The Dissertation Committee for Qiaohui Lin
certifies that this is the approved version of the following dissertation:

Inference and Uncertainty Characterization in Complex Structures

Committee:

Purnamrita Sarkar, Co-Supervisor

Peter Müller, Co-Supervisor

Abhra Sarkar

Yuan Ji

Inference and Uncertainty Characterization in Complex Structures

by

Qiaohui Lin

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2022

Acknowledgments

I wish to acknowledge and give my warmest thanks to people who guided and helped me through my pursuit of the doctorate degree. I would like to express sincere gratitude to my supervisors, Dr. Purnamrita Sarkar and Dr. Peter Müller, for their advice and support. The completion of this work could not be possible without their invaluable guidance. I would also like to thank my committee members, Dr. Abhra Sarkar and Dr. Yuan Ji for their encouragement and their brilliant comments and suggestions.

Inference and Uncertainty Characterization in Complex Structures

Qiaohui Lin, Ph.D.
The University of Texas at Austin, 2022

Co-Supervisors: Purnamrita Sarkar
Peter Müller

The common theme of the projects in this thesis is statistical inference and characterizing uncertainty for complex structures, including networks and separately exchangeable data matrices.

In the first two projects, we focus on uncertainty quantification of network subgraph count statistics. In the first project, we study the network jackknife procedure to consistently estimate the variance of subgraph counts under the sparse graphon model. In the second project, we develop a family of network multiplier bootstraps for subgraph counts using linear and quadratic weights. In both projects, we complement our theoretical proofs with simulation studies and real data analysis on social networks.

In the final third project we consider the more elementary questions of how investigators arrive at certain model assumptions, focusing on commonly used

symmetry assumptions known as various forms of exchangeability. In particular, we argue for a more common use of separate exchangeability as a modeling principle. We show how this notion is still ignored in some recent work, but could easily be included.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Graphon models and subgraph counts

Many applications involve data that are naturally characterized as networks. Examples include social network data such as Facebook, and biological data such as brain networks and protein-protein interaction networks. The increasing use of such data calls for more statistical inference tools.

Subgraph count statistics (count functionals) are of great importance in characterizing networks. For example, in biological networks, certain subgraphs may represent functional subunits in the larger system (Milo et al., 2002; Chen and Yuan, 2006; Daudin et al., 2008; Kim et al., 2014). In social networks, the frequency of triangles provides information about mutual friendships (Newman, 2001; Myers et al., 2014; Ugander et al., 2011). Although many results have been established to estimate subgraph counts on very large networks, quantifying the uncertainty of such estimates remains less studied. Such work is critical for statistical inference and has recently attracted much attention. Quantifying the variability of subgraph count estimators not only reveals information about the data generating process, but also is key to construct confidence intervals and conduct tests to compare networks. Such inference will be discussed in details in Chapter 2 and Chapter 3.

Resampling methods including jackknife, bootstrap, and subsampling have been well studied for quantifying uncertainty in independent and identically distributed data, and some recent literatures (reviewed in Section 3.1) have started to study and implement resampling methods on network data. In Chapter 2 and Chapter 3, we develop our new family of network resampling methods in quantifying subgraph counts, including network jackknife procedure and network multiplier bootstrap family. We study the properties of these procedures, compare with current methods and provide application examples on real world data.

1.1.1 The Sparse Graphon Model

Our network resampling methods are developed under sparse graphon models for networks. Sparse graphon models are a very rich family of network models. Many widely used network models such as stochastic block models (Holland et al., 1983) and random dot product graphs (Young and Scheinerman, 2007) are all part of this family. To formally define sparse graphon model, we use the parameterization introduced by Bickel and Chen (2009). Let a size n graph be represented by a $n \times n$ binary adjacency matrix $\{A^{(n)}\}_{n\in\mathbb{N}}$ and latent positions $X_1, \ldots, X_n \sim \text{Unif}[0,1]$. We assume that $A^{(n)}$ is generated by the following model: for $i = j$, $A^{(n)}_{ii} = 0$; for all $i \neq j$, let $\eta_{ij} \sim \text{Uniform}[0,1]$,

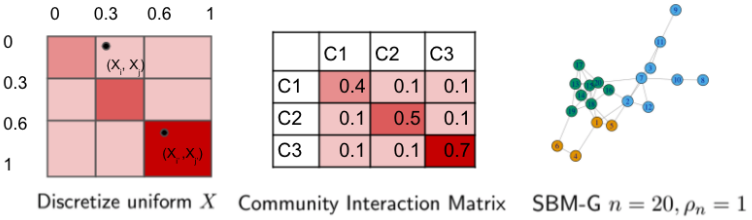
$$A^{(n)}_{ij} = A^{(n)}_{ji} = \mathbb{1}(\eta_{ij} \leq \rho_n w(X_i, X_j)) \sim \text{Bernoulli}(\rho_n w(X_i, X_j)), \tag{1.1}$$

where w is a graphon function that satisfies $w : [0,1]^2 \mapsto \mathbb{R}$ and is a symmetric measurable function such that $\int_0^1 \int_0^1 w(u,v)\, du\, dv = 1$ and $w(u,v) \leq C$ for some $1 \leq C < \infty$. We refer to w as a graphon. The parameter $\rho_n = P(A_{ij} = 1)$ determines the sparsity level of the sequence $\{A^{(n)}\}_{n\in\mathbb{N}}$. Many real world graphs are thought to be sparse, with $o(n^2)$ edges; $\rho_n \to 0$ is needed for graphs generated by (1.1) to exhibit this behavior.

Bounded graphons arise as a limiting object in the theory of graph limits; see Lovász (2012). Alternatively, graphons are a natural representation for (infinite-dimensional) jointly exchangeable arrays, where this notion of exchangeability corresponds to invariance under vertex permutation; see for example, Diaconis and Janson (2008). Bounded graphons subsume many other commonly studied network models, including stochastic block models (Holland et al., 1983) and random dot product graphs (Young and Scheinerman, 2007) as we mentioned above.

While boundedness of the graphon is a common assumption in the statistics literature (see, for example, the review article by Gao and Ma (2019)), it should be noted that unbounded graphons are known to be more expressive. As noted by Borgs et al. (2019), unboundedness allows graphs that exhibit power-law degree distributions, a property that bounded graphons fail to capture. For mathematical expedience, in the present article, we focus on the bounded case, but we believe that our analysis may be extended to sufficiently light-tailed unbounded graphons as well.

Here we illustrate two examples of sparse graphon models. They are also the examples we use in simulation studies in Chapter 2. Chapter 3 also uses similar examples, for which we do not provide details here to avoid repetition. The first example for illustration is a Stochastic Block Model (SBM) (Holland et al., 1983), a commonly used model to study networks with communities. In a SBM with sparsity level ρ_n and m number of communities, let binary matrix $Z \in \{0, 1\}^{n\times m}$ denote



	C1	C2	C3
C1	0.4	0.1	0.1
C2	0.1	0.5	0.1
C3	0.1	0.1	0.7

Figure 1.1: Illustration of a three-community Stochastic Block Model as an example of sparse graphon model: discretizing uniform latent positions and partitioning vertex set (left), community-community interaction matrix (middle), and a realization of size 20 $\rho_n = 1$ graph formed by the model (right).

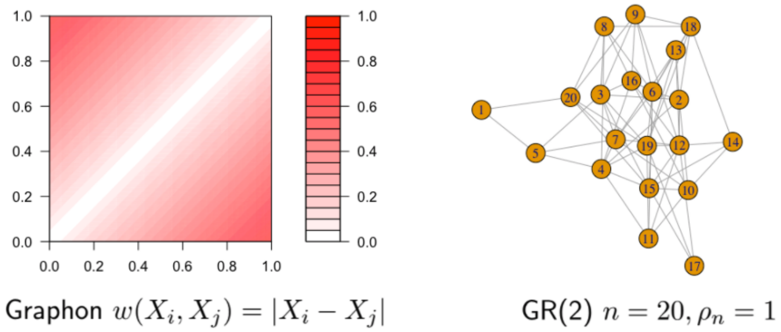


Figure 1.2: Illustration of a continuous graphon model (GR(2)) as an example of a sparse graphon model: graphon function (left), and a realization of size 20 $\rho_n = 1$ graph formed by the model (right).

cluster membership, and B denote a community-community interaction matrix. Let a,b denote community label. Conditioned on $Z_{ia} = 1$ and $Z_{jb} = 1$, nodes i and j form an edge with probability $\rho_n B_{ab}$:

$$P(A_{ij} = 1 \mid Z_{ia} = 1, Z_{jb} = 1) = \rho_n B_{ab}.$$

Here we present in Figure 1.1 an example of three-community SBM, where membership probability of three communities are $(0.3, 0.3, 0.4)$, $B = ((0.4, 0.1, 0.1), (0.1, 0.5, 0.1), (0.1, 0.1, 0.7))$ and a small size $n = 20$, $\rho_n = 1$ graph generated from it.

For the second example of sparse graphon model, we consider a continuous graphon which we call GR(2), where

$$P(A_{ij} = 1 \mid X_i = u, X_j = v) = \rho_n |u - v|.$$

We present in Figure 1.2 the illustration of this graphon model and a size $n = 20$ $\rho_n = 1$ graph simulated from the model.

1.1.2 Subgraph Count Functionals

In this subsection, we present the definitions of subgraph counts. The count functionals that we consider were first studied in Bickel et al. (2011). We first introduce some notation needed to define these functionals. Let G_n denote a graph with vertex set $V(G_n) = \{1, 2, \ldots, n\}$ and edge set $E(G_n) \subset V(G_n) \times V(G_n)$. Let $R \subseteq E(G_n)$ denote the subgraph of interest parameterized by its edge set. For convenience, we assume $V(R) = \{1, 2, \ldots, r\}$. Furthermore, let $G_n[R]$ denote the subgraph induced by the vertices of R.

We consider two different types of count functionals. The first one counts exact matches and has the following probability under the sparse graphon model: (Eq 1.1):

$$P(R) = P(G_n[R] = R) = E\left[\prod_{(i,j)\in R} \rho_n w(X_i, X_j) \prod_{(i,j)\in \overline{R}} (1 - \rho_n w(X_i, X_j))\right] \quad (1.2)$$

We also consider the functional $Q(R)$, which provides the probability of an induced subgraph containing the subgraph R:

$$Q(R) = P(R \subseteq G_n(R)) = E\left[\prod_{(i,j)\in R} \rho_n w(X_i, X_j)\right] \quad (1.3)$$

Note that $Q(R)$ is agnostic to the presence of additional edges. When the graph sequence is sparse, $P(R)$ and $Q(R)$ are uninformative, as $P(R),\ Q(R) \to 0$. Let $s = |E(R)|$ and $r = |V(R)|$. Instead, define the following normalized subgraph frequency:

$$\tilde{P}(R) = \rho_n^{-s}\, P(R) \qquad \tilde{Q}(R) = \rho_n^{-s}\, Q(R) \tag{1.4}$$

Furthermore, let $Iso(R)$ denote the class of graphs isomorphic to R, and $|Iso(R)|$ its cardinality. Our estimator of $\tilde{P}(R)$ is given by:

$$\hat{P}(R) = \rho_n^{-s} \frac{1}{\binom{n}{r}\,|\mathrm{Iso}(R)|} \sum_{S \sim R} \mathbb{1}(S = G_n[S]) \tag{1.5}$$

Similarly, define $\hat{Q}(R)$ as:

$$\hat{Q}(R) = \rho_n^{-s} \frac{1}{\binom{n}{r}\,|\mathrm{Iso}(R)|} \sum_{S \sim R} \mathbb{1}(S \subseteq G_n[S]) \tag{1.6}$$

Due to magnification by ρ_n^{-s}, (1.4), (1.5), and (1.6) are not necessarily upper bounded by 1; nevertheless, they are still meaningful quantities related to subgraph frequencies.

It is not hard to see that $P(R) = Q(R)(1 + O(\rho_n))$.

Examples of Count Functionals As examples of count functionals, we introduce the edge, triangle, Two-star (V-star) density. which we explicitly define below.

$$\hat{P}(\mathrm{Edge}) = \hat{Q}(\mathrm{Edge}) := \frac{\sum_{i<j} A_{ij}}{\binom{n}{2}\rho_n}$$

$$\hat{P}(\text{Triangle}) = \hat{Q}(\text{Triangle}) := \frac{\sum_{i<j<k} A_{ij}A_{jk}A_{ki}}{\binom{n}{3}\rho_n^3}$$

$$\hat{P}(\text{Two-star}) := \frac{\sum_{i,j<k,j,k\neq i} A_{ij}A_{ik}(1-A_{ik})}{\binom{n}{3}\rho_n^2(1-\rho_n)}$$

$$\hat{Q}(\text{Two-star}) := \frac{\sum_{i,j<k,j,k\neq i} A_{ij}A_{ik}}{\binom{n}{3}\rho_n^2}$$

We have developed shortcuts to efficiently calculate these count statistics in large size graphs. Let $C = A^2$ be the matrix square of A. $\hat{P}(\text{Triangle})$ can be expressed as $\sum_{ij}(C_{ij}\cdot A_{ij})/6$. $\hat{P}(\text{Two-star})$ can be expressed as $\sum_{ij}(C_{ij}\cdot(1-A_{ij}))/2$, while $\hat{Q}(\text{Two-star})$ is $\sum_i \binom{d_i}{2}/2$ where d_i is degree of node i. We have also developed approximated count functionals with its algorithm and properties introduced in details in Chapter 3.

1.2 Exchangeability and other paradigms in Bayesian model choices

In statistical inference we usually start with an assumed inference model. In particular, in Bayesian inference many commonly used models take the form of hierarchical models. While investigators usually do not extensively discuss the motivation for such model choices, there are actually broadly applicable theoretical results and representation theorems that one can rely on. We review such principles in some detail in Section 4.1, and provide here just a brief summary. The perhaps most widely invoked principle is exchangeability. In many problems it is natural to require that any inference model should be symmetric with respect to arbitrary re-labeling and change of indices for experimental units (patients, proteins, etc.).

That is, the probability model should be invariant under a permutation of indices of the experimental units.

The classic de Finetti's theorem (de Finetti, 1930) states that exchangeability of an extendable sequence $x_1, \ldots, x_n$ is equivalent to the assumption that it can be expressed as conditionally independently identically distributed from a random probability P, where $P \sim \mathcal{L}$. The de Finetti measure $\mathcal{L}$ can be interpreted as the prior in the Bayes-Laplace paradigm, thus calling for a Bayesian hierarchical model. Partial exchangeability satisfies invariance for permutations of observations only within sub-populations. Thus, the order of observations is only irrelevant when the membership of observations in sub-populations are preserved. Partial exchangeability allows meta analysis for modeling data from related populations and borrowing information across populations. However, it is common in many studies to have more than one type of experimental units. The need of separate exchangeability arises from a more complex structure of data. It allows two types of experimental units with the data usually recorded as a data matrix. For example, in one of the examples we studied different microbiomes observed in different subjects. Separate exchangeability indicates invariance under separate permutations of rows and columns, respectively. Mathematically, using $\overset{d}{=}$ to denote equality in distribution, a data matrix Z is separately exchangeable if

$$Z_{1:n,1:J} \overset{d}{=} Z_{\pi_1(1:n),\pi_2(1:J)}$$

for separate permutations π_1 and π_2 of rows and columns. Under such constructions, the Aldous-Hoover representation theorem (Aldous, 1981; Hoover, 1979) for separately exchangeable arrays allows the data to be modelled in a hierarchical model in terms of latent quantities. This result has motivated Bayesian statistical inference for such array data. We discuss these principles in more details in Chapter 4.

Exchangeability in graphon models Sparse graphon models exhibit vertex exchangeability. The distribution of the random graph is unchanged when node labels are permuted. It is natural to see that the adjacency matrix is thus jointly exchangeable:

$$A_{1:n,1:n} \overset{d}{=} A_{\pi(1:n),\pi(1:n)}$$

using the same permutations of π over rows and columns. Jointly exchangeability is close to separately exchangeability and is also subject to the Aldous-Hoover representation of extendable (infinitely-dimensional) exchangeable arrays. Many models studying graphs and relational data, including the sparse graphon model we study here, are motivated by Aldous-Hoover representation theorems for exchangeable arrays (Bickel and Chen, 2009; Lloyd et al., 2012; Caron and Fox, 2017).

1.3 Nonparametric Bayesian inference

Nonparametric Bayesian (BNP) models are Bayesian models with infinite-dimensional parameters (Ghosh and Ramamoorthi, 2003; Müller et al., 2015). Performing Bayesian inference on such models requires a prior probability model on the infinite-dimensional parameters. Such prior models are known as Bayesian nonparametric priors. The Dirichlet Process (DP), introduced by Ferguson (1973, 1974), an infinite-dimensional priors over distributions with discrete sample draws, as an analogue to the finite-dimensional Dirichlet Prior, is the most popular BNP Priors and remains the cornerstone of many BNP models. For more currently available BNP tools, extensive reviews of recent BNP priors beyond the Dirichlet and related processes can be found in Phadia (2015) and Hjort et al. (2010).

The essence and one of the biggest advantages of BNP models as pointed out by Hjort (2003), is flexibility. The data is not restrictively modelled by a fixed number or a low number of parameters. BNP models allow a growing number of parameters with increasing sample size, and in some cases, even growing number of candidate models.

Over the past few decades, nonparametric Bayesian methods have found a wide range of applications on many problems such regression, survival analysis, hierarchical models, clustering and feature allocation (Müller and Quintana, 2004). In our motivating examples in the third project, we focus on regression and nested random partition problems using BNP models, and discuss the use of separate exchangeability as a modeling principle under these models.

1.4 Contributions

The three project in this thesis include several original contributions to the theory of random graph models as well as to basic modeling principles in Bayesian statistics.

Network Jackknife In Chapter 2, we propose a leave-node-out jackknife procedure for network data and study its properties. Under the sparse graphon model, we prove an Efron-Stein type inequality, showing that the network jackknife is always conservative in expectation as an estimate of the variance for subgraph counts. We also establish consistency of the network jackknife. We complement our theoretical analysis with a range of simulated and real world data examples and show that the network jackknife offers competitive performance in cases where other resampling methods are known to be valid. In fact, for several network statistics, we see that the jackknife provides more accurate inferences compared to related methods such as subsampling.

Network multiplier bootstrap In Chapter 3, we propose a new class of multiplier bootstraps for subgraph counts. Based on first and second-order terms of Hoeffding decomposition of the bootstrapped statistic from multiplier bootstrap respectively, we propose bootstrap procedures with linear and quadratic weights. We show that the quadratic bootstrap procedure achieves higher-order correctness for appropriately sparse graphs. The linear bootstrap procedure requires fewer estimated network statistics, leading to improved accuracy over its higher-order correct counterpart in sparser regimes. To improve the computational properties of the linear bootstrap further, we consider fast sketching methods to conduct approximate subgraph counting and establish consistency of the resulting bootstrap procedure. We complement our theoretical results with a simulation study and real data analysis and verify that our procedure offers state-of-the-art performance for several functionals.

Separate exchangeability. In Chapter 4, we introduce the notion of separate exchangeability as a modeling principle. The main contributions in that chapter are:

(i) recognizing and clarifying the need of respecting separate exchangeability in model constructions and (ii) discussing two specific models that implement separate exchangeability. In many parametric models such structure is naturally respected. However, this is not the case in many nonparametric Bayesian models. We identify one example in recent literature in an analysis of microbiome data, and study one other example with original data from a protein study of a neurodegeneration disease. In both cases we discuss how to use separate exchangeability as a modeling principle. Methodologically, in the latter application we show how separate exchangeability is naturally respected in a nonparametric regression implemented as a popular dependent Dirichlet process model with appropriate choices. In the earlier application we modify a common atoms model for dependent (subject-specific) random probability measures to respect separate exchangeability by carefully introducing parameters in the model to reflect the desired symmetry.

