let $\alpha \in I_r$.

(a) If $k = m$, then

(i) For $b(\alpha, m)$ which does not depend on n:

$$|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = m\}| = b \tag{C1}$$

(ii) There exists α and m such that $b(\alpha, m) = 0$, and there exists α and m such that $b(\alpha, m) = |\eta(\alpha)|$.

(b) For $k < m$, $\nu \in (\frac{2l}{n}, 1)$ then:

$$|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}| \geq \frac{1}{2}(r-k+1)(m-k)(n(1-\nu))^{m-k}. \quad (C2)$$

(c) Furthermore, we have $|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}| \leq \binom{r}{k}\frac{m!n^{n-k}}{(m-k)!}$, thus for $k < m$:

$$|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}| = \Theta(n^{m-k}) \quad (C3)$$

Proof

(a) Let us first address $(C1)$, in which $k = m$. For $\beta \in I_m$ with $|\eta(\alpha) \cap \eta(\beta)| = m$ it follows that $\eta(\beta) \subseteq \eta(\alpha)$. As α is fixed and represents a simple path, β must be a contiguous subsequence of α as every node can only appear once in α and once in β. It is possible that the sequence is reversed in α, and therefore β must be a contiguous subsequence of α where the first and last node belong to different groups i.e. the first node in V_1 and the second in V_2 or the reverse. Therefore, for fixed $\alpha \in I_r$

$$|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = m\}| = \quad (5.248)$$
$$|\{i : 1 \leq i \leq r-m+1 \text{ and } (\alpha_i, \alpha_{i+m}) \in V_1 \times V_2 \cup V_2 \times V_1\}|$$

which for fixed α does not depend on n but does depend on m and α, thus proving that $|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = m\}|$ is a function of α and m only. We will provide a case where $b = 0$ and one where $b(\alpha, m) = |\eta(\alpha)|$. Let us consider the following α:

$$\alpha = (x_1, y_1, x_2, y_2, ..., x_z, y_z) \quad (5.249)$$

where $x_j \in V_1$ and $y_j \in V_2$. If we let $m = 2$, then $(\alpha_i, \alpha_{i+m}) = (\alpha_i, \alpha_{i+2}) \in V_1 \times V_1$ or $V_2 \times V_2$, therefore $|\{\beta \in I_2 : |\eta(\alpha) \cap \eta(\beta)| = 2\}| = 0$. Conversely, if

$m = 1$, clearly we can select any edge from α, therefore, $|\{\beta \in I_1 : |\eta(\alpha) \cap \eta(\beta)| = 1\}| = |\eta(\alpha)|$. Hence, $b \geq 0$.

(b) We will now address (*C2*). To construct a lower bound of the quantity of interest we split this set into disjoint subsets and sum over them as follows:

$$|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}| = \sum_{c=1}^{k} |\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k, C(\eta(\alpha) \cap \eta(\beta)) = c\}|, \tag{5.250}$$

where $C(x)$ is the number of contiguous subsequences that are generated from the edge set x. Note, for fixed $x = \eta(\alpha) \cap \eta(\beta)$, $C(\eta(\alpha) \cap \eta(\beta))$ is unique as the paths α and β are simple, and therefore every edge can only appear once. Clearly,

$$|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}| \geq |\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k,\ C(\eta(\alpha) \cap \eta(\beta)) = 1\}|. \tag{5.251}$$

For the remainder of this part of the proof we will focus on bounding the right-hand side of (5.251), which then gives us a lower bound for the left-hand side. Thus we will now focus on the right-hand side of (5.251), i.e.:

$$|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k,\ C(\eta(\alpha) \cap \eta(\beta)) = 1\}|. \tag{5.252}$$

This is equivalent to the number of ways in which we can construct β such that it contains one contiguous subsequence of length k from α. A contiguous subsequence of length k is possible as $k < m \leq r$ and the restrictions on the first and last node of the path do not prohibit this.

A path of length k has $k + 1$ vertices. Let us call these vertices $d_1, \ldots, d_{k+1}$. Let us first count the number of different ways we can select the path $\mathbf{d} =$

$(d_1, \ldots, d_{k+1})$ from $\alpha = (\alpha_1, ..., \alpha_{r+1})$, so that

$$\alpha = (\alpha_1, \ldots, \alpha_i, d_1, \ldots, d_{k+1}, \alpha_{i+k+2}, \ldots, \alpha_{r+1}). \tag{5.253}$$

Now the first starting point for the path $\mathbf{d}$ in α is 1 and the last starting point is $r - k + 1$, and hence there are

$$r - k + 1 \tag{5.254}$$

ways of selecting $\mathbf{d}$ from α.

Let us now consider the number of ways in which we could place this $\mathbf{d}$ into β. Importantly, we must consider the orientation of the contiguous set of edges as η does not preserve ordering i.e. $\eta((x_1, x_2, x_3)) = \eta((x_3, x_2, x_1))$, ($\eta$ is only injective over paths between V_1 and V_2, see Lemma 3). Therefore we consider each orientation of $\mathbf{d}$ separately. Finally, we must also consider the restrictions on the first and last node of β, i.e. that $\beta_1 \in V_1$ and $\beta_{|\beta|} \in V_2$. We first consider $\mathbf{d}$ in the same orientation as in α i.e. $\beta = (\beta_1, \ldots, \beta_i, d_1, \ldots, d_{k+1}, \beta_{i+k+2}, \ldots, \beta_{m+1})$. The following restrictions apply:

- If $d_1 \in V_2$ then we cannot place the overlapping subsequence at the beginning of β.
- If $d_{k+1} \in V_1$ then we cannot place the overlapping sequence at the end of β.

Therefore, the first starting point for $\mathbf{d}$ is given by $1 + \mathbb{1}(d_1 \in V_2)$ and the last possible starting point is $m - k + \mathbb{1}(d_{k+1} \in V_2)$. For $\mathbf{d}$ in the opposite orientation as in α, we have $\beta = (\beta_1, \ldots, \beta_g, d_{k+1}, \ldots, d_1, \beta_{g+k+2}, \ldots, \beta_{m+1})$. Using restrictions that are similar to the previous case, the first starting point for $\mathbf{d}$ is given by $1 + \mathbb{1}(d_{k+1} \in V_2)$, and the last possible starting point is

$m-k+\mathbb{1}(d_1 \in V_2)$. Summing these terms we obtain the number of ways of placing $\mathbf{d}$ in α to be:

$$\Big((m-k+\mathbb{1}(d_{k+1} \in V_2)) - (1+\mathbb{1}(d_1 \in V_2)) + 1\Big) + \Big((m-k+\mathbb{1}(d_1 \in V_2)) - (1+\mathbb{1}(d_{k+1} \in V_2)) + 1\Big) = 2(m-k). \tag{5.255}$$

To complete our bound on $|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k,\ C(\eta(\alpha) \cap \eta(\beta)) = 1\}|$, we need to select the $m-k$ remaining nodes in β.

To select them we impose an ordering on the selection. We arbitrarily decide to fill the nodes in the following order, $\beta_1,\ \beta_{|\beta|}, \beta_2, \beta_3, ..., \beta_{|\beta|-1}$, where we skip nodes which have already been fixed by the placement of $\mathbf{d}$. There are three restrictions on this choice:

1. We must restrict β_1 and $\beta_{|\beta|}$ such that $\beta_1 \in V_1$ and $\beta_{|\beta|} \in V_2$.
2. We must guarantee that no additional edges from α are introduced into β by the selection of the remaining nodes.
3. The simplicity of the path must be preserved, i.e. no node can appear twice in β.

The nodes on the path with the largest amount of *a priori* restrictions are β_1 and $\beta_{|\beta|}$, as all three conditions apply to them. Depending on the placement of $\mathbf{d}$ into β, we could have to fill one or two of β_1 and $\beta_{|\beta|}$.

The second restriction can be addressed by simply noting that if we exclude all nodes in α from the set of possible nodes at each location, then we also exclude the possibility of an additional edge. We can restrict the set of possible nodes as such because we are interested in a lower bound. This leaves $n-r-1$ nodes from which to select.

The final restriction can be satisfied by simply using a falling factorial, to encode that no node is repeated.

First consider if the case that the overlapping segment does not cover either β_1 or $\beta_{|\beta|}$ then we can bound the number of ways to select the number of choices for β from below by

$$\frac{(n-r-1)!}{(n-r-m+k-1)!} \geq (n-r-m+k)^{m-k}. \tag{5.256}$$

Now consider the case where we have to fill one of $\beta_1 \in V_1$ and $\beta_{|\beta|} \in V_2$. We can give a lower bound for the choice for this position as follows. The number of choices is then bounded from below by $|V_1 \setminus \alpha| = |V_1| - |V_1 \cap (\alpha_1, ..., \alpha_{r+1})| = |V_1| - |V_1 \cap (\alpha_1, ..., \alpha_r)| \leq |V_1| - r = |V_2| - r \geq |V_2 \setminus \alpha|$. This leaves us with $n-r-2$ possibilities for each node other than β_1 and $\beta_{|\beta|}$, resulting in an overall contribution of:

$$\left(\frac{n}{2} - r\right) \frac{(n-r-2)!}{(n-r-m+k-1)!} \geq \left(\frac{n}{2} - r\right) (n-r-m+k)^{m-k-1}. \tag{5.257}$$

We now address the case where we need to select nodes for both β_1 and $\beta_{|\beta|}$. Note, if $m = k-1$ then it is impossible to have both β_1 and $\beta_{|\beta|}$ to fill as in this case there is only one node to select. Again we can bound the number of choices from below by $|V_1| - r = |V_2| - r$. Note that the choice of β_1 does not affect the choice of $\beta_{|\beta|}$ as they are selecting from different sets. Therefore, we are left a number of choices of:

$$\left(\frac{n}{2} - r\right)^2 \frac{(n-r-3)!}{(n-r-m+k-1)!} \geq \left(\frac{n}{2} - r\right)^2 (n-r-m+k)^{m-k-2}, \tag{5.258}$$

where we have used the fact that α has $r+1$ nodes, β has $m+1$ nodes, and a contiguous overlap of size k fixes $k+1$ nodes.

Note, by assumption in this Lemma $r \leq l < \frac{n}{2}$, thus this term is well defined. Further as $m \leq r \leq l < \frac{n}{2}$, we have $n - r - m + k > n - r - \frac{n}{2} + k > \frac{n}{2} - r$, thus the greater the number of edge nodes (β_1 and $\beta_{|\beta|}$) that need to be filled, the smaller the number of options.

Thus, as we are interested in a lower bound, we focus on the cases where there is at least one of β_1 and $\beta_{|\beta|}$ to be filled. We note that for any $\nu \in [0, 1]$ such that $2l < \nu n$, we have $r \leq l < \frac{\nu n}{2} < \frac{n}{2}$ and $r + m - k < r + m \leq 2l < \nu n$. Therefore:

Case: One node to be filled.

$$\left(\frac{n}{2} - r\right)(n - r - m + k)^{m-k-1} \geq \left(\frac{n}{2} - \frac{\nu n}{2}\right)(n - \nu n)^{m-k-1} \geq \frac{(n(1-\nu))^{m-k}}{4}. \tag{5.259}$$

Case: Two nodes to be filled. ($m \neq k - 1$)

$$\left(\frac{n}{2} - r\right)^2 (n - r - m + k)^{m-k-2} \geq \left(\frac{n}{2} - \frac{\nu n}{2}\right)^2 (n - \nu n)^{m-k-2} = \frac{(n(1-\nu))^{m-k}}{4}. \tag{5.260}$$

Thus we bound both cases from below by $\frac{(n(1-\nu))^{m-k}}{4}$.

Combining the contribution from selecting the overlapping section from α from (5.254), placing the overlap in β from (5.255) and selecting the remaining edges from (5.259) and (5.260), we obtain:

$$\begin{aligned} |\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}| \geq & |\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k, C(\eta(\alpha) \cap \eta(\beta)) = 1\}| \\ \geq & 2(r - k + 1)(m - k)\frac{(n(1-\nu))^{m-k}}{4} \qquad (5.261) \\ = & \frac{1}{2}(r - k + 1)(m - k)(n(1-\nu))^{m-k}, \qquad (5.262) \end{aligned}$$

which is the lower bound in the Lemma, thus proving (*C2*) from the Lemma.

(c) Finally, we have to demonstrate that we can bound $|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}|$ from above by a polynomial in n of order $m-k$. In combination with the lower bound this will prove $(C3)$ of the Lemma i.e. that $|\{\beta \in I_m : |\eta(\alpha) \cap \eta(\beta)| = k\}| = \Theta(n^{m-k})$

To do so we will again split this calculation into three components, selecting the overlapping edges from α ($\mathfrak{B}_1$), their position in β ($\mathfrak{B}_2$) and the remaining nodes in β ($\mathfrak{B}_3$). The number of ways to select k edges from α is:

$$\mathfrak{B}_1 = \binom{r}{k}. \tag{5.263}$$

Further, we can find an upper bound of the number of ways to place those edges in β as follows. We first fix the order of the overlapping edges in β; there are at most $k!$ ways of doing this such that $\beta \in I_m$. We then select the locations for this ordered list, resulting in a contribution of at most:

$$\mathfrak{B}_2 = k!\binom{m}{k} = \frac{m!}{(m-k)!}. \tag{5.264}$$

We can find an upper bound for the number of ways of placing the remaining nodes in β as follows. The number of nodes fixed by placing k edges into β is a function of the number of contiguous subsequences formed by the overlapping edges. This can be seen as follows: every simple path with s edges requires $s+1$ vertices. If the number of edges in each of the contiguous subsequences is given by $s_1, s_2, ...s_{C(\eta(\alpha)\cap\eta(\beta))}$, then:

$$\mathfrak{B}_3 = \sum_{i=1}^{C(\eta(\alpha)\cap\eta(\beta))} (s_i + 1) = k + C(\eta(\alpha) \cap \eta(\beta)), \tag{5.265}$$

where $C(x)$ as above is the number of contiguous subsequences formed by edge set x. Therefore, there are $m + 1 - k - C(\eta(\alpha) \cap \eta(\beta))$ unfilled nodes in β.

Clearly, $m+1-k-C(\eta(\alpha)\cap\eta(\beta)) \leq m-k$. Therefore we can bound the selection of the remaining nodes from above by n^{m-k}. Combining this result with the number of ways to select k edges from α ($\mathfrak{B}_1$) from (5.263) and the number of ways of placing k edges into β ($\mathfrak{B}_2$) from (5.264) we obtain:

$$|\{\beta \in I_m : |\eta(\alpha)\cap\eta(\beta)| = k\}| \leq \mathfrak{B}_1\mathfrak{B}_2\mathfrak{B}_3 = \binom{r}{k}\frac{(m!)n^{m-k}}{(m-k)!}. \tag{5.266}$$

demonstrating an upper bound. Combining this with a lower bound constructed before in (5.262), we simplify the expression using ν, and we obtain that for all $\nu \in (\frac{2l}{n}, 1)$:

$$(r-k+1)(m-k)(n(1-\nu))^{m-k} \leq |\{\beta \in I_m : |\eta(\alpha)\cap\eta(\beta)| = k\}| \leq \binom{r}{k}\frac{(m!)n^{m-k}}{(m-k)!}. \tag{5.267}$$

For $k < m$ clearly the upper and lower bound has a positive constant. Thus we have demonstrated that $|\{\beta \in I_m : |\eta(\alpha)\cap\eta(\beta)| = k\}| = \Theta(n^{m-k})$ as it bounded above and below by a polynomial of order n^{m-k}.

Thus we have the final result in the Lemma, QED.

Lemma 22. For $0 < r \leq m \leq l < \frac{n}{2}$ and ζ_r, ζ_m as defined in (5.239):

$$\mathrm{Cov}(\zeta_r, \zeta_m) = \lambda^{r+m}\sum_{\alpha\in I_r}\sum_{k=1}^{m} p^{m+r-k}(1-p^k)\sum_{\beta\in I_m}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = k). \tag{5.268}$$

If $\lambda \geq 0$, then $\mathrm{Cov}(\zeta_r, \zeta_m) \geq 0$ and if $\lambda > 0$ and $0 < p < 1$, then $\mathrm{Cov}(\zeta_r, \zeta_m) > 0$.

Proof Recall that $\mathrm{Cov}(\zeta_r, \zeta_m) = E[(\zeta_r - E[\zeta_r])(\zeta_m - E[\zeta_m])]$. As $m \leq r$ we can rewrite this as:

$$\mathrm{Cov}(\zeta_r, \zeta_m) = \lambda^{r+m}\sum_{\alpha\in I_r}\sum_{k=0}^{m}\sum_{\beta\in I_m}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = k)E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])]. \tag{5.269}$$

By construction, if $\eta(\alpha) \cap \eta(\beta) = \emptyset$, then X_α and X_β are independent, and $E[(X_\alpha - E[X_\alpha])(X_\beta - E[X_\beta])] = 0$.

Further, for $\alpha = (\alpha_1, ..., \alpha_{|\alpha|}) \in I_{|\alpha|-1}$,

$$E[X_\alpha] = \prod_{i=1}^{|\alpha|-1} E[\mathfrak{A}_{\alpha_i \alpha_{i+1}}] = (\lambda p)^{|\alpha|-1} = (\lambda p)^r. \tag{5.270}$$

Finally, by direct calculation:

$$E[(X_\alpha - (\lambda p)^r)(X_\beta - (\lambda p)^m)] = E[X_\alpha X_\beta] - (\lambda p)^{m+r} = \lambda^{m+r}(p^{|\eta\alpha \cup \eta(\beta)|} - p^{m+r}) \tag{5.271}$$

$$= \lambda^{m+r} p^{|\eta(\alpha) \cup \eta(\beta)|}(1 - p^{|\eta(\alpha) \cap \eta(\beta)|}). \tag{5.272}$$

Putting this all together we obtain:

$$\mathrm{Cov}(\zeta_r, \zeta_m) = \lambda^{r+m} \sum_{\alpha \in I_r} \sum_{k=1}^{m} p^{m+r-k}(1 - p^k) \sum_{\beta \in I_m} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k). \tag{5.273}$$

It remains for us to show that $\mathrm{Cov}(\zeta_r, \zeta_m) \geq 0$, with strict inequality if $\lambda > 0$ and $0 < p < 1$. We know that for $0 \leq p \leq 1$, $p^{m+r-k}(1 - p^k) \geq 0$ and for $0 < p < 1$, $p^{m+r-k}(1 - p^k) > 0$. Therefore each term in the summation is greater than or equal to zero and the sum is greater than or equal to zero.

To show strict inequality we must demonstrate that there is at least one indicator of the form $\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k)$ that is non-zero. To show this we consider two cases. First, if $r = m$ then consider $\alpha = \beta$ therefore $|\eta(\alpha) \cap \eta(\beta)| = m$, therefore setting, $k = m$ produces the desired non-zero indicator.

If $r < m$, for arbitrary $\alpha \in I_r$ let $\beta = (\alpha_1, ..., \alpha_{|\alpha|}, \beta_{|\alpha|+1}, ..., \beta_{|\beta|})$. As $r, m \leq l < \frac{n}{2}$ there are sufficient nodes to fill β, however we must also show that β is still a simple path and $\beta_{|\beta|} \in V_2$. As by assumption $r \leq l < \frac{n}{2}$ therefore $\frac{n}{2} - r \geq \frac{n}{2} - l$. clearly as $\frac{n}{2} > l$, $\frac{n}{2} - l > 0$, as n is even and l is an integer therefore $\frac{n}{2} - l \geq 1$. By definition

$\alpha_1 \in V_1$ therefore there is a maximum of r nodes in α in V_2, as $|V_2 \setminus \alpha| \geq \frac{n}{2} - r \geq \frac{n}{2} - l \geq 1$, therefore we can construct β. Thus, $|\eta(\alpha) \cap \eta(\beta)| = r$, thus the desired non-zero factor.

The assertion follows. QED.

Proposition 23. For $l \geq r \geq m > 2$ and $l < \frac{n}{2}$ let $\nu \in (\frac{2l}{n}, 1)$. Assume that $\lambda > 0$ and $p = \frac{c}{n}$ such that $0 < p < 1$ and with $c > 0$. The covariance between ζ_r and ζ_m for fixed r and m is $\Theta(n)$. Further,

$$\mathrm{Cov}(\zeta_r, \zeta_m) \geq \frac{1}{2} \frac{r(m-1)|I_r|}{cn^r} (c\lambda)^{r+m} (1-\nu)^{m-1} (1-p). \qquad (P1)$$

Remark: In particular, the case $r = m$ in Proposition 23 gives that $Var(\zeta_r) = \Theta(n)$.

Proof By Lemma 22,

$$\mathrm{Cov}(\zeta_r, \zeta_m) = \lambda^{r+m} \sum_{\alpha \in I_l} \sum_{k=1}^{m} p^{m+r-k} (1-p^k) \sum_{\beta \in I_m} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k). \qquad (5.274)$$

We know from Lemma 21 that for $k < m$, $\sum_{\beta \in I_m} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k)$ is $\Theta(n^{m-k})$. Thus there are constants $d(r,k,m,\nu)$ and $D(r,k,m,\nu)$ such that $d(r,k,m,\nu)n^{m-k} \leq \sum_{\beta \in I_m} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k) \leq D(r,k,m,\nu)n^{m-k}$. To show $Cov(\zeta_r, \zeta_m)$ is $\Theta(n)$ we need to demonstrate that $Cov(\zeta_r, \zeta_m) \geq a_1 n$ and $Cov(\zeta_r, \zeta_m) \leq a_2 n$ for some a_1 and a_2. We consider the first inequality in detail, the second can be treated similarly.

Using this the inequality from Lemma 21 and (5.274) we obtain:

$$\begin{aligned}\mathrm{Cov}(\zeta_r,\zeta_m) \geq & \lambda^{r+m}\sum_{\alpha\in I_r}\Bigg(\Big(\sum_{k=1}^{m-1}p^{m+r-k}(1-p^k)d(r,m,k,\nu)n^{m-k}\Big) \qquad (5.275)\\ & + p^r(1-p^m)\sum_{\beta\in I_m}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)|=m)\Bigg)\\ = & \lambda^{r+m}\Bigg(|I_r|\Big(\sum_{k=1}^{m-1}p^{m+r-k}(1-p^k)d(r,m,k,\nu)n^{m-k}\Big) \qquad (5.276)\\ & + p^r(1-p^m)\sum_{\alpha\in I_r}\sum_{\beta\in I_m}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)|=m)\Bigg)\end{aligned}$$

Let us consider the summation over $\alpha \in I_r$, $\beta \in I_m$. We know from Lemma 21 that $\sum_{\beta\in I_m}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)|=m) = b(\alpha,m) \geq 0$, for any α, here we need a stronger result. To compute this we reverse the sum and following Lemma 21 we use a constructive argument and consider:

$$p^r(1-p^m)\sum_{\beta\in I_m}|\{\alpha\in I_r : \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)|=m)\}|. \qquad (5.277)$$

The expression $|\{\alpha \in I_r : \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)|=m)\}|$ can be seen as counting the number of ways of constructing an α such that the contiguous path β is contained in it. We know that the path must be contiguous as every edge of β is part of α and as α is simple every node only appears once and thus all of the edges of β must be adjacent to the next edge in the sequence. Following the argument in Lemma 21, α could be of the form $(\alpha_1,...,\alpha_i,\beta_1,...,\beta_{m+1},\alpha_{i+m+2},...,\alpha_{r+1})$ or could be of the form $(\alpha_1,...,\alpha_j,\beta_{m+1},...,\beta_1,\alpha_{j+m+2},...,\alpha_{r+1})$. We know that $(\beta_1,\beta_{m+1}) \in V_1\times V_2$, therefore following Lemma 21, for the first form the smallest starting point of β is 1, and the largest is $r-m+1$. Similarly, for the second form the smallest starting point is 2 (as $\beta_{m+1} \in V_2$), and the largest is $r-m$. However, in this case we cannot simply sum the contributions, as it might not be possible to place the sequence in reverse order. The number of possible start locations for the second form is $r-m-1$. However,

clearly for $m = r$ this result fails to hold, as the smallest starting point goes above the supposed largest. A simple solution to this issue is to use the function:

$$\omega(r, m) = r - m + 1 + \max(r - m - 1, 0). \tag{5.278}$$

Further, as by assumption $r \geq m$, the first term cannot suffer from this problem. Also, $\omega(r, m) > 0$. Therefore, the number of ways to place β in α is $\omega(r, m)$, which for $r > m$ is equivalent to $2(m - k)$. This echoes the similar result from Lemma 21.

As we know that the path must be contiguous, $\omega(r, m)$ is an exact solution for the number of ways of placing β into α. Thus we can complete the calculation by bounding from above and below the selection of the remaining nodes.

The number of ways to pick the remaining nodes can bounded from above by n^{r-m}, as there are at most n choices for each remaining node. For the lower bound we use the argument from Lemma 21 seen in (5.257), which we do not repeat here as they are identical. Putting this all together, and using the lower bound from (5.259), we obtain:

$$\sum_{\beta \in I_m} \omega(r, m) \frac{(n(1-\nu))^{r-m}}{4} \leq \sum_{\beta \in I_m} \sum_{\alpha \in I_r} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = m) \leq \sum_{\beta \in I_m} \omega(r, m) n^{r-m}, \tag{5.279}$$

and hence,

$$\frac{n^{2+r-m}(n-2)!\omega(r, m)(1-\nu)^{r-m}}{16(n-m-1)!} \leq \sum_{\beta \in I_m} \sum_{\alpha \in I_r} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = m) \tag{5.280}$$

$$\leq \frac{n^{r-m+2}(n-2)!\omega(r, m)}{4(n-m-1)!}, \tag{5.281}$$

where we have used the fact that $\sum_{\beta \in I_m} 1 = \frac{n^2}{4} \frac{(n-2)!}{(n-m-1)!}$. Clearly $\frac{n^{2+r-m}(n-2)!}{(n-m-1)!}$ is to leading order n^{r+1} in n. As $l > 2$ and $l < \frac{n}{2}$ therefore $n > 4$, thus the coefficient on the lower bound is positive. Therefore $\sum_{\beta \in I_m} \sum_{\alpha \in I_r} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = m)$ has an

upper and lower bound of order $r+1$ in n, with coefficients that are positive and are a function of m, r, ν, so there exists $b(m,r,\nu)$ and $B(m,r,\nu)$ are such that:

$$b^*(m,r,\nu)n^{r+1} \leq \sum_{\beta\in I_m}\sum_{\alpha\in I_r} \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = m) \leq B^*(m,r,\nu)n^{r+1}. \tag{5.282}$$

Therefore $\sum_{\beta\in I_m}\sum_{\alpha\in I_r}\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = m) = \Theta(n^{r+1})$. Substituting this result into (5.276), and using $p = \frac{c}{n}$, we obtain:

$$\begin{aligned}\mathrm{Cov}(\zeta_r,\zeta_m) &\geq \lambda^{r+m}\Bigg(|I_r|\Big(\sum_{k=1}^{m-1} p^{m+r-k}(1-p^k)d(r,m,k,\nu)n^{m-k}\Big) \qquad (5.283)\\ &\qquad + p^r(1-p^m)b^*(m,r,\nu)n^{r+1}\Bigg)\\ &= \lambda^{r+m}\Bigg(\frac{n^2}{4}\frac{(n-2)!}{(n-r-1)!}\Big(\sum_{k=1}^{m-1} c^{m+r-k}(1-p^k)d(r,m,k,\nu)n^{-r}\Big) \qquad (5.284)\\ &\qquad + c^r(1-p^m)b^*(m,r,\nu)n^1\Bigg)\end{aligned}$$

We can bound $|I_r|$ as follows, we note that for $n > 4$ that $n-r-1 > n-l-1 > \frac{n}{2}-1 > \frac{n}{4}$. Therefore $\frac{(n-2)!}{(n-r-1)!} \geq (n-r-1)^{r-1} \geq \left(\frac{n}{4}\right)^{r-1}$. Thus,

$$\begin{aligned}\mathrm{Cov}(\zeta_r,\zeta_m) \geq& \lambda^{r+m}\Bigg(\left(\frac{n}{4}\right)^{r-1}\frac{n^2}{4}\Big(\sum_{k=1}^{m-1} c^{m+r-k}(1-p^k)d(r,m,k,\nu)n^{-r}\Big) \qquad (5.285)\\ &\qquad + c^r(1-p^m)b^*(m,r,\nu)n^1\Bigg)\\ \geq& n\lambda^{r+m}\Bigg(\frac{1}{4^r}\Big(\sum_{k=1}^{m-1} c^{m+r-k}(1-p^k)d(r,m,k,\nu)\Big) \qquad (5.286)\\ &\qquad + c^r(1-p^m)b^*(m,r,\nu)\Bigg)\end{aligned}$$

where $c > 0$ and for $p < 1$ therefore terms in the summation are positive. Therefore $Cov(\zeta_r,\zeta_m)$ is bounded from below by a polynomial of leading order n multiplied by a constant which is positive as by assumption $0 < p < 1$, we will write as $\tilde{f}(r,m,\nu)$. Note that there is a dependence on n in p ($p = \frac{c}{n}$), so substituting and expanding

would lead to additional terms with smaller powers of n.

Similarly, we can show that $Cov(\zeta_r, \zeta_m) \leq \tilde{F}(r, m, \nu)$, and thus $Cov(\zeta_r, \zeta_m) = \Theta(n)$. Thus proving the leading order section of this Proposition.

The bound of this Proposition $(P1)$ is achieved as follows. Recall from Lemma 22:

$$\mathrm{Cov}(\zeta_r, \zeta_m) = \lambda^{r+m} \sum_{\alpha \in I_l} \sum_{k=1}^{m} p^{m+r-k}(1-p^k) \sum_{\beta \in I_m} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k). \tag{5.287}$$

Note that from Lemma 21 for $k < m$ and ν as described in the Lemma we have:

$$\sum_{\beta \in I_m} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k) \geq \frac{1}{2}(r-k+1)(m-k)(1-\nu)^{m-k} n^{m-k}. \tag{5.288}$$

By simply excluding the case where $m = k$, which is non-negative as $p^{m+r-k}(1-p^k)$ is non-negative, and by Lemma 21 $\sum_{\beta \in I_m} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = k)$ is also non zero. Thus, we obtain:

$$\begin{aligned}\mathrm{Cov}(\zeta_r, \zeta_m) \geq \frac{1}{2}\lambda^{r+m}|I_r| \sum_{k=1}^{m-1} p^{m+r-k}(1-p^k)(r-k+1)(m-k) \\ \times (1-\nu)^{m-k} n^{m-k}.\end{aligned} \tag{5.289}$$

Therefore as $1 - p^k > 1 - p$ for $0 \leq p \leq 1$ and $k \geq 1$:

$$\mathrm{Cov}(\zeta_r, \zeta_m) \geq \frac{1}{2}\lambda^{r+m}|I_r| \sum_{k=1}^{m-1} p^{m+r-k}(1-p^k)(r-k+1)(m-k)n^{m-k}(1-\nu)^{m-k} \tag{5.290}$$

$$\geq \frac{1}{2}\lambda^{r+m}(1-\nu)^m(1-p)c^{m+r}|I_r|n^{-r} \sum_{k=1}^{m-1} \frac{(r-k+1)(m-k)}{(c(1-\nu))^k} \tag{5.291}$$

$$\geq \frac{1}{2}\lambda^{r+m}(1-\nu)^m(1-p)c^{m+r}|I_r|n^{-r} \frac{(r)(m-1)}{(c(1-\nu))} \tag{5.292}$$

$$= \frac{1}{2}\frac{r(m-1)|I_r|}{cn^r}(c\lambda)^{r+m}(1-\nu)^{m-1}(1-p). \tag{5.293}$$

Note, we bound the summation by observing that all terms in the summation are

positive and bounding from below by the first term. The final expression is the assertion $(P1)$. QED.

Now we have the ingredients to prove Theorem 20.

Proof of Theorem 20 To compute the variance of $\zeta^{(l)} = \sum_{r=1}^{l} \zeta_r$, we use the sum of correlated variables formula:

$$Var(\zeta^{(l)}) = \sum_{i=1}^{l} Var(\zeta_i) + 2 \sum_{1 \le i < j \le l} \mathrm{Cov}(\zeta_i, \zeta_j). \tag{5.294}$$

We know from Proposition 23 that the covariance $Cov(\zeta_i, \zeta_j)$ is $\Theta(n)$. Thus, $Var(\zeta_i) = Cov(\zeta_i, \zeta_i) = \Theta(n)$. As l is fixed and by definition $Var(\zeta^{(l)}) = \sum_{i=1}^{l} Var(\zeta_i) + 2\sum_{i<j} Cov(\zeta_i, \zeta_j)$, then $Var(\zeta^{(l)})$ is the sum of $l + \binom{l}{2}$ terms which are $\Theta(n)$ and positive (Lemma 22). Therefore it follows that $Var(\zeta^{(l)})$ is $\Theta(n)$.

This is as stated in the Theorem 20, thus showing the first claim. The proof of the lower bound (5.247) is as follows. From (5.294) and as $\mathrm{Var}(\zeta_i) \ge 0$, $\mathrm{Cov}(\zeta_i, \zeta_j) \ge 0$, by Proposition 23:

$$\mathrm{Var}(\zeta^{(l)}) \ge \mathrm{Var}(\zeta_l) = \mathrm{Cov}(\zeta_l, \zeta_l) \ge \frac{1}{2} l(l-1) \frac{(c\lambda)^{2l}}{c} (1-\nu)^{l-1} (1-p) \frac{|I_l|}{n^l}. \tag{5.295}$$

Let us bound $|I_l|$:

$$|I_l| = \frac{n^2}{4} \frac{(n-2)!}{(n-l-1)!} > \frac{n^2}{4} (n-l)^{l-1}, \tag{5.296}$$

Let ν be as in the formulation of the Theorem for $l > 0$

$$\nu n > 2l > l \implies n - l - 1 > (1-\nu)n \implies |I_l| > \frac{n^2}{4} ((1-\nu)n)^{l-1}. \tag{5.297}$$

Putting all of this together:

$$\mathrm{Var}(\zeta^{(l)}) \ge \mathrm{Var}(\zeta_l) \ge \frac{n}{8} l(l-1)(1-p) \frac{(c\lambda)^{2l}}{c} (1-\nu)^{2l-2}, \tag{5.298}$$

which is (5.247). QED.

5.4.2.2 Normal approximation

We will now compute the bound for the distance between a standard normal distribution and our standardised weighted sum of paths for an arbitrary fixed l as n goes to infinity.

As in (5.67) on page 169, we define the random variable Y_α as follows:

$$Y_\alpha = \frac{X_\alpha - E[X_\alpha]}{\mathrm{Var}(\zeta^{(l)})}. \tag{5.299}$$

Recall, $X_\alpha = \lambda^{|\alpha|-1} \prod_{i=1}^{|\alpha|-1} \mathfrak{A}_{\alpha_i,\alpha_{i+1}}$. Again, notice that the definition of Y_α is consistent between the $l = 2$ case and the general case. We also define the sum:

$$\tilde{\zeta}^{(l)} = \sum_{\alpha \in I^{(l)}} Y_\alpha. \tag{5.300}$$

To use with the bound (2.48) we use the dependency neighbourhood:

$$\mathbb{K}_\alpha(I^{(l)}) = \{\beta \in I^{(l)} : |\eta(\alpha) \cap \eta(\beta)| > 0\}. \tag{5.301}$$

Note this formulation is consistent with the $l = 2$ case after fixing $l = 2$ from (5.27) on page 161. Again for notational convenience we drop the argument $(I^{(l)})$.

Similar to the $l = 2$ case $E[\tilde{\zeta}^{(l)}] = 0$ and $\mathrm{Var}(\tilde{\zeta}^{(l)}) = 1$. Moreover $\tilde{\zeta}^{(l)}$ has a dissociated decomposition with the dependency neighbourhood of $\alpha \in I^{(l)}$ given by $\mathbb{K}_\alpha$.

We then apply the bound from (2.48) which is derived from Theorem 2 (page 55) resulting in the following expression:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq 4 \sum_{\alpha \in I} \sum_{\beta,\gamma \in \mathbb{K}_\alpha} \left(E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|] E[|Y_\gamma|] \right). \tag{5.302}$$

Lemma 24. For $\alpha, \beta, \gamma \in I^{(l)}$,

$$E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|]E[|Y_\gamma|] \leq \frac{8}{\mathrm{Var}(\zeta^{(l)})^3}\lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|}p^{|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|}. \quad (5.303)$$

Proof Let us consider these terms from right to left:

$$E[|Y_\gamma|] = \frac{E[|X_\gamma - (\lambda p)^{|\eta(\gamma)|}|]}{\mathrm{Var}(\zeta^{(l)})^{0.5}} = 2\frac{\lambda^{|\eta(\gamma)|}}{\mathrm{Var}(\zeta^{(l)})}p^{|\eta(\gamma)|}(1 - p^{|\eta(\gamma)|}) \quad (5.304)$$

$$E[|Y_\alpha Y_\beta|] = \frac{E[|(X_\alpha - (\lambda p)^{|\eta(\alpha)|})(X_\beta - (\lambda p)^{|\eta(\beta)|})|]}{\mathrm{Var}(\zeta^{(l)})} \quad (5.305)$$

$$= \frac{E[|X_\alpha X_\beta - X_\alpha(\lambda p)^{|\eta(\beta)|} - X_\beta(\lambda p)^{|\eta(\alpha)|} + (\lambda p)^{|\eta(\alpha)|+|\eta(\beta)|}|]}{\mathrm{Var}(\zeta^{(l)})} \quad (5.306)$$

$$\leq \frac{\max(E[X_\alpha X_\beta + (\lambda p)^{|\eta(\alpha)|+|\eta(\beta)|}], E[X_\alpha(\lambda p)^{|\eta(\beta)|} + X_\beta(\lambda p)^{|\eta(\alpha)|}])}{\mathrm{Var}(\zeta^{(l)})} \quad (5.307)$$

$$= \frac{\lambda^{|\eta(\alpha)|+|\eta(\beta)|}}{\mathrm{Var}(\zeta^{(l)})}\max(p^{|\eta(\alpha)\cup\eta(\beta)|} + p^{|\eta(\alpha)|+|\eta(\beta)|}, 2p^{|\eta(\alpha)|+|\eta(\beta)|}) \quad (5.308)$$

$$= \frac{\lambda^{|\eta(\alpha)|+|\eta(\beta)|}}{\mathrm{Var}(\zeta^{(l)})}\left(p^{|\eta(\alpha)\cup\eta(\beta)|} + p^{|\eta(\alpha)|+|\eta(\beta)|}\right) < 2\frac{\lambda^{|\eta(\alpha)|+|\eta(\beta)|}}{\mathrm{Var}(\zeta^{(l)})}p^{|\eta(\alpha)\cup\eta(\beta)|}, \quad (5.309)$$

where we use the fact that for $x, y \geq 0$, $|x - y| \leq x$. Putting this together we obtain:

$$E[|Y_\alpha Y_\beta|]E[|Y_\gamma|] \leq 4\frac{\lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|}}{\mathrm{Var}(\zeta^{(l)})^{1.5}}p^{|\eta(\alpha)\cup\eta(\beta)|+|\eta(\gamma)|}(1 - p^{|\eta(\gamma)|}). \quad (5.310)$$

As $p \leq 1$ therefore $(1 - p^{|\eta(\gamma)|}) \leq 1$ and $|\eta(\alpha) \cup \eta(\beta)| + |\eta(\gamma)| \leq |\eta(\alpha) \cup \eta(\beta) \cup \eta(\gamma)|$, and so

$$E[|Y_\alpha Y_\beta|]E[|Y_\gamma|] \leq 4\frac{\lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|}}{\mathrm{Var}(\zeta^{(l)})^{1.5}}p^{|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|}. \quad (5.311)$$

Finally, we perform the same operation on the first term of the summand in the error term:

$$
\begin{aligned}
E[|Y_\alpha Y_\beta Y_\gamma|] =& \frac{1}{\mathrm{Var}(\zeta^{(l)})^{1.5}} E\Big[\Big|X_\alpha X_\beta X_\gamma - X_\alpha X_\beta (\lambda p)^{|\eta(\gamma)|} - X_\alpha X_\gamma (\lambda p)^{|\eta(\beta)|} && (5.312)\\
& - X_\beta X_\gamma (\lambda p)^{|\eta(\alpha)|} + X_\alpha (\lambda p)^{|\eta(\beta)|+|\eta(\gamma)|} + X_\beta (\lambda p)^{|\eta(\alpha)|+|\eta(\gamma)|} \\
& + (\lambda p)^{|\eta(\alpha)|+|\eta(\beta)|}(X_\gamma - (\lambda p)^{|\eta(\gamma)|})\Big|\Big] \\
\leq& \frac{1}{\mathrm{Var}(\zeta^{(l)})^{1.5}} \max\Bigg(E\Big[X_\alpha X_\beta X_\gamma + X_\alpha (\lambda p)^{|\eta(\beta)|+|\eta(\gamma)|} && (5.313)\\
& + X_\beta (\lambda p)^{|\eta(\alpha)|+|\eta(\gamma)|} + X_\gamma (\lambda p)^{|\eta(\alpha)|+|\eta(\beta)|})\Big], E\Big[X_\alpha (X_\beta (\lambda p)^{|\eta(\gamma)|} + \\
& X_\gamma (\lambda p)^{|\eta(\beta)|}) + (\lambda p)^{|\eta(\alpha)|}(X_\beta X_\gamma + (\lambda p)^{|\eta(\beta)|+|\eta(\gamma)|})\Big]\Bigg) \\
\leq& 4 \frac{\lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|}}{\mathrm{Var}(\zeta^{(l)})^{1.5}} p^{|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|}, && (5.314)
\end{aligned}
$$

where the omitted steps are very similar to the $E[|Y_\alpha Y_\beta|]$ case. Summing the two bounds leads to the assertion. QED.

Remarks:

1. Lemma 24 is related to Theorem 2 in Ref. [14], most notably Equation (3.8) from the paper, where the first term in the minimum bound is clearly similar to the bounds we derive here. Converting to notation of this document, and factoring out the constant (which differs between the bounds we are using) their bound is as follows:

$$\epsilon \leq \frac{16}{\sigma^3} \sum_{\alpha \in I} \sum_{\beta,\gamma \in K_i} E[X_\alpha X_\beta X_\gamma]. \tag{5.315}$$

   However, [14] is simply attempting to obtain the functional form of such a bound.

2. We note that we do not obtain a factor of $(1-p)$ unlike in the $l = 2$ case (see for example Lemma 15). This makes it difficult to compare the bounds between

the two cases.

3. It should be noted that the bound in Lemma 24 is not optimised, and better bounds could be constructed.

To help in computing the bound (5.302) we define the following functions. For, $(\alpha, \beta, \gamma) \in (I^{(l)})^3$ let:

$$\begin{aligned}\mathfrak{O}(\alpha,\beta,\gamma) =&(|\eta(\alpha)|, |\eta(\beta)|, |\eta(\gamma)|, |\eta(\alpha) \cap \eta(\beta)|, \\ &|\eta(\alpha) \cap \eta(\gamma)|, |\eta(\beta) \cap \eta(\gamma)|, |\eta(\alpha) \cap \eta(\beta) \cap \eta(\gamma)|).\end{aligned} \tag{5.316}$$

and,

$$\Psi_O(\alpha,\beta,\gamma) = \{(\psi_1,\psi_2,\psi_3) \in (I^{(l)})^3 : \quad \mathfrak{O}(\alpha,\beta,\gamma) = \mathfrak{O}(\psi_1,\psi_2,\psi_3)\}$$

Lemma 25. For $\Psi_O(\alpha, \beta, \gamma)$ as just defined:

$$\begin{aligned}|\Psi_O(\alpha,\beta,\gamma)| \leq & \sum_{k_g=1}^{K_g^{\max}(\alpha,\beta,\gamma)} \sum_{k_b=1}^{K_b^{\max}(\alpha,\beta,\gamma)} \frac{\mathfrak{C}(\alpha,\beta,\gamma,k_b)\mathfrak{D}(\alpha,\beta,\gamma)(k_\beta! k_\gamma!)}{4} \\ & \times \left(\frac{2}{n}\right)^{k_b+k_g} n^{3+|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|},\end{aligned} \tag{5.317}$$

where, $K_b^{\max}(\alpha,\beta,\gamma) = \min(|\eta(\alpha) \cap \eta(\beta)|, \left\lceil \frac{|\eta(\beta)|}{2} \right\rceil)$, $K_g^{\max}(\alpha,\beta,\gamma) = \min(|\eta(\gamma) \cap (\eta(\alpha) \cup \eta(\beta))|, \left\lceil \frac{|\eta(\gamma)|}{2} \right\rceil)$, and

$$\mathfrak{C}(\alpha,\beta,\gamma,k_b) = \binom{|\eta(\alpha)\cap\eta(\beta)|-1}{k_b-1}\binom{|\eta(\alpha)\setminus\eta(\beta)+1}{k_b}\binom{|\eta(\beta)\setminus\eta(\alpha)+1}{k_b} \tag{5.318}$$

$$\begin{aligned}\mathfrak{D}(\alpha,\beta,\gamma,k_g) =&\binom{|\eta(\alpha)\cap\eta(\beta)|}{|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)|}\binom{|\eta(\alpha)\setminus\eta(\beta)|}{|\eta(\alpha)\cap\eta(\gamma)\setminus\eta(\beta)|}\binom{|\eta(\beta)\setminus\eta(\alpha)|}{|\eta(\beta)\cap\eta(\gamma)\setminus\eta(\alpha)|} \\ &\times \binom{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1}{k_g}.\end{aligned} \tag{5.319}$$

Proof To compute $|\Psi_O(\alpha,\beta,\gamma)|$ we will use the following summation

$$|\Psi_O(\alpha,\beta,\gamma)| = |\{(\psi_1,\psi_2,\psi_3) \in (I^{(l)})^3 : \quad \mathfrak{O}(\alpha,\beta,\gamma) = \mathfrak{O}(\psi_1,\psi_2,\psi_3)\}| \tag{5.320}$$

$$= \sum_{\psi_1,\psi_2,\psi_3 \in (I^{(l)})^3} \mathbb{1}(\mathfrak{O}(\alpha,\beta,\gamma) = \mathfrak{O}(\psi_1,\psi_2,\psi_3)). \tag{5.321}$$

Rewriting the indicator variable in terms of its components we obtain:

$$\begin{aligned}|\Psi_O(\alpha,\beta,\gamma)| = \sum_{(\psi_1,\psi_2,\psi_3)\in(I^{(l)})^3} \Bigg(& \mathbb{1}\Big(|\eta(\alpha)| = |\eta(\psi_1)|\Big)\mathbb{1}\Big(|\eta(\beta)| = |\eta(\psi_2)|\Big) \\ & \times \mathbb{1}\Big(|\eta(\gamma)| = |\eta(\psi_3)|\Big)\mathbb{1}\Big(|\eta(\alpha)\cap\eta(\beta)| = |\eta(\psi_1)\cap\eta(\psi_2)|\Big) \\ & \times \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_3)|\Big) \\ & \times \mathbb{1}\Big(|\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_2)\cap\eta(\psi_3)|\Big) \\ & \times \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_2)\cap\eta(\psi_3)|\Big)\Bigg).\end{aligned} \tag{5.322}$$

We recall from (5.236), that I_a is the index set of all paths of length a. Consider:

$$\sum_{\psi_1\in I} \mathbb{1}(|\eta(\psi_1)| = |\eta(\alpha)|) = \sum_{\psi_1\in I\setminus I_{|\eta(\alpha)|}} \mathbb{1}(|\eta(\psi_1)| = |\eta(\alpha)|) + \sum_{\psi_1\in I_{|\eta(\alpha)|}} \mathbb{1}(|\eta(\psi_1)| = |\eta(\alpha)|) \tag{5.323}$$

$$= \sum_{\psi_1\in I_{|\eta(\alpha)|}} 1. \tag{5.324}$$

Thus, we can fulfil the first three indicator conditions by restricting the summation over $I_{|\eta(\alpha)|}$, $I_{|\eta(\beta)|}$ and $I_{|\eta(\gamma)|}$, giving

$$\begin{aligned}|\Psi_O(\alpha,\beta,\gamma)| = \sum_{\psi_1\in I_{|\eta(\alpha)|}} \sum_{\psi_2\in I_{|\eta(\beta)|}} \sum_{\psi_3\in I_{|\eta(\gamma)|}} \Bigg(& \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\beta)| = |\eta(\psi_1)\cap\eta(\psi_2)|\Big) \\ & \times \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_3)|\Big)\mathbb{1}\Big(|\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_2)\cap\eta(\psi_3)|\Big) \\ & \times \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_2)\cap\eta(\psi_3)|\Big)\Bigg).\end{aligned} \tag{5.325}$$

We know that a subset of edges from a simple path forms a unique set of contiguous

subsequences up to reversibility. Following the proof to Lemma 21, we let the function $C(x)$ be the number of contiguous subsequences formed by the edges in the set x.

Let us first focus on the summation over ψ_3. Let us condition on the number of contiguous subsequences in the edges of ψ_3 that are shared with ψ_1 and/or ψ_2. This is given by $C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3))$. We can bound $C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3))$ as follows for $(\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3) \neq \emptyset$:

$$1 \leq C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)) \leq \min\left(|(\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)|, \left\lceil \frac{|\eta(\psi_3)|}{2} \right\rceil\right) \tag{5.326}$$

$$= \min\left(|(\eta(\alpha) \cup \eta(\beta)) \cap \eta(\gamma)|, \left\lceil \frac{|\eta(\gamma)|}{2} \right\rceil\right). \tag{5.327}$$

The justifications for the bounds are as follows. For the lower bound, as there is more than one edge, there is at least one subsequence, hence the lower bound of 1. For the first upper bound: the number of subsequences can be at most the number of overlapping edges, hence the bound $|\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)| = |\eta(\alpha) \cup \eta(\beta)) \cap \eta(\gamma)|$. Finally for the second upper bound: the minimum length of a subsequence is 1, further the minimum length of a gap between subsequences is 1, therefore we obtain an upper bound of $\left\lceil \frac{|\eta(\psi_3)|}{2} \right\rceil = \left\lceil \frac{|\eta(\gamma)|}{2} \right\rceil$.

We recall from the Lemma that $K_g^{\max}(\alpha, \beta, \gamma) = \min(|(\eta(\alpha) \cup \eta(\beta)) \cap \eta(\gamma)|, \left\lceil \frac{|\eta(\gamma)|}{2} \right\rceil)$ which is equal to the upper bound on $C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3))$. Therefore we can

rewrite $|\Psi_O(\alpha,\beta,\gamma)|$ as

$$|\Psi_O(\alpha,\beta,\gamma)| = \sum_{k_g=1}^{K_g^{\max}} \sum_{\psi_1 \in I_{|\eta(\alpha)|}} \sum_{\psi_2 \in I_{|\eta(\beta)|}} \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\beta)| = |\eta(\psi_1)\cap\eta(\psi_2)|\Big) \tag{5.328}$$
$$\times \sum_{\psi_3 \in I_{|\eta(\gamma)|}} \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_3)|\Big)$$
$$\times \mathbb{1}\Big(|\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_2)\cap\eta(\psi_3)|\Big)$$
$$\times \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_2)\cap\eta(\psi_3)|)$$
$$\times \mathbb{1}\Big(C((\eta(\psi_1)\cup\eta(\psi_2))\cap\eta(\psi_3)) = k_g\Big).$$

We shall consider the summations in turn, first over ψ_3, then over ψ_2 and then over ψ_1. We will now focus on the summation over ψ_3 assuming that ψ_1 and ψ_2 and k_g are fixed. We obtain:

$$\sum_{\psi_3 \in I_{|\eta(\gamma)|}} \mathbb{1}\Big(|\eta(\alpha)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_3)|\Big)\mathbb{1}\Big(|\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_2)\cap\eta(\psi_3)|\Big) \tag{5.329}$$
$$\times \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_2)\cap\eta(\psi_3)|)\mathbb{1}\Big(C((\eta(\psi_1)\cup\eta(\psi_2))\cap\eta(\psi_3)) = k_g\Big).$$

We will bound the summation by giving a bound on the number of ψ_3 which satisfy all of the indicator variables. Following the procedure used in Lemma 21, we will derive a bound in three stages as follows. We will first bound the number of ways of selecting the overlapping edges. Second we will bound the number of ways of placing those edges into ψ_3. Finally we will bound the number of ways of placing the remaining nodes in ψ_3.

Due to the way we are bounding the sum, we do not want to assume any features of ψ_1 and ψ_2 other than those enforced by by the indicator variables in (5.328). Thus, we know that $|\eta(\psi_1)| = |\eta(\alpha)|$, $|\eta(\psi_2)| = |\eta(\beta)|$ and $|\eta(\psi_1)\cap\eta(\psi_2)| = |\eta(\alpha)\cap\eta(\beta)|$.

Let us consider the components to constructing a ψ_3 which matches the conditions in turn:

1. We bound the number of ways of selecting the overlapping edges from ψ_1 and ψ_2 as follows. The indicator variable $\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_2)\cap\eta(\psi_3)|)$ enforces that the number of edges shared by ψ_1, ψ_2 and ψ_3 is given by $|\eta(\alpha) \cap \eta(\beta)\cap\eta(\gamma)|$. Further we know that there are $|\eta(\alpha)\cap\eta(\beta)|$ edges shared by ψ_1 and ψ_2, therefore there are $\binom{|\eta(\alpha)\cap\eta(\beta)|}{|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)|}$ ways of selecting these edges. Similarly, combining the indicator variable $\mathbb{1}\Big(|\eta(\alpha)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_3)|\Big)$ with $\mathbb{1}(|\eta(\alpha)\cap \eta(\beta)\cap\eta(\gamma)| = |\eta(\psi_1)\cap\eta(\psi_2)\cap\eta(\psi_3)|)$ fixes the number of edges in ψ_1 and ψ_3 but not in ψ_2 to $|(\eta(\psi_1)\cap\eta(\psi_3))\setminus\eta(\psi_2)| = |\eta(\psi_1)\cap\eta(\psi_3)| - |\eta(\psi_1)\cap\eta(\psi_2)\cap\eta(\psi_3)| = |\eta(\alpha) \cap \eta(\gamma)| - |\eta(\alpha) \cap \eta(\beta) \cap \eta(\gamma)| = |(\eta(\alpha) \cap \eta(\gamma)) \setminus \eta(\beta)|$. Further, there are $|\eta(\psi_1)\setminus\eta(\psi_2)| = |\eta(\psi_1)| - |\eta(\psi_1)\cap\eta(\psi_2)| = |\eta(\alpha)| - |\eta(\alpha)\cap\eta(\beta)| = |\eta(\alpha)\setminus\eta(\beta)|$ edges shared between ψ_1 and ψ_2 to choose from. Thus, there are $\binom{|\eta(\alpha)\setminus\eta(\beta)|}{|(\eta(\alpha)\cap\eta(\gamma))\setminus\eta(\beta)|}$ ways to select the edges in ψ_1 and ψ_3 but not in ψ_2. Applying a similar argument to the edges in ψ_2 and ψ_3 but not in ψ_1 we obtain the following expression for the number of ways of selecting the overlapping edges from ψ_1 and ψ_2:

$$\binom{|\eta(\alpha) \cap \eta(\beta)|}{|\eta(\alpha) \cap \eta(\beta) \cap \eta(\gamma)|}\binom{|\eta(\alpha) \setminus \eta(\beta)|}{|(\eta(\alpha) \cap \eta(\gamma)) \setminus \eta(\beta)|}\binom{|\eta(\beta) \setminus \eta(\alpha)|}{|(\eta(\beta) \cap \eta(\gamma)) \setminus \eta(\alpha)|} \tag{5.330}$$

   However, this is an upper bound as we have not accounted for the indicator variable $\mathbb{1}\Big(C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)) = k_g\Big)$, which enforces the selection of edges in a fixed number of blocks, and thus many of these selections will not preserve the number of contiguous subsequences. However we include all of the selections that do, hence upper bound.