Chapter 2

On the Theoretical Properties of Network Jackknife

This chapter is published in the Proceedings of Machine Learning Research (Lin et al., 2020a). Contribution Statement: My contribution for this chapter includes performing the research, developing the theories and their proofs, developing analytic tools and R code for this work, analyzing data (simulated data and real-world data). I have also contributed to the writing of this chapter.

2.1 Introduction

Network-structured data are now everywhere. The internet is a giant, directed network of webpages pointing to other webpages. Facebook is an undirected network built via friendships between users. The ecological web is a directed network of different species with edges specified by 'who-eats-whom' relationships. Protein-protein interactions are undirected networks consisting of pairs of bait-prey proteins that bind to each other during coaffinity purification experiments arising in mass spectrometry analysis.

In these application areas, it is often of interest to characterize a network using statistics such as the clustering coefficient, triangle density, or principal eigenvalues. There has been a substantial amount of work on approximating these quantities with small error on massive networks Feige (2006); Goldreich and Ron (2008); Assadi et al. (2018); Eden et al. (2017); Gonen et al. (2010); Kallaugher et al. (2019). However, comparatively little attention has been paid to assessing the variability of these statistics with a few exceptions that we will discuss shortly. Quantifying the uncertainty of these estimators is of utmost importance, as it gives us information about the underlying variability of the data generating process. Take for example the problem of comparing two networks, which is a key question in many biological applications and in social network analysis. A natural direction would be to first obtain resamples of networks to construct distributions of different summary statistics and then compare these distributions. While there has been some recent interest in two-sample tests for networks Kim et al. (2014); Durante and Dunson (2018); Ghoshdastidar et al. (2017); Tang et al. (2017), very few works use resampling to compare networks.

Resampling methods have a long and celebrated history in statistics, with the bootstrap, jackknife,and subsampling being the three main forms. There is a now vast literature developing these methods for IID data; for seminal works in this area, see Quenouille (1949); Efron and Tibshirani (1986); Bickel et al. (1997); Politis et al. (1999); Shao and Wu (1989). Even when the data are not independent, resampling methods have been shown to yield asymptotically valid inferences for a wide range of functionals under various dependence structures. For weakly dependent time series, for example, the key innovation is to resample contiguous blocks of data instead of individual observations. Under mild conditions on the block length and nature of dependence, blocked variants of resampling methods, including the block bootstrap

(Künsch, 1989), block subsampling (Politis and Romano, 1994), and the blockwise jackknife (Künsch, 1989) have been shown to asymptotically capture the dependence structure of the data, leading to theories that closely resemble the corresponding theories for IID data.

Recently, some work has started to emerge involving resampling procedures for networks. Levin and Levina (2019) propose two bootstrap procedures for random dot product graphs that involve estimating the latent positions and resampling the estimated positions to conduct inference for the functional of interest. The authors establish bootstrap consistency for functionals that are expressible as U-statistics of the latent positions, which encompasses many important classes of functionals including subgraph counts.

Lunde and Sarkar (2019) consider a procedure that involves subsampling nodes and computing functionals on the induced subgraphs. This procedure is shown to be asymptotically valid under conditions analogous to the IID setting; that is, the subsample size must be $o(n)$ and the functional of interest must converge to a non-degenerate limit distribution. Previously, Bhattacharyya and Bickel (2015) had shown the validity of subsampling for count functionals. By proving a central limit theorem for eigenvalues, Lunde and Sarkar (2019) also establish subsampling validity for these functionals under certain conditions. Finally, in Green and Shalizi (2017), the authors propose sieve and nonparametric bootstrap procedures for networks.

We would like to note that that both the sieve approach of Green and Shalizi (2017) and the latent position estimation approaches of Levin and Levina (2019) depend on accurately estimating the underlying graphon. The nonparametric bootstrap procedure described in Green and Shalizi (2017) requires resampling much larger networks from a size n network, leading to computational inefficiency. Even subsampling requires weak convergence and a known rate of convergence; it turns out that the latter may be estimated (Bertail et al., 1999), but doing so entails a substantial increase in computation and is likely to adversely affect the finite-sample performance of the procedure. While asymptotically valid under general conditions, the finite-sample performance of subsampling methods is known to be sensitive to the choice of tuning parameters; see for example, Kleiner et al. (2014).

2.1.1 Our Contribution

In the present work, we study the properties of a network jackknife introduced by Frank and Snijders (1994) under the sparse graphon model. On the theoretical side, we make two primary contributions. First, analogous to the IID setting, we show that the network jackknife produces variance estimates that are conservative in expectation under general conditions. Our result here justifies the network jackknife as a rough-and-ready tool that produces reasonable answers (erring on the side of caution) even when the asymptotic properties of the functional of interest are poorly understood.

While the upward bias of the network jackknife is a favorable property, it does not provide information as to how the jackknife compares to other resampling procedures for more well-understood functionals. As another theoretical contribution, we establish consistency of the jackknife for a general class of count statistics studied in Bickel et al. (2011). We also extend our result to smooth functions of counts, which encompasses widely used measures such as the transitivity coefficient.

We complement our theoretical results with an empirical investigation of the network jackknife on both simulated and real datasets. In our simulation study, we study the rate of convergence of the jackknife variance estimate for two sparse graphon models under a range of choices for the network functional. Our results suggest that by and large, the jackknife has better finite-sample properties than subsampling. For real data, we conduct network comparisons of Facebook networks constructed from a number of different colleges such as Caltech, Berkeley, Stanford, Wellesley, etc.

The paper is organized as follows. In Section 2.2, we do problem setup and introduce notation. In Section 2.3, we present our theoretical results and some proof sketches. Finally in Section 3.6 we present experimental results on simulated and real networks.

2.2 Background

In this section, we first briefly recall the original Jackknife for IID data. Then we describe the network jackknife procedure under sparse graphon models. Recall that the sparse graphon models are introduced in Section 1.1.1.

2.2.1 The Jackknife for IID Data

The jackknife, attributed to Quenouille (1949) and Tukey (1958), is a resampling procedure that involves aggregating leave-one-out estimates. More precisely, let $Y_1, \ldots, Y_n \sim P$ and let S_n be a permutation-invariant function of n variables. Furthermore, let $S_{n,i}$ denote the functional computed on the dataset with Y_i removed and let $\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n} S_{n,i}$. The jackknife estimate of the variance of $S_{n-1} = S(Y_1, \ldots, Y_{n-1})$ is given by:

$$\widehat{\mathrm{Var}}_{\mathrm{JACK}}\, S_{n-1} := \sum_{i=1}^{n} (S_{n,i} - \bar{S}_n)^2 \tag{2.1}$$

For appropriately smooth functionals, it is well-known that the jackknife consistently estimates the variance; see for example, Shao and Tu (1995). The bootstrap, introduced by Efron (1979), typically requires weaker regularity conditions than the jackknife for consistency. In fact, it is well-known that the jackknife is inconsistent for the median (Miller, 1974) while the bootstrap variance remains consistent under reasonable conditions (Ghosh et al., 1984)[1].

However, for more complicated functionals, it may often be the case that both the bootstrap and the jackknife are inconsistent[2]. Even in these cases, the jackknife still provides reasonable answers. The remarkable inequality of Efron and Stein (1981) asserts that the jackknife is always upwardly biased, ensuring a conservative estimate of the variance.

Since networks are inherently high-dimensional objects, asymptotic results are often harder to come by compared to the IID setting. The theory of the jackknife for IID processes suggests that, even in this challenging regime, a network analogue of the jackknife may have some advantageous properties.

[1] It should be noted that the delete-d jackknife is valid under more general conditions; see Shao and Wu (1989).

[2] Recent work by Fang and Santos (2019) suggests that Hadamard differentiability of g is both necessary and sufficient for bootstrap consistency of $g(\hat{\theta}_n)$ whenever $\hat{\theta}_n$ is asymptotically Gaussian.

2.2.2 The Network Jackknife Procedure

Let $f : \{0, 1\}^{n-1 \times n-1} \mapsto \mathbb{R}$ denote a function that takes as input a $n-1 \times n-1$ adjacency matrix and let $Z_{n,i}$ denote the random variable formed by applying f to an induced subgraph with node i removed. Under the model (1.1), observe that each induced subgraph formed by leaving a node out is identically distributed as a consequence of vertex exchangeability. Therefore, functionals calculated on these induced subgraphs are similar in spirit to the the leave-one-out estimates for the jackknife in the IID setting. Following Frank and Snijders (1994), a natural generalization of the jackknife to the sparse graphon setting is given by:

$$\widehat{\mathrm{Var}}_{\mathrm{JACK}}\, Z_{n-1} := \sum_{i=1}^{n} (Z_{n,i} - \bar{Z}_n)^2 \tag{2.2}$$

where $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_{n,i}$ and $\widehat{\mathrm{Var}}_{\mathrm{JACK}}\, Z_{n-1}$ is an estimate of Var Z_{n-1}, the variance with respect to an induced subgraph with node set $\{1, \ldots, n-1\}$. We would like to note that letting $Z_{n-1} := Z_{n,n}$ constitutes a slight abuse of notation since ρ_{n-1} need not equal ρ_n, but doing so substantially improves the readability of our proofs.

2.3 Theoretical Results

2.3.1 The Network Efron-Stein Inequality

The first result we state here is our generalization of the Efron-Stein inequality to the network setting. Intuitively, the Efron-Stein inequality may be thought of as a general property of functions of independent random variables. While edges in the adjacency matrix are dependent through the latent positions, the fact that they are functions of independent random variables allow us to prove the following:

Theorem 1 (Network Efron-Stein Inequality)**.** *If Z_{n-1} is a permutation invariant statistic, then we have,*

$$\operatorname{Var} Z_{n-1} \leq E(\widehat{\operatorname{Var}}_{\text{JACK}}\, Z_{n-1}) \tag{2.3}$$

The main ingredients in our proof are an adaptation of a martingale argument due to Rhee and Talagrand (1986) and an appropriate filtration for graphon models inspired by Borgs et al. (2008). We provide a proof sketch below; for details, see Appendix A.1.

Proof Sketch. As discussed in the Supplementary Material, for $1 \leq i \leq n$, we may express $Z_{n,i}$ as a measurable function of latent positions $X_i \sim \text{Unif}[0,1]$ for $1 \leq i \leq n$ and $\eta_{ij} \sim \text{Unif}[0,1]$ for $1 \leq i < j \leq n$. More precisely, $Z_{n,i}$ is a function of the variables that are not shaded below:

$$Z_{n,i} = g\left(\begin{array}{ccccccc} X_1 & \eta_{12} & \eta_{13} & \dots & \eta_{1i} & \dots & \eta_{1n} \\ & X_2 & \eta_{23} & \dots & \eta_{2i} & \dots & \eta_{2n} \\ & & X_3 & \dots & \eta_{3i} & \dots & \eta_{3n} \\ & & & \dots & \dots & \dots & \dots \\ & & & X_i & \eta_{i,i+1} & .. & \eta_{in} \\ & & & & & X_{n-1} & \eta_{n-1,n} \\ & & & & & & X_n \end{array}\right). \tag{2.4}$$

We design a martingale difference sequence d_i,

$$d_i = E(Z_{n-1}|\Sigma_i) - E(Z_{n-1}|\Sigma_{i-1}), \tag{2.5}$$

based on filtration Σ_i:

$$\Sigma_i = \sigma\{X_1, , X_i, \eta_{12}, , \eta_{1i}, \eta_{23}, , \eta_{2i}, , , \eta_{i-1,i}\}$$

$$= \sigma \left\{ \begin{array}{ccccc} X_1 & \eta_{12} & \dots & \eta_{1,i-1} & \eta_{1i} \\ & X_2 & \dots & \eta_{2,i-1} & \eta_{2i} \\ & & \dots & .. & \\ & & X_{i-2} & \eta_{i-2,i-1}, & \eta_{i-2,i} \\ & & & X_{i-1} & \eta_{i-1,i} \\ & & & & X_i \end{array} \right\}. \tag{2.6}$$

Then we can show that,

$$\text{Var } Z_{n-1} = \sum_{i=1}^{n-1} E d_i^2.$$

On the other hand, the expectation of Jackknife estimate is:

$$E \sum_{i=1}^{n} (Z_{n,i} - \bar{Z}_n)^2 = (n-1) \frac{E(Z_{n,1} - Z_{n,2})^2}{2}. \tag{2.7}$$

Now, we construct another filtration $\mathcal{A}$ such that $E(Z_{n,1}|\mathcal{A}) = E(Z_{n,2}|\mathcal{A})$. In particular, we use:

$$\mathcal{A} = \sigma\{X_3, \dots, X_{i+1}, \eta_{34}, \dots, \eta_{3,i+1}, \dots, \eta_{i,i+1}\}. \tag{2.8}$$

This is essentially Σ_{i+1}, with the first and second row and columns removed. Define

$$U = E(Z_{n,1}|\Sigma_{i+1}) - E(Z_{n,1}|\mathcal{A})$$

$$V = E(Z_{n,2}|\Sigma_{i+1}) - E(Z_{n,2}|\mathcal{A})$$

Using the fact that $E(X^2) = E(E(X^2|\Sigma)) \geq E(E[X|\Sigma)^2)$ for some random variable X which is measurable w.r.t to some Sigma field Σ, we get:

$$E(Z_{n,1} - Z_{n,2})^2 \geq E(U - V)^2 = 2E(d_i^2) \tag{2.9}$$

The result follows from plugging in Eq 2.9 to Eq 2.7. □

Remark 1. Using the aforementioned filtration for graphon models, is also possible to prove another network variant of the Efron-Stein inequality following arguments in Boucheron et al. (2004). This alternative procedure does not require the functional to be invariant to node permutation and allows flexibility with the leave-one-out estimates. However, the resulting estimate is often not sharp. See the Appendix A.6 for more details.

2.3.2 Beyond the Efron-Stein inequality

While the Efron Stein inequality in Theorem 1 is surprising and useful for estimating uncertainty for network statistics, it would be much more satisfying if indeed the jackknife estimate of variance in fact coincided with the true underlying variance, at least asymptotically. We want to draw the attention of the reader to leftmost panel in Figure 2.1. The solid black line shows the mean and standard error of the ratio between the jackknife estimate of the variance and the true variance for edge density for a blockmodel and a smooth graphon (details in Section 3.6), as graph size grows. This figure shows the surprising trend that, in fact, the jackknife estimate is not only an upper bound on the true variance; it is in fact asymptotically unbiased. Our next proposition establishes exactly that. In what follows, let Z_n denote the edge density (see Section 3.6).

Proposition 1. *Suppose that* $\int_0^1 \int_0^1 w^2(u,v)\ du\ dv < \infty$ *and* $n\rho_n \to \infty$. *Let* $\sigma^2 = \lim_{n\to\infty} n \cdot \mathrm{Var}(Z_{n-1})$. *Then,*

$$n \cdot E(\widehat{\mathrm{Var}}_{\mathrm{JACK}}\ Z_{n-1}) \to \sigma^2 \tag{2.10}$$

The proof of the above result involves tedious combinatorial arguments and is deferred to the Appendix A.4. The above proposition says that, the jackknife estimate of variance of the edge density of a sparse graphon model (see Eq 1.1), in expectation, converges to the true variance. This is a somewhat weak result, since it does not say anything about the jackknife estimate obtained from one network. However, it begs the question, whether a stronger result is true. In fact, in the next section, we prove that for a broad class of count functionals, the jackknife estimate is in fact consistent. This paves the way to the next section, which we start by introducing count functionals.

2.3.3 Jackknife Consistency for Count Functionals

In this section, we study the properties of the jackknife for subgraph counts, which are an important class of functionals in network analysis. In graph limit theory, convergence of a sequence of graphs can be defined as the convergence of appropriate subgraph frequencies (Lovász, 2012). More practically, subgraph counts have been used to successfully conduct two-sample tests in various settings. In social networks, for example, the frequency of triangles provides information about the likelihood of mutual friendships/connections and is therefore a useful summary statistic.

Recall the definitions of count functionals in Section 1.1.2. Bickel et al. (2011) establish a central limit theorem for these functionals under general conditions on the sparsity level and structure of the subgraph. Under analogous conditions, we establish the following consistency result:

Theorem 2 (Jackknife Consistency for Counts)**.** *Suppose that R is acyclic graph* *and* $n\rho_n \longrightarrow \infty$ *or* R *is a simple* r*-cycle and* $n^{r-1}\rho_n^r \longrightarrow \infty$*. Furthermore, suppose that* $\int_0^1 \int_0^1 w^{2s}(u,v)\,du\,dv < \infty$*. Let* $\sigma^2 = \lim_{n\to\infty} n \cdot \mathrm{Var}\,\hat{P}(R)$*. Then,*

$$n \cdot \widehat{\mathrm{Var}}_{\mathrm{JACK}}\,\hat{P}(R) \xrightarrow{P} \sigma^2 \tag{2.11}$$

Our proof relies on a signal-noise decomposition of the jackknife variance. Bickel et al. (2011) establish that the variance of a count functional is largely driven by the variance of a U-statistic related to the edge structure of the subgraph. For this U-statistic component of the decomposition, results for jackknifing U-statistics due to Arvesen (1969) may be used to show convergence in probability towards the variance of the corresponding U-statistic. Since the jackknife is a sum of squares, we are able to decouple the effects of a remainder term and show that it is negligible. We provide a sketch below, and defer the details to Appendix A.2.

Proof Sketch. Define density (normalized counts) of R when leaving ith node out is $Z_{n,i} = \binom{n-1}{r}^{-1}\rho_n^{-s}(T - T_i)$, where T is total counts of R in G_n, r and s are number of vertices and edges in R. Define $Z_n = \binom{n}{r}^{-1}\rho_n^{-s}T$. Then $\widehat{\mathrm{Var}}_{\mathrm{JACK}} = \sum_i (Z_{n,i} - \overline{Z_n})^2$, $\mathrm{Var}\,\hat{P}(R) = \mathrm{Var}\,Z_n$.

Theorem 1 of Bickel et al. (2011) establishes that $n\mathrm{Var}(Z_n)$ converges to a positive constant. Thus we scale $\widehat{\mathrm{Var}}_{\mathrm{JACK}}$ by n, and decompose $n\widehat{\mathrm{Var}}_{\mathrm{JACK}}$ into

$$n\left[\sum_i (\alpha_i - \bar{\alpha}_n)^2 - 2\sum_i (\alpha_i - \bar{\alpha}_n)(\beta_i - \bar{\beta}_n) + \sum_i (\beta_i - \bar{\beta}_n)^2\right], \tag{2.12}$$

where $\alpha_i = Z_{n,i} - E(Z_{n,i}|\boldsymbol{X}_n)$, $\beta_i = E(Z_{n,i}|\boldsymbol{X}_n)$, $\bar{\alpha}_n = \frac{1}{n}\sum_{i=1}^n \alpha_i$, $\bar{\beta}_n = \frac{1}{n}\sum_{i=1}^n \beta_i$ and $\boldsymbol{X}_n = (X_1, \ldots X_n)$. The term $n\sum_i (\beta_i - \bar{\beta}_n)^2$ corresponds to the signal component discussed before the theorem statement.

We show in the Supplement that $E\sum_i(\alpha_i - \bar{\alpha}_n)^2$ can be further written into $E\left[\sum_{S,T} cov(S,T|\boldsymbol{X}_n)\right], \forall S,T \sim R$. By Bickel et al. (2011), $E\left[\sum_{S,T} cov(S,T|\boldsymbol{X}_n)\right] = o(\frac{1}{n})$. $n\sum_i(\alpha_i - \bar{\alpha}_n)^2$ is thus negligible by Markov Inequality. The cross term $n\sum_i(\alpha_i - \bar{\alpha}_n)(\beta_i - \bar{\beta}_n)$ is also negligible by the Cauchy-Schwartz Inequality.

□

Remark 2. Our theoretical results hold for both notions of subgraph frequencies. However, note that $\tilde{Q}(R)$ is independent of n, but $\tilde{P}(R)$ depends on n and approaches $\tilde{Q}(R)$. While $\sqrt{n}[\hat{P}(R) - \tilde{P}(R)]$ and $\sqrt{n}[\hat{Q}(R) - \tilde{Q}(R)]$ have the same limiting variance, inference for a fixed target using $\hat{P}(R)$ requires stronger sparsity conditions; namely $\rho_n = o(1/\sqrt{n})$. See Section 2.3.4 for a related discussion.

Remark 3. Central limit theorems and jackknife consistency can also be shown for more general (cyclic) graphs. However, in these cases, more stringent sparsity conditions are needed.

Now, let $f(G_n)$ denote a function of the vector $(\hat{P}(R_1), \ldots, \hat{P}(R_d))$. Furthermore, let ∇f denote the gradient of f and $\mu \in \mathbb{R}^d$ the limit of $(\tilde{P}(R_1), \ldots, \tilde{P}(R_d))$ as $n \longrightarrow \infty$; it turns out that μ corresponds to an integral parameter of the graphon related to the edge structure of the subgraph. We have the following result.

Theorem 3 (Jackknife Consistency for Smooth Functions of Counts)**.** *Suppose that $(R_1, \ldots, R_d)$ are simple cycles and $n^{r_i - 1}\rho_n^{r_i} \longrightarrow \infty$ for $i \in \{1, \ldots, d\}$ corresponding to simple cycles, or acyclic graphs and $n\rho_n \longrightarrow \infty$. Let $s^* = \max\{|E(R_1),$ $\ldots E(R_d)\}$ and suppose that $\int_0^1 \int_0^1 w^{2s^*}(u,v)\, du\, dv < \infty$. Furthermore, suppose* *that ∇f exists in a neighborhood of μ, $\nabla f(\mu) \neq 0$, and that ∇f is continuous at μ. Let σ_f^2 is the asymptotic variance of $\sqrt{n}[f(G_n) - f(E(G_n))]$. Then,*

$$n \cdot \widehat{\mathrm{Var}}_{\mathrm{JACK}}\, f(G_n) \xrightarrow{P} \sigma_f^2$$

Proof Sketch. Let $Z_{n,i} = (Z_{n,i}(1), \dots Z_{n,i}(d))$, where d is a constant w.r.t n and each entry corresponds to a count functional with node i removed. Let $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_{n,i}$. We use a Taylor expansion of $f(Z_{n,i})$ around $\bar{Z}_n$.

$$f(Z_{n,i}) = f\big(\bar{Z}_n\big) + \nabla f(\mu)^T(Z_{n,i} - \bar{Z}_n) + \underbrace{(\nabla f(\zeta_i) - \nabla f(\mu))^T(Z_{n,i} - \bar{Z}_n)}_{E_i},$$

where $\zeta_i = (\zeta_{i1}, \dots, \zeta_{id}) = c_i Z_{n,i} + (1 - c_i)\bar{Z}_n$ for some $c \in [0, 1]$. Thus, we also have:

$$f(Z_{n,i}) - \overline{f(Z_{n,i})} = \underbrace{\nabla f(\mu)^T(Z_{n,i} - \bar{Z}_n)}_{I_i} + \underbrace{E_i - \frac{1}{n}\sum_i E_i}_{II_i} \tag{2.13}$$

We bound $n\sum_i (I_i)^2$ and $n\sum_i (II_i)^2$ separately. Let Σ denotes the covariance matrix of a multivariate U-statistic with kernels $(h_1, \dots, h_d)$, where each h_j is the kernel corresponding to the count functional in the j^{th} coordinate of the vector Z_n (see Eq A.7 for detailed definition). We can show that

$$\left| n\sum_i (I_i)^2 - \nabla f(\mu)^T \Sigma \nabla f(\mu) \right| = o_p(1).$$

We can also show that $n\sum_i (II_i)^2$ is also $o_p(1)$. Then, let $\mu_n = E[Z_n]$. Note that if one counts subgraphs by an exact match as in Bickel et al. (2011) $\mu_n \to \mu$. If one counts subgraphs via edge matching, $\mu_n = \mu$. Thus, both these types of subgraph densities, which asymptotically have the same limit, can be handled by our theoretical results. By Theorem 3.8 in Van der Vaart (2000),

$$\sqrt{n}(f(Z_n) - f(\mu_n)) \rightsquigarrow N(0, \nabla f(\mu)^T \Sigma \nabla f(\mu))$$

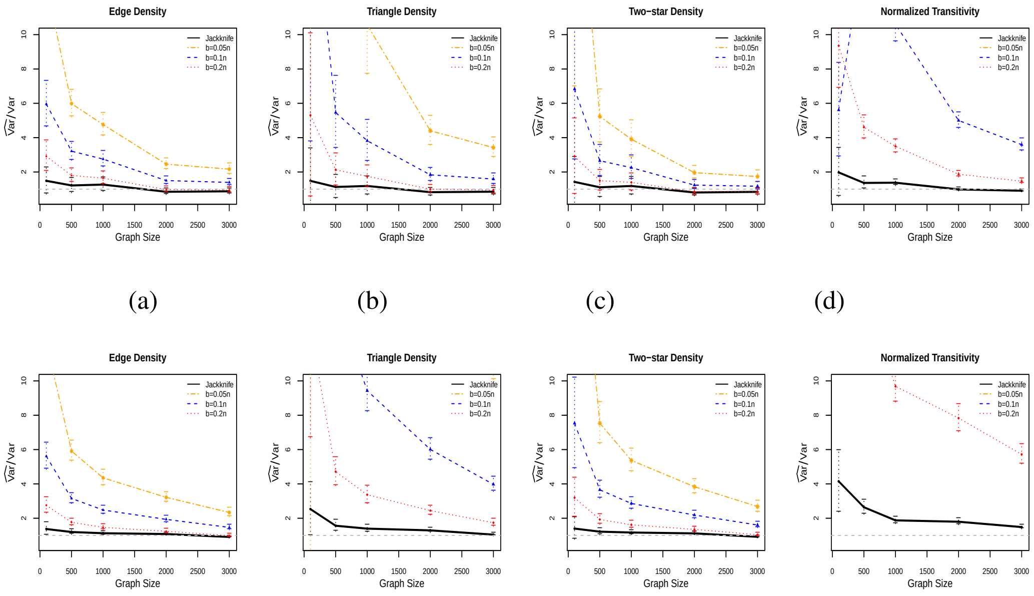
This shows that the jackknife estimate of variance converges to the asymptotic variance of $f(Z_n)$. The proof is deferred to the Appendix A.3. $\square$

2.3.4 A Remark on the use of the Network Jackknife for Two-Sample Testing

In principle, the jackknife variance provides a quantification of uncertainty that may be used for many inference tasks. When the limiting distribution is Normal, one may use a Normal approximation; otherwise, one may use Chebychev's inequality. However, in these cases, the centering is $\theta_{n-1} = E(Z_{n-1})$, which depends on n. Inferences about θ_{n-1} are often useful for a single graph, but for two-sample testing, issues may arise when comparing networks of different sizes. Probability statements involving a fixed population parameter θ are needed. To ensure that the jackknife yields valid inferences for an appropriate population parameter, we will need to impose some additional assumptions. In what follows, let $\{\tau_n\}_{n\in\mathbb{N}}$ denote a sequence of normalizing constants and let $U_{n-1} = \hat{\theta}_{n-1} - \theta$ for some $\theta \in \mathbb{R}$. We have the following result:

Proposition 2. *Suppose that $\tau_n \to \infty$ and $\tau_n U_n \rightsquigarrow U$ for some non-degenerate U with mean 0 and variance σ^2 and $\{(\tau_n U_n)^2\}_{n\in\mathbb{N}}$ is uniformly integrable. Then,*

$$\tau_n(\hat{\theta}_n - E(\hat{\theta}_n)) \rightsquigarrow U, \ \frac{\operatorname{Var} U_n}{\operatorname{Var} \hat{\theta}_n} \to 1 \tag{2.14}$$



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.1: Ratio of Jackknife estimate $\widehat{\mathrm{Var}}_{\mathrm{JACK}}$ to true variance Var for edge density, triangle density, two-star density and transitivity in size $n = 100, 500, 1000, 2000, 3000$ graphs simulated from the SBM (top) and GR2 (bottom), compared to subsampling with $b = 0.05n$, $b = 0.1n$, $b = 0.2n$ variance estimation on the same graphs.

As a consequence of Proposition 2, if a central limit theorem is known for $\sqrt{n}\, U_n$ and a uniform integrability condition is satisfied, then one may use the jackknife variance in conjunction with a Normal approximation to conduct (possibly conservative) inference for θ. For count functionals, we have mentioned when this condition holds, so in this case it does not need to be checked.

2.4 Experiments

In this section, we present simulation experiments and experiments on real data. For simulations, we compare our variance estimate with that estimated using subsampling. We present our results on two graphons. For real data, we compare networks based on C.I.'s constructed using jackknife estimates of variance of network statistics like edge or triangle density and normalized transitivity.

Count functionals used In this chapter, we consider the edge, triangle, two star density introduced in Section 1.1.2. We use $\hat{P}(\text{Edge})$ (same as $\hat{Q}(\text{Edge})$), $\hat{P}(\text{Triangle})$ (same as $\hat{Q}(\text{Triangle})$) and $\hat{Q}(\text{Two-star})$ (asymptotically same as $\hat{P}(\text{Two-star})$). As a smooth function of count statistics, we use:

$$\text{Normalized transitivity} := \frac{\hat{P}(\text{Triangle})}{\hat{Q}(\text{Two-star})}.$$

Here we use $\hat{Q}(\text{Two-star})$ instead of $\hat{P}(\text{Two-star})$ for the purpose of computation simplicity and for the fact that $\hat{Q}$ is independent of n. In a hypothesis test, it is more natural to use a statistic whose expectation does not depend on n, since one may wish to compare two networks of different sizes. See Remark and Section 2.3.4 for details.

2.4.1 Simulated Data

We simulate graphs from two different graphons. The first is a Stochastic Block Model (SBM) Holland et al. (1983), which is a widely used model for networks with communities. A SBM is characterized by a binary cluster membership matrix

$Z \in \{0, 1\}^{n \times r}$, where r is the number of communities, and a community-community interaction matrix B. Conditioned on $Z_{ia} = 1$ and $Z_{jb} = 1$, nodes i and j form a link with probability B_{ab}. We use $B = ((0.4, 0.1, 0.1), (0.1, 0.5, 0.1), (0.1, 0.1, 0.7))$ and generate a Z from a Multinomial$(0.3, 0.3, 0.4)$.

For the other graphon, we consider the following parameterization:

$$h_n(u, v) = P(A_{ij} = 1 \mid X_i = u, X_j = v) = \nu_n |u - v| \tag{GR2}$$

where ν_n is a sparsity parameter. We use $\nu_n = n^{-1/3}$. We denote this graphon by GR2.

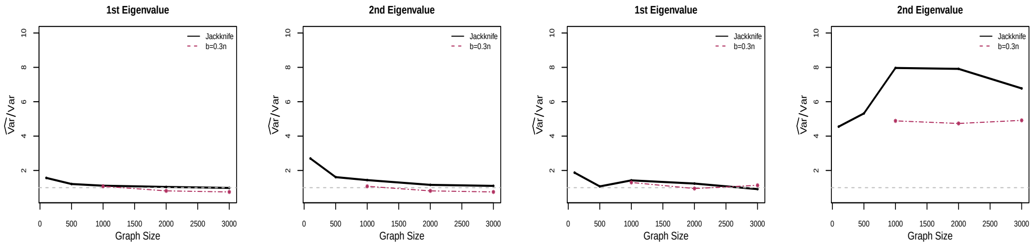
From these two graphons, we consider graph size n of $n = 100, 500, 1000, 2000,$ 3000. For each n, we simulated 100 graphs to calculate the approximate true variance of edge density, triangle density, two-star density and normalized transitivity among these graphs.

Computation: For each simulated network, we remove one node at a time, recalculate a statistic $Z_{n,i}$ on the graph with $(n-1)$ nodes left. Next we compute the jackknife estimate of the variance $\widehat{\text{Var}}_{\text{JACK}} := \sum_i (Z_{n,i} - \bar{Z}_n)^2$, where $\bar{Z}_n$ is the average of the $Z_{n,i}$'s. It should be noted that for some statistics, jackknife, owing to its leave-one-node-out characteristic, can be implemented to reduce computation. For example, in calculating triangles, we calculate the number of triangle on the whole graph once and the number of triangles each node is involved in. This can be done by keeping track of the number of common neighbors between a node and its neighbors.

For each statistic mentioned, we report the mean of the ratio $\widehat{\text{Var}}_{\text{JACK}}/\text{Var}\, Z_n$ among 100 graphs of each n and a 95% confidence interval from the standard deviation from a normal approximation of these 100 ratios. We also plot a dotted line to denote 1. Closer to this line a resampling procedure is, the better. In Figure 2.1, we plot this on the Y axis with n on the X axis. We also plot the same for subsampling with $b = 0.05n$, $b = 0.1n$, $b = 0.2n$ performed on the same graphs. Figure 2.1 a,b,c, and d contain results for the SBM, whereas the rest are for the smooth graphon.

We see that for both graphons, $\widehat{\mathrm{Var}}_{\mathrm{JACK}}/\mathrm{Var}$ converges to 1 much more quickly in comparison to subsampling and has much smaller variance. These figures also show how susceptible the performance of subsampling is to the choice of b. For $b = 0.05n$, subsampling overestimates the variance, and exceeds the upper bound on Y axis of some of the figures. In Figure 2.1 (h) we see that $\widehat{\mathrm{Var}}_{\mathrm{JACK}}$ for the normalized transitivity converges slowly for GR2, and subsampling with all choices of b are worse as well.

Eigenvalues: Here we examine the performance of jackknife on assessing the variance of eigenvalues, to which we have not yet extended our theoretical guarantees. In Figure 2.2 we show the $\widehat{\mathrm{Var}}_{\mathrm{JACK}}/\mathrm{Var}$ for the two principal eigenvalues of the SBM ((a) and (b)) and two graphons described before. Here we only compared with subsampling with $b = 0.3n$ and $n = 1000, 2000, 3000$ as subsampling for eigenvalues in sparse graphs only works asymptotically for very large n Lunde and Sarkar (2019). For smaller n and b in our sparsity setting, we saw that subsamples of adjacency matrices often were too sparse leading to incorrect estimate of the variance. Let us first look at Figure 2.2 (a) and (b) for the SBM setting. For both the eigenvalues in this case, $\widehat{\mathrm{Var}}_{\mathrm{JACK}}/\mathrm{Var}$ converges to 1, whereas subsampling consistently underestimates



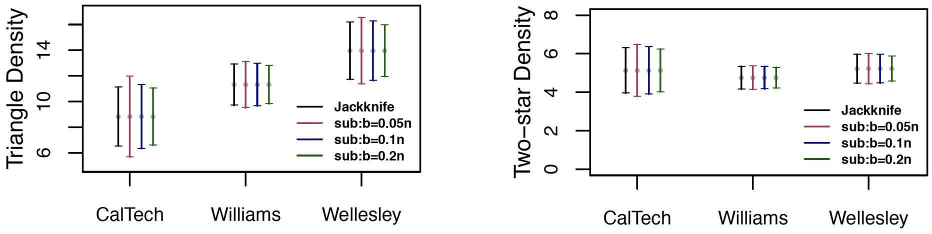
(a) (b) (c) (d)

Figure 2.2: Ratio of Jackknife estimate $\widehat{\text{Var}}_{\text{JACK}}$ to true variance Var for first and second eigenvalues in size $n = 100, 500, 1000, 2000, 3000$ graphs simulated from stochastic block model in (a) and (b) and the graphon GR2 in (c) and (d), compared to subsampling with $b = 0.3n$ variance estimation on the same graphs.

the true variance. For graphon GR2, we see from Figure 2.2 (c) that both jackknife and subsampling estimate the true variance well, whereas for the second eigenvalue (see (d)) they both perform extremely poorly. These preliminary results of jackknife estimates show tentative evidence that our theory can be applied to statistics beyond count statistics, like eigenvalues, which we aim to investigate in future work.

2.4.2 Real-world Data

We present two experiments using Facebook network data Rossi and Ahmed (2015). In the first experiment, we compared three colleges: Caltech, Williams and Wellesley. While Caltech is known for its strength in natural sciences and engineering, Williams and Wellesley are strong liberal arts colleges. They all have relatively small number of students (800-3000), but with different demographics. For example, Wellesley is a women's liberal arts college, whereas the other have a mixed population. We present the 95% confidence intervals (CI) obtained using a normal approximation



(A) (B)

Figure 2.3: (A) Triangle density , and (B) two-star density (bottom) and their CI's based on jackknife and subsampling variance estimates.

with the estimated variances for two-star and triangle densities for these networks.

We see that while all three have similar two-star density, Wellesley has significantly higher triangle density. We also see that CI's from jackknife and subsampling with $b = 0.1n$ and $b = 0.2n$ are comparable. Subsampling with $b = 0.05n$ tends have a wider CI, as it overestimates the variance. It is interesting to note that, for triangles, subsampling with $b = 0.2n$ took nearly 10 times as much time as jackknife, since we used the leave-one-node-out structure. In comparison, for both methods, two-star counting is overall much faster than counting triangles.

In the second experiment, we look at three college pairs: Berkeley and Stanford, Yale and Princeton, Harvard and MIT. First we decide which statistic differentiates between a given pair. For this, we split each college data set in half, into a training set and test set. On each of training set, we estimated their triangle density, two-star density, normalized transitivity and their variances estimated by jackknife, demonstrated in Table 2.1. Interestingly, in Table 2.1, the triangle density is large for all colleges, owing to the sparsity of the networks. From Table 2.1 we can see normalized transitivity estimates have relatively smaller variance and well separates

Table 2.1: Triangle, two-star density and normalized transitivity and their variances estimated in college training sets

College	Triangle		Two-star		Norm. Trans.	
	Est	$\widehat{Var}$	Est	$\widehat{Var}$	Est	$\widehat{Var}$
Berkeley	77.95	18.10	6.31	0.27	37.05	5.57
Stanford	36.62	5.12	5.90	0.11	18.61	0.16
Yale	24.20	2.40	5.22	0.09	13.90	0.06
Princeton	20.87	2.34	5.25	0.11	11.91	0.06
Harvard	38.56	5.11	6.28	0.10	18.43	0.10
MIT	30.20	7.89	6.11	0.24	14.82	0.15

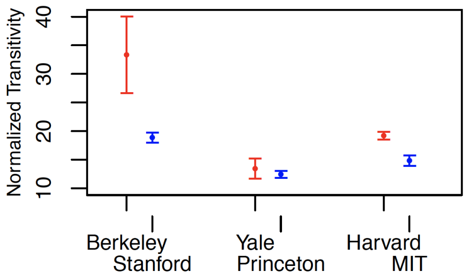


Figure 2.4: For 3 pairs of colleges, 97.5% CI constructed using $\widehat{\text{Var}}_{\text{JACK}}$ on normalized transitivity

each of the pairs in training sets. Thus we choose normalized transitivity as the test statistic. We now obtain jackknife estimate of variance of normalized transitivity using the the test sets.

Figure 2.4 presents 97.5% CI's for normalized transitivity for each college. Thus, two disjoint CI's are equivalent to rejecting a level 0.05 test. Figure 2.4 basically shows that transitivity can in fact separate Berkeley and Stanford Facebook networks, as well as Harvard and MIT Facebook networks, giving us interesting information about the inherent differences between the network structures of these colleges.