2. The second stage is a bound on the number of ways of placing those edges into ψ_3 and is constructed as follows. The first stage fixes the overlapping edge selections, giving us the identity of the overlapping edges. By assumption ψ_3 is a simple path (every node is involved in at most two edges), therefore we can

construct a unique set of contiguous paths from the known overlapping edges. The indicator variable $\mathbb{1}\Big(C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)) = k_g\Big)$ enforces that there are k_g such segments. To construct ψ_3 we must decide on their orientation (i.e. which end of the subsequence is closest to $\psi_3(1)$) and the order in which they appear in ψ_3. The number of different orientations and orderings of k_g blocks is given by:

$$(k_g!)2^{k_g} \tag{5.331}$$

where the first term accounts for the order of the blocks and the second the orientation of each block (reversability). The final contribution to placing the overlapping edges is the position of the remaining nodes in ψ_3, i.e. how many nodes separate each contiguous subsequence. We know the total number of nodes is given by $|\eta(\gamma)| + 1$, and from the proof of Lemma 21 we know that the number of nodes fixed by the overlapping edges is given by $|(\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)| + C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)) = |(\eta(\alpha) \cup \eta(\beta)) \cap \eta(\gamma)| + k_g$, leaving $|\eta(\psi_3) \setminus (\eta(\psi_1) \cup \eta(\psi_2))| + 1 - k_\gamma = |\eta(\gamma) \setminus (\eta(\alpha) \cup \eta(\beta))| + 1 - k_\gamma$ nodes to place in between the subsequences. To calculate the number of ways to place these nodes, we use *compositions*. A composition in combinatorics is the set of solutions $x_1, x_2, \ldots, x_q$ to the equation:

$$x_1 + x_2 + \cdots + x_q = d \tag{5.332}$$

such that $x_i \in \mathbb{N}$ and $d \in \mathbb{N}$. A so called weak composition relaxes this constraint to let $x_i \in \mathbb{N} \cup \{0\}$.

There are $\binom{d-1}{q-1}$ such decompositions [176] and the number of weak decompositions is given by $\binom{d+q-1}{q-1}$ [176].

For ψ_3 we let the number of nodes placed before the first common subsequence be x_1, the number of nodes between the first and the second common subse-

quence be x_2 etc, giving a total of $k_g + 1$ variables. Note, x_i can be 0, as this implies that the common subsequences are next to each other but the edge between them is not present in ψ_1 or ψ_2. This is an upper bound as we do not exclude the case that this edge is part of ψ_1 or ψ_2, but we do include the case where this condition is satisfied. Further, if we let the size of the ith common subsequence be y_i, each common subsequence has $y_i + 1$ nodes. Using the formula for weak compositions we obtain:

$$\binom{|\eta(\gamma) \setminus (\eta(\alpha) \cup \eta(\beta))| + 1}{k_g}. \tag{5.333}$$

3. The final term in the bound is the number of ways of selecting the nodes to fill the gaps which we just specified which then uniquely defines ψ_3. There are $|\eta(\psi_3) \setminus (\eta(\psi_1) \cup \eta(\psi_2))| + 1 - k_\gamma$ nodes to fill. Therefore we can bound by $n^{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1-k_\gamma}$.

Combining the final bound with the bound on the number of ways of selecting the overlap from (5.330), the number of ways of placing the overlaps in γ from (5.331) and (5.333) we obtain:

$$\begin{aligned}
&\sum_{\psi_3 \in I_{|\eta(\gamma)|}} \mathbb{1}\Big(|\eta(\alpha) \cap \eta(\gamma)| = |\eta(\psi_1) \cap \eta(\psi_3)\Big) \mathbb{1}\Big(|\eta(\beta) \cap \eta(\gamma)| = |\eta(\psi_2) \cap \eta(\psi_3)\Big) && (5.334)\\
&\quad \times \mathbb{1}(|\eta(\alpha) \cap \eta(\beta) \cap \eta(\gamma)| = |\eta(\psi_1) \cap \eta(\psi_2) \cap \eta(\psi_3))\\
&\quad \times \mathbb{1}\Big(C((\eta(\psi_1) \cup \eta(\psi_2)) \cap \eta(\psi_3)) = k_g\Big)\\
&\leq \binom{|\eta(\alpha) \cap \eta(\beta)|}{|\eta(\alpha) \cap \eta(\beta) \cap \eta(\gamma)|} \binom{|\eta(\alpha) \setminus \eta(\beta)|}{|(\eta(\alpha) \cap \eta(\gamma)) \setminus \eta(\beta)|} \binom{|\eta(\beta) \setminus \eta(\alpha)|}{|(\eta(\beta) \cap \eta(\gamma)) \setminus \eta(\alpha)|} && (5.335)\\
&\quad \times \binom{|\eta(\gamma) \setminus (\eta(\alpha) \cup \eta(\beta))| + 1}{k_g} (k_g!) 2^{k_g} n^{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1-k_\gamma}\\
&= \mathfrak{D}(\alpha, \beta, \gamma, k_g)(k_\gamma!) 2^{k_\gamma} n^{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1-k_\gamma} && (5.336)
\end{aligned}$$

where $\mathfrak{D}(\alpha, \beta, \gamma, k_g) = \binom{|\eta(\alpha)\cap\eta(\beta)|}{|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)|} \binom{|\eta(\alpha)\setminus\eta(\beta)|}{|(\eta(\alpha)\cap\eta(\gamma))\setminus\eta(\beta)|} \binom{|\eta(\beta)\setminus\eta(\alpha)|}{|(\eta(\beta)\cap\eta(\gamma))\setminus\eta(\alpha)|}$

$\times \binom{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1}{k_g}$. Therefore,

$$|\Psi_O(\alpha,\beta,\gamma)| \leq \sum_{k_g=1}^{K_g^{\max}} \mathfrak{D}(\alpha,\beta,\gamma,k_g)(k_g!)2^{k_g} n^{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1-k_\gamma} \tag{5.337}$$
$$\times \sum_{\psi_1 \in I_{|\eta(\alpha)|}} \sum_{\psi_2 \in I_{|\eta(\beta)|}} \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = |\eta(\psi_1)\cap\eta(\psi_2)|)$$

We now bound the sum over ψ_2. We assume that ψ_1 is fixed, using a similar approach to ψ_3, we obtain:

$$\sum_{\psi_2 \in I_{|\eta(\beta)|}} \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = |\eta(\psi_1)\cap\eta(\psi_2)|) \tag{5.338}$$
$$= \sum_{k_b=1}^{K_b^{\max}} \sum_{\psi_2 \in I_{|\eta(\beta)|}} \mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = |\eta(\psi_1)\cap\eta(\psi_2)|)\mathbb{1}(C(\eta(\psi_1)\cap\eta(\psi_2)) = k_b), \tag{5.339}$$

where following the result for K_g we let $K_b^{\max} = \min(|\eta(\alpha)\cap\eta(\beta)|, \left\lceil \frac{|\eta(\beta)|}{2} \right\rceil)$.

We will again bound this sum by assuming that ψ_1 is fixed and bounding the number of ways to construct ψ_2 which matches the conditions. Again we first bound by the number of ways of selecting the overlapping edges from ψ_1, second we bound the number of ways of placing these edges in ψ_2, third we bound the number of ways of filling the remaining nodes.

As we are only selecting from one sequence we can construct a sharper bound for the selection than the ψ_3 case. We number the k_b overlapping subsequences between ψ_1 and ψ_2 from the left in ψ_1, then we let y_i be the number of edges in the ith subsequence. Clearly, $\sum_{i=1}^{k_b} y_i = |\eta(\psi_1)\cap\eta(\psi_2)| = |\eta(\alpha)\cap\eta(\beta)|$, where the equality is enforced by the indicator variable $\mathbb{1}(|\eta(\alpha)\cap\eta(\beta)| = |\eta(\psi_1)\cap\eta(\psi_2)|)$. Further, $y_i > 0$, thus the number of ways of distributing the overlapping edges into k_b subsequences can be calculated using compositions, giving $\binom{|\eta(\alpha)\cap\eta(\beta)|-1}{k_b-1}$ ways of selecting y_i.

By specifying the number of nodes between the overlapping subsequences in ψ_1,

we uniquely define the overlapping edges. The number of ways of specifying these nodes is a weak composition as in the bound of the sum over ψ_3 earlier in the proof. Let x_1 be the number of nodes before the first common subsequence, x_2 the number of nodes between the first and second common subsequence etc, giving $k_b + 1$ variables. The number of nodes to be placed is given by $|\eta(\psi_1)| + 1 - \sum_{i=1}^{k_b}(y_i + 1) = |\eta(\psi_1)| - |\eta(\psi_1) \cap \eta(\psi_2)| - k_b = |\eta(\alpha)| - |\eta(\alpha) \cap \eta(\beta)| + 1 - k_b$. Applying the weak composition formula with $q = k_b + 1$ and simplifying we obtain, $\binom{|\eta(\alpha)\setminus\eta(\beta)|+1}{k_b}$. Combining the two contributions we obtain a bound on the number of ways of selecting the overlaping edges from ψ_1:

$$\binom{|\eta(\alpha) \cap \eta(\beta)| - 1}{k_b - 1}\binom{|\eta(\alpha) \setminus \eta(\beta)| + 1}{k_b} \tag{5.340}$$

The number of ways of placing the overlapping subsequences into ψ_2 is bounded as follows. Following the calculation for ψ_3, we have to accord for the order of the sequences in ψ_2, their orientation in ψ_2 and the placement of the remaining nodes. There are $k_b!$ ways of reordering the sequences and 2^{k_b} ways of orienting them. To count the number of ways of placing the remaining nodes, we use a composition argument which we omit as it is identical to the argument for placing the remaining nodes in ψ_1, giving a contribution of $\binom{|\eta(\beta)\setminus\eta(\alpha)|+1}{k_b}$. Thus the number of ways of placing the overlapping subsequences into ψ_2 is bounded from above by

$$2^{k_b} k_b! \binom{|\eta(\beta) \setminus \eta(\alpha)| + 1}{k_b}. \tag{5.341}$$

Using (5.340) and (5.341) and bounding the number of ways to select the $|\eta(\psi_2) \setminus$

$\eta(\psi_1)|+1-k_b|\eta(\beta)\setminus\eta(\alpha)|+1-k_b$ unfixed nodes in ψ_2 by $n^{|\eta(\beta)\setminus\eta(\alpha)|+1-k_b}$ we obtain:

$$
\begin{aligned}
&|\Psi_O(\alpha,\beta,\gamma)| && (5.342)\\
&\leq \sum_{k_g=1}^{K_g^{\max}} \mathfrak{D}(\alpha,\beta,\gamma,k_g)(k_g!)\left(\frac{2}{n}\right)^{k_g} n^{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1-k_g} && (5.343)\\
&\quad\times \sum_{\psi_1\in I_{|\eta(\alpha)|}}\sum_{\psi_2\in I_{|\eta(\beta)|}} \mathbb{1}(|\eta(\psi_1)\cap\eta(\psi_2)|=|\eta(\alpha)\cap\eta(\beta)|)\\
&\leq \sum_{k_g=1}^{K_g^{\max}}\sum_{k_b=1}^{K_b^{\max}} \mathfrak{D}(\alpha,\beta,\gamma,k_g)(k_g!)\left(\frac{2}{n}\right)^{k_g} n^{|\eta(\gamma)\setminus(\eta(\alpha)\cup\eta(\beta))|+1-k_g}\sum_{\psi_1\in I_{|\eta(\alpha)|}}\binom{|\eta(\alpha)\cap\eta(\beta)|-1}{k_b-1} && (5.344)\\
&\quad\times\binom{|\eta(\alpha)\setminus\eta(\beta)|+1}{k_b}2^{k_b}k_b!\binom{|\eta(\beta)\setminus\eta(\alpha)|+1}{k_b}n^{|\eta(\beta)\setminus\eta(\alpha)|+1-k_b}\\
&\leq \sum_{k_g=1}^{K_g^{\max}}\sum_{k_\beta=1}^{K_b^{\max}} \mathfrak{D}(\alpha,\beta,\gamma,k_g)(k_b!k_g!)\left(\frac{2}{n}\right)^{k_a+k_g} n^{2+|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|-|\eta(\alpha)|} && (5.345)\\
&\quad\times\binom{|\eta(\alpha)\cap\eta(\beta)|-1}{k_b-1}\binom{|\eta(\alpha)\setminus\eta(\beta)|+1}{k_b}\binom{|\eta(\beta)\setminus\eta(\alpha)|+1}{k_b}\sum_{\psi_1\in I_{|\eta(\alpha)|}}1\\
&\leq \sum_{k_\gamma=1}^{K_\gamma^{\max}}\sum_{k_\beta=1}^{K_b^{\max}} \frac{\mathfrak{C}(\alpha,\beta,\gamma,k_b)\mathfrak{D}(\alpha,\beta,\gamma,k_g)(k_\beta!k_\gamma!)}{4}\left(\frac{2}{n}\right)^{k_\alpha+k_\gamma} n^{3+|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|}, && (5.346)
\end{aligned}
$$

where we let $\mathfrak{C}(\alpha,\beta,\gamma,k_b)=\binom{|\eta(\alpha)\cap\eta(\beta)|-1}{k_b-1}\binom{|\eta(\alpha)\setminus\eta(\beta)|+1}{k_b}\binom{|\eta(\beta)\setminus\eta(\alpha)|+1}{k_b}$ and we use the fact that $\sum_{\alpha\in I_{|\eta(\alpha)|}}1=|I_{|\eta(\alpha)|}|=\frac{n^2}{4}\frac{(n-2)!}{(n-|\eta(\alpha)|-1)!}<\frac{n^{|\eta(\alpha)|+1}}{4}$. Thus giving the bound in the Lemma, QED.

Finally, we bring these results together:

Theorem 26. *For $c>1$, and $c\lambda\geq 1$. The Wasserstein distance between a normal distribution and the normalised weighted sum of paths of length $\leq l$ ($\tilde{\zeta}^{(l)}$) between the two sets of nodes is bounded as follows:*

$$
d_{H_W}(\tilde{\zeta}^{(l)},Z)\leq\frac{n}{(n-c)^{\frac{3}{2}}}\mathfrak{E}\left((l+1)\Big(1+\frac{6l^3}{5n}\Big)+\frac{4l^8e^{(\frac{2l^3}{n})}}{n^2(1-\nu)}\right), \tag{5.347}
$$

where

$$\mathfrak{E} = \frac{(1-\nu)^{3-3l} l^{\frac{5}{2}}}{(l-1)^{\frac{3}{2}}} \frac{256\sqrt{2} e^{2c^{-\frac{2}{3}}} c^{\frac{11}{6}}}{(c^{\frac{2}{3}}-1)(c-1)} \left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}} \right)^2 \tag{5.348}$$

for $l > 2$ *and for* $0 < \nu < 1$ *such that* $\nu n > 2l$.

Remark The leading order term on the bound in n is of order $\frac{1}{\sqrt{n}}$ which is the same order as in the $l = 2$ case.

Proof Let us recall the normal bound from (5.302) which was derived from the results in Ref. [14], namely

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq 4 \sum_{\alpha \in I} \sum_{\beta,\gamma \in \mathbb{K}_\alpha} \left(E[|Y_\alpha Y_\beta Y_\gamma|] + E[|Y_\alpha Y_\beta|] E[|Y_\gamma|] \right). \tag{5.349}$$

We bound the summand using Lemma 24, resulting in the following form:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq 4 \sum_{\alpha \in I} \sum_{\beta,\gamma \in \mathbb{K}_\alpha} \frac{8}{\mathrm{Var}(\zeta^{(l)})^{\frac{3}{2}}} \lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|} p^{|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|}. \tag{5.350}$$

In order to use Lemma 25 we have to express this summation in terms of Ψ_O. To do so we sum over each of the components of $\mathfrak{O}(\alpha, \beta, \gamma)$ as follows:

$$\begin{aligned} d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq & \frac{32}{\mathrm{Var}(\zeta^{(l)})^{\frac{3}{2}}} \sum_{O_1=0}^{l} \sum_{O_2=0}^{l} \sum_{O_3=0}^{l} \sum_{O_4=0}^{l} \sum_{O_5=0}^{l} \sum_{O_6=0}^{l} \sum_{O_7=0}^{l} \sum_{\alpha \in I} \sum_{\beta,\gamma \in \mathbb{K}_\alpha} \Bigg(\\ & \mathbb{1}(|\eta(\alpha)| = O_1) \mathbb{1}(|\eta(\beta)| = O_2) \mathbb{1}(|\eta(\gamma)| = O_3) \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4) \\ & \times \mathbb{1}(|\eta(\alpha) \cap \eta(\gamma)| = O_5) \mathbb{1}(|\eta(\beta) \cap \eta(\gamma)| = O_6) \\ & \times \mathbb{1}(|\eta(\beta) \cap \eta(\beta) \cap \eta(\gamma)| = O_7) \lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|} p^{|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|} \Bigg) \end{aligned} \tag{5.351}$$

We restrict the upper bound on the summation as follows. Consider $|\eta(\alpha) \setminus \eta(\beta)|$. If we choose O_4 and O_1 such that $O_4 > O_1$, then the indicator variables $\mathbb{1}(|\eta(\alpha)| =$

$O_1)\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4)$ imply that:

$$|\eta(\alpha) \setminus \eta(\beta)| = |\eta(\alpha)| - |\eta(\alpha) \cap \eta(\beta)| = O_1 - O_4 < 0, \tag{5.352}$$

which is clearly impossible. To avoid this we note that for sets X and Y by definition:

$$|X \cap Y| \leq |X|. \text{ Therefore, } \mathbb{1}(|X \cap Y| = a)\mathbb{1}(|X| = b) = 0 \text{ if } a > b. \tag{5.353}$$

Therefore:

$$\sum_{O_1=0}^{l} \sum_{O_2=0}^{l} \sum_{O_4=0}^{l} \mathbb{1}(|\eta(\alpha)| = O_1)\mathbb{1}(|\eta(\beta)| = O_2)\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4) \tag{5.354}$$

$$= \sum_{O_1=0}^{l} \sum_{O_2=0}^{l} \Bigg(\sum_{O_4=0}^{\min(O_1,O_2)} \Big(\mathbb{1}(|\eta(\alpha)| = O_1)\mathbb{1}(|\eta(\beta)| = O_2)\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4) \Big) \tag{5.355}$$

$$+ \sum_{O_4=\min(O_1,O_2)+1}^{l} \Big(\mathbb{1}(|\eta(\alpha)| = O_1)\mathbb{1}(|\eta(\beta)| = O_2)\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4) \Big) \Bigg)$$

$$= \sum_{O_1=0}^{l} \sum_{O_2=0}^{l} \sum_{O_4=0}^{\min(O_1,O_2)} \Big(\mathbb{1}(|\eta(\alpha)| = O_1)\mathbb{1}(|\eta(\beta)| = O_2)\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4) \Big). \tag{5.356}$$

Applying this argument to O_5, O_6 and O_7 results in upper bounds of $\min(O_1, O_3)$, $\min(O_2, O_3)$ and $\min(O_4, O_5, O_6)$ respectively.

In a similar manner we can also restrict the lower bound of the summations. We note that for all $x \in I$, $|x| > 1$, which implies that $|\eta(x)| \geq 1$. Therefore, $\sum_{O_1=0}^{l} \mathbb{1}(|\eta(\alpha)| = 0) = 0$. Thus, we can restrict the lower bound to 1 rather than 0. Repeating this argument for O_2 and O_3 we obtain a lower bound of 1 for O_2 and O_3.

Finally, we can restrict the lower bound of O_4 and O_5 as follows. For, $\alpha \in I$ $\beta \in \mathbb{K}_\alpha = \{x \in I : |\eta(\alpha) \cap \eta(x)| > 0\}$ therefore, $|\eta(\alpha) \cap \eta(\beta)| > 0$. Thus, $\sum_{\alpha \in I} \sum_{\beta \in \mathbb{K}_\alpha} \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = 0) = 0$. Applying the indicator $\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4)$ we can restrict the summation to $O_4 > 0$, and applying the same argument to $\mathbb{1}(|\eta(\alpha) \cap \eta(\gamma)| = O_5)$ we obtain $O_5 > 0$. However we cannot apply the same result

to O_6 as there is no condition on $|\eta(\beta) \cap \eta(\gamma)|$.

Applying these restrictions we can rewrite the summation as follows:

$$\begin{aligned} d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq & \frac{32}{\mathrm{Var}(\zeta^{(l)})^{\frac{3}{2}}} \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} \sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \sum_{O_7=0}^{\min(O_4,O_5,O_6)} \\ & \times \sum_{\alpha \in I} \sum_{\beta,\gamma \in \mathbb{K}_\alpha} \Bigg(\mathbb{1}(|\eta(\alpha)| = O_1)\mathbb{1}(|\eta(\beta)| = O_2)\mathbb{1}(|\eta(\gamma)| = O_3) \\ & \times \mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4)\mathbb{1}(|\eta(\alpha) \cap \eta(\gamma)| = O_5)\mathbb{1}(|\eta(\beta) \cap \eta(\gamma)| = O_6) \\ & \times \mathbb{1}(|\eta(\alpha) \cap \eta(\beta) \cap \eta(\gamma)| = O_7)\lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|} p^{|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|} \Bigg). \end{aligned} \tag{5.357}$$

We observe that for any non-zero terms in the summation the powers of λ and p will be uniquely defined by the value of $O_1, ..., O_7$ which can be seen as follows. We note that:

$$\begin{aligned} |\eta(\alpha) \cup \eta(\beta) \cup \eta(\gamma)| = & |\eta(\alpha)| + |\eta(\beta)| + |\eta(\gamma)| - |\eta(\alpha) \cap \eta(\beta)| \\ & - |\eta(\alpha) \cap \eta(\gamma)| - |\eta(\beta) \cap \eta(\gamma)| + |\eta(\alpha) \cap \eta(\beta) \cap \eta(\gamma)|. \end{aligned} \tag{5.358}$$

Therefore we can rewrite the summand as follows:

$$\begin{aligned} & \mathbb{1}(|\eta(\alpha)| = O_1)\mathbb{1}(|\eta(\beta)| = O_2)\mathbb{1}(|\eta(\gamma)| = O_3)\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4) \\ & \times \mathbb{1}(|\eta(\alpha) \cap \eta(\gamma)| = O_5)\mathbb{1}(|\eta(\beta) \cap \eta(\gamma)| = O_6) \\ & \times \mathbb{1}(|\eta(\beta) \cap \eta(\beta) \cap \eta(\gamma)| = O_7)\lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|} p^{|\eta(\alpha)\cup\eta(\beta)\cup\eta(\gamma)|} \end{aligned} \tag{5.359}$$

$$\begin{aligned} = & \mathbb{1}(|\eta(\alpha)| = O_1)\mathbb{1}(|\eta(\beta)| = O_2)\mathbb{1}(|\eta(\gamma)| = O_3)\mathbb{1}(|\eta(\alpha) \cap \eta(\beta)| = O_4) \\ & \times \mathbb{1}(|\eta(\alpha) \cap \eta(\gamma)| = O_5)\mathbb{1}(|\eta(\beta) \cap \eta(\gamma)| = O_6) \\ & \times \mathbb{1}(|\eta(\beta) \cap \eta(\beta) \cap \eta(\gamma)| = O_7)\lambda^{O_1+O_2+O_3} p^{(\sum_{i=1}^{3} O_i)-(\sum_{i=4}^{6} O_i)+O_7} \end{aligned} \tag{5.360}$$

Further, recall that

$$\mathfrak{O}(\alpha,\beta,\gamma) =(|\eta(\alpha)|,|\eta(\beta)|,|\eta(\gamma)|,|\eta(\alpha)\cap\eta(\beta)|,|\eta(\alpha)\cap\eta(\gamma)|, \tag{5.361}$$

$$|\eta(\beta)\cap\eta(\gamma)|,|\eta(\alpha)\cap\eta(\beta)\cap\eta(\gamma)|).$$

We note that the product of the indicator variables is equivalent to $\mathbb{1}(\mathfrak{O}(\alpha,\beta,\gamma) = (O_1,...,O_7))$. Thus we can rewrite the summation as:

$$\begin{aligned} d_{H_W}(\tilde{\zeta}^{(l)},Z) \leq & \frac{32}{\mathrm{Var}(\zeta^{(l)})^{\frac{3}{2}}} \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} \sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \sum_{O_7=0}^{\min(O_4,O_5,O_6)} \\ & \times (\lambda p)^{O_1+O_2+O_3} p^{-O_4-O_5-O_6+O_7} \sum_{\alpha\in I} \sum_{\beta,\gamma\in\mathbb{K}_\alpha} \mathbb{1}(\mathfrak{O}(\alpha,\beta,\gamma) = (O_1,...,O_7)) \end{aligned} \tag{5.362}$$

We recall for, $(\alpha,\beta,\gamma) \in (I^{(l)})^3$ that:

$$\Psi_O(\alpha,\beta,\gamma) = \{(\psi_1,\psi_2,\psi_3) \in (I^{(l)})^3 : \quad \mathfrak{O}(\alpha,\beta,\gamma) = \mathfrak{O}(\psi_1,\psi_2,\psi_3)\}. \tag{5.363}$$

By the definition of Ψ_O we observe that:

1. $(\alpha,\beta,\gamma) \in \Psi_O(\alpha,\beta,\gamma)$
2. If $(\alpha^*,\beta^*,\gamma^*) \in \Psi_O(\alpha,\beta,\gamma)$ then $(\alpha,\beta,\gamma) \in \Psi_O(\alpha^*,\beta^*,\gamma^*)$.

The first observation comes straight from the definition of Ψ_O, the second observation comes from noticing that $(\alpha^*,\beta^*,\gamma^*) \in \Psi_O(\alpha,\beta,\gamma)$ implies that $\mathfrak{O}(\alpha^*,\beta^*,\gamma^*) = \mathfrak{O}(\alpha,\beta,\gamma)$ which in turn implies that $(\alpha,\beta,\gamma) \in \Psi_O(\alpha^*,\beta^*,\gamma^*)$.

It is easy to see that if for all $x \in S_1$ then $x \in S_2$, and for all $x \in S_2$ then $x \in S_1$, it follows that $S_1 = S_2$. Therefore, if $(\alpha^*,\beta^*,\gamma^*) \in \Psi_O(\alpha,\beta,\gamma)$ then $\Psi_O(\alpha,\beta,\gamma) = \Psi_O(\alpha^*,\beta^*,\gamma^*)$. Further, by trivial extension of the previous result if $(\alpha^*,\beta^*,\gamma^*) \notin \Psi_O(\alpha,\beta,\gamma)$ then $\Psi_O(\alpha,\beta,\gamma) \cap \Psi_O(\alpha^*,\beta^*,\gamma^*) = \emptyset$.

Therefore:

$$\sum_{\alpha\in I}\sum_{\beta,\gamma\in\mathbb{K}_\alpha}\mathbb{1}(\mathfrak{O}(\alpha,\beta,\gamma)=(O_1,...,O_7))= \tag{5.364}$$

$$|\{(\psi_1,\psi_2,\psi_3):\psi_1\in I\ \psi_2,\psi_3\in\mathbb{K}_\alpha \text{ and } \mathfrak{O}(\psi_1,\psi_2,\psi_3)=(O_1,...,O_7)\}| \tag{5.365}$$

$$\leq|\{(\psi_1,\psi_2,\psi_3):\psi_1,\psi_2,\psi_3\in(I^{(l)})^3 \text{ and } \mathfrak{O}(\psi_1,\psi_2,\psi_3)=(O_1,...,O_7)\}| \tag{5.366}$$

We consider two cases, the first where (5.366) is empty and the second where it is not.