2.5 Discussion

In the present work, we have shown that the network jackknife is a versatile tool that may be used in a wide variety of situations. For poorly understood functionals, the Network Efron-Stein inequality ensures that the jackknife produces conservative estimates of the variance in expectation. For a general class of functionals related to counts, we establish consistency of the jackknife. Our empirical investigation is encouraging regarding the finite sample properties of the procedure, as the network jackknife outperforms subsampling in many simulation settings.

Chapter 3

Trading off Accuracy for Speedup: Multiplier Bootstraps for Subgraph Counts

A paper based partly on the contents of this chapter is under revision (Lin et al., 2020b).

3.1 Introduction

From social networks like Twitter and Facebook to biological networks like protein-protein interaction networks and brain networks, network data has become ubiquitous in a broad range of real-world applications.

Count functionals play a pivotal role in the analysis of network data. In biological networks, it is believed that certain subgraphs may represent functional subunits within the larger system (Milo et al., 2002; Chen and Yuan, 2006; Daudin et al., 2008; Kim et al., 2014). In social networks, the frequency of triangles provides information about the likelihood of mutual friendships (Newman, 2001; Myers et al., 2014; Ugander et al., 2011). At a more theoretical level, count functionals may be viewed as network analogs of the moments of a random variable. Thus, a method of moments approach may be used to estimate the underlying model under suitable conditions (Bickel et al., 2011). Furthermore, the convergence of graph sequences (at least dense sequences) may be stated in terms of the convergence of a collection of subgraph frequencies (Borgs et al., 2008).

Given their practical and theoretical importance, quantifying the uncertainty of count functionals is naturally of substantial interest. While real-world networks share many qualitative features (see for example, Newman (2003)), they often vary substantially in terms of size, given by the number of vertices in the network, and sparsity level, given by the number of edges relative to the number of vertices. For networks of small to moderate size, inferential methods that are highly accurate are advantageous; for sparse, massive networks, one needs to simultaneously consider computational tractability and accuracy.

To meet these diverse needs in real-world applications, we develop a new family of bootstrap procedures for count functionals of networks, ranging from a very fast and consistent randomized linear bootstrap to a fast quadratic bootstrap procedure that offers improved accuracy for moderately sparse networks. Both procedures may be viewed as approximations[1] to a multiplier bootstrap method in which each potential subgraph in the network is perturbed by the product of independent multiplier random variables. This multiplier bootstrap is closely related to a bootstrap method for U-statistics (see for example, Bose and Chatterjee (2018)). Under the sparse graphon model (see Section 1.1.1), subgraph counts may be viewed as U-statistics perturbed by asympotically negligible noise, allowing the adaptation of bootstrap methods for U-statistics to the network setting.

[1] More precisely, the linear and quadratic bootstraps may be viewed as first and second-order terms of a Hoeffding decomposition for the multiplier bootstrap, respectively.

One of the main theoretical contributions of our paper is deriving (uniform) Edgeworth expansions for the quadratic bootstrap. Edgeworth expansions may be viewed as a refinement of the Normal approximation that accounts for skewness of the distribution of interest; an excellent treatment of this topic in the IID setting is given by Hall (2013). By establishing an Edgeworth expansion for the quadratic bootstrap and showing that it is very close to the Edgeworth expansion of the sampling distribution, we show that the bootstrap is higher-order correct under certain sparsity conditions, meaning that it offers a faster convergence rate in the Kolmogorov distance than the Berry-Esseen bound. Establishing higher-order correctness is often key to justifying the (typically) computationally intensive bootstrap over a Normal approximation.

Edgeworth expansions of network moments were first studied by Zhang and Xia (2020). The authors show that the network noise has a smoothing effect that allows them to bypass the typical Cramér's condition, which is restrictive in the network setting. We also bypass the Cramér's condition for the bootstrap, but by using a different approach. We choose a continuous multiplier that matches the first three moments of the data; it is well-known that continuous random variables satisfy Cramér's condition. To derive our Edgeworth expansion for the bootstrap, we also build upon results from Wang and Jing (2004) for order two U-statistics. It turns out that network noise, particularly when the graph is sparse, causes certain terms related to our Edgeworth expansion to blow up. While the details are technical, we show that a valid Edgeworth expansion is still possible, with the sparsity level directly affecting the convergence rate.

On the other hand, the linear bootstrap is not higher-order correct in any sparsity regime as it only aims to approximate the leading term of the Hoeffding decomposition. However for sparser networks, we observe an interesting phenomenon; in this case, the linear bootstrap outperforms the higher-order correct variants in terms of accuracy. In essence, the extra terms in the quadratic bootstrap cannot be estimated accurately enough for sparse graphs and consequently these terms hurt more than they help in sparse regimes. For sparser graphs, we propose speeding up the linear bootstrap further by randomizing the precomputation of count functionals. By randomizing to an extent that is appropriate for the sparsity level of the network, we sacrifice very little statistical performance for vastly reduced computation. Thus, the approximate linear bootstrap is well-suited for scalable inference on large, sparse graphs.

In addition to obtaining Edgeworth expansions for count statistics, we also obtain Edgeworth expansions for smooth functions of U-statistics. We show that, under suitable sparsity assumptions, the cumulative distribution function of smooth functions arising from the quadratic bootstrap match this asymptotic expansion and are therefore higher-order correct. In this setting, obtaining analytical expressions for Edgeworth expansions are cumbersome, whereas the bootstrap is automatic and user-friendly.

We will now provide a roadmap for the rest of the paper. In Section 3.2, we discuss related work, focusing on the emerging area of resampling methods for network data. The problem setting and our bootstrap proposal is introduced in Section 3.3. In Section 3.5, we present our main results, which establish higher-order correctness for our bootstrap procedures. In Section 3.6, we present a simulation study, which shows that our procedure exhibits strong finite-sample performance in various settings. Finally, in Section 3.7, we use our bootstrap methods to compare networks representing the voting similarities of U.S. Congress from 1949 to 2012.

3.2 Related Work

The first theoretical result for resampling network data was attained by Bhattacharyya and Bickel (2015). Their subsampling proposals involve expressing the variance of a count functional in terms of other count functionals and estimating the non-negligible terms through subsampling. Lunde and Sarkar (2019) show that it is also possible to conduct inference using quantiles of the subsampling distribution as in Politis and Romano (1994). Green and Shalizi (2017) propose a bootstrap based on the empirical graphon. Lin et al. (2020a) establish the validity of the network jackknife for count functionals.

Levin and Levina (2019) study a two-step procedure that is closely related to our linear bootstrap procedure. The above authors propose estimating the latent positions with the adjacency spectral embedding in the first step (see, for example, Athreya et al. (2018)) and resampling the corresponding U-statistic with the estimated positions in the second step. They derive theoretical results under the assumption that the rank is known of the random dot product graph model is known and finite. In contrast, our procedures do not impose assumptions on the spectral properties of the underlying graphon.

Zhang and Xia (2020) establish conditions under which the empirical graphon bootstrap exhibits higher order correctness. They require Cramér's condition for the leading term of the Hoeffding projection, which is restrictive for network models. The empirical Edgeworth expansion proposal, which has been considered in other settings (see, for example, Putter and Van Zwet (1998) and Maesono (1997)), involves studentizing by a variance estimate and plugging in estimated moments into an Edgeworth expansion. While our rates are less sharp than existing work, we see that multiplier bootstraps can handle functions of subgraph counts more easily than empirical Edgeworth corrections. We show that even for smooth functions, our proposed bootstrap procedures exhibit higher-order correctness. While computationally more demanding, work in other settings suggests that the bootstrap may have some favorable properties over empirical Edgeworth expansions (see, for example, Hall (1990)).

On the mathematical side, the analysis of our multiplier bootstrap involves Edgeworth expansions for weighted sums. Prior work (c.f. Bai and Zhao (1986) and Liu (1988)) suggests that establishing sharp rates of convergence for the independent but non-identically distributed sequences is more difficult, with the above references establishing a $o(n^{-1/2})$ error bound instead of the $O(n^{-1})$ bound for i.i.d. sequences. Edgeworth expansions for multiplier bootstraps of (degree 2) U-statistics are also considered in Wang and Jing (2004).

From the computational standpoint, (Chen and Kato, 2019) presented a randomized algorithm to estimate high dimensional U-statistics from a subsample of subsets from the set of all subsets of a given size. We propose a different sampling method that exploits the structure of U-statistics and draws random permutations instead of subsets. Empirically, we show that this method provides faster computation over subset sampling.

3.3 Problem Setup and Notation

3.3.1 Count Functionals

Under the sparse graphon models, recall the definitions of count functionals in Section 1.1.2. For notational convenience, in this chapter we re-express them as follows. Let R denote the adjacency matrix of a subgraph of interest, with r vertices and s edges. Let $A^{(n)}_{i_1,\dots,i_r}$ denote the adjacency matrix formed by the node subset $\{i_1,\dots,i_r\}$ and for each such r-tuple, define the following function:

$$H(A^{(n)}_{i_1,\dots,i_r}) := \mathbb{1}(A^{(n)}_{i_1,\dots,i_r} \cong R)$$

where we say that $A^{(n)}_{i_1,\dots,i_r} \cong R$ if there exists a permutation function π such that $A_{\pi(i_1),\dots,\pi(i_r)} = R$. Our count functional, which we denote $\hat{T}_n(R)$, or $\hat{T}_n$ when there is no ambiguity, is formed by averaging over all r-tuples in the graph.

$$\hat{T}_n := \frac{1}{\binom{n}{r}} \sum_{1\le i_1<i_2<\dots<i_r\le n} H(A^{(n)}_{i_1,\dots,i_r}) \tag{3.1}$$

Recall the definition of $\hat{P}(R)$ in Eq 1.5. Note that $\hat{P}(R) = \rho_n^{-s}\hat{T}_n$. Define the following kernel:

$$h_n(X_{i_1},\dots X_{i_r}) := \mathrm{E}\{H(A^{(n)}_{i_1,\dots,i_r}) \mid X_{i_1},\dots X_{i_r}\}. \tag{3.2}$$

For readability, we will suppress the n in h_n in what follows. Now, define the following (conventional) U-statistic:

$$T_n := \frac{1}{\binom{n}{r}} \sum_{1\le i_1<i_2<\dots<i_r\le n} h(X_{i_1},\dots X_{i_r})$$

For notational convenience we will refer to $h(X_{i_1}, \dots, X_{i_r})$ by $h(X_S)$, where S is the subset $\{i_1, \dots, i_r\}$. Denote $\theta_n := \mathrm{E}\{h(X_S)\}$. We see that $\theta_n/\rho_n^s \to \mu$. This can be thought of as a normalized subgraph density that we want to infer. The normalization by ρ_n^s is to ensure that our functional converges to an informative non-zero quantity.

By a central limit theorem for U-statistics (Hoeffding, 1948), it can be shown that $(T_n - \theta_n)/\sigma_n$ is asymptotically Gaussian. Here we have:

$$\tau_n^2 = \mathrm{var}[\mathrm{E}\{h(X_S) \mid X_1\}], \quad \sigma_n^2 = r^2\tau_n^2/n \tag{3.3}$$

Furthermore, Bickel et al. (2011) show that, $(\hat{T}_n - T_n)/\sigma_n = o_P(1)$ under mild sparsity conditions for a wide range of subgraphs. Thus, we may view $(\hat{T}_n - \theta_n)/\sigma_n = (\hat{T}_n - T_n)/\sigma_n + (T_n - \theta_n)/\sigma_n$ as a U-statistic perturbed by asymptotically negligible noise.

3.3.2 Preliminaries of proposed bootstrap procedures

In order to estimate the subgraph density, we will consider the following multiplier bootstrap procedures. In what follows let $\xi_1, \dots \xi_n$ be i.i.d. continuous random variables with mean $\mu = 1$ and central moments $\mu_2 = 1$, and $\mu_3 = 1$. An example of such a random variable is the product Z of two independent Normal random variables X and Y, defined below:

$$X \sim N(1, 1/2) \qquad Y \sim N(1, 1/3) \qquad Z = XY \tag{3.4}$$

Let $\xi_{i_1 \cdots i_r}$ denote $\xi_{i_1} \times \dots \times \xi_{i_r}$ and define the following multiplicative bootstrap:

$$\hat{T}^*_{n,M} = \hat{T}_n + \frac{1}{\binom{n}{r}} \sum_{1\leq i_1<i_2<\ldots i_r} \xi_{i_1\cdots i_r} \cdot \left\{H(A^{(n)}_{i_1,\ldots,i_r}) - \hat{T}_n\right\} \tag{3.5}$$

Our multiplicative bootstrap is motivated by Hoeffding's decomposition (see Supplementary Section B.1 for details). The first two terms of the decomposition for $T_n - \theta_n$ are given by:

$$\begin{aligned} g_1(X_i) &= E\{h(X_i, X_{i_2} \ldots X_{i_r}) \mid X_i\} - \theta_n \\ g_2(X_i, X_j) &= E\{h(X_i, X_j, X_{i_3} \ldots X_{i_r}) \mid X_i, X_j\} - g_1(X_i) - g_1(X_j) - \theta_n, \end{aligned}$$

leading to the representation:

$$T_n - \theta_n = \frac{r}{n}\sum_{i=1}^{n} g_1(X_i) + \frac{r(r-1)}{n(n-1)}\sum_{i<j} g_2(X_i, X_j) + o_p\left(\frac{\rho_n^s}{n}\right) \tag{3.6}$$

Similarly, conditional on the data, it can be shown that we have the following bootstrap analog. Let:

$$\hat{g}_1(i) = \frac{1}{\binom{n-1}{r-1}} \sum_{1\leq i_2<\ldots i_r\leq n, i_u\neq i} \left\{H(A_{i,i_2,\ldots i_r}) - \hat{T}_n\right\} \tag{3.7}$$

$$\tilde{g}_2(i, j) = \frac{1}{\binom{n-2}{r-2}} \sum_{1\leq i_3<\ldots i_r\leq n, i_u\neq i, i_u\neq j} \left\{H(A_{i,i_2,\ldots i_r}) - \hat{T}_n\right\} \tag{3.8}$$

$$\hat{g}_2(i, j) = \tilde{g}_2(i, j) - \hat{g}_1(i) - \hat{g}_1(j) \tag{3.9}$$

Furthermore, Eq 3.7 can be used to standardize the bootstrap replicates using the following estimate of τ_n (Eq 3.3):

$$\hat{\tau}_n^2 = \sum_i \frac{\hat{g}_1(i)^2}{n} \tag{3.10}$$

We now present the Hoeffding decomposition for our bootstrap statistic. The proof is deferred to Supplement Section B.1.

Lemma 4. *We have the following decomposition:*

$$\hat{T}^*_{n,M} - \hat{T}_n = \frac{r}{n}\sum_{i=1}^{n}(\xi_i - 1)\cdot \hat{g}_1(i) + \frac{r(r-1)}{n(n-1)}\sum_{i<j}(\xi_i\,\xi_j - \xi_i - \xi_j + 1)\cdot \tilde{g}_2(i,j)$$
$$+ O_P\left(\rho_n^s n^{-1/2}\delta(n,\rho_n,R)\right), \tag{3.11}$$

where $\delta(n,\rho_n,R)$ is defined as follows:

$$\delta(n,\rho_n,R) = \begin{cases} \dfrac{1}{n\rho_n} & R \text{ is acyclic} \\ \dfrac{1}{n\rho_n^{3/2}} & R \text{ is a simple cycle.} \end{cases}$$

Although the quadratic term in the above expansion may seem different from Eq 3.6, Some manipulation yields that $\sum_{i<j}(\,\xi_i\,\xi_j - \xi_i - \xi_j + 1)\cdot\tilde{g}_2(i,j)$ is equivalent to $\sum_{i<j}(\xi_i\xi_j - 1)\tilde{g}_2(i,j) - (\xi_i - 1)\cdot\hat{g}_1(i) - (\xi_i - 1)\cdot\hat{g}_1(j)$, which is similar to the corresponding term in the Hoeffding decomposition of the U statistic (see Eq 3.6).

Viewing $\hat{g}_1(i)$ and $\hat{g}_2(i,j)$ as estimates of $g_1(X_i)$ and $g_2(X_i, X_j)$, respectively, it is clear that that our weighted bootstrap version encapsulates important information about $\hat{T}_n - \theta_n$. The above decomposition also suggests that one may approximate the non-negligible terms more directly. Ignoring the remainder term, we arrive at the linear and quadratic bootstrap estimates:

$$\hat{T}^*_{n,L} = \hat{T}_n + \frac{r}{n}\sum_{i=1}^{n}(\xi_i - 1)\cdot\hat{g}_1(i) \tag{3.12}$$

$$\hat{T}^*_{n,Q} = \hat{T}^*_{n,L} + \frac{r(r-1)}{n(n-1)}\sum_{i<j}(\xi_i\,\xi_j - \xi_i - \xi_j + 1)\cdot\tilde{g}_2(i,j). \tag{3.13}$$

Now that we have introduced the main concepts, we are ready to present the our bootstrap procedures. We first present the results on our fast linear bootstrap method.

3.4 Proposed algorithms

In this section, we present a fast linear bootstrap method using Eq 3.12. Recall that the multiplicative bootstrap requires to precompute $\hat{T}_n$ and $\hat{g}_1(i)$ for all i. This computation is $O(n^r)$ in the worst case. In addition to this, the computation complexity for MB-L is $O(Bn)$. Therefore, in what follows, our goal is to reduce the precomputation time.

3.4.1 Fast linear bootstrap

We propose a randomized approximation for $\hat{T}_n$ and $\hat{g}_1(i)$. The main idea is that an average over all size r subset can be written as an average over $n!$ permutations (see Hoeffding (1948); Lunde and Sarkar (2019)).

For any $i \in \{1, \dots, n\}$, denote the set of all subsets of size $r-1$ taken from $\{1, \dots, i-1, i+1, \dots n\}$ as $\mathbb{S}_{-i}$. Denote $H(A_{i,i_2,\dots i_r})$ for $S = \{i_2, \dots i_r\} \in \mathbb{S}\{-i\}$ as $H(A_{S\cup i})$.

Denote

$$H_1(i) = \frac{1}{\binom{n-1}{r-1}} \sum_{S \in \mathbb{S}\{-i\}} H(A_{S\cup i}).$$

One can also write $H_1(i)$ as follows:

$$H_1(i) = \frac{1}{(n-1)!} \sum_{\pi} H_\pi(i).$$

Here $H_\pi(i) = \frac{\sum_{S\in\mathbb{S}_\pi} H(A_{S\cup i})}{\frac{n-1}{r-1}}$, where $\mathbb{S}_\pi$ denotes the set of all disjoint subsets $\{\pi_{(i-1)(r-1)+1}, \dots, \pi_{i(r-1)}\}$, $i = 1, \dots, \frac{n-1}{r-1}$ obtained from permutation π. Now let π_j

be a permutation picked with replacement and uniformly at random from the set of all permutations of $\{1, \ldots, n\} \setminus i$.

Our randomized algorithm makes use of this structure and draws $j = 1, \ldots, N$ independent permutations π_j. We compute

$$\tilde{H}_1(i) = \frac{\sum_j H_{\pi_j}(i)}{N} \qquad \tilde{T}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{H}_1(i) \tag{3.14}$$

To calculate $\tilde{H}_1(i)$, for each i, we permute the node set excluding i for N times and for each of these permutations π we check the disjoint set $\mathbb{S}_\pi$ for count functionals. Thus, the complexity for calculating $\tilde{H}_1(i)$ is now $O\left(N\frac{n}{r}\right)$. From $\{\tilde{H}_1(i)_{i=1}^{n}\}$, $\tilde{T}_n$ is calculated from their mean and $\tilde{\tau}_n$ is defined as

$$\tilde{\tau}_n^2 = \frac{\sum_{i=1}^{n}\{\tilde{H}_1(i) - \tilde{T}_n\}^2}{n^2}. \tag{3.15}$$

The linear bootstrap uses $\tilde{T}_{n,L}$ by plugging in $\tilde{g}_1(i) = \tilde{H}_1(i) - \tilde{T}_n$ and $\tilde{T}_n$ in Eq 3.12.

$$\tilde{T}^*_{n,L} = \tilde{T}_n + \frac{r}{n}\sum_{i=1}^{n}(\xi_i - 1)\{\tilde{H}_1(i) - \tilde{T}_n\}, \tag{3.16}$$

We denote this algorithm by `MB-L-apx` and explicitly provide the algorithm in Algorithm 1.