Case 1 By assumption the set in (5.366) is empty. Therefore, there does not exist $(\alpha^*,\beta^*,\gamma^*)$ such that $\mathfrak{O}(\alpha^*,\beta^*,\gamma^*)=(O_1,...,O_7)$. Therefore,

$$\sum_{\alpha\in I}\sum_{\beta,\gamma\in\mathbb{K}_\alpha}\mathbb{1}(\mathfrak{O}(\alpha,\beta,\gamma)=(O_1,...,O_7))=0. \tag{5.367}$$

Case 2 By assumption the set in (5.366) is non-empty. Therefore there exists $\alpha^*,\beta^*,\gamma^*\in(I^{(l)})^3$ such that $\mathfrak{O}(\alpha^*,\beta^*,\gamma^*)=(O_1,...,O_7)$. Thus, by definition of Ψ_O and Lemma 25

$$\sum_{\alpha\in I}\sum_{\beta,\gamma\in\mathbb{K}_\alpha}\mathbb{1}(\mathfrak{O}(\alpha,\beta,\gamma)=(O_1,...,O_7))\leq|\Psi_O(\alpha^*,\beta^*,\gamma^*)| \tag{5.368}$$

$$\leq\sum_{k_g=1}^{K_g^{\max}(\alpha^*,\beta^*,\gamma^*)}\sum_{k_b=1}^{K_b^{\max}(\alpha^*,\beta^*,\gamma^*)}\frac{\mathfrak{C}(\alpha^*,\beta^*,\gamma^*,k_b)\mathfrak{D}(\alpha^*,\beta^*,\gamma^*)(k_b!k_g!)}{4}\times\left(\frac{2}{n}\right)^{k_b+k_g}n^{3+|\eta(\alpha^*)\cup\eta(\beta^*)\cup\eta(\gamma^*)|}. \tag{5.369}$$

We note that the summation over $O_1,...,O_7$ may include combinations of the O's that are not present in the original summation over α,β and γ, however the inequality holds as the bound from Lemma 25 is non negative.

We first focus on the second case, i.e. where for a given $O_1,...,O_7$ there exists $\alpha^*,\beta^*,\gamma^*\in(I^{(l)})^3$ such that $\mathfrak{O}(\alpha^*,\beta^*,\gamma^*)=(O_1,...,O_7)$. To perform the sum-

mations over $O_1, ..., O_7$ we require the summand to be parameterised by $O_1, ...O_7$ rather than α^*, β^* and γ^*. Using the same argument that we used when rewriting $\lambda^{|\eta(\alpha)|+|\eta(\beta)|+|\eta(\gamma)|}$, i.e. that $\mathbb{1}(\mathfrak{O}(\alpha^*, \beta^*, \gamma^*) = (O_1, ..., O_7))$ fixes the value of the set intersection, we can express the final line in terms of $O_1, ..., O_7$ using simple set operations. We use the following relations to do so:

$$\begin{aligned}
K_g^{\max}(\alpha^*, \beta^*, \gamma^*) =& \min\left(|\eta(\gamma^*) \cap (\eta(\alpha^*) \cup \eta(\beta^*)|, \left\lceil \frac{|\eta(\gamma^*)|}{2} \right\rceil\right) && (5.370)\\
=& \min\Big(|\eta(\alpha^*) \cap \eta(\gamma^*)| + |\eta(\beta^*) \cap \eta(\gamma^*)| && (5.371)\\
& - |\eta(\gamma^*) \cap \eta(\alpha^*) \cap \eta(\beta^*)|, \left\lceil \frac{|\eta(\gamma^*)|}{2} \right\rceil\Big)\\
=& \min\left(O_5 + O_6 - O_7, \left\lceil \frac{O_3}{2} \right\rceil\right) && (5.372)\\
K_b^{\max}(\alpha^*, \beta^*, \gamma^*) =& \min\left(|\eta(\alpha^*) \cap \eta(\beta^*)|, \left\lceil \frac{|\eta(\beta^*)|}{2} \right\rceil\right) = \min\left(O_4, \left\lceil \frac{O_2}{2} \right\rceil\right) && (5.373)
\end{aligned}$$

$$\begin{aligned}
\mathfrak{C}(\alpha^*, \beta^*, \gamma^*, k_b) =& \binom{|\eta(\alpha^*) \cap \eta(\beta^*)| - 1}{k_b - 1}\binom{|\eta(\alpha^*) \setminus \eta(\beta^*) + 1}{k_b} && (5.374)\\
& \times \binom{|\eta(\beta^*) \setminus \eta(a) + 1}{k_b}\\
=& \binom{O_4 - 1}{k_b - 1}\binom{O_1 - O_4 + 1}{k_b}\binom{O_2 - O_4 + 1}{k_b} && (5.375)\\
\mathfrak{D}(\alpha^*, \beta^*, \gamma^*, k_g) =& \left(\binom{|\eta(\alpha^*) \cap \eta(\beta^*)|}{|\eta(\alpha^*) \cap \eta(\beta^*) \cap \eta(\gamma^*)|}\binom{|\eta(\alpha^*) \setminus \eta(\beta^*)|}{|\eta(a) \cap \eta(\gamma^*) \setminus \eta(\beta^*)|}|\eta(\beta^*) \setminus \eta(\alpha^*)|\right) && (5.376)\\
& \times |\eta(\beta^*) \cap \eta(\gamma^*) \setminus \eta(\alpha^*)|\binom{|\eta(\gamma^*) \setminus (\eta(\alpha^*) \cup \eta(\beta^*))| + 1}{k_g}\\
=& \binom{O_4}{O_7}\binom{O_1 - O_4}{O_5 - O_7}\binom{O_2 - O_4}{O_6 - O_7}\binom{O_3 - O_5 - O_6 + O_7 + 1}{k_g} && (5.377)
\end{aligned}$$

Putting all of this together and using the expression for $|\eta(\alpha) \cup \eta(\beta) \cup \eta(\gamma)|$ we obtain the following bounds for the second case.

If there exists $\alpha^*, \beta^*, \gamma^* \in (I^{(l)})^3$ such that $\mathfrak{O}(\alpha^*, \beta^*, \gamma^*) = (O_1, ..., O_7)$ then

$$\sum_{\alpha \in I} \sum_{\beta,\gamma \in \mathbb{K}_\alpha} \mathbb{1}(\mathfrak{O}(\alpha, \beta, \gamma) = (O_1, ..., O_7)) \leq \sum_{k_g=1}^{\min(O_5+O_6-O_7, \lceil \frac{O_3}{2} \rceil)} \sum_{k_b=1}^{\min(O_4, \lceil \frac{O_2}{2} \rceil)} \tag{5.378}$$
$$\frac{(k_b! k_g!)}{4} \binom{O_4 - 1}{k_b - 1} \binom{O_1 - O_4 + 1}{k_b} \binom{O_2 - O_4 + 1}{k_b} \binom{O_4}{O_7} \binom{O_1 - O_4}{O_5 - O_7} \binom{O_2 - O_4}{O_6 - O_7}$$
$$\times \binom{O_3 - O_5 - O_6 + O_7 + 1}{k_g} \left(\frac{2}{n}\right)^{k_b + k_g} n^{3 + (\sum_{i=1}^{3} O_i) - (\sum_{i=4}^{6} O_i) + O_7}.$$

Every term in the summation in the second case is positive and therefore the bound is greater than 0. Thus, as the first case is 0 we can bound both cases by the bound on the second case. Substituting this bound into (5.362) we obtain:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq \frac{32}{\mathrm{Var}(\zeta^{(l)})^{\frac{3}{2}}} \sum_{O_1, O_2, O_3 = 1}^{l} \sum_{O_4=1}^{\min(O_1, O_2)} \sum_{O_5=1}^{\min(O_1, O_3)} \sum_{O_6=0}^{\min(O_2, O_3)} \sum_{O_7=0}^{\min(O_4, O_5, O_6)} \tag{5.379}$$
$$\times (\lambda p)^{O_1 + O_2 + O_3} p^{-O_4 - O_5 - O_6 + O_7} \sum_{k_g=1}^{\min(O_5+O_6-O_7, \lceil \frac{O_3}{2} \rceil)} \sum_{k_b=1}^{\min(O_4, \lceil \frac{O_2}{2} \rceil)} \frac{(k_b! k_g!)}{4}$$
$$\times \binom{O_4 - 1}{k_b - 1} \binom{O_1 - O_4 + 1}{k_b} \binom{O_2 - O_4 + 1}{k_b} \binom{O_4}{O_7} \binom{O_1 - O_4}{O_5 - O_7} \binom{O_2 - O_4}{O_6 - O_7}$$
$$\times \binom{O_3 - O_5 - O_6 + O_7 + 1}{k_g} \left(\frac{2}{n}\right)^{k_b + k_g} n^{3 + O_1 + O_2 + O_3 - O_4 - O_5 - O_6 + O_7}.$$

Recalling the result from Theorem. 20

$$\mathrm{Var}(\zeta_l) \geq \frac{n}{8} l(l-1)(1-p) \frac{(c\lambda)^{2l}}{c} (1-\nu)^{2l-2} \tag{5.380}$$

for $l > 2$ and for $0 < \nu < 1$ such that $\nu n > 2l$. For notational convenience, we let:

$$\mathbb{F}_V = ((1-\nu)^{2l-2}(1-p))^{-\frac{3}{2}} \qquad \mathbb{F}_W = \left(\frac{8}{l(l-1)}\right)^{\frac{3}{2}} \tag{5.381}$$

Therefore:

$$\frac{1}{\mathrm{Var}(\zeta^{(l)})^{\frac{3}{2}}} \leq n^{-\frac{3}{2}} \frac{c^{\frac{3}{2}}}{(c\lambda)^{3l}} \mathbb{F}_V \mathbb{F}_W \tag{5.382}$$

Substituting this into the expression and using the fact that $p = \frac{c}{n}$ to simplify we obtain:

$$\begin{aligned} d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq & 8n^{\frac{3}{2}}c^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} \sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \sum_{O_7=0}^{\min(O_4,O_5,O_6)} \\ & \times (\lambda c)^{O_1+O_2+O_3-3l} c^{-O_4-O_5-O_6+O_7} \sum_{k_g=1}^{\min(O_5+O_6-O_7, \lceil \frac{O_3}{2} \rceil)} \sum_{k_b=1}^{\min(O_4, \lceil \frac{O_2}{2} \rceil)} (k_b! k_g!) \\ & \times \binom{O_4-1}{k_b-1}\binom{O_1-O_4+1}{k_b}\binom{O_2-O_4+1}{k_b}\binom{O_4}{O_7}\binom{O_1-O_4}{O_5-O_7} \\ & \times \binom{O_2-O_4}{O_6-O_7}\binom{O_3-O_5-O_6+O_7+1}{k_g}\left(\frac{2}{n}\right)^{k_b+k_g}. \end{aligned} \tag{5.383}$$

Replacing the upper limit of k_g by $\left\lceil \frac{O_3}{2} \right\rceil$, which is greater than the previous bound of $\min(O_5 + O_6 - O_7, \left\lceil \frac{O_3}{2} \right\rceil)$, and rearranging the summations we obtain:

$$\begin{aligned} d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq & 8n^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} (\lambda c)^{O_1+O_2+O_3-3l} \sum_{k_g=1}^{\lceil \frac{O_3}{2} \rceil} \sum_{k_b=1}^{\min(O_4, \lceil \frac{O_2}{2} \rceil)} \\ & (k_b! k_g!)\binom{O_4-1}{k_b-1}\binom{O_1-O_4+1}{k_b}\binom{O_2-O_4+1}{k_b}\left(\frac{2}{n}\right)^{k_b+k_g} \\ & \times \sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \sum_{O_7=0}^{\min(O_4,O_5,O_6)} c^{\frac{3}{2}-O_4-O_5-O_6+O_7} \\ & \times \binom{O_4}{O_7}\binom{O_1-O_4}{O_5-O_7}\binom{O_2-O_4}{O_6-O_7}. \end{aligned} \tag{5.384}$$

We are now in a position to construct a simple bound in l, i.e. a bound which is concise in formulation but is not necessarily particularly close to the real value. Clearly:

$$\binom{n}{k} \leq \frac{n^k}{k!} \qquad \text{and} \qquad \binom{n}{k-a} \leq \frac{n^{k-a}}{(k-a)!} = \frac{k!}{n^a(k-a)!}\frac{n^k}{k!} \leq \frac{k^a}{n^a}\frac{n^k}{k!} \leq \frac{n^k}{k!} \tag{5.385}$$

where the final inequality comes from the fact that $n \geq k$. Further, using Vander-

monde's identity [185] for arbitrary $r \geq k$:

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k} \text{ and hence } \binom{n'}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n'-m}{r-k} \geq \binom{m}{k}\binom{n'-m}{r-k} \tag{5.386}$$

Therefore:

$$\binom{O_4}{O_7}\binom{O_1-O_4}{O_5-O_7} \leq \binom{O_1}{O_5} \tag{5.387}$$

$$\binom{O_2-O_4}{O_6-O_7} \leq \binom{O_2}{O_6-O_7} \leq \frac{O_2^{O_6}}{O_6!} \tag{5.388}$$

$$\binom{O_4}{O_7}\binom{O_1-O_4}{O_5-O_7}\binom{O_2-O_4}{O_6-O_7} \times \binom{O_3+1-O_5-O_6+O_7}{k_g} \leq \frac{O_1^{O_5}}{O_5!}\frac{O_2^{O_6}}{O_6!}\binom{O_3+1-O_5-O_6+O_7}{k_g}. \tag{5.389}$$

Let us consider the final three summations, substituting the bound we just derived we obtain:

$$\sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \sum_{O_7=0}^{\min(O_4,O_5,O_6)} \frac{O_1^{O_5}}{O_5!}\frac{O_2^{O_6}}{O_6!}\binom{O_3+1-O_5-O_6+O_7}{k_g} c^{\frac{3}{2}+O_7-\sum_4^6 O_j}. \tag{5.390}$$

We observe that $1-O_5-O_6+O_7 \leq 0$ as $O_6 \geq O_7$ and $O_5 \geq 1$ further, $\min(O_4,O_5,O_6) \leq \frac{O_4+O_5+O_6}{3}$ therefore,

$$\binom{O_3}{k_g} c^{\frac{3}{2}-O_4} \sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \frac{(\frac{O_1}{c})^{O_5}}{(O_5)!}\frac{(\frac{O_2}{c})^{O_6}}{(O_6)!} \sum_{O_7=0}^{\frac{O_4+O_5+O_6}{3}} c^{O_7} \tag{5.391}$$

$$= \binom{O_3}{k_g} c^{\frac{3}{2}-O_4} \sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \frac{(\frac{O_1}{c})^{O_5}}{(O_5)!}\frac{(\frac{O_2}{c})^{O_6}}{(O_6)!} \frac{c^{\frac{O_4+O_5+O_6}{3}+1}-1}{c-1} \tag{5.392}$$

$$\leq \binom{O_3}{k_g} \frac{c^{\frac{5}{2}-\frac{2}{3}O_4}}{c-1} \sum_{O_5=1}^{\min(O_1,O_3)} \sum_{O_6=0}^{\min(O_2,O_3)} \frac{(\frac{O_1}{c^{\frac{2}{3}}})^{O_5}}{(O_5)!}\frac{(\frac{O_2}{c^{\frac{2}{3}}})^{O_6}}{(O_6)!} \leq \binom{O_3}{k_g} \frac{lc^{\frac{11}{6}-\frac{2}{3}O_4}}{c-1} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}}. \tag{5.393}$$

The final inequalities are obtained by replacing the finite summation of positive terms

by an infinite summation of positive terms. Note, the use of the geometric sum formula is valid as $c > 1$. Substituting this bound into (5.384) we obtain:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq \frac{8n^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V l c^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} (\lambda c)^{O_1+O_2+O_3-3l} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} \tag{5.394}$$
$$\times \sum_{k_g=1}^{\lceil \frac{O_3}{2} \rceil} \sum_{k_b=1}^{\min(O_4, \lceil \frac{O_2}{2} \rceil)} (k_b! k_g!) \binom{O_4-1}{k_b-1} \binom{O_1-O_4+1}{k_b} \binom{O_2-O_4+1}{k_b} \left(\frac{2}{n}\right)^{k_b+k_g} c^{-\frac{2}{3}O_4} \binom{O_3}{k_g}.$$

We now split the right hand side into three parts: first we will find a bound for the summation when $k_b + k_g = 2$; second we will find a bound for the summation when $k_b + k_g = 3$; and third the remaining summation $k_b + k_g > 3$. Mathematically this is equivalent to adding an additional summation and indicator variable so that the bound (5.394) can now be written:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq \sum_{k_s=2}^{l+1} \frac{8n^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V l c^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} (\lambda c)^{O_1+O_2+O_3-3l} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} \tag{5.395}$$
$$\times \sum_{k_g=1}^{\lceil \frac{O_3}{2} \rceil} \sum_{k_b=1}^{\min\left(O_4, \lceil \frac{O_2}{2} \rceil\right)} \mathbb{1}(k_b + k_g = k_s)(k_b! k_g!) \binom{O_4-1}{k_b-1} \binom{O_1-O_4+1}{k_b}$$
$$\times \binom{O_2-O_4+1}{k_b} \left(\frac{2}{n}\right)^{k_b+k_g} c^{-\frac{2}{3}O_4} \binom{O_3}{k_g}.$$

Note, the limits on the summation over k_s are computed by considering the upper and lower bounds on the summations of k_b and k_g. The lower limit for k_s is computed by noting that the lower limit for k_b and k_g is 1, so that $\min(k_b + k_g) = 2$. The upper limit of k_s is computed by noting that $\max(k_b + k_g) = \left\lceil \frac{O_3}{2} \right\rceil + \min\left(O_4, \left\lceil \frac{O_2}{2} \right\rceil\right) \leq \left\lceil \frac{O_3}{2} \right\rceil + \left\lceil \frac{O_2}{2} \right\rceil \leq l + 1$. We then rewrite (5.395), in terms of a summation over three new variables representing the case where $k_g + k_b = 2$, $k_b + k_g = 3$, and $k_b + k_g > 3$ respectively:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq \mathfrak{I}_{k_g+k_b=2} + \mathfrak{I}_{k_g+k_b=3} + \mathfrak{I}_{k_g+k_b>3} \tag{5.396}$$

where

$$\mathfrak{I}_X = \frac{8n^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V l c^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} (\lambda c)^{O_1+O_2+O_3-3l} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} \sum_{k_g=1}^{\lceil \frac{O_3}{2} \rceil} \sum_{k_b=1}^{\min\left(O_4, \lceil \frac{O_2}{2} \rceil\right)} \tag{5.397}$$
$$\mathbb{1}(X)(k_b!k_g!)\binom{O_4-1}{k_b-1}\binom{O_1-O_4+1}{k_b}\binom{O_2-O_4+1}{k_b}\left(\frac{2}{n}\right)^{k_b+k_g} c^{-\frac{2}{3}O_4}\binom{O_3}{k_g}.$$

Note the indicator variable $\mathbb{1}(X)$ which is introduced to restrict the summation. Further, note that the decomposition in (5.396) is exhaustive as $k_b, k_g \geq 1$ and therefore $k_b + k_g \geq 2$. Since k_b and k_g are integer, the three conditions $k_b + k_g = 2$, $k_b + k_g = 3$ and $k_b + k_g > 3$ are exhaustive.

The case $\mathbf{k_b + k_g = 2}$ ($\mathfrak{I}_{k_b+k_g=2}$). This is only possible when $k_b = k_g = 1$. Therefore, after substituting this condition into (5.397) and simplifying, we obtain:

$$\mathfrak{I}_{k_b+k_g=2} = \frac{32\mathbb{F}_V\mathbb{F}_W}{n^{0.5}} \sum_{O_1,O_2,O_3=1}^{l} (c\lambda)^{\sum_1^3 O_i - 3l} \sum_{O_4=1}^{\min(O_1,O_2)} \binom{O_4-1}{0}\binom{O_1-O_4+1}{1} \tag{5.398}$$
$$\binom{O_2-O_4+1}{1} \times O_3 \frac{l c^{\frac{11}{6}-\frac{2}{3}O_4}}{c-1} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}}.$$

Note $\binom{O_4-1}{0}\binom{O_1-O_4+1}{1}\binom{O_2-O_4+1}{1} \leq O_1O_2$. Therefore, under the assumption that $c\lambda \geq 1$ and thus $(c\lambda)^{O_1-l} \leq 1$ as $O_1 \leq l$, we obtain:

$$\begin{aligned}
\mathfrak{I}_{k_b+k_g=2} \leq & \frac{32\mathbb{F}_V\mathbb{F}_W}{n^{0.5}} \sum_{O_1,O_2,O_3=1}^{l} (c\lambda)^{\sum_1^3 O_i-3l} \sum_{O_4=1}^{\min(O_1,O_2)} O_1O_2O_3 \frac{lc^{\frac{11}{6}-\frac{2}{3}O_4}}{c-1} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} && (5.399)\\
= & \frac{32l^3\mathbb{F}_V\mathbb{F}_W}{n^{0.5}} \frac{c^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} (c\lambda)^{\sum_1^3 O_i-3l} \sum_{O_4=1}^{\min(O_1,O_2)} O_3 c^{-\frac{2}{3}O_4} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} && (5.400)\\
\leq & \frac{32l^3\mathbb{F}_V\mathbb{F}_W}{n^{0.5}} \frac{c^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} O_3 e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} \frac{1-c^{-\frac{2}{3}l}}{c^{\frac{2}{3}}-1} && (5.401)\\
\leq & \frac{32l^3\mathbb{F}_V\mathbb{F}_W}{n^{0.5}(c^{\frac{2}{3}}-1)} \frac{c^{\frac{11}{6}}}{c-1} \left(\sum_{O_3=1}^{l} O_3\right)\left(\sum_{O_1=1}^{l} e^{\frac{O_1}{c^{\frac{2}{3}}}}\right)^2 && (5.402)\\
= & \frac{16l^4(l+1)\mathbb{F}_V\mathbb{F}_W}{n^{0.5}(c^{\frac{2}{3}}-1)} \frac{c^{\frac{11}{6}}e^{2c^{-\frac{2}{3}}}}{c-1} \left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2, && (5.403)
\end{aligned}$$

where we use the geometric sum formula in two places noting that we can use it as the ratios ($c^{-\frac{2}{3}}$ and $e^{c^{-\frac{2}{3}}}$) are not equal to 1 given that $c > 1$.

The case $\mathbf{k_b + k_g = 3}$ ($\mathfrak{I}_{k_b+k_g=3}$). This is possible in two ways, first $k_b = 1$ and $k_g = 2$ or the reverse. Using the notation we defined in (5.397) we represent these terms by $\mathfrak{I}_{(k_b=1)\cap(k_g=2)}$ and $\mathfrak{I}_{(k_b=2)\cap(k_g=1)}$ where $\mathfrak{I}_{k_b+k_g=3} = \mathfrak{I}_{(k_b=1)\cap(k_g=2)} + \mathfrak{I}_{(k_b=2)\cap(k_g=1)}$.

Using the fact that $\binom{O_3}{2} \leq \frac{O_3^2}{2}$ and $\binom{O_4-1}{0}\binom{O_1-O_4+1}{1}\binom{O_2-O_4+1}{1} \leq O_1O_2$ then we can then bound $\mathfrak{I}_{(k_b=1)\cap(k_g=2)}$ as follows:

$$\mathfrak{I}_{(k_b=1)\cap(k_g=2)} \leq \frac{64n^{-\frac{3}{2}}\mathbb{F}_V\mathbb{F}_W lc^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} (c\lambda)^{\sum_1^3 O_i-3l} \sum_{O_4=1}^{\min(O_1,O_2)} \frac{1}{c^{\frac{2}{3}O_4}} O_1O_2O_3^2 e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}}. \quad (5.404)$$

Observing that the summation over O_1, O_2, O_4 is equivalent to the summation in the $k_b + k_g = 2$ case, and that $\sum_{O_3} O_3^2 = \frac{1}{6}l(l+1)(2l+1)$ we obtain:

$$\mathfrak{I}_{(k_b=1)\cap(k_g=2)} \leq \frac{32l^4(l+1)(2l+1)\mathbb{F}_V\mathbb{F}_W}{3n^{1.5}(c^{\frac{2}{3}}-1)} \frac{c^{\frac{11}{6}}e^{2c^{-\frac{2}{3}}}}{c-1} \left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2. \quad (5.405)$$

For $\mathfrak{I}_{(k_b=2)\cap(k_g=1)}$ we use a similar approach and obtain the following bound from (5.397):

$$\mathfrak{I}_{(k_b=2)\cap(k_g=1)} \leq \frac{128n^{-\frac{3}{2}}\mathbb{F}_V\mathbb{F}_W l c^{\frac{11}{6}}}{c-1}\sum_{O_1,O_2,O_3=1}^{l}(c\lambda)^{\sum_1^3 O_i-3l}\sum_{O_4=1}^{\min(O_1,O_2)}\sum_{k_b=1}^{K_b^{\max}}\sum_{k_g=1}^{K_g^{\max}}\binom{O_4-1}{1} \quad (5.406)$$
$$\times\binom{O_1-O_4+1}{2}\binom{O_2-O_4+1}{2}\frac{O_3 e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}}}{c^{\frac{2}{3}O_4}}.$$

Note that $\binom{O_4-1}{1}\binom{O_1-O_4+1}{2}\binom{O_2-O_4+1}{2} \leq \frac{O_4(O_1O_2)^2}{4} \leq \frac{l^5}{4}$. Factoring out the constants we obtain:

$$\mathfrak{I}_{(k_b=2)\cap(k_g=1)} \leq \frac{32n^{-\frac{3}{2}}\mathbb{F}_V\mathbb{F}_W l^6 c^{\frac{11}{6}}}{c-1}\sum_{O_1,O_2,O_3=1}^{l}(c\lambda)^{\sum_1^3 O_i-3l}\sum_{O_4=1}^{\min(O_1,O_2)}\frac{O_3 e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}}}{c^{\frac{2}{3}O_4}}. \quad (5.407)$$

The expression for the summations over O_1, O_2, O_3 and O_4 takes the same form as in case where $k_b + k_g = 2$. Therefore:

$$\mathfrak{I}_{(k_b=2)\cap(k_g=1)} \leq \frac{16l^7(l+1)\mathbb{F}_V\mathbb{F}_W}{n^{1.5}(c^{\frac{2}{3}}-1)}\frac{c^{\frac{11}{6}}e^{2c^{-\frac{2}{3}}}}{c-1}\left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2. \quad (5.408)$$

Combining the cases of $\mathbf{k_b + k_g = 2}$ and $\mathbf{k_b + k_g = 3}$ We can then construct the case of $k_b + k_g \leq 3$ by summing (5.403), (5.405) and (5.408), to obtain:

$$\begin{aligned}\mathfrak{I}_{(k_b+k_g=2)} + \mathfrak{I}_{(k_b+k_g=3)} =& \mathfrak{I}_{(k_b+k_g=2)} + \mathfrak{I}_{(k_b=2)\cap(k_g=1)} + \mathfrak{I}_{(k_b=1)\cap(k_g=2)} \\ \leq& \frac{16l^4(l+1)c^{\frac{11}{6}}e^{2c^{-\frac{2}{3}}}\mathbb{F}_V\mathbb{F}_W}{(c^{\frac{2}{3}}-1)(c-1)n^{0.5}}\left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2\left(1+\frac{1}{n}(l^3+\frac{4}{3}l+\frac{2}{3})\right).\end{aligned} \quad (5.409)$$

The case $\mathbf{k_b + k_g > 3}$ ($\mathfrak{I}_{k_b+k_g>3}$). Substituting this condition into (5.397) we obtain:

$$\begin{aligned}\mathfrak{I}_{k_b+k_g>3} =& \frac{8n^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V lc^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} (\lambda c)^{O_1+O_2+O_3-3l} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} \\ &\times \sum_{k_g=1}^{\lceil \frac{O_3}{2} \rceil} \sum_{k_b=1}^{\min(O_4, \lceil \frac{O_2}{2} \rceil)} \mathbb{1}(k_b+k_g>3)(k_b!k_g!) \binom{O_4-1}{k_b-1} \binom{O_1-O_4+1}{k_b} \\ &\times \binom{O_2-O_4+1}{k_b} \left(\frac{2}{n}\right)^{k_b+k_g} c^{-\frac{2}{3}O_4} \binom{O_3}{k_g},\end{aligned} \tag{5.410}$$

where the indicator variable restricts the summation to the case where $k_b + k_g > 3$. Accounting for the restrictions on the summation over k_b and k_g and using the fact that $\binom{O_3}{k_g} \leq \frac{O_3^{k_g}}{k_g!}$, we can bound (5.410) from above by:

$$\begin{aligned}\mathfrak{I}_{k_b+k_g>3} \leq& \frac{8n^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V lc^{\frac{11}{6}}}{c-1} \sum_{O_1,O_2,O_3=1}^{l} \sum_{O_4=1}^{\min(O_1,O_2)} (\lambda c)^{O_1+O_2+O_3-3l} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} \\ &\times \sum_{k_b=1}^{\min(O_4, \lceil \frac{O_2}{2} \rceil)} (k_b!) \binom{O_4-1}{k_b-1} \binom{O_1-O_4+1}{k_b} \binom{O_2-O_4+1}{k_b} \left(\frac{2}{n}\right)^{k_b} c^{-\frac{2}{3}O_4} \\ &\times \sum_{k_g=\max(4-k_b,1)}^{\max(4-k_b,1,\lceil \frac{O_3}{2} \rceil)} \left(\frac{2O_3}{n}\right)^{k_g},\end{aligned} \tag{5.411}$$

where the upper limit on the final summation has been increased to guarantee that the summation is not empty. We will see in the following expression that this increase in the final summation does not effect the final outcome because of how we subsequently

bound the expression. Let us first focus on the summation over k_g in (5.411):

$$\sum_{k_g=\max(4-k_b,1)}^{\max(4-k_b,1\lceil\frac{O_3}{2}\rceil)}\left(\frac{2O_3}{n}\right)^{k_g}=\left(\frac{2O_3}{n}\right)^{\max(4-k_b,1)}\frac{1-\left(\frac{2O_3}{n}\right)^{\max(4-k_b,1,\lceil\frac{O_3}{2}\rceil)-\max(4-k_b,1)+1}}{1-(\frac{2O_3}{n})} \tag{5.412}$$

$$\leq\left(\frac{2l}{n}\right)^{\max(4-k_b,1)}\frac{1}{1-(\frac{2O_3}{n})}\leq\left(\frac{2l}{n}\right)^{\max(4-k_b,1)}\left(1-\frac{2l}{n}\right)^{-1}\leq\left(\frac{2l}{n}\right)^{\max(4-k_b,1)}(1-\nu)^{-1}, \tag{5.413}$$

where we use the fact that $2l<n$, and therefore $2\frac{O_3}{n}<1$, so that we can use the geometric sum formula. We also use the assumption $2l<\nu n$, therefore $\frac{2l}{n}<\nu$, and thus $(1-\nu)^{-1}>(1-\frac{2l}{n})^{-1}\geq(1-\frac{2O_3}{n})^{-1}$. Applying this to (5.411) we obtain:

$$\begin{aligned}\mathfrak{I}_{k_b+k_g>3}\leq&\frac{8n^{\frac{3}{2}}\mathbb{F}_W\mathbb{F}_V l c^{\frac{11}{6}}}{(c-1)(1-\nu)}\sum_{O_1,O_2,O_3=1}^{l}\sum_{O_4=1}^{\min(O_1,O_2)}(\lambda c)^{O_1+O_2+O_3-3l}e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}}\\&\times\sum_{k_b=1}^{\min(O_4,\lceil\frac{O_2}{2}\rceil)}(k_b!)\binom{O_4-1}{k_b-1}\binom{O_1-O_4+1}{k_b}\\&\times\binom{O_2-O_4+1}{k_b}\left(\frac{2}{n}\right)^{k_b}c^{-\frac{2}{3}O_4}\left(\frac{2l}{n}\right)^{\max(4-k_b,1)}.\end{aligned} \tag{5.414}$$

We know that $\binom{O_4-1}{k_b-1}\binom{O_1-O_4+1}{k_b}\binom{O_2-O_4+1}{k_b}k_b!\leq\frac{(O_4-1)^{k_b-1}}{(k_b-1)!}\frac{(O_1)^{k_b}}{(k_b)!}(O_2)^{k_b}\leq\frac{(O_1O_2O_4)^{k_b}}{k_b!}$. Therefore substituting this in to (5.414) and rearranging we obtain:

$$\begin{aligned}\mathfrak{I}_{k_b+k_g>3}\leq&\frac{8n^{\frac{3}{2}}\mathbb{F}_V\mathbb{F}_W l c^{\frac{11}{6}}}{(c-1)(1-\nu)}\sum_{O_1,O_2,O_3=1}^{l}(c\lambda)^{\sum_1^3 O_i-3l}e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}}\sum_{O_4=1}^{\min(O_1,O_2)}c^{-\frac{2}{3}O_4}\\&\times\sum_{k_b=1}^{\min(O_4,\lceil\frac{O_2}{2}\rceil)}\frac{(O_1O_2O_4)^{k_b}}{k_b!}\left(\frac{2}{n}\right)^{k_b}\left(\frac{2l}{n}\right)^{\max(4-k_b,1)}.\end{aligned} \tag{5.415}$$

Considering only the summation over k_b, and splitting the summation in two, we obtain:

$$\sum_{k_b=1}^{2} \frac{(O_1O_2O_4)^{k_b}}{k_b!}\left(\frac{2}{n}\right)^{k_b}\left(\frac{2l}{n}\right)^{4-k_b} + \sum_{k_b=3}^{\min(O_4,\lceil\frac{O_2}{2}\rceil)} \frac{(O_1O_2O_4)^{k_b}}{k_b!}\left(\frac{2}{n}\right)^{k_b}\left(\frac{2l}{n}\right) \tag{5.416}$$

$$=O_1O_2O_4\left(\frac{2}{n}\right)^4 l^3 + \frac{(O_1O_2O_4)^2}{2}l^2\left(\frac{2}{n}\right)^4 + \frac{2l}{n}(\frac{2O_1O_2O_4}{n})^3 \sum_{k_b=0}^{\min(O_4,\lceil\frac{O_2}{2}\rceil)-3} \frac{(\frac{2O_1O_2O_4}{n})^{k_b}}{(k_b+3)!}. \tag{5.417}$$

For $k_b \geq 0$, $(k_b+3)! = (k_b!)(k_b+1)(k_b+2)(k_b+3) \geq 6(k_b!)$, so therefore we can bound the summation over k_b as follows:

$$O_1O_2O_4\left(\frac{2}{n}\right)^4 l^3 + \frac{(O_1O_2O_4)^2}{2}l^2\left(\frac{2}{n}\right)^4 + \frac{l}{3n}(\frac{2O_1O_2O_4}{n})^3 \sum_{k_\beta=0}^{K_b^{\max}-3} \frac{(\frac{2O_1O_2O_4}{n})^{k_b}}{(k_b)!} \tag{5.418}$$

$$\leq l^4\left(\frac{2}{n}\right)^4\left(l^2 + \frac{l^4}{2} + \frac{l^6}{6}e^{(\frac{2O_1O_2O_4}{n})}\right) \tag{5.419}$$

$$\leq l^4\left(\frac{2}{n}\right)^4\left(l^2 + \frac{l^4}{2} + \frac{l^6}{6}\right)e^{(\frac{2O_1O_2O_4}{n})} \leq \left(\frac{2}{n}\right)^4 \frac{l^{10}}{2}e^{(\frac{2l^3}{n})}, \tag{5.420}$$

where the final inequality uses the fact that $l > 1$, so that $\frac{l^2}{3} > 1$ and $\frac{l^4}{6} > 1$. Substituting this bound on the summation over k_b into (5.415) we obtain:

$$\mathfrak{I}_{k_b+k_g>3} \leq \frac{8n^{\frac{3}{2}}\mathbb{F}_V\mathbb{F}_W l c^{\frac{11}{6}}}{(c-1)(1-\nu)}\left(\frac{2}{n}\right)^4 \frac{l^{10}}{2}e^{(\frac{2l^3}{n})} \sum_{O_1,O_2,O_3=1}^{l} (c\lambda)^{\sum_1^3 O_i - 3l} e^{\frac{O_1+O_2}{c^{\frac{2}{3}}}} \sum_{O_4=1}^{\min(O_1,O_2)} c^{-\frac{2}{3}O_4}. \tag{5.421}$$

We note that the summations over O_1, O_2, O_4 are now identical to the summations we saw in the case of $k_b = k_g = 1$. Thus:

$$\mathfrak{I}_{k_b+k_g>3} \leq \frac{8n^{\frac{3}{2}}\mathbb{F}_V\mathbb{F}_W l c^{\frac{11}{6}}}{(c-1)(1-\nu)}\left(\frac{2}{n}\right)^4 \frac{l^{10}}{2} e^{(\frac{2l^3}{n})} e^{2c^{-\frac{2}{3}}} \frac{\left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2}{(c^{\frac{2}{3}}-1)} \sum_{O_3=1}^{l} 1 \tag{5.422}$$

$$= \frac{64\mathbb{F}_V\mathbb{F}_W c^{\frac{11}{6}} l^{12} e^{(\frac{2l^3}{n})}}{n^{\frac{5}{2}}(c^{\frac{2}{3}}-1)(c-1)(1-\nu)} e^{2c^{-\frac{2}{3}}} \left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2. \tag{5.423}$$

Combining this with the case of $k_b + k_g < 4$ from (5.409) we obtain:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq \mathfrak{I}_{k_g+k_b=2} + \mathfrak{I}_{k_g+k_b=3} + \mathfrak{I}_{k_b+k_g>3} \tag{5.424}$$

$$\leq \frac{16\mathbb{F}_V\mathbb{F}_W e^{2c^{-\frac{2}{3}}} c^{\frac{11}{6}}}{n^{0.5}(c^{\frac{2}{3}}-1)(c-1)} \left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2 \left(l^4(l+1)\left(1+\frac{1}{n}\left(l^3+\frac{4}{3}l+\frac{2}{3}\right)\right)\right. \tag{5.425}$$

$$\left. + \frac{4l^{12}e^{(\frac{2l^3}{n})}}{n^2(1-\nu)}\right).$$

Further, we know that $\mathbb{F}_V\mathbb{F}_W = ((1-\nu)^{2l-2}(1-p))^{-\frac{3}{2}}(\frac{8}{l(l-1)})^{\frac{3}{2}} = 16\sqrt{2}(1-\nu)^{3-3l}(l(l-1)(1-p))^{-\frac{3}{2}}$. Further, $l \geq 3$ therefore $\frac{l^2}{9} \geq 1$ and $\frac{l^3}{27} \geq 1$, therefore $l^3 + \frac{4}{3}l + \frac{2}{3} \leq l^3(1 + \frac{4}{27} + \frac{2}{81}) \leq \frac{6}{5}l^3$. Applying these inequalities we obtain:

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq (1-\nu)^{3-3l} \frac{l^{\frac{5}{2}}}{(l-1)^{\frac{3}{2}}} \frac{n}{(n-c)^{\frac{3}{2}}} \frac{256\sqrt{2}e^{2c^{-\frac{2}{3}}} c^{\frac{11}{6}}}{(c^{\frac{2}{3}}-1)(c-1)} \tag{5.426}$$

$$\times \left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2 \left((l+1)\Big(1+\frac{6l^3}{5n}\Big) + \frac{4l^8 e^{(\frac{2l^3}{n})}}{n^2(1-\nu)}\right)$$

Therefore, let $\mathfrak{E} = (1-\nu)^{3-3l} \frac{l^{\frac{5}{2}}}{(l-1)^{\frac{3}{2}}} \frac{256\sqrt{2}e^{2c^{-\frac{2}{3}}} c^{\frac{11}{6}}}{(c^{\frac{2}{3}}-1)(c-1)} \left(\frac{1-e^{lc^{-\frac{2}{3}}}}{1-e^{c^{-\frac{2}{3}}}}\right)^2$ then,

$$d_{H_W}(\tilde{\zeta}^{(l)}, Z) \leq \frac{n\mathfrak{E}}{(n-c)^{\frac{3}{2}}} \left((l+1)\Big(1+\frac{6l^3}{5n}\Big) + \frac{4l^8 e^{(\frac{2l^3}{n})}}{n^2(1-\nu)}\right) \tag{5.427}$$

which is the result in the theorem. QED.

5.4.2.3 Performance of the general l bound.

To quantify the performance of the general l bound we plot the distributional distance against the number of nodes, for c fixed at 100. For each l we fix ν at $\frac{2l}{n} + 0.001$ which satisfies the requirements of theorem. Note, that by decreasing the constant factor we can obtain slightly better results. Note that because of the bounding approximations we used, the expression no longer depends on λ. Thus rather than showing the performance with respect to λ, we show the performance with respect to l. The results can be seen in Figure. 5.3.

We note, that the range of n where we can prove that the weighted sum of paths can be approximated by a normal distribution is much more constrained than in the case of $l = 2$. We are interested in a CDF bound, and thus we require the distributional distance to be less than 1 and preferably much smaller than that. For $c = 100$ the minimum size of network where this approximation makes sense is around 10^{11} for $l = 3$, although values around 10^{14} seem to be more reasonable for $l = 3$ with higher values of n required for higher values of l.

This is likely to be due at least in part a result of the cruder approximations we used in order to produce an expression general l ($l > 2$). However, this could also reflect the fact that with increasing l the approximation does simply get worse, for the same n, as the level of dependence increases with increasing path length.

We note that the error is dominated by the term with the largest power of n for much of the range of interest. In Figure. 5.4, we plot the percentage of the distributional distance bound which is accounted for by the leading order term. We note that for $n > 10^6$, that the leading order term dominates, up to $l = 15$. Further, we note that this show that the shift initial drop in the bound from $n = 10^3$ to $n = 10^5$, is clearly accounted for by the non-leading order terms.

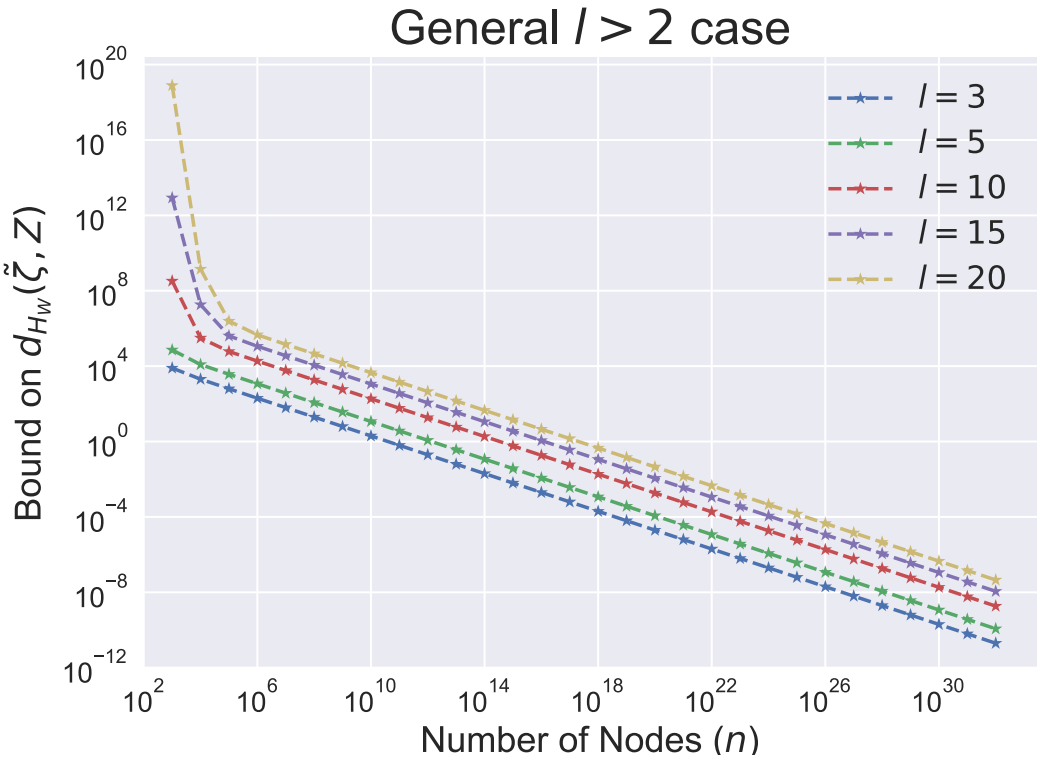

Figure 5.3: Plot shows the largest possible values for d_{H_W} (Wasserstein distance) using the result from Theorem 26 between the standard normal and the weighted sum of paths of length l, with $c = 100$

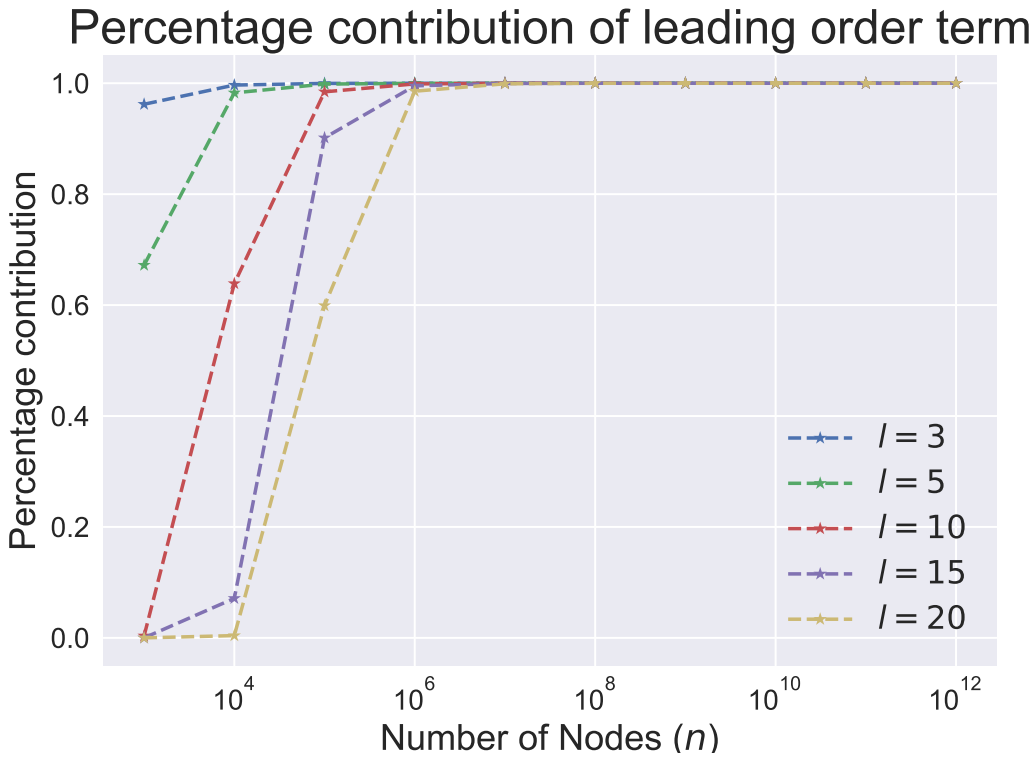

Figure 5.4: Percentage of the distributional distance (Wasserstein) between the standard normal and the weighted sum of paths of length l which is due to the leading order term with $c = 100$. **Note the range on X axis is different than the previous plot.**

In conclusion, we have shown that there is an order $\frac{1}{\sqrt{n}}$ bound and derived an expression for the constant. However, the bound is currently too crude to be useful in the analysis of real world networks, as we require networks that are prohibitively large before this bound will reduces to a practical level.

5.5 Application: Significance threshold of an given community structure.

We can use our bounds on the bisection cut in our measure of modularity to measure it's significance in the following way. Due to the superior bound we will focus on the case of $l = 2$.

First, let us recall the specific form of the bisection cut we are interested in:

$$Q_2^{Path-Diff}(g) = \sum_{\Upsilon=1}^{2} \lambda_1^{\Upsilon} \sum_{i,j} \left(A_{ij\Upsilon} - \lambda_2 \frac{(n-2)!}{(n-1-\Upsilon)!} p^{\Upsilon} \right) \delta(g_i \neq g_j). \tag{5.428}$$

We know from (5.23) that we can write this in terms of ζ by setting $\lambda = \lambda_1$, where ζ is our random variable for the weighted sum of paths. Note, as we are focusing on the case of $l = 2$ we use the notation from the $l = 2$ case, thus we use ζ rather than $\zeta^{(2)}$ (note that $\zeta = \zeta^{(2)}$).

$$Q_2^{Path-Diff}(g) = \zeta - \lambda_2 \left(\sum_{\Upsilon=1}^{2} \lambda^{\Upsilon} \sum_{i \in V_1} \sum_{j \in V_2} \frac{(n-2)!}{(n-1-\Upsilon)!} p^{\Upsilon} \right). \tag{5.429}$$

For simplicity let us focus on the case where $\lambda_2 = \lambda$. From (5.68) and Lemma 4 we know:

$$\tilde{\zeta} = \frac{\zeta - E[\zeta]}{(\mathrm{Var}(\zeta))^{0.5}} = \frac{1}{(\mathrm{Var}(\zeta))^{0.5}} \left(\zeta - \lambda \frac{n^2}{4} p - \lambda^2 \frac{n^2}{4} (n-2) p^2 \right). \tag{5.430}$$

Therefore, $\zeta = \tilde{\zeta}(\mathrm{Var}(\zeta))^{0.5} + \lambda \frac{n^2}{4} p + \lambda^2 \frac{n^2}{4} (n-2) p^2$. Observing that $E[\zeta] = \lambda \frac{n^2}{4} p + \lambda^2 p^2 \frac{n^2}{4} (n-2)$ is the expectation of ζ under null model in $Q_2^{Path-Diff}(g)$, we obtain:

$$Q_2^{Path-Diff}(g) = \tilde{\zeta}(\mathrm{Var}(\zeta))^{0.5}. \tag{5.431}$$

We can now use our bound on the Wasserstein distance between $\tilde{\zeta}$ and a standard normal random variable to compute a threshold for size of a bisection cut required for the result to be significant. Thus we are interested in determining a value $t = t(s^*)$ such that for a given s^*,

$$P\left(Q_2^{Path-Diff}(g) \geq t \right) \leq s^*. \tag{5.432}$$

For example if we set $s^* = 0.05$ then if we observe $Q_2^{Path-Diff}(g)$ greater than $t(s^*)$ than we know that the p-value of the test is at most 0.05 and therefore it is significant at the 5% level. Such a statement is known as a tail bound, and because of the relation we derived between $\tilde{\zeta}$ and the normal distribution we can use normal tail bounds.

Constructing the threshold is equivalent to bounding the CDF of a distribution. Recall from Section 2.7.1 that the Kolmogorov distance ($d_{H_K}(u, v)$) is the largest difference in absolute value between the CDF of two probability distributions.

Therefore we relate our bound in Wasserstein distance to a bound in Kolmogorov distance. To do so we use a result from Ref. [142] recall from (2.20)

$$d_{H_K}(X, Y) \leq \left(\frac{2}{\pi}\right)^{\frac{1}{4}} \sqrt{d_{H_W}(X, Y)}. \tag{5.433}$$

where X and Y are random variables. Note, following Ref. [142] we again use a somewhat loose form of notation, where for a random variable X with distribution u and a random variable Y with distribution v we let $d_H(u, v) = d_H(X, Y)$. Applying this result we obtain:

$$|P(Z \geq t) - P(\tilde{\zeta} \geq t)| \leq d_{H_K}(Z, \tilde{\zeta}) \leq \left(\frac{2}{\pi}\right)^{\frac{1}{4}} \sqrt{d_{H_W}(Z, \tilde{\zeta})} \tag{5.434}$$

$$P(\tilde{\zeta} \geq t) \leq P(Z \geq t) + \left(\frac{2}{\pi}\right)^{\frac{1}{4}} \sqrt{d_{H_W}(Z, \tilde{\zeta})} \tag{5.435}$$

To bound this equation we need a bound for $P(Z \geq t)$. We consider the bound from [140]:

$$P(Z \geq t) \leq e^{-\frac{t^2}{2}} \tag{5.436}$$

We want to find t such that $P(Z \geq t) + \left(\frac{2}{\pi}\right)^{\frac{1}{4}} \sqrt{d_{H_W}(Z, \tilde{\zeta})} \leq s^*$. We can solve for t by simple rearrangement. Applying the tail bound and the bound on $d_{H_W}(\tilde{\zeta}, Z)$ in

Theorem 18 to (5.435) we obtain:

$$P(\tilde{\zeta} \geq t) \leq e^{-\frac{t^2}{2}} + \left(\frac{2}{\pi}\right)^{\frac{1}{4}} \sqrt{8\frac{1 + 5c\lambda + 30c^2\lambda^2\max(1,\lambda) + 11(c\lambda)^3 + \frac{44c^3\lambda^2}{n}\left(\frac{2}{5}\lambda c + \frac{1}{3}\right)}{c^{0.5}(n-c)^{0.5}\Big(1 + c\lambda^2(2c-7)\Big)^{1.5}}}. \tag{5.437}$$

Therefore for an arbitrary significance level s^* we require $t^*(s^*, n, \lambda)$ such that:

$$e^{-\frac{t^2}{2}} + \left(\frac{2}{\pi}\right)^{\frac{1}{4}} \sqrt{8\frac{1 + 5c\lambda + 30c^2\lambda^2\max(1,\lambda) + 11(c\lambda)^3 + \frac{44c^3\lambda^2}{n}\left(\frac{2}{5}\lambda c + \frac{1}{3}\right)}{c^{0.5}(n-c)^{0.5}\Big(1 + c\lambda^2(2c-7)\Big)^{1.5}}} \leq s^*. \tag{5.438}$$

Note, t^* is well defined when the bounded Kolmogorov distance is less than the desired significance level. If the bound is larger than the significance level then it is clearly impossible to find a threshold as the resolution on the bound is lower than the desired significance.

Putting this all together we obtain:

$$P\left(\tilde{\zeta} \geq t^*(s^*, n, \lambda)\right) \leq s^*, \tag{5.439}$$

$$\text{so that } P\left((Var(\zeta))^{0.5})\tilde{\zeta} \geq (Var(\zeta))^{0.5})t^*\right) \leq s^*, \tag{5.440}$$

$$\text{and hence } P\left(Q_2^{Path-Diff} \geq (Var(\zeta))^{0.5})t^*(s^*, n, \lambda)\right) \leq s^*. \tag{5.441}$$

To demonstrate the result we show an example. We set $c = 100$ and $\lambda = 1$ and we compute the value of t^* such that (5.438) holds with equality for $s^* = 0.05$. Computing t^* from (5.438) uses the tail bounds described in (5.436), for comparison we also compute the t^* for the case where $\tilde{\zeta}$ is perfectly normally distributed, and an explicit CDF result (computed using the SciPy normal CDF function [89]) to indicate how far the tail bounds diverge. The results can be seen in Figure 5.5. Note, we plot t^* rather than $t^*Var(\zeta)$ as the variance term dominates the bound as it is linear in n (Theorem 7).