3.4.2 Higher order correct bootstrap procedures

In this section, we present our proposed Quadratic, and Multiplicative algorithms (`MB-Q`, and `MB-M`).

Algorithm 1. Construction of linear or approximated linear bootstrap estimate of CDF

```
Input: Network $A$, motif $R$, number of resamples $B$,
 approximate $\in \{True, False\}$, parameter $u$
If approximate $= True$
    Compute $\{\tilde{H}_1(i)\}_{i=1}^n$, $\tilde{T}_n$ (Eq 3.14) and $\tilde{\tau}_n$ (Eq 3.15)
Else
    Compute $\hat{T}_n$, (Eq 3.1), $\{\hat{g}_1(i)\}_{i=1}^n$ (Eq 3.7) and $\hat{\tau}_n$ (Eq 3.10)
End
for $j \in \{1, \ldots, B\}$ do
    Generate $n$ weights $\boldsymbol{\xi}^{(j)} = \{\xi_i^{(j)}, i = 1, \ldots, n\}_{j=1}^B$ using Eq 3.4
    If approximate $= True$
        $T_n^*(j) \leftarrow \tilde{T}_{n,L}^*$ (using Eq 3.16.)
    Else
        $T_n^*(j) \leftarrow \hat{T}_{n,L}^*$ (using Eq 3.12.)
    End
end
Return $\frac{1}{B} \sum_j \mathbb{1}\left(\frac{T_n^*(j) - \tilde{T}_n}{\frac{r}{n^{1/2}}\tilde{\tau}_n} \leq u\right)$
```

Algorithm 2. Construction of quadratic or multiplier bootstrap estimate of CDF

```
Input: Network $A$, motif $R$, number of resamples $B$, choice of bootstrap
  procedure $a \in \{M, Q\}$, parameter $u$
Compute $\hat{T}_n$ (Eq 3.1), $\{\hat{g}_1(i)\}_{i=1}^n$ (Eq 3.7), $\{\hat{g}_2(i,j)\}_{i=1}^n$ (Eq 3.9) and $\hat{\tau}_n$
  (Eq 3.10)
for $j \in \{1, \dots, B\}$ do
    Generate $n$ weights $\boldsymbol{\xi}^{(j)} = \{\xi_i^{(j)}, i = 1, \dots, n\}_{j=1}^B$ using Eq 3.4
    If $a = M$
        $T_n^*(j) \leftarrow \hat{T}_{n,M}^*$ (using Eq 3.5.)
    Else
        $T_n^*(j) \leftarrow \hat{T}_{n,Q}^*$ (using Eq 3.13.)
    End
end
Return $\frac{1}{B} \sum_j \mathbb{1}\left(\frac{T_n^*(j) - \hat{T}_n}{\frac{r}{n^{1/2}}\hat{\tau}_n} \leq u\right)$
```

For a given network, we first compute $\hat{T}_n$ and $\hat{\tau}_n$ (see Eqs 3.1, 3.10). For each algorithm, we generate B samples of n weights $\{\xi_i^{(j)}, i = 1, \dots, n\}_{j=1}^B$ from the `Gaussian Product` distribution (see beginning of Section 3.3.2). For each of these, `MB-M`, `MB-Q`, and `MB-L` respectively values $\hat{T}_{n,M}^*$, $\hat{T}_{n,Q}^*$ and $\hat{T}_{n,L}^*$. From the B values one then constructs the CDF of the statistic in question, after shifting and normalizing it appropriately. We present this in Algorithm 2.

While we divide by $\frac{r}{n^{1/2}}\hat{\tau}_n$, note that our statistic is not studentized, which is why our expansion differs from previous work. Conditioned on the data, $\hat{\tau}_n$ is constant for the bootstrap samples.

Note that `MB-M` is computationally expensive since it involves computing the expression in Eq 3.5 for each sample of the bootstrap. The worst-case complexity of evaluating all $\binom{n}{r}$ subsets of nodes is n^r. For B bootstrap samples, the worst-case timing of MB-M will be Bn^r. In comparison, for MB-L and MB-Q, we can precompute the $\hat{g}_1(i)$ and $\hat{g}_2(i, j)$ values in $O(n^r)$ time. After that, the time per bootstrap sample is linear for MB-L and quadratic for MB-Q. Thus worst-case computational complexity for a dense network for MB-M, MB-Q, and MB-L is $O(Bn^r)$, Bn^2 and Bn respectively, *excluding* precomputation time (which is O(n^r) in the worst case). In contrast, the approximate linear bootstrap algorithm MB-L-apx takes O(Nn/r) computation for each $\tilde{H}_1(i), i \leq n$. Note that we can easily parallelize this step. With C cores, that will lead to a computational cost of Nn^2/rC.

3.5 Main Results

3.5.1 Theoretical guarantees for approximate linear bootstrap

In this section, we show that the linear bootstrap statistic using the approximate moments in Eq 3.14 is indeed first-order correct under appropriate sparsity conditions as long as N is large enough. For theorem 5, we will use the following assumption:

Assumption 1. We assume the following:

(a) $\tau_n/\rho_n^s \geq c > 0$, for some constant c.

(b) $0 < w(u, v) < C$, for some constant C.

The first condition is a standard non-degeneracy assumption for U-statistics.

Theorem 5. *Suppose Assumption 1 is satisfied, the weights* $\xi_1, \ldots, \xi_n$ *are generated from a distribution such that* $E[\xi_1] = 1$, $E[(\xi_1 - 1)^2] = 1$, $E[(\xi_1 - 1)^3] < \infty$. *Further assume that* $\rho_n = \omega(1/n)$ *when R is acyclic or* $\rho_n = \omega(n^{-1/r})$ *when R is cyclic. Then,* *(a) The standardized bootstrap distribution converges at the Berry-Esseen rate to a standard Normal under the condition that $N \gg \frac{1}{n\rho_n^s}$,*

$$\sup_{u \in R} \left| P^* \left(\frac{\tilde{T}^*_{n,L} - \tilde{T}_n}{\{var(\tilde{T}^*_{n,L} \mid A, X)\}^{1/2}} \leq u \right) - \Phi(u) \right| = O_P\left(n^{-1/2}\right), \qquad (3.17)$$

where $P^*(\cdot)$ denotes the conditional measure conditioned on A and X.

(b) The variance of $\tilde{T}^*_{n,L}$ satisfies:

$$\frac{var(\tilde{T}^*_{n,L} \mid A, X)}{\sigma_n^2} = 1 + O_P\left(\frac{1}{n\rho_n}\right) + O_P\left(\frac{1}{Nn\rho^s}\right). \qquad (3.18)$$

(c) If $\tilde{T}^*_{n,L}$ in Eq 3.17 is replaced by the $T^*_{n,L}$ computed without approximate moments and $\tilde{T}_n$ is replaced by $\hat{T}_n$, then Eq 3.17 holds. We also have:

$$\frac{var(T^*_{n,L} \mid A, X)}{\sigma_n^2} = 1 + O_P\left(\frac{1}{n\rho_n}\right). \qquad (3.19)$$

Remark 4 (Comparison to existing work on approximating U- statistics). In Chen and Kato (2019), the authors draw $\omega(n)$ subsets of size r from all $\binom{n}{r}$ subsets with replacement to estimate an incomplete U-statistic. The total number of subsets we examine for approximating a local count statistic is also $\omega(n)$. Comparing our Theorem 5 with their result shows that both methods require similar computation to achieve consistency. However, practically, drawing Nn/r subsamples with replacement seems to be slower than drawing a N permutations and then dividing each into disjoint subsets (see Fig 3.3).

Remark 5 (Approximation quality). In the above theorem, if $Nn/r = \omega(\rho_n^{-s})$, then the ratio of the variance of the linear bootstrap statistic and that of the count statistic in question converges in probability to one. This shows that for sparse networks we need larger number of random permutations to estimate the moments.

Remark 6 (Broader Sparsity Regime). Eq 3.19 suggests that the linear bootstrap without approximation gives a consistent estimate of variance even when the average degree $n\rho_n$ goes to infinity. This shows a stark contrast to Theorem 6 and Corollary 6.1, where the average degree has to much larger to achieve higher-order correctness. It should be noted that the arguments in Zhang and Xia (2020) require $\rho_n = \omega(1/\sqrt{n})$ for acyclic graphs and therefore, their convergence rates for empirical Edgeworth expansions do not hold in sparser regimes.

Remark 7 (Conditions on subgraphs). While we state Theorem 5 for acyclic and general cyclic subgraphs, it should be noted that weaker sparsity conditions are possible for simple cycles. In particular, for simple cycles one only needs $n^{s-1}\rho_n^s \to \infty$.

3.5.2 Results on higher-order correct bootstrap procedures

Below, we establish an Edgeworth expansion normalized by the true standard deviation, which is more appropriate for our purposes. Since estimating the variance leads to a non-negligible perturbation, the polynomials in our expansion differ from those established by Zhang and Xia (2020). All proofs and details are deferred to Supplement Section B.3 and Section B.4. In what follows, let $F_n(u)$ denote the CDF of $\hat{T}_n$ and $G_n(u)$ denote the Edgeworth expansion of interest, given by:

$$G_n(u) = \Phi(u) - \phi(t)\frac{(u^2-1)}{6n^{1/2}\tau_n^3}\left[E\{g_1^3(X_1)\} + 3(r-1)E\{g_1(X_1)g_1(X_2)g_2(X_1,X_2)\}\right]. \tag{3.20}$$

Furthermore, recall $\tau_n^2 = \mathrm{var}[\mathrm{E}\{h(X_S) \mid X_1\}]$ denotes the asymptotic variance of the U-statistic. Throughout this section, we will impose the following condition:

Assumption 2. For acyclic R, $\rho_n = \omega(n^{-1/2})$ and for cyclic R, $\rho_n = \omega(n^{-1/r})$.

The above is a nontrivial sparsity assumption that we require for higher-order correctness. We have the following result:

Proposition 3. *Let G_n be the Edgeworth expansion defined in Eq 3.20 and let R be a fixed subgraph. Suppose that Assumptions 1 and 2 hold. Further suppose that $\rho_n = O(1/\log n)$ or Cramér's condition holds, i.e.* $\limsup_{t\to\infty} \left|E\left\{e^{itg_1(X_1)/\tau_n}\right\}\right| < 1$ *then we have,*

$$\sup_u |F_n(u) - G_n(u)| = O(\mathcal{M}(n, \rho_n, R)) \tag{3.21}$$

where F_n is the cumulative distribution function of

$$\mathcal{M}(n, \rho_n, R) = \begin{cases} \frac{1}{n\rho_n} & R \text{ is acyclic} \\ \frac{1}{n\rho_n^{r/2}} & R \text{ is cyclic} \end{cases} \tag{3.22}$$

Now, we will state our bootstrap approximation results. We will first show that conditioned on the network, and latent variables, the CDF of `MB-Q` matches the asymptotic expansion in Eq 3.20, where the true moments are replaced by their empirical versions. In what follows, let $\hat{E}_n(\cdot)$ denote the expectation operator with respect to the empirical measure of A and X. Define

$$\hat{G}_n(u) = \Phi(u) - \frac{(u^2-1)\phi(u)}{6n^{1/2}\widehat{\tau}_n^3}\left[\widehat{E}_n\left\{g_1(i)^3\right\} + 3(r-1)\widehat{E}_n\{g_1(i)g_1(j)g_2(i,j)\}\right], \tag{3.23}$$

where we have:

$$\begin{aligned} \widehat{E}_n\left\{g_1(i)^3\right\} &= \frac{1}{n}\sum_{i=1}^{n}\hat{g}_1(i)^3, \\ \widehat{E}_n\{g_1(i)g_1(j)g_2(i,j)\} &= \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\tilde{g}_2(i,j)\hat{g}_1(i)\hat{g}_1(j). \end{aligned} \tag{3.24}$$

Theorem 6. *If Assumptions 1 and 2 are satisfied, the weights* $\xi_1, \ldots, \xi_n$ *are generated from a non-lattice distribution (see Feller (1971) page 539) such that* $E[\xi_1] = 1$, $E[(\xi_1 - 1)^2] = 1$, $E[(\xi_1 - 1)^3] = 1$, *then*

$$\sup_u \left| P^*\left(\frac{\hat{T}^*_{n,Q} - \hat{T}_n}{\hat{\sigma}_n} \le u\right) - \hat{G}_n(u)\right| = o_P(n^{-1/2}) + O_P\left(\frac{\log n}{n^{2/3}\rho_n}\right),$$

where $P^*(\cdot)$ *denotes the conditional probability of event* $(\cdot)$ *conditioned on* A *and* X.

Remark 8. While the above theorem is for standardized bootstraps, our proof may be adapted to yield an analogous statement for bootstraps studentized by a variance estimator inspired by the Delta Method. In essence, the studentized bootstrap may also be expressed as a weighted U-statistic and a negligible remainder term, allowing the use of similar proof techniques.

Remark 9. While Zhang and Xia (2020) establish higher-order correctness under milder sparsity conditions for subsampling and the empirical graphon, our result here does not require Cramér's condition for $g_1(X_i)$, which is an important feature for network applications. Our simulation study suggests that our rate here can be improved, but we leave this to future work.

Combining Theorem 6 with the Hoeffding decomposition in Eq 3.11, we obtain the corollary below for the multiplicative bootstrap. Since the remainder term in the Hoeffding decomposition concentrates slowly for well-connected subgraphs of sparser networks, we impose additional assumptions on the subgraph to maintain the rate from the previous theorem.

Corollary 6.1. *Suppose Assumption 1 is satisfied and either R is acyclic and $\rho_n = \omega(1/\sqrt{n})$ or R is a simple cycle and $\rho_n = \omega(n^{-1/r})$. Further suppose that the weights $\xi_1, \ldots, \xi_n$ are generated from a non-lattice distribution with such that $E(\xi_1) = 1$, $E\{(\xi_1 - 1)^2\} = 1$, $E\{(\xi_1 - 1)^3\} = 1$, then,*

$$\sup_u \left| P^* \left(\frac{\hat{T}^*_{n,M} - \hat{T}_n}{\hat{\sigma}_n} \le u \right) - \hat{G}_n(u) \right| = o_P(n^{-1/2}) + O_P\left(\frac{\log n}{n^{2/3}\rho_n} \right),$$

where $P^*(\cdot)$ denotes the conditional probability of event $(\cdot)$ conditioned on A and X and $\hat{\sigma}_n = r\hat{\tau}_n/n^{1/2}$.

The proof of Theorem 6 build upon results from Wang and Jing (2004), which establish higher-order correctness of the weighted bootstrap for order-2 U-statistics. However, certain terms that appear as constants in their work blow up when perturbed by sparse network noise. To deal with this issue, we control various terms unique to the network setting and use different arguments to control the overall error rate.

As in Zhang and Xia (2020), it is also possible to consider an empirical Edgeworth expansion in which the expectations of interest are estimated. We state a result for this procedure below:

Lemma 7. *Under the assumptions in Assumption 1 and 2, we have*

$$\sup_u |\hat{G}_n(u) - F_n(u)| = O_P(\mathcal{M}(n, \rho_n, R))$$

The lemma above suggests that the empirical Edgeworth expansion achieves a better rate than the bootstrap procedures considered. In the experimental section, we see that the empirical Edgeworth expansion (`EW`) in fact achieves the smallest error when the network is dense enough. However, for smooth functions of counts, it is cumbersome to derive such expansions and the bootstrap emerges as a strong practical alternative that offers improved accuracy over a Normal approximation in certain regimes.

3.5.3 Smooth functions of count statistics

In network science, the transitivity coefficient, which may be defined as a smooth function of triangles and two-stars, is commonly used to quantify how much nodes in the network cluster together. Given the importance of such functions in applications, accurate inference for these parameters is naturally of substantial interest. Our results in this section establish the quadratic and multiplicative bootstraps as accurate and user-friendly methods for smooth functions of counts that sidestep the cumbersome computation of gradients and moments required by empirical Edgeworth expansions.

Our theorem below is the first result in the literature for Edgeworth expansions of smooth functions of count statistics. In fact, to the best of our knowledge, Edgeworth expansions for smooth functions of U-statistics were not derived previously. It turns out that arguments to derive Edgeworth expansions for smooth functions of IID means such as those in Hall (2013) depend heavily on the properties of cumulants of independent random variables and require multivariate Edgeworth expansions, complicating extensions even to U-statistics. In contrast, we adapt flexible Edgeworth expansion results of Jing and Wang (2010) to approximate non-negligible terms arising from a Taylor approximation of the smooth functional.

To state our result, we need to introduce some additional notation. Let u denote a d-dimensional vector of count functionals, let u^* be a vector of corresponding bootstrap statistics generated by either the multiplier bootstrap $\hat{T}^*_{n,M}$ or the the quadratic bootstrap $\hat{T}^*_{n,Q}$. Furthermore, let $f : \mathbb{R}^d \mapsto \mathbb{R}$ denote the function of interest. Consider the following smooth function of bootstrapped count frequencies:

$$S_n^* = n^{1/2}\{f(u^*) - f(u)\}/\tilde{\sigma}_f \tag{3.25}$$

where $\tilde{\sigma}_f$ is used to standardize the bootstrap version and will be defined shortly. The standard Delta Method involves a first-order Taylor expansion; to attain higher-order correctness, we need to consider a second-order expansion. We use the following notation to denote the derivatives of interest evaluated at the expectation $E(u) = \mu$:

$$a_i = \left.\frac{\partial f(x)}{\partial x^{(i)}}\right|_{x=\mu},\ a_{ij} = \left.\frac{\partial^2 f(x)}{\partial x^{(i)} \partial x^{(j)}}\right|_{x=\mu}, \tag{3.26}$$

Define corresponding gradients for the bootstrap evaluated at the count functional u:

$$\hat{a}_i = \left.\frac{\partial f(x)}{\partial x^{(i)}}\right|_{x=u},\ \hat{a}_{ij} = \left.\frac{\partial^2 f(x)}{\partial x^{(i)} \partial x^{(j)}}\right|_{x=u}. \tag{3.27}$$

Define the asymptotic variance of the smooth function as:

$$\sigma_f^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} a_i a_j E\left(\frac{r_i \hat{g}_1^{(i)}(l)}{\rho_n^{s_i}} \frac{r_j \hat{g}_1^{(j)}(l)}{\rho_n^{s_j}} \right). \tag{3.28}$$

and define the empirical analogue of the asymptotic variance as:

$$\tilde{\sigma}_f^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} \hat{a}_i \hat{a}_j \widehat{E}_n \left(\frac{r_i \hat{g}_1^{(i)}(l)}{\rho_n^{s_i}} \frac{r_j \hat{g}_1^{(j)}(l)}{\rho_n^{s_j}} \right). \tag{3.29}$$

We are now ready to state our Edgeworth expansion for the smooth function of the bootstrapped statistics. For simplicity, we state the Edgeworth expansion for u^* resulting from the quadratic bootstrap procedure `MB-Q`. A similar result holds for `MB-M`, albeit under stronger conditions on the subgraph like those imposed in Corollary 6.1.

Theorem 8. *Suppose that $\sigma_f > 0$, the function f has three continuous derivatives in a neighbourhood of μ and suppose that the weights $\xi_1, \ldots, \xi_n$ are generated from a non-lattice distribution such that $E(\xi_1) = 1$, $E\{(\xi_1 - 1)^2\} = 1$, $E\{(\xi_1 - 1)^3\} = 1$. Further suppose that Assumptions 1 and 2 are satisfied. Then, we have:*

$$P^*(S_n^* \leq x) = \Phi(x) + n^{-1/2}\phi(x)\{\tilde{A}_1 \tilde{\sigma}_f^{-1} + \frac{1}{6}\tilde{A}_2 \tilde{\sigma}_f^{-3}(x^2 - 1)\} + O_P\left(\frac{\log n}{n^{2/3}\rho_n}\right). \tag{3.30}$$

where:

$$\begin{aligned}
\tilde{A}_1 &= \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \hat{a}_{ij} \widehat{E}_n \left(\frac{r_i \hat{g}_1^{(i)}(l)}{\rho_n^{s_i}} \frac{r_j \hat{g}_1^{(j)}(l)}{\rho_n^{s_j}} \right), \\
\tilde{A}_2 &= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \hat{a}_i \hat{a}_j \hat{a}_k \widehat{E}_n \left(\frac{r_i \hat{g}_1^{(i)}(l)}{\rho_n^{s_i}} \frac{r_j \hat{g}_1^{(j)}(l)}{\rho_n^{s_j}} \frac{r_k \hat{g}_1^{(k)}(l)}{\rho_n^{s_k}} \right) \\
&\quad + 3\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{t=1}^{d} \hat{a}_i \hat{a}_j \hat{a}_{kt} \widehat{E}_n \left(\frac{r_i \hat{g}_1^{(i)}(l)}{\rho_n^{s_i}} \frac{r_k \hat{g}_1^{(k)}(l)}{\rho_n^{s_k}} \right) \widehat{E}_n \left(\frac{r_j \hat{g}_1^{(j)}(l)}{\rho_n^{s_j}} \frac{r_j \hat{g}_1^{(t)}(l)}{\rho_n^{s_t}} \right) \\
&\quad + 3\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \hat{a}_i \hat{a}_j \hat{a}_k \widehat{E}_n \left(\frac{r_i \hat{g}_1^{(i)}(l)}{\rho_n^{s_i}} \frac{r_j \hat{g}_1^{(j)}(m)}{\rho_n^{s_j}} \frac{r_k(r_k - 1)\tilde{g}_2^{(k)}(l,m)}{\rho_n^{s_k}} \right).
\end{aligned}$$

In the Supplementary Material, we derive Edgeworth expansions for smooth functions of U-statistics corresponding to the non-negligible component of the count functional in Proposition B.5.1 and show that our bootstrap version of the Edgeworth expansion is close to this expansion in Proposition B.5.2. To derive Edgeworth expansions for the U-statistic, we impose a non-lattice condition; however, it is likely that this assumption can be removed for count functionals if a smoothing argument used in Zhang and Xia (2020) is adapted.

3.6 Simulation Study

We consider two graphons in our simulation study. The first graphon we consider is a Stochastic Blockmodel (SBM), introduced by Holland et al. (1983). The SBM is a popular model for generating networks with community structure. The SBM may be parameterized by a $K \times K$ probability matrix B and a membership probability vector π that takes values in the probability simplex in $\mathbb{R}^K$. Let $Y_1, \ldots Y_n \in \{1, \ldots, K\}$ be random variables indicating the community membership of the corresponding node, with probability given by the entries of π. Under this model, we have that $P(A_{ij}^{(n)} = 1 \mid Y_i = u, Y_j = v) = \rho_n B_{uv}$. In our simulations, we consider a two block SBM (`SBM-G`) with $B_{ij} = 0.6$ for $i = 1, j = 1$ and 0.2 for the rest. $\pi = (0.65, 0.35)$

The second model we use is a smooth graphon model from Zhang et al. (2017) (`SM-G`) with $w(u, v) = (u^2 + v^2)/3 \times cos(1/(u^2 + v^2)) + 0.15$. This graphon is continuous and high rank in contrast to the first graphon, which is piece-wise constant and low rank.

Define $\text{err}(F, G)$ as the maximum of $|F(x) - G(x)|$ over a grid on $[-3, 3]$

with grid size 0.1; this will serve as an approximation to the Kolmogorov distance between F and G. In order to study this error, we first need an estimate of the true CDF. To this end, we conduct Monte Carlo simulations with M samples generated from each model. Note that, since our goal is to show that the error is better than the Normal approximation, we need $M = \omega((n\rho_n)^2)$, which ensures that the error from the Monte Carlo samples is $o(1/n\rho_n)$. To ease the computational burden, we perform simulations on small networks with $n = 160$ nodes. We generate $M = 10^6$ Monte Carlo simulations, so that the higher order correctness is not obscured by error from the simulations. In addition to this, we also compare the coverage for different resampling methods in Figure 3.2, where we use $n = 500$, since the true CDF does not have to be estimated. In this setting we estimate the true parameter of inferential interest from a 15000 node network. In the next subsection we show how to obtain higher-order correct confidence intervals.

3.6.1 Higher-order correct confidence intervals

In this paper, we have studied the properties of bootstrap methods for standardized count functionals of networks. While our Edgeworth expansions establish that the standardized bootstrap is higher-order correct in the Kolmogorov norm, it is well known (see Hall (1988)) that corresponding confidence intervals do not offer refined accuracy over those formed from a Normal approximation. In contrast, studentized bootstraps produce higher-order correct confidence intervals. While our theoretical results can be extended to studentized count functionals (see Remark 8), it is also well-known that for statistics such as correlation coefficients, studentization may not be suitable because the variance estimate can be unstable (Hall, 2013). Alternatively, bias-corrected standardized CIs have also been extensively investigated in other settings; see for example, Efron (1980), Efron (1987), and Hall (1988). In this section, we show how to correct standardized intervals to attain higher-order accuracy for coverage.

The (uncorrected) two-sided confidence interval for the standardized bootstrap two-sided confidence interval with nominal coverage α is given by: $\mathcal{I}_1 = \left(\hat{y}_{\frac{1-\alpha}{2}}, \hat{y}_{\frac{1+\alpha}{2}}\right)$, where we define function $\hat{L}(t) = P(\hat{T}_n^* < t \mid A, X)$ for bootstrap samples $\{\hat{T}_n^*\}$, and $\hat{y}_\alpha = \hat{L}^{-1}(\alpha)$. Let z_α denote standard normal critical point at α where $\Phi(z_\alpha) = \alpha$, then $\hat{y}_\alpha = \hat{L}^{-1}(\alpha)$. Equivalently, define $\hat{u}_\alpha$ as the critical point at α for standardized bootstrap sample $(\hat{T}_n^* - \hat{T}_n)/\hat{\sigma}_n$ distribution and $\hat{v}_\alpha$ as the critical point at α for studentized bootstrap sample $(\hat{T}_n^* - \hat{T}_n)/\hat{\sigma}_n^*$ distribution. The standardized bootstrap CI $\mathcal{I}_1$ can be written as

$$\mathcal{I}_1 = (\hat{T}_n + n^{-1/2}\hat{\sigma}_n\hat{u}_{\frac{1-\alpha}{2}}, \hat{T}_n + n^{-1/2}\hat{\sigma}_n\hat{u}_{\frac{1+\alpha}{2}}).$$

Since the population version of $\mathcal{I}_1$ involves the true σ_n instead of $\hat{\sigma}_n$, it follows that $\sigma_n - \hat{\sigma}_n$ is too large of a perturbation for higher-order correctness to hold. In order to make them higher order correct, i.e. make them identical to the studentized CI, one can correct the CI as follows (see Hall (2013) for more details):

$$\mathcal{I}_1' = \left(\hat{y}_{\frac{1-\alpha}{2}} + n^{-1}\hat{\sigma}_n\left\{\hat{p}_1(z_{\frac{1-\alpha}{2}}) + \hat{q}_1(z_{\frac{1-\alpha}{2}})\right\},\ \hat{y}_{\frac{1+\alpha}{2}} + n^{-1}\hat{\sigma}_n\left\{\hat{p}_1(z_{\frac{1+\alpha}{2}}) + \hat{q}_1(z_{\frac{1+\alpha}{2}})\right\}\right) \tag{3.31}$$

where for any $x \in \mathbb{R}$, $\hat{p}_1(x)$ and $\hat{q}_1(x)$ are estimates for $p_1(x)$ and $q_1(x)$. Recall that $p_1(x)$ and $q_1(x)$ are polynomial coefficients of the second order term in the Edgeworth Expansions of the standardized and studentized statistic.

When the statistic is a count functional like the triangle or two-star density, then we already know the form of $p_1(x)$ and $q_1(x)$. For smooth functions of count statistics like the transitivity, we derive the standardized and studentized Edgeworth expansion for the smooth function of the corresponding U statistics in the supplement Sections B.5.1 and B.6 respectively. We also show that the Edgeworth expansion of the bootstrapped smooth function converges to this population version in the supplement Section B.5.4.

3.6.2 Competing methods

We compare our algorithms, namely MB-M and MB-Q, with the network resampling procedures discussed in Section 3.2. In particular, we consider subsampling with subsample size $b_n = 0.5n$ (SS), the empirical graphon with resample size n (EG), the latent space bootstrap (LS), and the empirical Edgeworth expansion (EW). For the latent space bootstrap, we treat the latent dimension as known for SBM-G and estimate the latent dimension for SM-G using Universal Singular Value Thresholding (USVT) procedure of Chatterjee (2015). We provide a brief description of each algorithm below.

Empirical Graphon (EG). We draw B size n resamples S_i^* with replacement from $1, \ldots, n$. We compute the count functional $\hat{T}_{n,i}^*$ on $A^{(n)}(S_i^*, S_i^*)$. We also compute $\hat{T}_n$ and $\hat{\sigma}_n^2$ on the whole graph. Now for triangles and two-stars we compute the CDF of $\{(\hat{T}^*_{n,i} - \hat{T}_n)/\hat{\sigma}_n\}_{i=1}^B$. For functions of count functions, we compute the function for each resampled graph, center using the function computed on the whole network.

Subsampling (`SS`*).* We draw B size b subsamples S_i^* without replacement from $1, \ldots, n$. We compute the count functional $\hat{T}^*_{b,i}$ on $A^{(n)}(S_i^*, S_i^*)$. We also compute $\hat{T}_n$ and $\hat{\sigma}_n^2$ on the whole graph. We set $\hat{\sigma}_b^2 = n/b\hat{\sigma}_n^2$. Now for triangles and two-stars we compute the CDF of $\{(\hat{T}^*_{b,i} - \hat{T}_n)/\hat{\sigma}_b\}_{i=1}^B$. For functions of count functions, we compute the function for each subsampled graph, center using the function computed on the whole network.

Latent Space (`LS`*).* We first estimate the latent variables $\hat{X} := \{\hat{X}_1, \ldots, \hat{X}_n\}$ from the given network. For `SBM-G`, we use the true number of blocks, whereas for smooth graphon `SM-G`, we use the `USVT` algorithm to estimate the number of latent variables. We compute the count functional To be concrete, we compute $g_1(\hat{X}_i)$ for $i = 1 \ldots n$, and then compute $T_n(\hat{X}) = T_n(\hat{X}_1, \ldots, \hat{X}_n)$. Now we simply use the additive variant of bootstrap $T_n(\hat{X}) + \frac{r}{n} \sum_i (g_1(\hat{X}_i) - T_n(\hat{X}))$ (see Levin and Levina (2019); Bose and Chatterjee (2018)). For triangles and two-stars, we normalize by the square root of $r^2/n \sum_i g_1(\hat{X}_i)^2$. For functions of count functionals, we center using the function computed on $\hat{X}$.

We compare the performance of the resampling methods for two-stars, triangles and a variant of the transitivity coefficient defined in Example 3 of Bhattacharyya and Bickel (2015), which is essentially an appropriately defined ratio between triangle and two-star.

3.6.3 Results

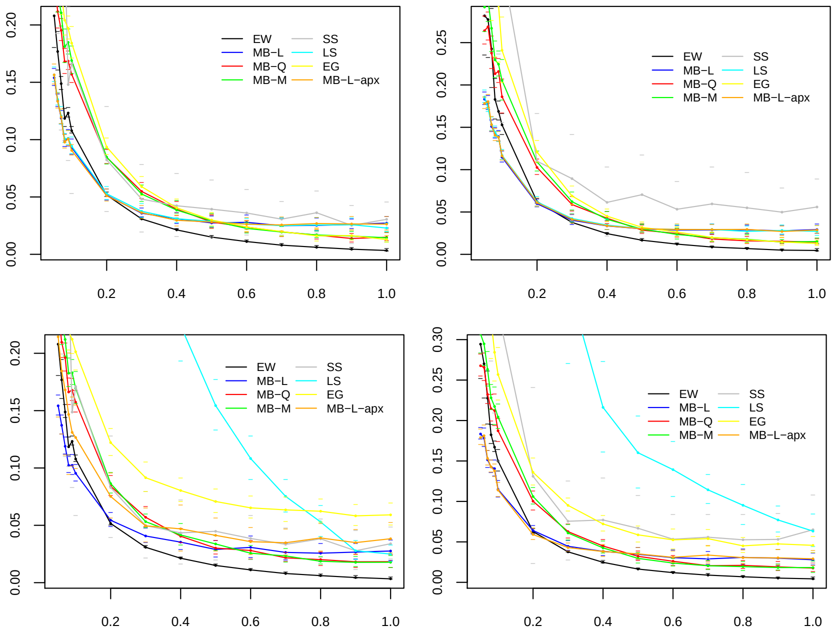
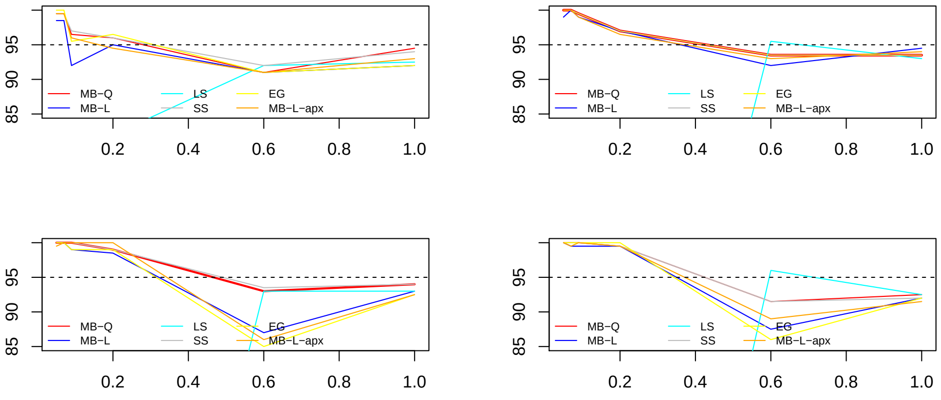


Figure 3.1: We plot err(F_n,F_n^*) for triangle density for all methods on the Y axis, where $F_n^*(t)$ corresponds to the appropriate resampling distribution. We vary the sparsity parameter ρ_n on the X axis. The networks in the left column are simulated from `SBM-G` and those in the right column are simulated from `SM-G`. The first row is centered at bootstrap mean and normalized by variance estimation from each method $\hat{\sigma}_n$ The second row is centered by triangles density estimated on the whole graph (`MB-L-apx` is centered at approximate triangle density estimated from the whole graph) and normalized by $\hat{\sigma}_n$.

In Figure 3.1, we plot the maximum of (absolute) difference of the bootstrap CDF F_n^* over the $[-3, 3]$ range (err(F_n,F_n^*)) for triangle density from the true CDF F_n for sparsity parameter ρ_n varying from 0.05 to 1. We show the average of the expected difference over 30 independent runs along with error-bars. In this figure we see an interesting phenomenon. For sparse networks with $\rho_n < 0.2$, the linear



(A) (B)

Figure 3.2: We present coverage of 95% Bootstrap Percentile CI with correction for triangles (top) and transitivity coefficient (bottom) of the `SBM-G` (A) and `SM-G` (B) models with ρ_n varying from 0.05 to 1

method outperforms the higher order correct methods. As the networks become denser, the higher order correct methods start performing better. For $\rho_n \leq 0.2$, we also see that the empirical Edgeworth expansion performs worse than the linear bootstrap method. We also compare the bootstrap samples centered at the bootstrap mean (first row) with bootstrap samples centered at the subgraph density computed on the whole network. We see that the latest space method (LS) and approximate linear bootstrap method (`MB-L-apx`) behave differently under these two centerings, with LS performing much worse. This suggests that while both suffer from bias, LS suffers from it to a higher degree. For all parameter settings, we used $N = 50 \log n$. It is possible that increasing N for sparser settings may lead to reduced bias of `MB-L-apx`.

In Figure 3.2, we show the coverage of 95% Bootstrap Percentile CI with

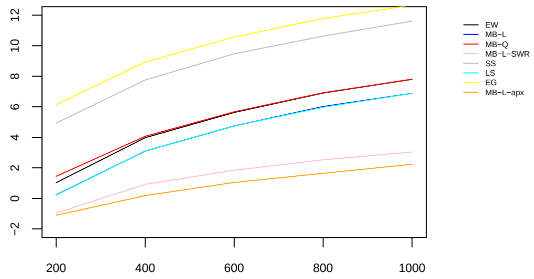


Figure 3.3: Logarithm of running time for four-cycles in SBM-G against sample size n.

correction for triangles (top) and transitivity coefficient (bottom) of the SBM-G and SM-G models in ρ_n from 0.05 to 1. We simulate 200 graphs for each ρ_n from SBM-G and SM-G models, construct CI from bootstrap percentiles and correct the CI using Eq 3.31 for triangles and transitivity respectively. For smooth functions, computing the bootstrap distribution is straightforward. One simply computes u^* which is now a vector of bootstrapped triangles and two-star densities. Now a standardized bootstrap replicate is given by $\{f(u^*) - f(u)\}/\tilde{\sigma}_f$, where u is the vector of triangles and two-star densities computed on the whole graph, and $\tilde{\sigma}_f$ is given by Eq 3.29. For transitivity, $f(x, y) = x/y$.

3.6.4 Computation time

In Figure 3.3 we show logarithm of running time for 4-cycles count against growing n for SBM-G model. (See the Supplement for timing of SM-G.) We compare our approximate linear method MB-L-apx with MB-L-SWR, which uses a randomized algorithm for approximating U-statistics proposed by Chen and Kato (2019), Section 2.2, for precomputation of the local network statistics. We see that among higher order correct methods, MB-Q offers strong computational performance, outperforming methods such as the EG and SS. We see that while EG has comparable performance to MB-Q, it requires recomputation of the count statistic for every bootstrap iteration, making it about 500 times slower than MB-Q for $n = 500$ for four-cycle counting. MB-M is the slowest one we do not show here as the weights make symmetric counting shortcuts not as simple to apply. EW is the fastest among higher order correct methods, but it cannot be readily adapted for smooth functions of count statistics and is much slower compared to additive methods LS, MB-L, MB-L-SWR, MB-L-apx. The four additive methods, i.e. LS, MB-L, MB-L-SWR, MB-L-apx are the fastest four of all methods, but they are not higher order accurate; among them MB-L-apx is the fastest method in all with our proposed approximate precomputation. The procedure MB-L-SWR draws around $N(n-1)/(r-1)$ size $r-1$ subsets with replacement, whereas we draw N permutations at random and then divide each into consecutive disjoint subsets of size $r-1$. While these two methods have similar computational complexity theoretically, we observe that MB-L-apx appears to be faster empirically.

For better presentation, additional experiments including the sup of (absolute) difference of the bootstrap CDFs for two-star density and timing for four-cycles for the SM-G model are deferred to Supplement Section B.7. The experiments are run on the Lonestar super computer (1252 Cray XC40 compute nodes, each with two 12-core Intel® Xeon® processing cores for a total of 30,048 compute cores) at the Texas Advance Computing Center.