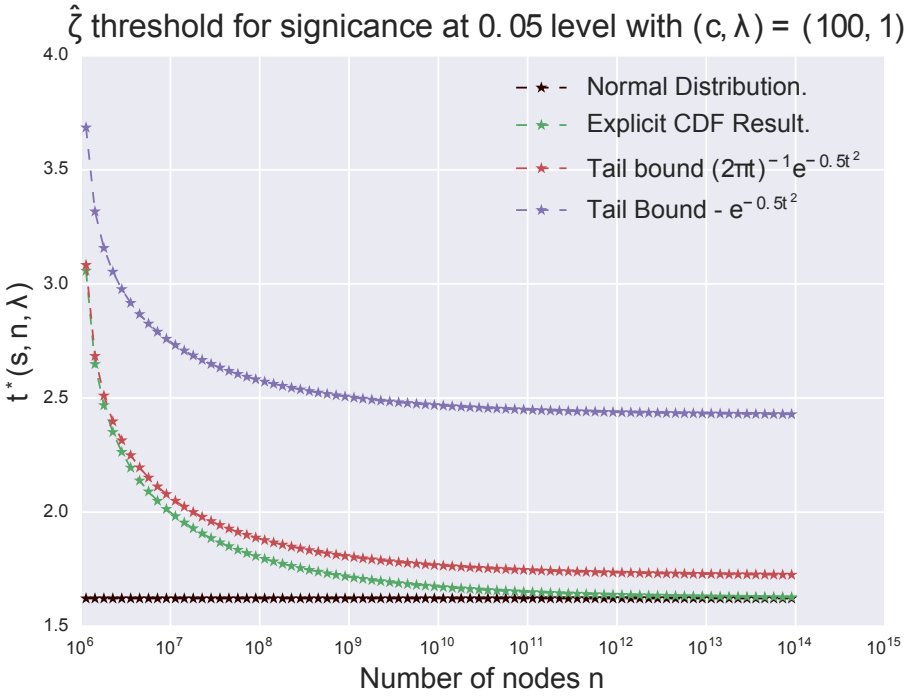

Figure 5.5: Threshold for $\tilde{\zeta}$ (standardised weighted cut) to be significant at the 5% level using the normal approximation and (5.438). The Normal distribution serves as a baseline for the case where $\tilde{\zeta}$ is perfectly normally distributed (black). The CDF bound (green) uses the normal CDF function from the SciPy package [89] to numerically compute the bound. The other bounds pair of tail bounds one of which is solved numerically (red) from [56] and the other analytically (blue) [140]. The underlying graph is assumed to be a sparse Erdős Rényi network with $p = \frac{100}{n}$. The parameters in the modified modularity are $\lambda_1 = \lambda_2 = 1$.

We observe that for $n > 10^6$ we can use our result to compute a t^*, and for $n > 10^{10}$ the difference between the bound on $\tilde{\zeta}$ using the explicit CDF and a standard normal is very small. A better bound on the normal from [56] tail is given by:

$$P(Z \geq t) \leq \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}t} \tag{5.442}$$

but in this case we cannot solve analytically. Replacing the tail bound in (5.438) with this bound and using numerical approximation gives substantially smaller values of t^*, see Figure 5.5.

5.6 Conclusion

In this chapter we generalised our use of path modularity to non sampled networks using the uniform null model from Refs. [101, 137, 169]. We analytically explored this null model in two ways.

In Section 5.2.1 we showed that paths of length 2 do contain useful information in stochastic block models by exploring the case of two even blocks and equal within block connection probability. We demonstrate that the expected difference between the number of paths between nodes in the same community and different communities is large for certain parameter sets. However, the expected difference is strongly dependent on the parameters of the model.

The bulk of the work in this chapter relates the new modularity to a normal distribution. To be more precise, for a fixed community structure of two blocks of size $\frac{n}{2}$ we show that the weighted sum of paths between the blocks is asymptotically normally distributed for $l = 2$ (Theorem 18) and for $3 \leq l < \frac{n}{2}$ (Theorem 26). The bound on the rate of convergence in Wasserstein distance in both cases is given by $\frac{1}{\sqrt{n}}$. We also constructed analytic bounds for the Wasserstein distance. In the case of $l = 2$ these bounds indicated that for realistic graph sizes we can use the normal distribution to approximate the summation after suitable adjustment.

In the general l case the bound is unrealistic large for reasonable graph sizes (Figure 5.3). After investigating whether this large bound stems from the leading order term or from another term with a large coefficient we discovered that indeed the bulk of the contribution to the bound was from the leading order term (Figure 5.4). As we took special care to minimise the size of the leading order term in Theorem 26 it is likely that better bounds may require a different strategy. One good place to start could be the bounds used in Lemma 25.

While we believe that this work is of interest in its own right, it can be used in a number of ways. Firstly this result can be used in a similar manner to the result

by Franke and Wolfe [63] to evaluate the significance of a predefined split. Further, the statistic can be used as test statistic for the null hypothesis that a graph comes from an Erdős Rényi graph, against the alternative that it comes from a simple stochastic block with two groups with equal within group probabilities. The normal approximation can then be used to approximate the distribution.

To allow us to apply this result to more problem sets, the next stage in this work will be to expand the analysis to groups of uneven size, and to explore the general l case further to attempt to obtain a better coefficient. Moreover we can prove results for the modified modularity directly not only for the weighted cut.

Chapter 6

Conclusion and open problems

The methods and results reported in this thesis lie within the emerging and interdisciplinary field of network science, and are motivated by an interest in network pharmacology. While some of the techniques developed here are relevant to the analysis of protein-protein interaction networks, and hence network pharmacology, they serve more broadly to enhance the set of tools for network science that can be applied in a variety of problem domains.

More specifically, the insights developed in this thesis have focused around the exploration of sampling and community detection techniques with a view to incorporating path information. The main contributions include: (1) constructing a significance test for sampled networks that accounts for a key factor (namely redundant seed nodes); (2) the use of this test to select between sampled networks for further study; (3) the use of a variant of a component of this test, and a modified modularity method, to detect communities in sampled networks; (4) analytic results based on this modified community detection method demonstrating that a statistic related to our modified community detection approach is asymptotically normally distributed.

We will now describe these contributions in more detail, before we sketch some ideas for future work.

6.1 Summary of work

Selecting a Sampling Method In Chapter 3 we identified a problem encountered in selecting sampled networks related to a disease from PPI networks. The problem is that there are many methods for constructing sampled networks, and it can be difficult to choose between them. To address this problem we provided a significance test for sampled networks, based on randomly selected seed lists, while adjusting for a key factor (i.e. namely redundant seed nodes).

Redundant seed nodes are nodes on a seed list that do not contribute to the sampled network, and thus can be removed without affecting the sampled network structure. While they do not affect the structure of the sampled network, they do affect the null model, as they artifically increase the size of the seed list without adding any structure to the sampled network. Thus, if we want to measure the significance of a network statistic, we must account for redundant seed nodes. We did this by forming the minimum seed list, which is the shortest list of nodes that would generate the same sampled network, and then using this seed list to compute significance.

We demonstrated the importance of the minimum seed list in three ways. First, we showed, based on exact results (extended from the results in Ref. [61]), that increasing or decreasing the size of the seed list can have a large impact on the expectation and variance of the number of nodes in a two-hop snowball-sampled network. Second, we showed that if we construct the longest seed list that generates the same network, and compare the apparent significance under this and the original seed list, then large differences in p-value are possible. Third, we showed that a real world example, namely a network constructed from a seed list related to Parkinson's disease, exhibits the same issue. Thus, we demonstrated that adjusting for the redundant seed nodes is important when measuring the significance of network features in these networks. We constructed a null model around this idea of a minimum seed list, constructing the smallest subset of nodes that generate the same network, and then finding an

ensemble by sampling the original network with randomly selected seed list with the same length as the original seed list.

Further, we showed that our null model compares favourably to the configuration model on sampled networks. We demonstrated that the distribution of p-values over graphs sampled with random seed lists for two commonly used network statistics is approximately uniform in most cases (with the exception of cases where we would not expect it to be uniform due to a non-continuous distribution). By contrast, the configuration model is non-uniform in all cases, indicating that it is a poor choice for this system.

We believe that the null model is interesting in its own right, as testing for the significance of a feature such as the clustering coefficient in sampled networks is a relevant problem. Our results are also useful in the context of the overall aim of this chapter, which is to provide a framework for selecting between different sampled networks.

We constructed a test based on a basket of network statistics, to identify whether networks can be distinguished from random. We hypothesised that networks that differ significantly from a suitably defined null model may also be of greater interest, in the sense that they should be prioritised in further analysis.

We verified this hypothesis using a benchmark based on the stochastic block model, demonstrating empirically that for the parameters tested lower p-values in our test were in almost all cases correlated with higher values of two commonly used statistics (purity and accuracy) to measure the how successful the identification of the groups is.

We illustrate with our benchmark that tradeoffs between purity and accuracy can inform the selection between sampled networks.

It should noted that strict comparisons were only possible within a given sampling technique. Although we showed how networks generated by different sampling

techniques could be compared heuristically, some caution is required since in some cases comparisons may be problematic.

Finally, we used this approach to select between twelve sampled network related to Parkinson's disease, which allowed us to identify two networks that should be prioritised for further study.

Community Detection in Path Sampled Networks A commonly used and powerful tool in network science is community detection. Community detection algorithms divide a network into groups of nodes that are more densely connected within groups, and more sparsely connected between groups, than would be expected at random. These algorithms group together nodes using the graph structure, and in certain cases on this basis also can group together nodes that are related to a biological process (see for example Ref. [108]), which is why they have the potential to inform network pharmacology.

Therefore, it is natural for us to be interested in community detection on our sampled networks. However, community detection approaches suffer from the same issue as we observed in the previous section, namely that they do not account for the sampling structure.

To construct a community detection method that accounts for the sampling structure we followed Refs. [53, 115, 147] and constructed a null model for modularity based community detection. Due to the good performance of the shortest-path sampling technique in Chapter 3, and the fact that computational experiments indicated that community detection was strongly affected by the bias introduced by shortest-path sampling, we decided to focus on shortest-path sampling. In Chapter 4, we constructed a community detection quality function related to modularity, that incorporates simple paths and ensembles of random sampled networks using a similar approach to Chapter 3. We found that this approach outperformed all others tested

on a benchmark constructed for the purpose of comparison.

We constructed this benchmark by following a procedure similar to the one considered in Chapter 3, where we generated a stochastic block model and sampled from it. It is worth noting that there is a key difference between these two benchmarks. In the case of Chapter 3, we have two communities, one of which is "correct", and therefore measuring success is simple. However, in Chapter 4, we are interested in the community structure of the sample after adjusting for the sampling procedure, and thus it is non-trivial to embed a community structure. To address the problem of inducing a community structure we made the following assumption: that for a sufficiently strong community structure, the community structure of the sample should reflect the community structure of the wider network. Under this assumption we can then test the success of a community detection method by varying the strength of the community structure of the wider network.

We found that our path-based method outperforms all other null models tested including: (1) the standard Newman-Girvan null model with a resolution parameter; (2) a variant of our novel null model restricted to paths of length one; (3) a further variant which multiplicatively varies the standard deviation rather than the mean of the connection probability.

We also compared our method against stability, a different but related quality function to modularity, which also incorporates longer path lengths. We found that stability performs better than the other approaches based on paths of length one (edges), with respect to recovering the community structure in the sampled network, that is induced by the community structure of the whole network. However, stability is outperformed by our method. This indicates that incorporating longer path lengths appears to be important in performing well on this benchmark.

Path Based Community detection We took the novel quality function from Chapter 4 and generalised it to apply to non-sampled networks, by replacing the null model for shortest path sampled networks with a uniform null model. With the new general community detection approach we demonstrated that the weighted cut (i.e. the change in path modularity between a community that contains all nodes and a fixed partition of the nodes) is asymptotically normally distributed over the ensemble of sparse Erdős Rényi graphs (with probability of connection $p = \frac{c}{n}$, where n is the number of nodes and c is a fixed constant) as $n \to \infty$.

We showed this by relating the statistic to a sum of weighted (dependent) paths. We employed Stein's method, an approach that is useful when assessing the limiting behaviour of sums of non-independent variables. Using this approach we demonstrated and explicitly bounded the Wasserstein distance between our statistic and the standard normal distribution.

For the proof we used a result from Barbour *et al.* [14], which bounds the Wasserstein distance between a sum of random variables and a normal distribution under certain conditions. Using this result we demonstrated that the Wasserstein distance between a standardised weighted sum of paths and a normal distribution goes to 0 as $n \to \infty$. We derived an explicit bound (in Wasserstein distance) for both the case of $l = 2$ and for general $2 < l < \frac{n}{2}$ and in both cases we obtain a bound of the order $\frac{1}{n^{0.5}}$.

For the specialised bound where $l = 2$, the bound performs well with a relatively small Wasserstein distance for moderate graph sizes. While the general case shows the same convergence, the result is weaker.

We believe that our result is interesting in its own right, that it and extensions thereof allow a few interesting applications. It can be used to compute bounds on the significance (p-value) of a given weighted sum over the graph ensemble, for a known split of the node set into two sets of equal size (following the more general work of

Franke and Wolfe [63]). Further, in cases where we wish to test if a network has a simple block model structure we can use this as a test statistic for the null hypothesis that the network is an Erdős Rényi graph against a general alternative. Again in this case the explicit bound to the normal distribution allows faster computation.

6.2 Outlook and future work

This thesis advances current methods in significance testing on sampled networks, and in community detection in both sampled and non-sampled networks. It also contributes important and useful tools for selecting between sampled networks for further study. The contributions are computational as well as analytical.

There are several possible extensions of the results presented in this thesis. In the following section we will explore some interesting and promising avenues for future research. We will first discuss particular strands of future work that arise from each chapter, and then consider some of the larger challenges in the field that potential extensions of our work may be able to help address.

6.2.1 Chapter 3: Selecting a sampling method

We group future directions for this chapter into two categories: biological applications and methodological advances.

Biological Applications One clear biological application that was not explored in this thesis is a more complete analysis of the Parkinson's Disease network that we constructed. This would involve exploring the significance of other network features to see if the networks highlighted in Chapter 3 had any other properties that appear to be special using the significance test we developed. It would also involve a more in depth study of the nodes involved, for example by looking at functional annotation of the nodes in these networks.

A further interesting analysis (inspired by [68]) could compare the results from different disease conditions to look for common patterns in significant networks.

Methodological Advances One of the features of PPI networks that we did not account for is error in the data. The interactions reported in PPI databases can suffer from large numbers of false positives and false negatives [181]. Building methods that can account for this uncertainty should be very valuable (see for example [117]).

One idea would be to construct probabilistic versions of the sampling techniques, and thus rather than having a single sampled network, obtain a probability distribution for the set of all possible sampled networks. We could then construct a significance test and therefore a prioritisation scheme in a similar manner to the error-free case by testing whether (say) the average of a summary statistic over the sampled ensemble is significant given the sampling scheme.

Another methodological issue is to reduce the computational complexity of the approaches we discussed in this work. Rather than implicitly using information from the full network by looking at the distribution of summary statistics of sampled networks, we could directly use information from the overall network. For example, a statistic in a sampled network could be estimated using the overall network. The analytic results in Chapter 3 are a first step towards this for the number of nodes in a snowball sampled network. It may be possible to extend this to other statistics and sampling techniques. This could speed up the assessment of significance, as Monte Carlo approach can be very computationally intensive.

6.2.2 Chapter 4 Community detection in sampled networks

In Chapter 4 we constructed a method to detect communities in networks sampled by shortest-path sampling. This again can be split into advances with respect to specific applications and methodological advances.

Applications Clearly the first step for this work is to take the method into an application domain using sampled networks, and attempt to validate the approach. This could involve sampled networks from a biological system (for example the networks we discussed in Chapter 3), or this could involve a different path-sampled network (for example trace-route networks), or other networks that are constructed from individual paths (for example FourSquare, a website that allows individuals to check into geographical locations).

Methodological Work As mentioned in Chapter 4, some other approaches do also use information contained in longer paths in order to perform community detection. It would be interesting to compare these approaches with our approach for sampled networks in a systematic study. This may help to identify why it appears that community detection methods which take into account paths of length greater than one appear to outperform other approaches on this benchmark.

6.2.3 Chapter 5: Path based community detection

In this chapter many of the contributions are methodological. There are several further advances we could make to the mathematics of the chapter.

First we could optimise the current bounds. In the case of the general bound ($2 < l < \frac{n}{2}$), it should be possible to gain large improvements in the explicit bounds obtained. Some areas where this would be easily possible include the variance and some of the counting arguments we used to compute a bound on the Wasserstein distance. In the case of $l = 2$ it also should be possible to improve the bound, although there is less scope, since we have already considered this bound in detail.

A further advance would be to extend the results to a case where we have more groups that could be differently sized. This should not be difficult given the results that we have already obtained. While it would require a stronger condition on some

of the variables, much of the mathematics would be very similar. This would also make the result much easier to apply to real world networks, as they rarely divide exactly into two groups of equal size.

Another interesting advance would be to consider the path modularity directly. This again would only require a minor extension to the proof, as this only changes the nodes on either side of the path.

6.2.4 Future work on general community detection methods.

Modularity maximisation is a very flexible framework that allows the inclusion of arbitrary null models. Thus, as we saw in Chapter 4, this allows the detection of communities while accounting for certain features of the network, as is the case for geographic networks [53], correlation networks [115], geographic correlation networks [147], or in the case of Chapter 4 shortest-path sampled networks. Attempting to do this with other types of objective functions is more difficult.

Modularity-based approaches have a problem known as the resolution limit, which is the inability of these methods to observe communities on multiple scales [58].

In response to this problem, so called multi-resolution methods were derived (for example [8, 137, 169]) which mitigate some of the limitations from Ref. [58]. However, issues still persist [74]. Further, Ref. [102] observed that while these methods may be able to change the size of the communities, they can be problematic because real-world networks consist of different sized communities, which will not all be in scope in multiresolution methods.

A reason why modularity is so powerful is that it is able to detect communities after adjusting for a null model, without specifying a complete model. Only the expectation of the number of edges is specified. By contrast, when fitting a stochastic block model (see for example [92]), we specify the complete distribution and attempt to compute a partition that maximises the likelihood over it. The cost of not specify-

ing the complete model means that modularity implicitly or explicitly makes certain assumptions. Different community detection approaches make different assumptions, and thus may result in different estimated community structures.

Thus, a good future direction is to use techniques that are not constrained by restrictions such as the resolution limit, but that allow us to detect communities with arbitrary null models. However, Ref. [102] claims that all global methods suffer from similar limitations, but to greater or lesser extent.

A possible solution could be to simply specify the community size distribution as a parameter of the model, and then maximise either a suitably adjusted modularity or (perhaps even better) a stochastic block model. Some evidence from Ref. [40] indicates that simply knowing the number of communities can vastly improve the performance of modularity-based methods.

Another possibility would be to attempt to understand the limitations of modularity with a generic null model of the form $\sum_{ij}(A_{ij} - P_{ij})$ for an arbitrary P_{ij}. Indeed using a similar approach to Ref. [102], we have worked with a very simple block model with three groups of size n_1, n_2 and n_3, where the weight of each edge is given by p_i within community i and $p_i p_j$ between communities i and j. In some preliminary work we have observed that it is possible to generate networks for which it is impossible to construct a null model for modularity which can detect it under the following assumption. Namely, that $P_{ij} \geq P_{i'j}$ if $k_i > k_{i'}$, where k_i is the degree of node i regardless of the resolution parameter(s). Note, this is a condition that all major null models adopt.

If we were able to fully understand these limitations, we could continue to use modularity until we succeed in developing an approach which does not have it's flaws but allows a generic null model.

6.3 Closing thoughts

The field of network science has moved at a remarkably quick pace and the work in the thesis was aimed to contribute to that.

In this thesis, we have developed several methods which we hope are useful to network pharmacology. However, this is just the first step, since computational approaches on their own can only go so far.

Without validation from biologists, and iterative improvement of algorithms and approaches in conjunction with new experimental findings guided by theory, these methods cannot reach their potential.

Appendix A

Appendix - Chapter 3

A.1 Seed Lists Tables

This section contains the tables containing the seed lists and the conversions into BioGrid identifiers. Note, some of the identifiers convert to more than one node in BioGrid, and some are not present.

The seeds for OMIM can be found in Table A.1. In the case of the table for the expression seed list copyright issues prevent us from combining the original table from Ref. [36] and conversions we have made from this table into gene names and subsequently into BioGrid IDs. Thus, we present the seed from the expression seed list in two tables, one a copy of the table from Ref. [36] (Table A.2). and the second our conversions (Table A.3).

Note that when constructing this list we were unable successfully to map two of the genes in the expression seed into gene names. Subsequently we have been able to do so and this will be incorporated into future. However, the results in this document are based on the list without these genes, as is reflected in the table.

Coord.	Up	Down	Name	Gene bank	RT-PCR
A02f	2.6		Galactosidase-binding lectin 3	M35368	ND

A08n	2.7		Yin & yang (YY1)	X70683	ND
A09a		2.8	Twist-related protein	X99268	84.8 ± 0.671
A14b		**2.7**	**Phosphoglyceride kinase 1 (PGK1)**	**V00572**	94.5 ± 1.88
A14e		7.8	RNA binding protein 3 (RMB3)	U28686	44.2 ± 1.50
A14j	2.0		U2 small nuclear ribonucleoprotein B	M15841	ND
B01j	2.0		CDC-like kinase 3 (CLK3)	L29220	ND
B02k	6.1		CDK inhibitor (p21)	U09579	161 ± 7.11
B02n	2.2		NEF2-related factor 1	U08853	ND
B04d	4.7		Asparagine sythetase	M27396	127 ± 3.23
B04e	2.6		Inhibitor of DNA-binding protein 3	X69111	ND
B04f	3.0		Inhibitor of DNA binding 1 protein	D13889	ND
B04i		2.0	Prothymosin alpha (PTMA)	M26708	93.6 ± 1.30
B08j	5.8		IGF binding protein 5	M65062	148 ± 3.19
B09j	5.3		Jun proto-oncogene	J04111	181 ± 27.3
B09k		**13.**	**5 Myc proto-oncogene**	**V00568**	49.5 ± 1.97
B12g		2.4	SYT-interacting protein (SIP)	AF080561	ND
B12k		2.3	v-erbA related protein 3 (EAR3)	X12795	ND
B14a		2.3	Heat shock 70-kD protein 1	M11717	ND
B14j	5.1		Mitochondrial stress-70 protein	L15189	132.7 ± 3.9
C06h	2.1		solute carrier family 12 member 4	U55054	ND
C14g	2.3		Membrane-bound and soluble COMT	M65212	ND
D01j	2.4		Transgluminase	M98252	ND
D01n	2.6		Glutamic oxaloacetic transaminase 1	M37400	ND

D02c	6.4		Bifunctional methylenetetrahydrofolate	X16396i	125 ± 5.95
D04m	2.4		T-complex protein 1 epsilon	D43950	ND
D05i	2.1		Ribophorin II	Y00282	ND
D07d	**4.5**		**GADD45**	**M60974**	190 ± 13.4
D10f		2.0	Glutathione S-transferase theta 2	L38503	ND
D10g	29.0		GADD153	S40706	332 ± 12.7
D10h	2.4		HSIAH2	U76248	ND
D11d	2.0		Structure-specific recognition protein 1	M86737	ND
D131		9.0	Interleukin 2 receptor alpha subunit	X01057	ND
E02e		4.7	Delta-like protein (DLK)	U15979	ND
E03d		2.5	Glucose-6-phosphate isomerase (GPI)	K03515	96.1 ± 2.22
E04d	4.8		Vascular endothelial growth factor	M32977	147 ± 4.71
E04l	2.0		Ribonuclease/angiogenin inhibitor	M36717	ND
E08m	2.2		Mitogen-activated protein-kinase 5	U25265	ND
E10h	2.7		Protein kinase C alpha polypeptide	M22199	ND
E12g	7.8		Protein phosphatase (PP2A)	M64930	ND
E13i	2.0		Ras-related C3 botulinum toxin	M29870	ND
E14l	2.4		14-3-3 protein	X57346	ND
F01a	2.1		Emsl oncogene	M98343	ND
F01b	2.0		Protein kinase C inhibitor 1 (PKCI1)	U51004	ND
F01f	3.3		STAT-induced STAT inhibitor 2	AB004903	ND
F01n	2.0		Mu-crystallin homolog (CRYM)	L02950	ND
F02a	2.4		Phospholipase C γ-binding protein	AB005216	ND

F02e	2.0		Tuberin	X75621	ND

Table A.2: Table reproduced with permission from Conn. et. al. *cDNA microarray analysis of changes in gene expression associated with MPP+ toxicity in SH-SY5Y cells* [36]. Replicated with permision from Springer via RightsLink. Table shows genes that are sufficiently differentially expressed (for details see [36]) in SH-SY5Y cells 72 hours after a 1mM dose of MPP+ [36]

No.	Gene Name	Mapped BioGrid Ids
1	LGALS3	110149
2	SOX4	112542
3	TWIST1	113142
4	PGK1	111251
5	NOT IDENTIFIED	NONE
6	SNRPB2	112513
7	CLK3	107609
8	CDKN1A	107460, 111099
9	NFE2L1	110851
10	ASNS	106932
11	ID3	109625
12	ID1	109623
13	PTMA	111724, 111728*
14	IGFBP5	109709
15	JUN	109928
16	MYC	110694
17	RBM14	115700
18	NR2F1	112883, 112884
19	HSPA1A	109535, 109536
20	HSPA9	109545
21	SLC12A4	112449*
22	COMT	107707
23	PLOD1	111366
24	GOT1	109067, 119232*
25	MTHFD2	116011*
26	CCT5	116603
27	RPN2	112100
28	GADD45A	108014
29	GSTT2	NONE
30	DDIT3	108016
31	SIAH2	112373
32	SSRP1	112627
33	IL2RA	109774*
34	DLK1	114316
35	GPI	115325, 109082*
36	VEGFA	113265
37	RNH1	111977*, 128882, 129787*
38	NOT IDENTIFIED	NONE
39	PRKCA	111564
40	PPP2R2B	111513
41	RAC1	111817

42	YWHAB	113361
43	CTTN	108332
44	HINT1	109341
45	SOCS2	114362
46	CRYM	107815
47	SOCS7	119053, 125796
48	TSC2	113100

Table A.3: Table contains our conversions of the genes in Table A.2 into gene names and the subsequent conversion into a BioGrid identifier using the BioGrids internal conversions. The ordering is the same in each table, but they can't be combined because of copyright restrictions. Proteins marked with a * are present in BioGrid but are not part of the largest connected component of the network composed of interactions from Yeast 2 Hybrid.

A.2 Algorithm to add additional redundant seed nodes.

For n-hop snowball sampling, we can construct a list of possible seed nodes using the following simplified procedure:

```
listofSeeds=[]
OutsideNodes=list of nodes in (n+1)-hop snowball sample but not in n-hop sample
for node in Subnetwork
    for outNode in OutsideNodes
        if shortest path length between node and outNode is less than n+1
            failed=1
            break
    if failed==0
        add node to listofSeeds
```

In the other two construction methods we cannot use such a simple construction due to interdependence between the seed nodes. Thus we must compare the size of the sampled network with the original sampled network (by comparing the number of nodes and edges) when any seed is added to the seed list.