3.7 Real Data Application

In this section, we apply our algorithms to compare networks representing the voting similarities of U.S. Congress. We use roll call vote data from the U.S. House of Representatives (Jeffrey B. et al., 2020) from 1949 (commencement of the 81^{st} Congress) to 2012 (adjournment of 112^{nd} Congress). Each Congress forms a network of representatives (nodes). An edge between a node pair is formed when the number of agreements, i.e. number of times they both vote *yay* or *nay* exceeds a threshold computed by (Andris et al., 2015) of this congress. The threshold is computed by constructing histograms of same-party pairs' number of agreements and cross-party pair's number of agreements and using the intersection point of the two histograms as the threshhold. We will denote same-party by SP and cross-party by CP. For example, the threshold value is 124 for 81st Congress and 766 for 109^{th} Congress as illustrated in Figure 3.4. For each network, we calculate the normalized cross-party edge density and cross-party triangle density, and perform our bootstrap methods on these quantities. We construct 95% second-order corrected Confidence

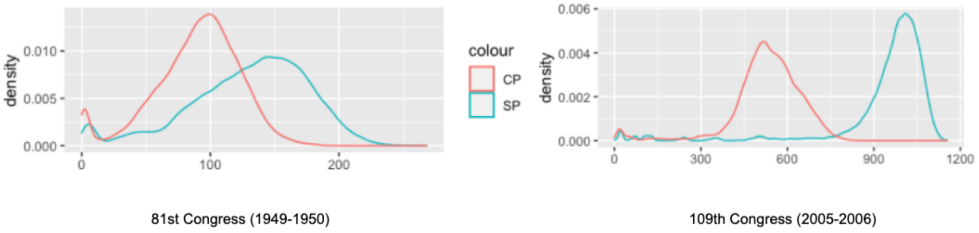


Figure 3.4: Threshold for forming edges between congress members calculated from histograms of same-party (SP) and cross-party (CP) agreements, illustrated with example of 81st Congress and 109th Congress

Intervals from the MB-Q method and present the CIs over 81st to 112nd Congress. Note that the CIs are adjusted by Bonferonni Correction where $\alpha = 0.05/32$ for 32 experiment congresses.

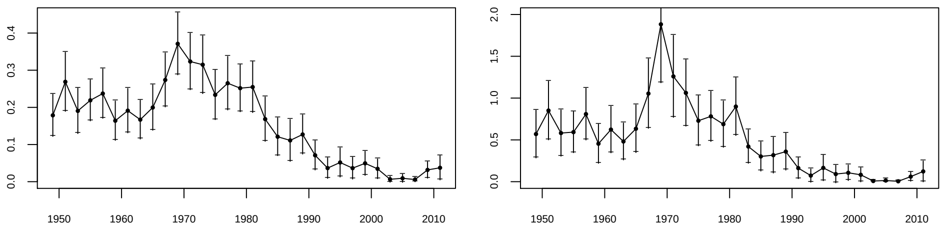


Figure 3.5: Bonferroni-adjusted 95% second-order corrected CI for cross party edge density (left) and cross party triangle density (right) from 1949 (commencement of the 81st Congress) to 2012 (adjournment of 112nd Congress).

In Figure 3.5, we can a significant decrease in cross party edge densities and cross party triangle densities over the years, suggesting a trend of decreasing bipartisan agreement.

3.8 Conclusion

In this paper, we propose multiplier bootstraps for network count functionals. Our multiplicative proposal involves perturbing a potential subgraph by the product of independent multiplier random variables. We also present the linear and quadratic bootstrap, which can be seen as different orders of approximations of the Hoeffding decomposition of the statistic arising from the multiplier bootstrap. We show that the quadratic bootstrap is higher-order correct for moderately sparse graphs whereas the linear bootstrap is first-order correct but faster. For the first time in the literature, we also derive Edgeworth expansions for smooth functions of counts. Empirically, we observe an interesting phenomenon in which the linear bootstrap, which is not higher-order correct in any regime, performs better than other methods for sparse graphs since the higher-order correct methods directly or indirectly involve estimation of higher-order moments that may not be accurately estimated under sparsity. To truly harness the computational power of the linear bootstrap, we also present and analyze an approximate bootstrap method which uses randomized sketching algorithms for estimating local network counts that are used by the linear method. Taken together, we establish the multiplier bootstrap as a user-friendly, automatic procedure that can be tailored to yield higher-order correctness or scalable and consistent inference.

Chapter 4

Separate Exchangeability as Modeling Principle in Bayesian Nonparametrics

This chapter is submitted for publication (Lin et al., 2021). It is currently under review for Statistical Science.

4.1 Introduction

We argue for the use of separate exchangeability as a modeling principle and unifying framework for data that involves multiple sets of experimental units. While exchangeability and partial exchangeability have proven to be powerful principles for statistical modeling (Bernardo and Smith, 2009), separate exchangeability has been curiously under-used as a modeling principle in some of the nonparametric Bayesian literature. In the context of two typical examples, we introduce two modeling frameworks that implement inference under separate exchangeability.

Contrary to partial exchangeability and exchangeability, separate exchangeability facilitates inference that maintains the identity of experimental units in situations like the two motivating applications with shared sets of one type of experimental units over subsets of the other type of experimental units or blocks. That is, if the data (or a design matrix) is a rectangular array, the nature of rows and columns as different experimental units is preserved. See the two motivating examples for illustrations.

Under parametric inference separate exchangeability is often naturally preserved by just introducing additive row- or column-specific effects. In contrast, this is not true in nonparametric Bayesian models. In the present article we discuss two general strategies to define Bayesian nonparametric (BNP) models to perform flexible inference and prediction under separate exchangeability. The first example concerns inference for microbiome data, with J patients and I OTUs (essentially different types of microbiomes) being two different experimental units (Denti et al., 2021). We discuss how to define separately exchangeable partition structures via nested partitions similar to Lee et al. (2013). The second example is about inference for protein activation for I proteins in J patients, under two different conditions. We build a separately exchangeable model for random effects, using simple additive structure.

The rest of this article is organized as follows. Section 4.2 reviews the notion of exchangeability and how it relates to modeling in Bayesian inference. Section 4.2.3 discusses the borrowing of information under separate exchangeability, focusing in particular on dependent random partitions as they arise under mixture models. Section 4.3 describes two motivating data sets. In the context of these two applications we then proceed to introduce two specific models to implement separate exchangeability in related problems. Section 4.4 presents one general strategy based on a construction of nested partitions that preserves the identity of two types of experimental units. The strategy is illustrated with inference for the microbiome data set. Section 4.5 presents another general strategy based on a BNP regression model with an additive structure of random effects related to the two types of experimental units. Section 4.6 contains concluding remarks and suggested future work. Additional details, including Markov chain Monte Carlo (MCMC) based posterior inference algorithms, are presented in the supplementary materials.

4.2 Exchangeability as a Modeling Principle

To perform inference and prediction we rely on some notion of homogeneity across observations that allows us to leverage information from the sample $x_{1:n} = (x_1, \ldots, x_n)$ to deduce inference about a set of future observations $x_{n+1:n+m}$. De Finetti refers to this as "analogy" (de Finetti, 1937). In Bayesian statistics, the assumptions are stated in the language of probability and learning is naturally performed via conditional probability.

4.2.1 Exchangeability and partial exchangeability.

Exchangeability. A fundamental assumption that allows such generalization in Bayesian learning is exchangeability, that is, invariance of the joint law of the data with respect to permutations of the observation indices. This entails that the order of the observations is irrelevant in the learning process and one can deduce inference for $x_{n+1:n+m}$ from observations $x_{1:n}$. More precisely, a sequence $x_{1:n}$ is judged exchangeable if

$$x_{1:n} \overset{d}{=} x_{\pi(1:n)} \tag{4.1}$$

for any permutation π of $[n] := \{1, \ldots, n\}$. Here $\overset{d}{=}$ denotes equality in distribution. If the observable $x_1, \ldots, x_n$ are considered a sample from an infinite exchangeable sequence $(x_i)_{i\geq 1}$, that is finite exchangeability holds for any sample size $n \geq 1$, de Finetti's theorem (de Finetti, 1930) states that such an extendable sequence $x_1, \ldots, x_n$ is exchangeable if and only if it can be expressed as conditionally independently identically distributed (i.i.d.) from a random probability P. The model is completed with a prior on P:

$$\begin{aligned} x_i \mid P &\overset{iid}{\sim} P, \quad i = 1, 2, \ldots \\ P &\sim \mathcal{L}. \end{aligned} \tag{4.2}$$

The characterization of infinite exchangeability as a mixture of i.i.d. sequences highlights the fact that the homogeneity assumption of exchangeability in Bayesian learning is equivalent to assuming an i.i.d. sequence in the frequentist paradigm. The de Finetti measure $\mathcal{L}$ can be interpreted as the prior in the Bayes-Laplace paradigm. The random measure P in (4.2) is known as the directing measure and its prior $\mathcal{L}$ is known as the de Finetti measure. If P is restricted to a parametric family, we can write P as P_θ, where θ denotes a finite dimensional random parameter and $\mathcal{L}$ reduces to a prior probability model on θ. However, if P is unrestricted we have the characterization in (4.2) and $\mathcal{L}$ can take the form of a BNP prior on P (Müller and Quintana, 2004).

Note that the unknown probability P arises from an assumption on observable quantities, justifying inference on the parameters/latent quantities in terms of measurable quantities. Eliciting assumptions in terms of observable, and thus testable, events is fundamental in science also if the main inference goal were inference on latent quantities. In general, a predictive approach of statistics is becoming increasingly popular in the statistics and machine learning community. See Fortini et al. (2000, 2012) or Fortini and Petrone (2016) for interesting discussions of the predictive approach and characterization results for the prior probability measure in terms of predictive sequences under exchangeability and partial exchangeability.

Partial exchangeability. In real world applications, the assumption of exchangeability is often too restrictive. To quote de Finetti (1937): *"But the case of exchangeability can only be considered as a limiting case: the case in which this 'analogy' is, in a certain sense, absolute for all events under consideration. [..] To get from the case of exchangeability to other cases which are more general but still tractable, we must take up the case where we still encounter 'analogies' among the events under consideration, but without attaining the limiting case of exchangeability."* Indeed, depending on the design of the experiment it is often meaningful to generalize exchangeability to less restrictive invariance assumptions that allow us to introduce more structure into the prior, and thus into the learning mechanism.

If the data are collected in different, related populations a simple generalization of exchangeability is partial exchangeability. Let $\boldsymbol{X} = (x_{ij} : i = 1, \ldots, I_j,\ j = 1, \ldots, J)$ denote a data array, where j is the label of the population from which x_{ij} is collected. We say that $\boldsymbol{X}$ is partially exchangeable if the joint law is invariant under different permutations of the observations within each population

$$(x_{\pi_j(i),j} : i = 1, \ldots, I_j;\ j = 1, \ldots, J) \overset{d}{=} (x_{\pi_j(i),j} : i = 1, \ldots, I_j;\ j = 1, \ldots, J), \tag{4.3}$$

for any family of permutations $\{\pi_1, \ldots, \pi_J\}$. Partial exchangeability entails that the order of the observations is irrelevant in the learning mechanism up to preserving the information of the population memberships. If partial exchangeability holds for any sample sizes $(I_1, \ldots, I_J)$ the "analogy" assumption of partial exchangeability can be characterized in terms of latent quantities (e.g. parameters) via de Finetti's theorem (de Finetti, 1937),

$$\begin{aligned} x_{ij} \mid (P_1, \ldots, P_J) &\overset{ind}{\sim} P_j \quad j = 1, 2, \ldots;\ i = 1, 2, \ldots \\ (P_1, \ldots, P_J) &\sim \mathcal{L}. \end{aligned} \tag{4.4}$$

Therefore, partially exchangeable extendable arrays can be thought of as decomposable into different conditionally independent exchangeable populations. As in (4.2) the characterization in (4.4) does not restrict the distributions associated with the different populations to any parametric family. The P_j would usually be assumed to be dependent, allowing borrowing of strength across blocks.

Note that exchangeability is a degenerate special case of partial exchangeability which corresponds to ignoring the information on the specific populations j from which the data are collected, i.e., ignoring the known heterogeneity. The opposite, also degenerate, extreme case corresponds to modeling data from each population independently, i.e., ignoring similarities between populations. See Aldous (1985) or Kallenberg (2006) for detailed probabilistic accounts on different exchangeability assumptions and Foti and Williamson (2013) and Orbanz and Roy (2014) for insightful discussions on the topic in the context of non-parametric Bayesian models (BNP).

The way how partial exchangeability preserves heterogeneity is perhaps easiest seen in considering marginal correlations between pairs of observations. As desired, partial exchangeability allows increased dependence between two observation arising from the same experimental condition, compared to correlation between observations arising from different populations. For instance, under partial exchangeability (4.3) it is possible that

$$\mathrm{Corr}(x_{ij}, x_{i'j}) > \mathrm{Corr}(x_{ij}, x_{i'j'}), \quad j \neq j',\ i \neq i' \tag{4.5}$$

while by definition of exchangeability, under (4.1)

$$\mathrm{Corr}(x_{ij}, x_{i'j}) = \mathrm{Corr}(x_{ij}, x_{i'j'}), \tag{4.6}$$

entailing that in the learning process we can borrow information across groups, but to learn about a specific population j the observations recorded in the same population j are more informative than observations from another population j'. Here and in the following we assume that x_{ij} are real valued, square integrable random variables such that the stated correlations are well defined. See Appendix A for a (simple) proof of (4.5).

Going beyond correlations between individual observations, one can see a similar pattern for groups of observations under populations j and j'. See Section 4.2.3 for more discussion.

Statistical inference. From a statistical perspective partial exchangeability is a framework that allows to borrow information across related populations while still preserving heterogeneity of populations. Partially exchangeable models are widely used in many applications, including meta-analysis, topic modeling and survival analysis when the design matrix records different statistical units (e.g. patients) under related experimental conditions (e.g. hospitals).

Flexible learning can be achieved assuming dependent non-parametric priors for the vector of random probabilities P_j in (4.4). An early proposal appeared in Cifarelli and Regazzini (1978), but the concept was widely taken up in the literature only after the seminal paper of MacEachern (1999) introduced the dependent DP (DDP) as a prior over families $\mathcal{F} = \{F_x,\ x \in X\}$, an instance of which could be used for $\mathcal{L}$ in (4.4). See Quintana et al. (2020) for a recent review of DDP models.

Finally, in anticipation of the later application examples we note that while in a stylized setup it is convenient to think of x_{ij} in (4.4) as the observable data, in many applications the discussed symmetry assumptions are used to construct prior probability models for (latent) parameters. The x_{ij} might be, for example, cluster membership indicators for bi-clustering models, or any other latent quantities in a larger hierarchical model. The same comment applies for the upcoming discussion of separate exchangeability.

4.2.2 Separate exchangeability.

We discussed earlier that exchangeability can be too restrictive to model the case when observations are arranged in different populations. Similarly, partial exchangeability can prove to be too restrictive when the same experimental unit is recorded across different blocks of observations (e.g., the same patient is recorded in different hospitals). A similar case arises when two types of experimental units are involved, and observations are recorded for combinations of the two types of units, as is common in many experimental designs (e.g., in the first motivating study in Section 4.3 the same type of microbiome is observed across different subjects).

In such a case a simple, but effective, homogeneity assumption that preserves the information on the experimental design is separate exchangeability, that is, invariance of the joint law under different permutations of indices related to the two types of units (or blocks). If data is arranged in a data matrix with rows and columns corresponding to two different types of units (or blocks), this reduces to invariance with respect to arbitrary permutations of rows and column indices. More precisely, a data matrix is separately exchangeable if

$$x_{1:n,1:J} \overset{d}{=} x_{\pi_1(1:n),\pi_2(1:J)} \tag{4.7}$$

for separate permutations π_1 and π_2 of rows and columns, respectively.

The notion of separate exchangeability formally reflects the known design in the learning mechanism. That is, it introduces more dependence between two values recorded from the same statistical unit than between values recorded in different statistical units. Note also that partial exchangeability of $x_{1:I,1:J}$ grouped w.r.t. columns plus exchangeability of the columns is a stronger homogeneity assumption than separate exchangeability in a similar way as exchangeability is a degenerate case of partial exchangeability (i.e. we loose some structure). Similar to (4.5) and (4.6), under separate exchangeability it is possible that

$$\mathrm{Corr}(x_{ij}, x_{ij'}) > \mathrm{Corr}(x_{ij}, x_{i'j'}), \quad j \neq j',\ i \neq i' \tag{4.8}$$

while by the definition of partial exchangeability, under (4.3) we always have

$$\mathrm{Corr}(x_{ij}, x_{ij'}) = \mathrm{Corr}(x_{ij}, x_{i'j'}). \tag{4.9}$$

In fact, inequality (4.8) is always true with ≥ (for extendable separately exchangeable array), as can be shown similar to a corresponding argument for partial exchangeability in Appendix A. See Section 4.2.3 for a more detailed discussion on the borrowing of information under partial and separate exchangeability.

Finally, we conclude this section reviewing a version of de Finetti's theorem for separately exchangeable arrays. If $x_{1:I,1:J}$ is extendable, that is, it can be seen as a projection of $(x_{ij} : i = 1, 2, \ldots;\ j = 1, 2, \ldots)$, a representation theorem in terms of latent quantities was proven independently by Aldous (1981) and Hoover (1979). See also Kallenberg (1989) and reference therein. More precisely, an extendable matrix $x_{1:n,1:J}$ is separately exchangeable if and only if

$$x_{ij} = f(\theta, \xi_i, \eta_j, \zeta_{ij}), \tag{4.10}$$

for some measurable function $f : [0, 1]^4 \rightarrow \mathbb{R}$ and i.i.d. Unif$(0, 1)$ random variables θ, ξ_i, η_j and $\zeta_{ij},\ i, j \in \mathbb{N}$. The representation theorem in (4.10) implies a less strict representation theorem in which the uniform distributions are replaced by any distributions p_ξ, p_η and p_ζ, as long as independence is maintained (together with a corresponding change of the domain for f). This representation in turn can alternatively be written as

$$p(x_{1:I,1:J}) = \int p_\theta(\theta) \prod_{i=1}^{n} p_\xi(\xi_i) \prod_{j=1}^{J} p_\eta(\eta_j) \prod_{ij} p(x_{ij} \mid \theta, \xi_i, \eta_j)\, d\theta\, d\xi_{1:n}\, d\eta_{1:J}. \tag{4.11}$$

Similar to (4.2) or (4.4) model (4.11) can be stated as a hierarchical model

$$\begin{aligned} &x_{ij} \mid \theta, \eta_j, \xi_i \overset{ind}{\sim} P_{\theta,\xi_i,\eta_j}, \quad i = 1, \ldots;\ j = 1, 2, \ldots \\ &\theta \sim p_\theta \perp \xi_i \overset{iid}{\sim} p_\xi \perp \eta_j \overset{iid}{\sim} p_\eta,\ i = 1, 2, \ldots;\ j = 1, 2, \ldots \end{aligned} \tag{4.12}$$

where P_{θ,ξ_i,η_j} is the law of $p(x_{ij} \mid \theta, \xi_i, \eta_j)$ in (4.11). Finally, in both, (4.11) and (4.12), the probability models can still be indexed with unknown hyperparameters.

4.2.3 Separately exchangeable random partitions

We discuss in more detail the nature of borrowing of information under partial and separate exchangeability. For easier exposition we focus on a matrix $\boldsymbol{X} = [x_{ij}, j = 1, \ldots, J,\ i = 1, \ldots, I]$, with two columns, i.e. $J = 2$. Note that the assumptions of partial and separate exchangeability imply marginal exchangeability within columns. Moreover, we exclude the degenerate cases of (marginal) independence between $x_{1:ij}$ and $x_{1:I,j'}$, when no borrowing of strength occurs under Bayesian learning, as well as the fully exchangeable case when we loose heterogeneity across columns.