OMIM Entry	Mapped BioGrid Ids
Parkinson disease 1, 168601 (3)|SNCA, NACP, PARK1, PARK4|163890|4q22.1	112506
Parkinson disease 10 (2)|PARK10, AAOPD|606852|1p32	NONE
Parkinson disease 11, 607688 (3)|GIGYF2, KIAA0642, PARK11|612003|2q37.1	117520
Parkinson disease 12 (2)|PARK12|300557|Xq21-q25	NONE
Parkinson disease 13, 610297 (3)|HTRA2, OMI, PARK13, PRSS25|606441|2p13.1	118165
Parkinson disease 14, 612953 (3)|PLA2G6, IPLA2, INAD1, NBIA2B, NBIA2A, PARK14|603604|22q13.1	113986*
Parkinson disease 15, autosomal recessive, 260300 (3)|FBXO7, FBX7, FBX, PKPS, PARK15|605648|22q12.3	117326
Parkinson disease 17, 614203 (3)|VPS35, MEM3, PARK17|601501|16q11.2	120855
Parkinson disease 18, 614251 (3)|EIF4G1, EIF4G, PARK18|600495|3q27.1	108296
Parkinson disease 3 (2)|PARK3|602404|2p13	NONE
Parkinson disease 4, 605543 (3)|SNCA, NACP, PARK1, PARK4|163890|4q22.1	112506
Parkinson disease 6, early onset, 605909 (3)|PINK1, PARK6|608309|1p36.12	122376
Parkinson disease 7, autosomal recessive early-onset, 606324 (3)|DJ1, PARK7|602533|1p36.23	116446
Parkinson disease 8, 607060 (3)|LRRK2, PARK8|609007|12q12	125700
Parkinson disease 9, 606693 (3)|ATP13A2, PARK9, KRPPD|610513|1p36.13	116973
Parkinson disease, juvenile, type 2, 600116 (3)|PRKN, PARK2, PDJ, LPRS2|602544|6q26	111105
Parkinson disease 16 (2)|PARK16|613164|1q32	NONE
Parkinson disease 5, susceptibility to, 613643 (3)|UCHL1, PARK5|191342|4p13	113192
Parkinson disease, late-onset, susceptibility to, 168600 (3)|GBA|606463|1q22	108899*
Parkinson disease, susceptibility to, 168600 (3)|ADH1C, ADH3|103730|4q23	NONE
Parkinson disease, susceptibility to, 168600 (3)|MAPT, MTBT1, DDPAC, MSTD|157140|17q21.31	110308*
Parkinson disease, susceptibility to, 168600 (3)|TBP, SCA17, HDL4|600075|6q27	112771

Table A.1: OMIM Seed List including mappings to BioGrid. Conversion is performed using the mapping present in the BioGrid database. Proteins marked with a * are present in BioGrid but are not part of the largest connected component of the network composed of interactions from Yeast 2 Hybrid.

A.2.0.1 Optimised algorithm for finding redundant seed nodes

The algorithm presented in Section 3.2.3 is the following:

1. Remove each seed in turn and check if the number of nodes and edges in the subnetwork do not change. If not, then add the node to the list of redundant seeds.
2. Form a list of the remaining seeds.
3. Define a dummy variable L and set $L = 0$
4. For lists of redundant seeds of length L
   (a) Test if sampling with the list of the remaining seeds and the selected redundant seeds produce the same network.
   (b) Store all seed lists which pass the test.
5. If there are no seed lists which pass the test, set $L \to L + 1$ and go to step 4.
6. Return the smallest seed list(s) that produce the same network.

The major problem in this procedure is the large number of options that may need to be checked to find the minimum seed list. As stated in chapter finding the minimum seed list for snowball sampled networks is equivalent to the set cover problem which is NP Hard.

Thus in cases where we require additional speed we could simply reformulate this as a set cover problem and use state of the art algorithms for this problem. This would involve for each redundant seed node computing the set of nodes which are sampled by this node. Then removing from each set all nodes that are sampled by the non-redundant seed nodes. We could then apply exact set cover algorithms to this problem set.

It may also be possible to convert the path based techniques into another related NP Complete or NP Hard problem and use a similar technique. However, as we do not require the speed for the work we are doing here we have not attempted to do so.

A further optimisation to speed up computation of the minimum seed list, which is sampling technique dependent, can be performed on sampling techniques that scale with number of nodes in the whole network and that only depend on the information in the subnetwork. Sampling in the subnetwork rather than in the wider network can be more efficient while still guaranteeing the result. One example of this procedure is shortest path sampling. All of the information about shortest paths is included in the subnetwork. Therefore sampling with the reduced seed list in the subnetwork saves time (as shortest path scales with number of nodes and edges depending on implementation) and guarantees that the result is correct as long as the seed list is a subset of the original seed list.

A.3 Empirical seed lists: additional results

To test whether the results we see Chapter 3 are robust with respect to the bins that we use to generate the random seed lists, we recalculate the results using a minimum bin size of 5, 10, 20, 30 and 50. The smaller the bin size, the closer the degree sequence will match the test sequence, whereas the larger minimum bin sizes produce seed lists which have degree sequences which are further from the test list, but have a lower likelihood of selecting the same small set of seed nodes. Figure A.1 shows the results for the OMIM seed list and Figure A.2 shows the results for the expression seed list.

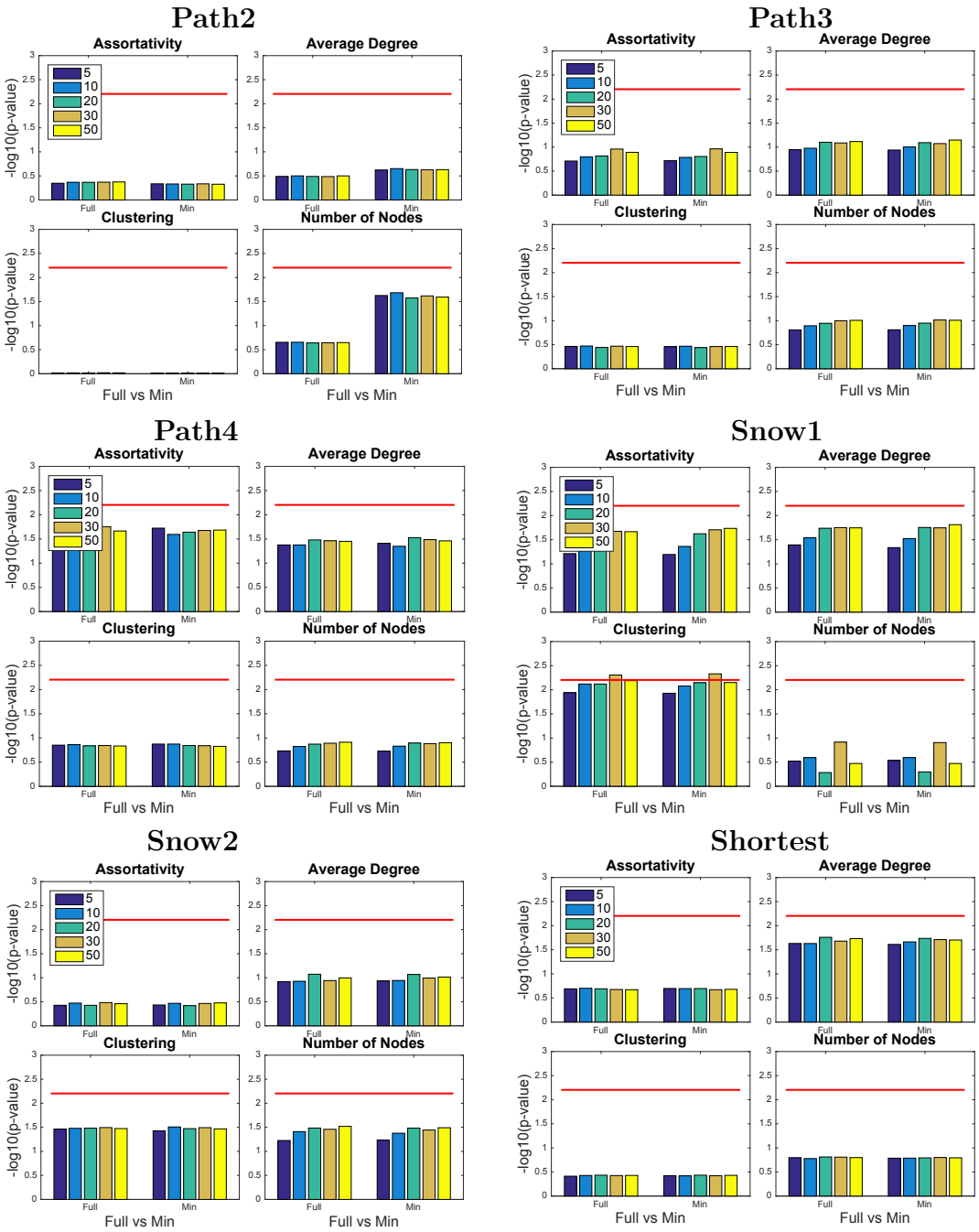

Figure A.1: Effect of bin size on test results for the OMIM seed list: smallest p-value, on a negative log scale under our null model. Multiple bars demonstrate the robustness of the method to the minimum number of items in each bin chosen when accounting for degree. Red: significance level (0.05/8). Left: Full seed list, right minimum seed list. Note negative logarithmic scale, e.g. only results that are above the Red significance line are considered significant. Only in Snowball 1 sampling is the result significant and only for the clustering coefficient for one tested parameter.

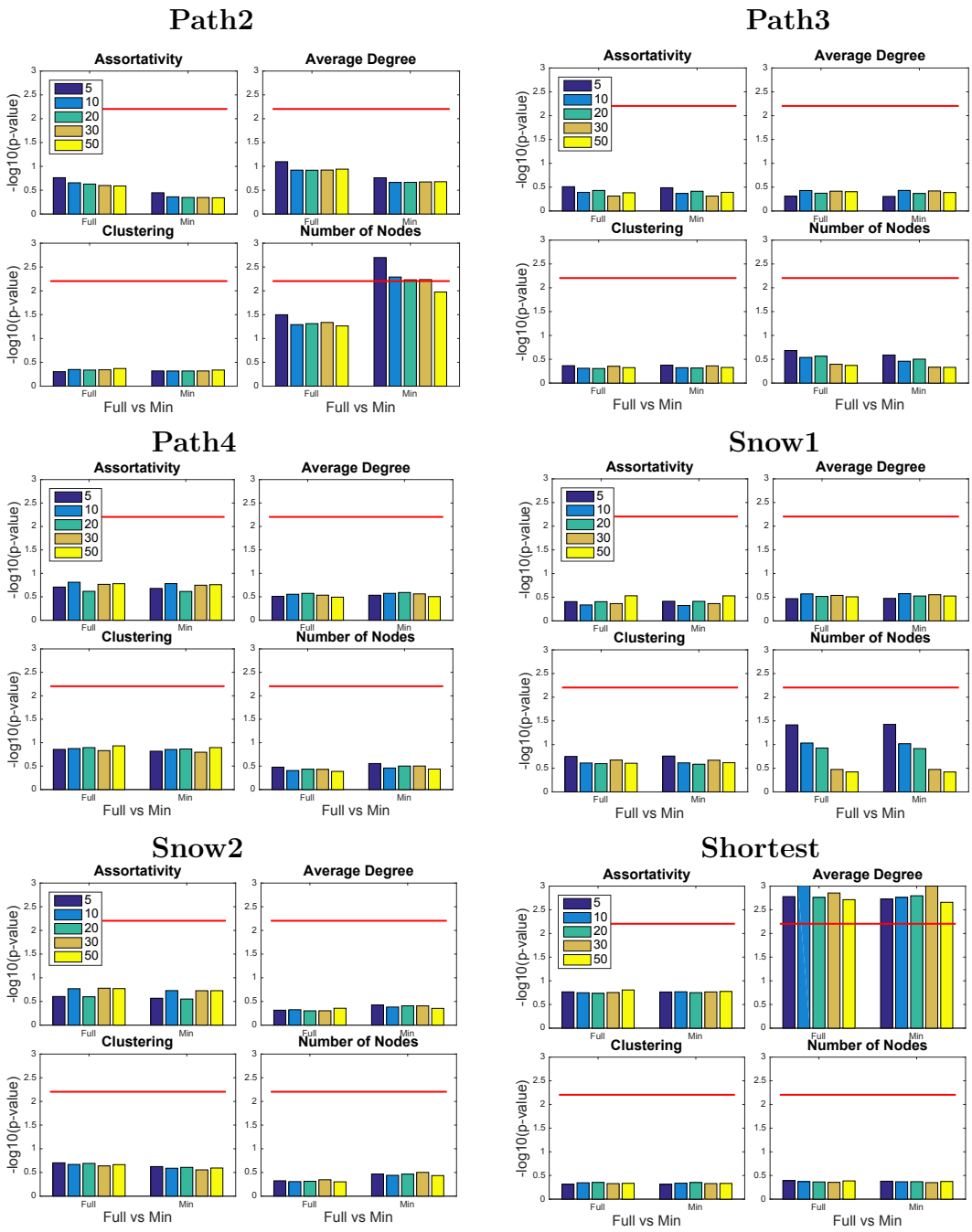

Figure A.2: Effect of bin size on test results for the Expression seed list: smallest p-value, on a negative log scale under our null model. Multiple bars demonstrate the robustness of the method to the minimum number of items in each bin chosen when accounting for degree. Red: significance level (0.05/8). Left: Full seed list, right minimum seed list. Note negative logarithmic scale, e.g. only results that are above the Red significance line are considered significant. Significant results are observed in number of nodes in Shortest path sampling and in number of nodes in Path2 with the minimum seed list.

Appendix B

Appendix - Chapter 4

B.1 Constructing a multi-resolution modularity of sampled networks based on the variance.

The issue that we encounter in constructing communities from sampled networks could be related to the resolution parameter. In this section we explore an alternative formulation to the resolution parameter which can be applied to the context of sampled networks.

The resolution parameter was in its conception an ad-hoc fix to a problem of community size, and while further studies have validated both its usefulness [108] and its relation to community detection by Markov random walks [101], it can be inappropriate if the elements of P have very different values. To illustrate this issue, consider the following situation we wish to detect communities in a given sampled network. Following the procedure in Section 4.3.1 we estimate a null model and suppose that for degrees a and b that the estimated connection probabilities are as follows: $\hat{\mathfrak{P}}_{aa}^{(f_{samp}-rescale)} = 0.75$ and $\hat{\mathfrak{P}}_{ab}^{(f_{samp}-rescale)} = 0.1$. When attempting to scale this system in the same way as Ref. [137], a problem arises. For a resolution parameter (λ) of 1.5, the null model term in modularity (λP_{ij}) for a pair of nodes of

degree a becomes 1.125, i.e. regardless of whether there is an edge, it is considered disadvantageous if these nodes are in the same community, whereas the null model term for edges between a node of degree a and a node of degree b is 0.15, so dense regions of edges between a node of degree a and a node of degree b may be considered a community.

A solution to this problem for our use case is to use the distribution of $\mathfrak{P}_{ab}^{(N_{et})}$ over the network ensemble $\mathfrak{N}_{ab}(s)$ Thus, rather than scaling the mean of this distribution multiplicatively, we could scale using the inverse CDF of this distribution. If we let $q_{ab}(\lambda)$ for $0 \leq \lambda \leq 2$ be the smallest solution to the following equation then:

$$\frac{\sum_{N_{et} \in \mathfrak{N}_{ab}(S)} \mathbb{1}(\mathfrak{P}_{ab}^{(N_{et})} \leq q_{ab}(\lambda))}{|\mathfrak{N}_{ab}(S)|} = \frac{\lambda}{2} \tag{B.1}$$

where $\mathbb{1}(\mathfrak{P}_{ab}^{(N_{et})} \leq q_{ab}(\lambda))$ is an indicator function, such that it equals 1 if $\mathfrak{P}_{ab}^{(N_{et})} \leq q_{ab}(\lambda)$ and 0 otherwise.

Using this as a null model in modularity (Equation (2.1)) we obtain:

$$Q^{dist} = \sum_{ij} \Big(A_{ij} - q_{k_i k_j}(\lambda)\Big)\delta(c_i, c_j) \tag{B.2}$$

We note that by construction $q(1)$ is the median of the distribution and $q(1.5)$ is the 75th percentile of this distribution. Thus, for $\lambda = 1$ we obtain similar behaviour to Equation. (4.20), except that we remove the median rather than the mean. Note that $q_{ab}(\lambda)$ cannot go above 1, as the underlying set of $\mathfrak{P}_{ab}^{(N_{et})}$ does not. Thus, this addresses the problem posed at the start of this section. Further, unlike in Eq. (2.4), in this formulation λ is restricted to a closed interval, making it easier to test a representative sample of the space of possible λ.

In general the computational cost of producing enough sampled networks to estimate the CDF to a high accuracy is large. A simpler approach is to use the mean and standard deviation giving more information about the distribution without requiring

a large number of samples. The standard deviation is then varied rather than the mean. The standard deviation can be computed using the following formula:

$$\hat{\sigma}_{ab}^{(f_{samp})} = \sqrt{\frac{1}{|\mathfrak{N}_{ab}|}\Big(\Big(\sum_{N_{et}\in\mathfrak{N}_{ab}}(\hat{\mathfrak{P}}_{ab}^{(Net)})^2\Big) - (\hat{\mathfrak{P}}_{ab}^{(f_{samp})})^2\Big)} \tag{B.3}$$

Note, for simplicity we use the uncorrelated standard deviation. This would lead to be following expression for modularity:

$$Q^{dist1} = \sum_{ij}(A_{ij} - \hat{\mathfrak{P}}_{ij}^{(f_{samp})} + \gamma\hat{\sigma}_{ij}^{(f_{samp})})\delta(c_i, c_j) \tag{B.4}$$

where $\gamma \in \mathbb{R}$ (whereas $\lambda \in \mathbb{R}^+$). However, as with Q_1^{Path} (Eq. (4.20)) we must account for the average number of edges, giving the following expression for modularity:

$$Q^{var} = \sum_{ij}\left(A_{ij} - (\hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp})} + \gamma\hat{\sigma}_{k_ik_j}^{(f_{samp})})\left(\frac{\sum_{i,j}A_{ij}}{\sum_{i,j}\hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp})}}\right)\right)\delta(c_i, c_j) \tag{B.5}$$

We note that:

$$\frac{\hat{\sigma}_{ab}^{(f_{samp})}\sum_{i,j}A_{ij}}{\sum_{i,j}\hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp})}} = \sqrt{\frac{1}{|\mathfrak{N}_{ab}|}\left(\left(\sum_{N_{et}\in\mathfrak{N}_{ab}}\left(\frac{\hat{\mathfrak{P}}_{ab}^{(Net)}\sum_{i,j}A_{ij}}{\sum_{i,j}\hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp})}}\right)^2\right) - \left(\frac{\hat{\mathfrak{P}}_{ab}^{(f_{samp})}\sum_{i,j}A_{ij}}{\sum_{i,j}\hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp})}}\right)^2\right)} \tag{B.6}$$

$$= \hat{\sigma}_{ab}^{(f_{samp}-rescale)} \tag{B.7}$$

Therefore, $\hat{\sigma}_{k_ik_j}^{(f_{samp}-rescale)}$ is the standard deviation over the rescaled distribution. Putting this all together we obtain:

$$Q^{var} = \sum_{ij}(A_{ij} - \hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp}-rescale)} + \gamma\hat{\sigma}_{k_ik_j}^{(f_{samp}-rescale)})\delta(c_i, c_j) \tag{B.8}$$

Note, after the rescaling, for $\gamma = 0$ the $\sum_{ij}(A_{ij} - \hat{\mathfrak{P}}_{k_ik_j}^{(f_{samp}-rescale)} + \gamma\hat{\sigma}_{k_ik_j}^{(f_{samp}-rescale)}) = 0$, therefore a value of $\gamma = 0$ is comparable to $\lambda = 1$.

B.2 Testing Q_1^{Path} and Q^{var}

We test the ability to detect communities in sampled networks of Q_1^{Path} as described in Section 4.3.1.3 and Q^{var} as described in Section B.1. To assess the performance we use the benchmark we derived in Section 4.2.2. Note, this is the same procedure which was used to test the configuration model in Section 4.2.4.

We let $\lambda_2 = 0.1, 0.3, ..., 1.9$ and $\lambda_1 = 1$ and for the variance model we let $\gamma = -2, -1.6, ..., 1.6, ..., 2$ giving approximately the same number of observations as the configuration model (we allow the variance one more observation simply to preserve symmetry). We allow γ to become negative as this is effectively equivalent to allowing λ to be greater than 1.

The results are displayed in Figure B.1 and Figure. B.2. For Q_1^{path} the results are very similar to the result we observed in the configuration model. The VOI result is very similar to the trivial value computed using a single community, and similarly the λ values chosen seem to indicate either a single community or a small number of communities. The NMI results are also similar to the results we observed from the configuration model with the trivial partition outperforming in the geometric and with the detected partition outperforming for higher values of the ratio in the arithmetic. The behaviour on the Zrand score is also similar.

The variance model Figure B.2 is broadly on all statistics but arithmetic NMI. On arithmetic NMI, it outperforms Q_1^{Path} and the configuration model and outperforms the trivial partition for ratios greater than 40. However, it does not outperform the trivial partition in VOI and geometric NMI and the performance in Zrand is comparable to Q_1^{Path}. The distribution of the γ corresponding to the best similarity for VOI appears to indicate that many of the results are occur when $\gamma = 2$. Thus some of the issues in the case of VOI may stem from the constaints on the parameter regime. This is also the parameter range which is more likely to produce a single partition which may explain the behaviour.

Thus in conclusion the variance model outperform the configuration model on one statistic but both models are outperformed by trivial partitions on two statistics and thus these approaches do not in general outperform the configuration model.

Hence this approach does not solve the problem of detecting communities in shortest path sampled networks.

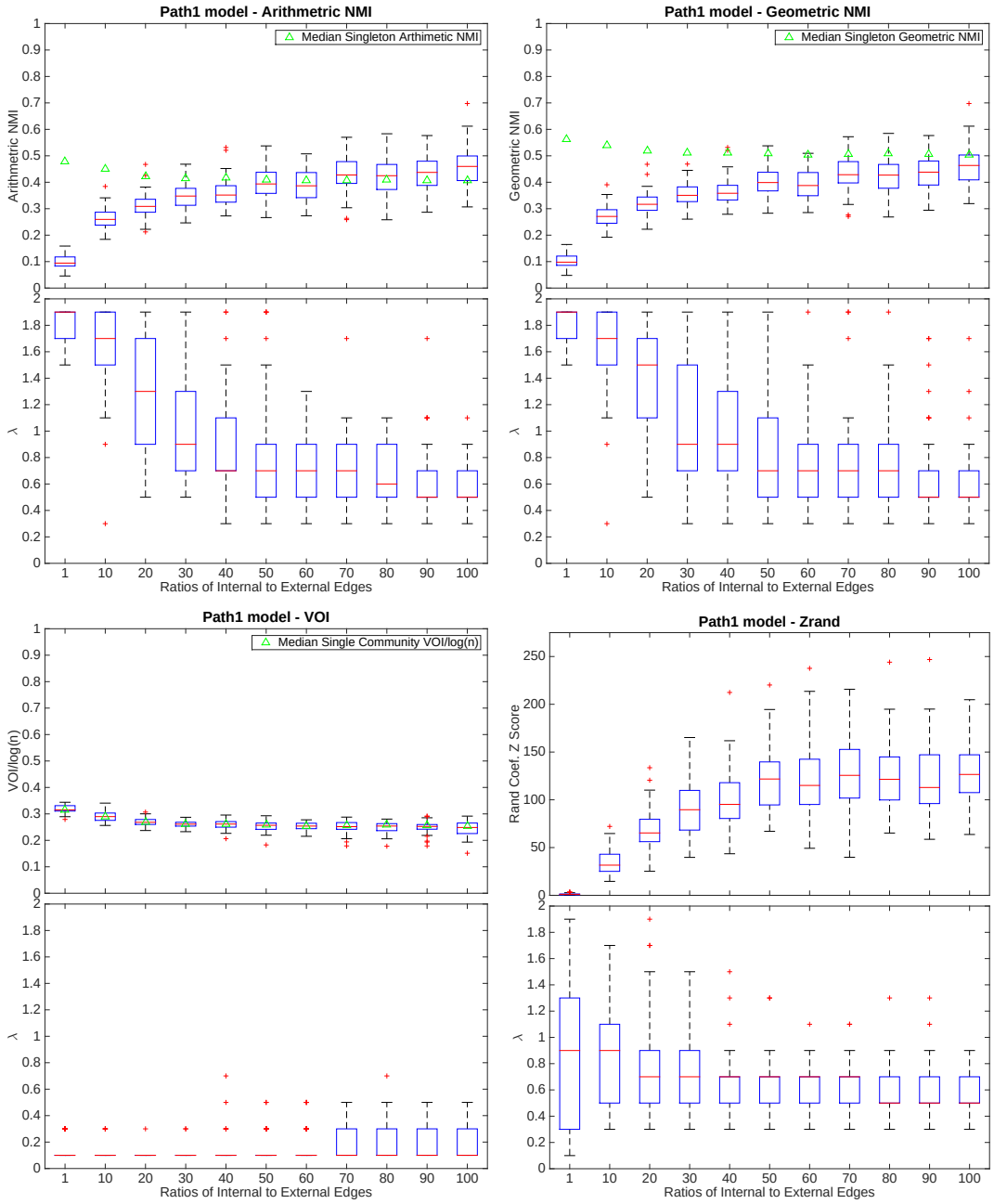

Figure B.1: Performance of community detection using Q_1^{Path} modularity at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $\lambda_1 = 1$ and $\lambda_2 = 0.1, 0.3, ..., 1.9$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

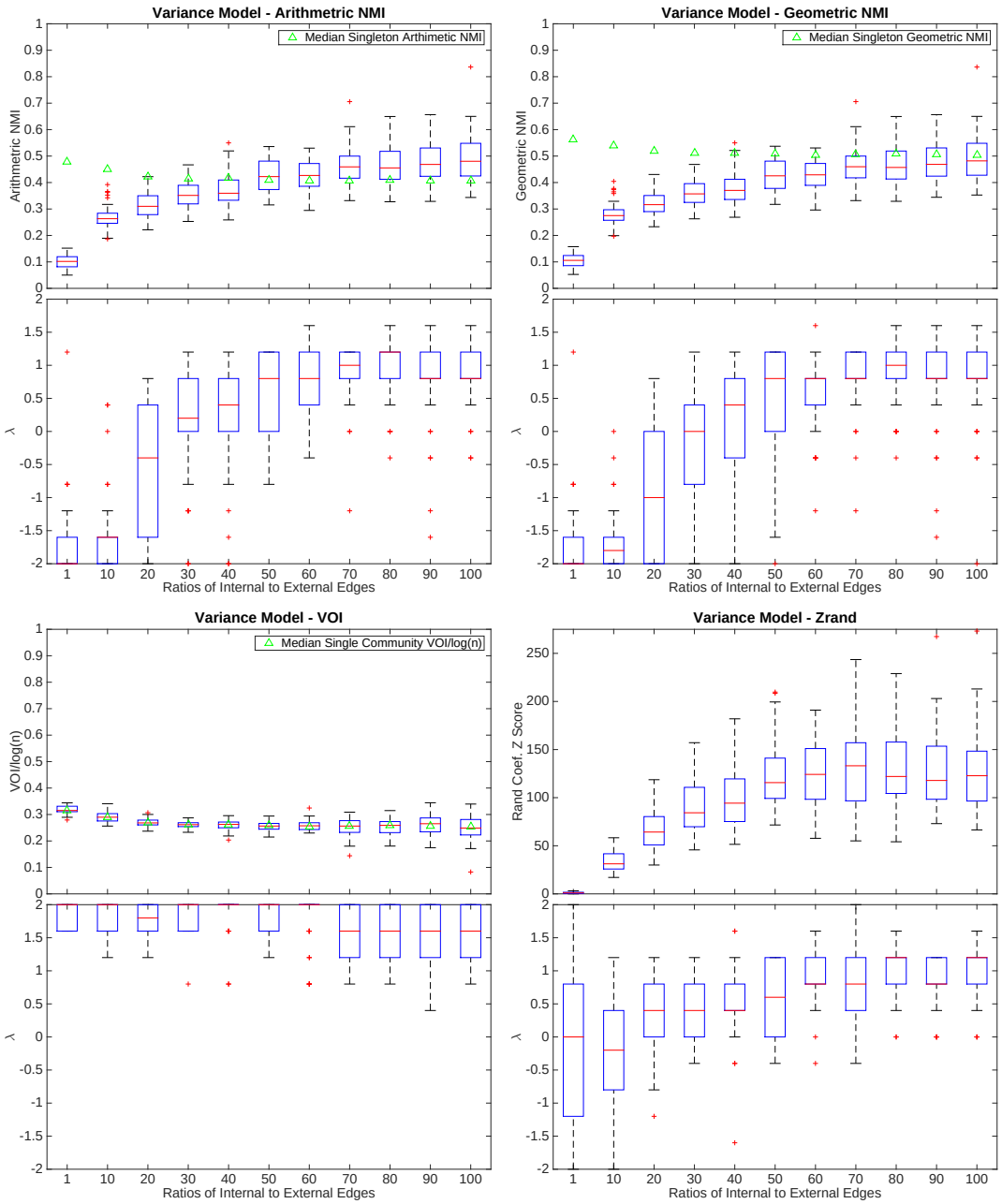

Figure B.2: Performance of community detection using Q^{var} modularity at detecting the embedded partition of a sampled network with respect to the ratio of internal to external edge probability in the unsampled network ($\frac{c_{in}}{c_{out}}$) with 50 replicates for each ratio. (for details see 4.2.2). Performance is measured using VOI, Arith. NMI, Geom. NMI and Zrand (see 4.2.3). Community detection is performed with $\gamma = -2, -1.6, ..., 2$, and the best performance is reported in the upper panels. The green triangles indicate the median score by a trivial partition (see 4.2.3.5). The distribution of best performing resolutions is reported in the lower panels.

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science Taylor & Francis Group, 4 edition, 2002.

[2] W. Ali, C. Deane, and G. Reinert. Protein interaction networks and their statistical analysis. In *Handbook of Statistical Systems Biology*, pages 200–234. John Wiley & Sons, Ltd, 2011.

[3] W. Ali, T. Rito, G. Reinert, F. Sun, and C. M. Deane. Alignment-free protein interaction network comparison. *Bioinformatics*, 30(17):i430–i437, 2014.

[4] D. Alvarez-Fischer, C. Henze, C. Strenzke, J. Westrich, B. Ferger, G. U. Höglinger, W. H. Oertel, and A. Hartmann. Characterization of the striatal 6-OHDA model of Parkinson's disease in wild type and α-synuclein-deleted mice. *Experimental Neurology*, 210(1):182 – 193, 2008.

[5] L. N. F. Ana and A. K. Jain. Robust data clustering. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–128–II–133 vol.2, June 2003.

[6] A. Anastasiou and G. Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 23(1):191–218, 2017.

[7] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. The intact molecular interaction database in 2010. *Nucleic Acids Research*, 38(suppl 1):D525–D531, 2010.

[8] A. Arenas, A. Fernández, and S. Gómez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5):053039, 2008.

[9] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000.

[10] P. K. Auluck, G. Caraveo, and S. Lindquist. α-Synuclein: Membrane Interactions and Toxicity in Parkinson's Disease. *Annual Review of Cell and Developmental Biology*, 26(1):211–233, 2010.

[11] G. D. Bader, D. Betel, and C. W. V. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[12] G. D. Bader, M. P. Cary, and C. Sander. Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(suppl 1):D504–D506, 2006.

[13] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[14] A. D. Barbour, M. Karoński, and A. Ruciński. A central limit theorem for decomposable random variables with applications to random graphs. *Journal of Combinatorial Theory, Series B*, 47(2):125–145, 1989.

[15] M. Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[16] D. S. Bassett, N. F. Wymbs, M. P. Rombach, M. A. Porter, P. J. Mucha, and S. T. Grafton. Task-based core-periphery organization of human brain dynamics. *PLOS Computational Biology*, 9(9):1–16, 09 2013.

[17] M. Bazzi, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Modeling & Simulation*, 14(1):1–41, 2016.

[18] M. Beguerisse-Díaz, G. Garduño-Hernández, B. Vangelov, S. N. Yaliraki, and M. Barahona. Interest communities and flow roles in directed networks: the Twitter network of the UK riots. *Journal of The Royal Society Interface*, 11(101):20140940, 2014.

[19] S. Berger, J. Posner, and A. Ma'ayan. Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, 8:372, 2007.

[20] H. R. Bernard, T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelniciuc, G. A. Shelley, P. Sirinirund, S. Weir, and D. F. Stroup. Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections*, 86(Suppl 2):ii11–ii15, 2010.

[21] R. Betarbet, T. B. Sherer, G. MacKenzie, M. Garcia-Osuna, A. V. Panov, and J. T. Greenamyre. Chronic systemic pesticide exposure reproduces features of Parkinson's disease. *Nature Neuroscience*, 3(12):1301–1306, 2000.

[22] P. Blohm, G. Frishman, P. Smialowski, F. Goebels, B. Wachinger, A. Ruepp, and D. Frishman. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic acids research*, page gkt1079, 2013.

[23] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[24] D. Bredesen, R. Rao, and P. Mehlen. Cell death in the nervous system. *Nature*, 443:796–802, 2006.

[25] R. L. Brennan and R. J. Light. Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, 27(2):154–163, 1974.

[26] L. Breydo, J. W. Wu, and V. N. Uversky. α-Synuclein misfolding and Parkinson's disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(2):261–285, 2012.

[27] T. N. Bui. On bisecting random graphs. Master's thesis, Department of Electrical Engineering and Computer Science, MIT, 1983.

[28] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[29] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. Mint, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(suppl 1):D532–D539, 2010.

[30] F. Cerina, V. De Leo, M. Barthelemy, and A. Chessa. Spatial correlations in attribute communities. *PLoS One*, 7(5):1–10, 2012.

[31] A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, T. Reguly, A. Breitkreutz, A. Sellam, D. Chen, C. Chang, J. Rust, M. Livstone, R. Oughtred, K. Dolinski, and M. Tyers. The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D816–D823, 2013.

[32] Z. Cheung and N. Ip. The emerging role of autophagy in Parkinson's disease. *Molecular Brain*, 2(1):29, 2009.

[33] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):140, 2007.

[34] F. Chung, L. Lu, and V. Vu. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 100(11):6313–6318, 2003.

[35] F. Chung and M. Radcliffe. On the spectra of general random graphs. *The Electronic Journal of Combinatorics*, 18(1):215–229, 2011.

[36] K. J. Conn, M. D. Ullman, M. J. Larned, P. B. Eisenhauer, R. E. Fine, and J. M. Wells. cDNA microarray analysis of changes in gene expression associated with MPP+ toxicity in SH-SY5Y cells. *Neurochemical Research*, 28(12):1873–1881, 2003.

[37] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction To Algorithms*. MIT Press, 2001.

[38] P. Csermely, V. Ágoston, and S. Pongor. The efficiency of multi-target drugs: the network approach might help drug design. *Trends in Pharmacological Sciences*, 26(4):178–182, 2005.

[39] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. R. Carvunis, N. Simonis, J.-F. F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal. Literature-curated protein interaction datasets. *Nature Methods*, 6(1):39–46, 2009.

[40] R. K. Darst, Z. Nussinov, and S. Fortunato. Improving the performance of algorithms to find communities in networks. *Phys. Rev. E*, 89:032809, Mar 2014.

[41] W. Dauer and S. Przedborski. Parkinson's disease: Mechanisms and models. *Neuron*, 39(6):889–909, 2003.

[42] C. A. Davie. A review of Parkinson's disease. *British Medical Bulletin*, 86(1):109–127, 2008.

[43] J. De Las Rivas and C. Fontanillo. Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, 6(6):e1000807, 2010.

[44] L. M. de Lau, P. C. Giesbergen, M. C. de Rijk, A. Hofman, P. J. Koudstaal, and M. M. Breteler. Incidence of parkinsonism and Parkinson disease in a general population. *Neurology*, 63(7):1240–1244, 2004.

[45] M. C. de Rijk, C. Tzourio, M. M. Breteler, J. F. Dartigues, L. Amaducci, S. Lopez-Pousa, J. M. Manubens-Bertran, A. Alprovitch, and W. A. Rocca. Prevalence of parkinsonism and Parkinson's disease in Europe: the EUROPARKINSON Collaborative Study. European Community Concerted Action on the Epidemiology of Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 62(1):10–15, 1997.

[46] C. M. Deane, Ł. Salwiński, I. Xenarios, and D. Eisenberg. Protein interactions. *Molecular and Cellular Proteomics*, 1(5):349–356, 2002.

[47] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Reviev Letters*, 107:065701, 2011.

[48] A. Delmotte and M. Schaub. Partitionstability. `https://github.com/michaelschaub/PartitionStability`.

[49] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755–12760, 2010.

[50] R. E. Drolet, B. Behrouz, K. J. Lookingland, and J. L. Goudreau. Mice lacking α-Synuclein have an attenuated loss of striatal dopamine following prolonged chronic MPTP administration. *NeuroToxicology*, 25(5):761–769, 2004.

[51] J. A. Dunne, R. J. Williams, and N. D. Martinez. Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, 10 2002.

[52] A. Elliott, E. A. Leicht, A. Whitmore, G. Reinert, and F. Reed-Tsochas. A nonparametric significance test for sampled networks. https://arxiv.org/abs/1511.00182.

[53] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[54] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[55] J. Faugier and M. Sargeant. Sampling hard to reach populations. *Journal of Advanced Nursing*, 26(4):790–797, 1997.

[56] W. Feller. *An Introduction to Probability Theory and Its Applications*. Number v. 1 in Wiley publications in statistics. John Wiley and Sons, 3rd ed. edition, 1964.

[57] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[58] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[59] S. Fortunato and D. Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

[60] T. M. Fountaine, L. L. Venda, N. Warrick, H. C. Christian, P. Brundin, K. M. Channon, and R. Wade-Martins. The effect of α-synuclein knockdown on MPP+ toxicity in models of human neurons. *European Journal of Neuroscience*, 28(12):2459–2473, 2008.

[61] O. Frank. Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1(3):235–264, 1977.

[62] O. Frank and T. Snijders. Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics - Stockholm*, 10:53–53, 1994.

[63] B. Franke and P. J. Wolfe. Network modularity in the presence of covariates. https://arxiv.org/abs/1603.01214, 2016.

[64] M. Ganji, A. Seifi, H. Alizadeh, J. Bailey, and P. J. Stuckey. Generalized modularity for community detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 655–670. Springer, 2015.

[65] J. T. Gao, R. Guimerà, H. Li, I. M. Pinto, M. Sales-Pardo, S. C. Wai, B. Rubinstein, and R. Li. Modular coherence of protein dynamics in yeast cell polarity system. *Proceedings of the National Academy of Sciences*, 108(18):7647–7652, 2011.

[66] J. Garcia, E. Guney, R. Aragues, J. Iglesias, and B. Oliva. Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, 11(1):56, 2010.

[67] D. J. Gelb, E. Oliver, and S. Gilman. Diagnostic criteria for Parkinson disease. *Archives of Neurology*, 56(1):33–39, 1999.

[68] S. D. Ghiassian, J. Menche, and A.-L. Barabási. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Computational Biology*, 11(4):e1004120, 2015.

[69] S. D. Ghiassian, J. Menche, D. I. Chasman, F. Giulianini, R. Wang, P. Ricchiuto, M. Aikawa, H. Iwata, C. Müller, T. Zeller, et al. Endophenotype network models: Common core of complex diseases. *Scientific reports*, 6:27414, 2016.

[70] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[71] H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm, A. Droege, K. S. Lindenberg, M. Knoblich, C. Haenig, M. Herbst, J. Suopanki, E. Scherzinger, C. Abraham, B. Bauer, R. Hasenbank, A. Fritzsche, A. H. Ludewig, K. Buessow, S. H. Coleman, C.-A. Gutekunst, B. G. Landwehrmeyer, H. Lehrach, and E. E. Wanker. A protein interaction network links GIT1, an enhancer of Huntingtin aggregation, to Huntington's disease. *Molecular Cell*, 15(6):853–865, 2004.

[72] S. Gómez, P. Jensen, and A. Arenas. Analysis of community structure in networks of correlated data. *Physical Review E*, 80(1):016114, 2009.

[73] C. Gomez-Santos, I. Ferrer, J. Reiriz, F. Vinals, M. Barrachina, and S. Ambrosio. MPP+ increases α-synuclein expression and ERK/MAP-kinase phosphorylation in human neuroblastoma SH-SY5Y cells. *Brain Research*, 935(1-2):32–39, 2002.

[74] B. H. Good, Y.-A. de Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[75] R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[76] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 05 2005.

[77] R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106:22073–22078, 2009.

[78] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3):036102, 2007.

[79] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell. Protein-protein interaction networks and biology—what's the connection? *Nature Biotechnology*, 26(1):69–72, 2008.

[80] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl 1):D514–D517, 2005.

[81] G. T. Hart, A. Ramani, and E. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120, 2006.

[82] M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St. Onge, M. Tyers, D. Koller, R. B. Altman, R. W. Davis, C. Nislow, and G. Giaever. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science*, 320(5874):362–365, 2008.

[83] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[84] A. L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690, 2008.

[85] F. Hormozdiari, R. Salari, V. Bafna, and S. S. Cenk. Protein-protein interaction network evaluation for identifying potential drug targets. *Journal of Computational Biology.*, 17(5):669–684, 2010.

[86] W.-C. Hwang, A. Zhang, and M. Ramanathan. Identification of information flow-modulating drug targets: A novel bridging paradigm for drug discovery. *Clinical Pharmacology and Therapeutics*, 84(5):563–572, 2008.

[87] M. Jayapandian, A. Chapman, V. G. Tarcea, C. Yu, A. Elkiss, A. Ianni, B. Liu, A. Nandi, C. Santos, P. Andrews, B. Athey, D. States, and H. V. Jagadish. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic Acids Research*, 35(suppl 1):D566–D571, 2007.

[88] L. G. S. Jeub, M. Bazzi, I. S. Jutla, and P. J. Mucha. A generalized Louvain method for community detection implemented in MATLAB. `http://netwiki.amath.unc.edu/GenLouvain`, (2011-2016).

[89] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2017-01-11].

[90] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[91] B. Karrer, E. Levina, and M. E. Newman. Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.

[92] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[93] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[94] H. Keane, B. J. Ryan, B. Jackson, A. Whitmore, and R. Wade-Martins. Protein-protein interaction networks identify targets which rescue the MPP+ cellular model of Parkinson's disease. *Scientific Reports*, 5:17004, 2015.

[95] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database—2009 update. *Nucleic Acids Research*, 37(suppl 1):D767–D772, 2009.

[96] C. Klein and A. Westenberger. Genetics of Parkinson's Disease. *Cold Spring Harbor Perspectives in Medicine*, 2(1):a008888, 2012.

[97] G. Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, 2006.

[98] Y. Kraytsberg, E. Kudryavtseva, A. C. McKee, C. Geula, N. W. Kowall, and K. Khrapko. Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nature Genetics*, 38(5):518–520, 2006.

[99] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[100] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. V. Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[101] R. Lambiotte, J.-C. Delvenne, and M. Barahona. Laplacian dynamics and multiscale modular structure in networks. https://arxiv.org/abs/0812.1770, 2008.

[102] A. Lancichinetti and S. Fortunato. Limits of modularity maximization in community detection. *Physical Review E*, 84:066122, 2011.

[103] J. W. Langston, P. Ballard, J. W. Tetrud, and I. Irwin. Chronic parkinsonism in humans due to a product of meperidine-analog synthesis. *Science*, 219(4587):979–980, 1983.

[104] E. A. Leicht, P. Holme, and M. E. J. Newman. Vertex similarity in networks. *Physical Review E*, 73:026120, 2006.

[105] J. Leskovec and J. J. Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[106] S. Letovsky and S. Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(suppl 1):i197–i204, 2003.

[107] O. Levy, C. Malagelada, and L. Greene. Cell death pathways in Parkinson's disease: proximal triggers, distal effectors, and final steps. *Apoptosis*, 14:478–500, 2009.

[108] A. C. F. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100, 2010.

[109] A. C. F. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane. What evidence is there for the homology of protein-protein interactions? *PLoS Computational Biology*, 8(9):e1002645, 2012.

[110] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou. Identification of colorectal cancer related genes with mRMR and shortest path in Protein-Protein Interaction network. *PLoS One*, 7(4):e33393, 2012.

[111] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[112] J. Lim, T. Hao, C. Shaw, A. J. Patel, G. Szabó, J.-F. Rual, C. J. Fisk, N. Li, A. Smolyar, D. E. Hill, A.-L. Barabási, M. Vidal, and H. Y. Zoghbi. A Protein–Protein Interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 125(4):801–814, 2006.

[113] G. Lima-Mendez and J. van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–1493, 2009.

[114] Y. Liu, G. Fiskum, and D. Schubert. Generation of reactive oxygen species by the mitochondrial electron transport chain. *Journal of Neurochemistry*, 80(5):780–787, 2002.

[115] M. MacMahon and D. Garlaschelli. Community detection for correlation matrices. *Physical Review X*, 5:021006, 2015.

[116] A. Martin, M. E. Ochagavia, L. C. Rabasa, J. Miranda, J. Fernandez-de Cossio, and R. Bringas. BisoGenet: a new tool for gene network building, visualization and analysis. *BMC bioinformatics*, 11(1):91, 2010.

[117] T. Martin, B. Ball, and M. Newman. Structural inference for uncertain networks. *Physical Review E*, 93(1):012306, 2016.

[118] J. Masel and M. V. Trotter. Robustness and evolvability. *Trends in Genetics*, 26(9):406–414, 2010.

[119] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

[120] S. Mathivanan, B. Periaswamy, T. Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, Y. Ramachandra, and A. Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7(Suppl 5):S19, 2006.

[121] D. Mehrle, A. Strosser, and A. Harkin. Walk-modularity and community structure in networks. *Network Science*, 3(3):348–360, 2015.

[122] M. Meilă. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

[123] G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K. S. Arun, S. Sharma, K. N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekha, R. Nayak, G. Vishnupriya, H. G. M. Kumar, M. Nagini, G. S. S. Kumar, R. Jose, P. Deepthi, S. S. Mohan, T. K. B. Gandhi, H. C. Harsha, K. S. Deshpande, M. Sarker, T. S. K. Prasad, and A. Pandey. Human protein reference database—2006 update. *Nucleic Acids Research*, 34(suppl 1):D411–D414, 2006.

[124] R. R. Nadakuditi and M. E. J. Newman. Graph spectra and the detectability of community structure in networks. *Physical Review Letters*, 108:188701, 2012.

[125] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, 2010.

[126] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[127] 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.

[128] T. Pan, S. Kondo, W. Le, and J. Jankovic. The role of autophagy-lysosome pathway in neurodegeneration associated with Parkinson's disease. *Brain*, 131(8):1969–1978, 2008.

[129] A. Patil, K. Nakai, and H. Nakamura. Hitpredict: a database of quality assessed protein–protein interactions in nine species. *Nucleic Acids Research*, 39(suppl 1):D744–D749, 2011.

[130] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. Shivashankar, B. Rashmi, M. Ramya, Z. Zhao, K. Chandrika, N. Padma, H. Harsha, A. Yatish, M. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, 2003.

[131] A. Pujol, R. Mosca, J. Farrés, and P. Aloy. Unveiling the role of network and systems biology in drug discovery. *Trends in Pharmacological Sciences*, 31(3):115–123, 2010.

[132] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[133] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[134] H. P. Rang, editor. *Drug Discovery and Development*. Churchill Livingstone / Elsevier, 2006.

[135] O. Ratmann, C. Andrieu, C. Wiuf, and S. Richardson. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, 106(26):10576–10581, 2009.

[136] S. Razick, G. Magklaras, and I. M. Donaldson. iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, 2008.

[137] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[138] G. Reinert. A short introduction to Stein's method. *Lecture Notes*, 2011.

[139] T. Rito, Z. Wang, C. M. Deane, and G. Reinert. How threshold behaviour affects the use of subgraphs for network comparison. *Bioinformatics*, 26(18):i611–i617, 2010.

[140] O. Rivasplata. Subgaussian random variables: an expository note. *Internet publication, PDF*, 2012.

[141] P. Ronhovde and Z. Nussinov. Local resolution-limit-free Potts model for community detection. *Physical Review E*, 81:046114, 2010.

[142] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

[143] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte. Memory in network flows and its effects on spreading dynamics and community detection. *Nature communications*, 5, 2014.

[144] M. J. Salganik. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health*, 83(1):98–112, 2006.

[145] V. Salnikov, M. T. Schaub, and R. Lambiotte. Using higher-order Markov models to reveal flow-based communities in networks. *Scientific Reports*, 6:23194, 2016.

[146] F. Sams-Dodd. Target-based drug discovery: is something wrong? *Drug Discovery Today*, 10(2):139–147, 2005.

[147] M. Sarzynska, E. A. Leicht, G. Chowell, and M. A. Porter. Null models for community detection in spatially embedded, temporal networks. *Journal of Complex Networks*, 4:363–406, 2015.

[148] E. E. Schadt, S. H. Friend, and D. A. Shaywitz. A network view of disease and compound screening. *Nature Reviews Drug Discovery*, 8(4):286–295, 2009.

[149] M. T. Schaub, J.-C. Delvenne, S. N. Yaliraki, and M. Barahona. Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit. *PLoS One*, 7(2):1–11, 2012.

[150] A. Schober. Classic toxin-induced animal models of Parkinson's disease: 6-OHDA and MPTP. *Cell and Tissue Research*, 318:215–224, 2004.

[151] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[152] A. Sharma, J. Menche, C. C. Huang, T. Ort, X. Zhou, M. Kitsak, N. Sahni, D. Thibault, L. Voung, F. Guo, S. D. Ghiassian, N. Gulbahce, F. Baribaud, J. Tocker, R. Dobrin, E. Barnathan, H. Liu, R. A. Panettieri, K. G. Tantisira, W. Qiu, B. A. Raby, E. K. Silverman, M. Vidal, S. T. Weiss, and A.-L. Barabási. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Human Molecular Genetics*, 24(11):3005–3020, 2015.

[153] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[154] S. Shi, Y. Cai, X. Cai, X. Zheng, D. Cao, F. Ye, and Z. Xiang. A network pharmacology approach to understanding the mechanisms of action of traditional medicine: bushenhuoxue formula for treatment of chronic kidney disease. *PloS One*, 9(3):e89123, 2014.

[155] S. Shimohama, H. Sawada, Y. Kitamura, and T. Taniguchi. Disease model: Parkinson's disease. *Trends in Molecular Medicine*, 9(8):360–365, 2003.

[156] J. Simon-Sanchez, C. Schulte, J. M. Bras, M. Sharma, J. R. Gibbs, D. Berg, C. Paisan-Ruiz, P. Lichtner, S. W. Scholz, D. G. Hernandez, R. Kruger, M. Federoff, C. Klein, A. Goate, J. Perlmutter, M. Bonin, M. A. Nalls, T. Illig, C. Gieger, H. Houlden, M. Steffens, M. S. Okun, B. A. Racette, M. R. Cookson, K. D. Foote, H. H. Fernandez, B. J. Traynor, S. Schreiber, S. Arepalli, R. Zonozi, K. Gwinn, M. van der Brug, G. Lopez, S. J. Chanock, A. Schatzkin, Y. Park, A. Hollenbeck, J. Gao, X. Huang, N. W. Wood, D. Lorenz, G. Deuschl, H. Chen, O. Riess, J. A. Hardy, A. B. Singleton, and T. Gasser. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genetics*, 41(12):1308–1312, 2009.

[157] A. B. Singleton, M. Farrer, J. Johnson, A. Singleton, S. Hague, J. Kachergus, M. Hulihan, T. Peuralinna, A. Dutra, R. Nussbaum, S. Lincoln, A. Crawley, M. Hanson, D. Maraganore, C. Adler, M. R. Cookson, M. Muenter, M. Baptista, D. Miller, J. Blancato, J. Hardy, and K. Gwinn-Hardy. α-synuclein locus triplication causes Parkinson's disease. *Science*, 302(5646):841, 2003.

[158] T. A. Snijders and C. Baerveldt. A multilevel network study of the effects of delinquent behavior on friendship evolution. *Journal of mathematical sociology*, 27(2-3):123–151, 2003.

[159] T. A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

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

[161] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539, 2006.

[162] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. Van den Broeck, et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC medicine*, 9(1):1, 2011.

[163] M. B. Stern, A. Lang, and W. Poewe. Toward a redefinition of Parkinson's disease. *Movement Disorders*, 27(1):54–60, 2012.

[164] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, Mar. 2003.

[165] M. P. H. Stumpf and M. A. Porter. Critical truths about power laws. *Science*, 335(6069):665–666, 2012.

[166] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102(12):4221–4224, 2005.

[167] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(D1):D71–D75, 2012.

[168] T. Thorne and M. P. H. Stumpf. Graph spectral analysis of protein interaction network evolution. *Journal of The Royal Society Interface*, 9(75):2653–2666, 2012.

[169] V. A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84:016114, 2011.

[170] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

[171] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[172] C.-J. Tsai, B. Ma, and R. Nussinov. Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends in Biochemical Sciences*, 34(12):594–600, 2009.

[173] United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. National ambulatory medical care survey (2009). 2011.

[174] S. Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

[175] M. van den Heuvel, D. Clark, R. Fielder, P. Koundakjian, G. Oliver, D. Pelling, N. Tomlinson, and A. Walker. The international validation of a fixed-dose procedure as an alternative to the classical LD50 test. *Food and Chemical Toxicology*, 28(7):469–482, 1990.

[176] J. van Lint and R. Wilson. *A Course in Combinatorics*. Cambridge University Press, 2001.

[177] A. Vazquez. Optimal drug combinations and minimal hitting sets. *BMC Systems Biology*, 3(1):81, 2009.

[178] L. L. Venda, S. J. Cragg, V. L. Buchman, and R. Wade-Martins. α-Synuclein and dopamine at the crossroads of Parkinson's disease. *Trends in Neurosciences*, pages 559–568, 2010.

[179] T. Vogiatzi, M. Xilouri, K. Vekrellis, and L. Stefanis. Wild type α-synuclein is degraded by chaperone-mediated autophagy and macroautophagy in neuronal cells. *Journal of Biological Chemistry*, 283(35):23542–23556, 2008.

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

[181] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[182] T. V. Votyakova and I. J. Reynolds. $\Delta\Psi_m$-dependent and -independent production of reactive oxygen species by rat brain mitochondria. *Journal of Neurochemistry*, 79(2):266–277, 2001.

[183] S. Wagner and D. Wagner. Comparing clusterings: an overview, 2007.

[184] W. Wang and W. N. Street. Modeling influence diffusion to uncover influence centrality and community structure in social networks. *Social Network Analysis and Mining*, 5(1):15, 2015.

[185] E. W. Weisstein. Chu-Vandermonde Identity. `http://mathworld.wolfram.com/Chu-VandermondeIdentity.html`.

[186] E. W. Weisstein. Bell Number. `http://mathworld.wolfram.com/BellNumber.html`, 2016.

[187] A. White and A. Ma'ayan. Connecting seed lists of mammalian proteins using steiner trees. In *Signals, Systems and Computers, 2007. ACSSC 2007.*, pages 155–159, 2007.

[188] S. J. Wodak, S. Pu, J. Vlasblom, and B. Séraphin. Challenges and rewards of interaction proteomics. *Molecular & Cellular Proteomics*, 8(1):3–18, 2009.

[189] WolframAlpha. `http://www.wolframalpha.com/input/?i=parkinsons+disease`. Accessed 11/02/2014.

[190] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, 2002.

[191] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang. Finding multiple target optimal intervention in disease-related molecular network. *Molecular Systems Biology*, 4:228, 2008.

[192] Q. Yang, H. She, M. Gearing, E. Colla, M. Lee, J. J. Shacka, and Z. Mao. Regulation of neuronal survival factor MEF2D by chaperone-mediated autophagy. *Science*, 323(5910):124–127, 2009.

[193] M. P. Young, S. Zimmer, and A. V. Whitmore. Drug molecules and biology: Network and systems aspects. In J. R. Morphy and C. J. Harris, editors, *Designing Multi-Target Drugs*, RSC Drug Discovery, chapter 3, pages 32–49. Royal Society of Chemistry, 2012.

[194] A. Zanzoni, M. Soler-López, and P. Aloy. A network medicine approach to human disease. *FEBS Letters*, 583(11):1759–1765, 2009.

[195] P. Zhang. Evaluating accuracy of community detection using the relative normalized mutual information. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(11):P11006, 2015.