In the previous section we discussed how separate exchangeability, contrary to partial exchangeability, allows to respect the identity of an experimental unit i by allowing for increased dependence between $x_{i,1}$ and $x_{i,2}$ compared to dependence between $x_{i,1}$ and $x_{i',2}$. [1]

We now extend the discussion to borrowing of information also on other functionals of interest of $x_{1:I,1}$ and $x_{1:I,2}$, potentially even while assuming $x_{i,1}$ and $x_{i',2}$ independent for any $i, i' \in [I]$, but $x_{1:I,1}$ and $x_{1:I,2}$ dependent. In the following we focus on a fundamental example related to borrowing of information about a partition of $x_{1:I,1}$ and $x_{1:I,2}$. For this discussion, we assume discrete $P_{\theta,\xi,\eta}$. In that case, ties of $(x_{ij},\ i = 1, 2, \ldots)$ define a partition of $\{1, \ldots, I\}$ for each column j (and similarly for rows – but for simplicity we only focus on columns). Let Ψ_j denote this partition in column j. A common context when such situations arise is when x_{ij} are latent indicators in a mixture model for observable data y_{ij}, with a top level sampling models $p(y_{ij} \mid x_{ij} = \ell, \varphi_\ell)$, where φ_ℓ are additional cluster-specific parameters.

Similarly, a partially exchangeable model (4.4) with discrete probability measures P_j defines a random partition Ψ_j in column j. If we then induce dependence between Ψ_1 and Ψ_2 (by way of dependent P_j), it is possible to borrow information about the law of the random partitions, for example, the distribution of the number of clusters. See Franzolini et al. (2021) for definition and probabilistic characterizations of partially exchangeable partitions. However, by definition of partial exchangeability (4.3), with only partial exchangeability it is not possible to borrow information about the actual realizations of the random partitions Ψ_1, Ψ_2. For example, under a partially exchangeable partition

$$p(\{x_{1,j'} = x_{2,j'}\} \mid \{x_{1,j} = x_{2,j}\}) = p(\{x_{1,j'} = x_{3,j'}\} \mid \{x_{1,j} = x_{2,j}\}). \qquad (4.13)$$

[1] Note that sometimes it might be desirable to introduce negative dependence (repulsion), e.g. as $\mathrm{Corr}(x_{i,1}, x_{i',2}) < 0$.

In contrast, under separate exchangeability it is possible to have

$$p(\{x_{1,j'} = x_{2,j'}\} \mid \{x_{1,j} = x_{2,j}\}) > p(\{x_{1,j'} = x_{3,j'}\} \mid \{x_{1,j} = x_{2,j}\}). \qquad (4.14)$$

From a statistical perspective this means that under separate exchangeability the fact that two observations are clustered together (e.g. 1 and 2) in one column can increase the probability that the same two observations are clustered together in another column. This difference in the probabilistic structures, and thus in the borrowing of information under Bayesian learning, is particularly relevant when observations indexed by i have a meaningful identity in a particular application. For example, in one of the motivating examples, units $i = 1, 2, 3$ refer to three different types of microbiomes (OTU). In words, (4.14) implies that seeing OTUs 1 and 2 clustered together in subject j increases the probability of seeing the same OTUs co-cluster in subject j'. In the actual application x_{ij} will refer to parameters in a hierarchical model.

In Section 4.4 we discuss an effective way to define flexible, but analytically and computationally tractable, non-degenerate separately exchangeable partitions. We first cluster columns, and then set up nested partitions of the rows, nested within column clusters. That is, all columns in the same column-cluster share the same nested partition of rows.

4.3 Two Examples

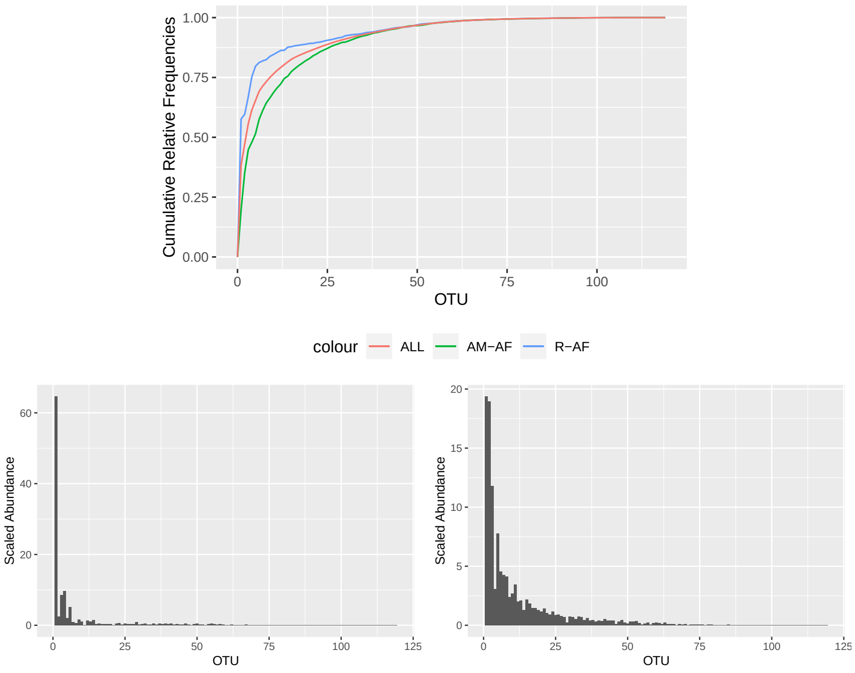
We use two motivating examples to illustrate the notion of separate exchangeability, and to introduce two modeling strategies that provide specific implementations respecting separate exchangeability.

In the first example we assume separate exchangeability for data y_{ij}. In short, we construct a separately exchangeable model for a data matrix $[y_{ij}]$ by setting up a partition of columns and, nested within column clusters, partitions of row. Such nested partitions are created by assuming separate exchangeability for parameters x_{ij} in a hierarchical prior model. In the second example we assume separate exchangeability for the prior on model parameters θ_{it} which index semi-parametric regression models. The separately exchangeable model is set up in a straightforward way by defining additive structure with exchangeable priors on terms indexed by i and t. In both cases the model is separately exchangeable without reducing to the special cases of partial exchangeability or marginal exchangeability.

Microbiome data - random partitions. The first example is inference for microbiome data for $J = 38$ subjects and $I = 119$ OTUs (operational taxonomic units). The data report the frequencies y_{ij} of OTU i for subject j (after suitable normalization). In building a model for these data we are guided by separate exchangeabilty with respect to subject indices j and OTU indices i. We use a publicly available microbiome data set from a diet swap study (O'Keefe et al., 2015). The same data was analyzed by Denti et al. (2021). The dataset includes subjects from the US and rural Africa. The interest is to investigate the different patterns of microbial diversity (OTU counts) across subjects and how the distributions of OTU abundancy vary across subgroups of subjects. Focusing on inference for microbial diversity they restricted attention to inference about identifying clusters of subjects with similar distributions of OTU frequencies, considering OTU counts y_{ij} as partially exchangeable. While this focus is in keeping with the tradition in the literature, it is ignoring the shared identity of the OTUs i across subjects j. In Section 4.4 we will set up an alternative model that respects OTU identity by modeling the data as separately exchangeable.

We show some summary plots of the data, trying to motivate the proposed inference. First we sort OTUs by overall abundancy (across all subjects). Figure 4.1 (top) shows the cumulative relative frequencies of OTUs in rural Africans (RU), African Americans (AA) and all subjects, highlighting a difference in distribution of OTU frequencies between RU and AA. Throughout, The OTU frequencies in this data have been scaled by average library size, i.e, y_{ij} is the absolute count of OTU z_{ij} normalized by the totals $\gamma_j = \sum_{i=1}^{n} z_{ij}$. The two barplots at the bottom of the same figure show OTU abundancies in the two groups, suggesting that subjects might meaningfully group by distribution OTU frequencies.

In Figure 4.2, we show hierarchical clustering for subjects based on the 10 OTUs with highest empirical variance. Note how the clusters correlate well with the two groups, RA and AA, suggesting that in grouping subjects by distribution of OTUs frequencies we should proceed in a way that maintains and respects OTU identities (as is the case in the hierarchical clustering). Importantly, any inference on such groupings has to account for substantial uncertainty. Observing these features in the figures motivates us to formalize inference on grouping subjects by OTU abundancies using model-based inference. We will set up a separately exchangeable model, exchangeable with respect to subject and OTU indices.



Rural-African

African-American

Figure 4.1: Top: Cumulative Relative Frequencies of OTU for average Rural Africans (RU), average African Americans (AA) and average of all subjects. Bottom: Histogram of OTU abundance in RA and AA (scaled as described in the text).

Protein expression - nonparametric regression. In a second example we analyze protein measurement data in a study of ataxia, a neurodegenerative disease. The same data is studied in Lee et al. (2021). The data measures the abundancy of 4350 potentially disease-related proteins among two groups of subjects, one control group and one patient group. Each group includes 16 subjects with ages between 5 and 50 years. The subjects' ages define time for our observations. There are $T = 16$ unique times τ_t, $t = 1, \ldots, T$, with one control and one patient for each unique time τ_t. That is, τ refers to time in calendar years, and t is an index in the list of the T unique time points. The data are protein activation data y_{ij}, $j = 1, \ldots, J$, $i = 1, \ldots, I$, for

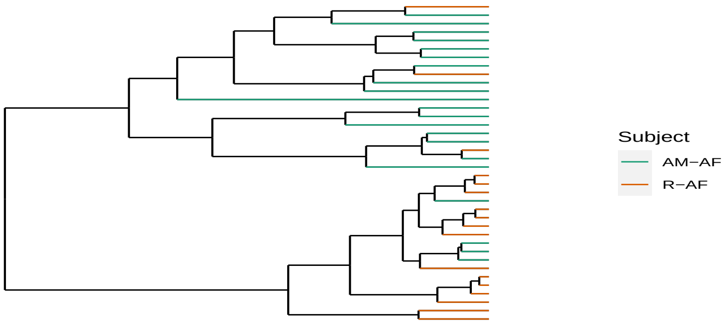


Figure 4.2: Agglomerative Hierarchical Clustering with Euclidean Distance and Complete method subjects using 10 OTUs with highest cross-subject variance.

the $J = 32$ subjects and $I = 4350$ proteins. For each subject, z_j is an indicator for being a patient ($z_j = 1$) or control (0), and $t_j \in \{1, \ldots, T\}$ denotes the age. The inference goal is to identify proteins that are related with the disease, defined as proteins with a large difference between patients and controls in change of protein expression over age. We set up a nonparametric regression for y_{ij} versus age t_j. The regression mean function is constructed using a cubic B-splines with 2 interior knots, an offset for protein i, and an interaction of treatment and B-spline, to allow for a difference in spline coefficients for patients versus controls. See the later discussion for more model details. In this case, the separate exchangeability assumption is made for protein-and-age specific effects θ_{it} that appear in the regression mean function.

Figure 4.3 shows a random subset of the data. The figure shows protein expression (on a logarithmic scale) over time for 10 randomly selected proteins, separately for patient and control groups. The figure highlights the high level of noise in this data. Recall the primary inference goal of identifying proteins with the largest

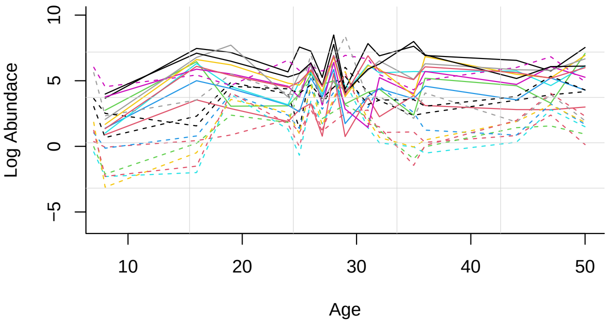


Figure 4.3: Randomly selected 10 proteins: log abundancy y_{ij} against age t_j. Solid lines correspond to patients ($z_j = 1$), and dotted lines indicate controls ($z_j = 0$). Lines corresponding to the same protein share the same color.

Table 4.1: Top 10 selected proteins using simple data summaries. Notice the lack of overlap. See the text for details.

Naive method 1: Empirical difference of differences					
1	Q9H6R3	Q9Y3E1	P49591	Q9NQ66-1	P01859
6	O60256-1	P10915	P10768	P53041	Q9H6U6-8
Naive method 2: separate regression for each protein					
1	Q13634-1	Q9Y6U3	P04275	Q5VSL9-1	Q9NYY8
6	Q9UPU9-1	P06454-1	P24844-1	Q8WZA9	P48651

difference in slopes between patients versus controls. Table 4.1 shows the selected top 10 proteins using two simple data summaries. Let j_{01}, j_{0T}, j_{11} and j_{1T} denote the indices of four subjects, with j_{zt} indicating the subject with $z_j = z$ and $t_j = t$. The first set of 10 proteins are the proteins with the largest empirical difference

$$\hat{\gamma}_i = \frac{y_{ij_{1T}} - y_{ij_{11}}}{T-1} - \frac{y_{ij_{0T}} - y_{ij_{01}}}{T-1}, \tag{4.15}$$

that is, the 10 proteins with largest observed difference between patients and controls in change over time. One problem with $\hat{\gamma}_i$ is that it is based on only the 4 subjects with minimum and maximum age, and does not borrow any strength from data for subjects with ages in between, or other proteins. The second set of 10 proteins are the proteins with largest fitted difference, fitting for each protein two separate smoothing splines, one to all patients and a second one to all controls, and evaluating $\hat{\gamma}_i$ replacing the data by the fitted values $\hat{y}_{ij}$ under these smoothing splines. The second set therefore includes borrowing of strength across all subjects, but still no borrowing of strength across proteins. And both summaries ignore uncertainty of $\hat{\gamma}_i$. Note that the top 10 proteins based on these two data summaries include no overlap, highlighting again the high level of noise in these data, and the need for more principled inference and characterization of uncertainties. These observations motivate us to develop a hierarchical model to borrow strength across proteins and time, and to allow for a full description of uncertainties. The hierarchical model across subjects and proteins is set up using separate exchangeability on parameters $\boldsymbol{\theta}_{it}$. Note that in this case separate exchangeability is not on the data, but on parameters (including slope etc.) of the fitted curves. See later for details. The model includes a random partition of proteins to group proteins with similar patterns into common clusters, allowing us to identify a group of interesting proteins as the cluster with the highest cluster-specific change in protein expression over time.

4.4 Separate Exchangeability through Nested Partitions

4.4.1 Separate versus partial exchangeability

The microbiome data records frequencies y_{ij} of $I = 119$ OTUs $i = 1, \ldots, I$, for $J = 38$ subjects, $j = 1, \ldots, J$. The measurement y_{ij} reports the scaled abundancy of OTU i in subject j. The main inference goal is to identify subgroups C_k, $k = 1, \ldots, K^+$, of subjects with different patterns of OTU frequencies. The subsets C_k define a partition as $\bigcup_{k=1}^{K^+} C_k = [J]$ with $C_k \cap C_{k'} = \emptyset$ for $k \neq k'$. We refer to the C_k as subject clusters. Alternatively we represent C_k using cluster membership indicators $S_j = k$ if and only if $j \in C_k$.

Partially exchangeable model. We assume mixture of normal distributions. Let $\text{GEM}(\alpha)$ denote a stick-breaking prior for a sequence of weights (Sethuraman, 1994). Marginally for each subject j (i.e., marginalizing over other subjects, $j' \neq j$), we assume

$$\begin{aligned}
& y_{ij} \mid S_j = k \overset{iid}{\sim} G_k,\ i = 1, 2, \ldots \\
& G_k = \sum_{\ell=1}^{\infty} w_{k\ell} N(\mu_\ell, \sigma_\ell^2) \text{ and } \theta_\ell = (\mu_\ell, \sigma_\ell^2) \overset{iid}{\sim} G_0,\ \boldsymbol{w}_k \sim \text{GEM}(\alpha) \\
& p(S_j = k \mid \boldsymbol{\pi}) = \pi_k \text{ and } \boldsymbol{\pi} = (\pi_1, \ldots) \sim \text{GEM}(\beta) \qquad (4.16)
\end{aligned}$$

with independent sampling across $i = 1, \ldots, I$. Marginally, for one subject j, the first two lines of (4.16) imply a Dirichlet process (DP) mixture of normal model for y_{ij}, $i = 1, 2, \ldots$. That is, the marginal law for each j is a DP mixture

$$y_{ij} \mid G \overset{iid}{\sim} G,\ i = 1, \ldots, I,$$

$$G = \int N(\mu, \sigma^2)\, dF(\mu, \sigma^2) \text{ and } F = \sum_{\ell} w_{k\ell} \delta_{\mu_\ell, \sigma_\ell} \sim \text{DP}(\alpha, G_0). \tag{4.17}$$

We keep using notation $w_{k\ell}$ to highlight the link with (4.16). Here (μ_ℓ, σ_ℓ) are the atoms of a discrete random probability measure F and $\text{DP}(\alpha, G_0)$ defines a DP prior with total mass α and base measure G_0. See, for example, (Müller et al., 2015, Chapter 2) for a review of such DP mixture models. However, the fact that in (4.16) multiple subjects can share the same G_k introduces dependence across subjects. Additionally, note that the normal moments $\theta_\ell = (\mu_\ell, \sigma_\ell^2)$ are indexed by ℓ only, implying common atoms of the G_k across k. The model construction is completed by assuming that y_{ij} in (4.16) are sampled independently also across j given the vector of random probabilities. Below we will introduce a variation of this final assumption, motivated by the following observation.

The use of common atoms (μ_ℓ, σ_ℓ) across G_k allows us to define clusters of observations across OTUs and subjects. This is easiest seen by replacing the mixture of normal model in the first line of (4.16) by a hierarchical model with latent indicators $\tilde{M}_{ij}$ as

$$p(y_{ij} \mid \tilde{M}_{ij} = \ell) = N(\mu_\ell, \sigma_\ell^2) \text{ and } p(\tilde{M}_{ij} = \ell \mid S_j = k) = w_{k\ell}.$$

Interpreting $\tilde{M}_{ij}$ as cluster (of OTUs) membership indicators, the model defines clusters $\tilde{R}_{j\ell} = \{(ij) : \tilde{M}_{ij} = \ell\}$. The model also defines a random partition $\tilde{\Psi}_j$ of OTUs for each subject j with clusters defined by $\tilde{R}_{j\ell} = \{(ij) : \tilde{M}_{ij} = \ell\}$, i.e., $\tilde{\Psi}_j = \{\tilde{R}_{j\ell},\ \ell = 1, 2, \ldots\}$. In this construction subjects j with shared $S_j = k$ share the same *prior* $p(\tilde{\Psi}_j)$ on the partition of OTUs, implied by $y_{ij} \mid P_j \sim P_j$ with $P_j = G_{S_j}$, defining partial exchangeability as in (4.4), with the random partition

$(S_1, \ldots, S_J)$ defining a dependent prior on $(P_1, \ldots, P_J)$. Conditional of $S_1, \ldots, S_J$ the latter is characterized by only K^+ distinct G_k.

A separately exchangeable prior. Recognizing the described construction with $\tilde{M}_{ij}$ as reducing to partial exchangeability when non-degenerate separate exchangeability is implied by the nature of the experiment, we introduce a modification. While the change is minor in terms of notation, it has major consequences for interpretation and inference as we will show later. We replace the indicators $\tilde{M}_{ij}$ by M_{ik} (note the subindex $_k$) specific to each OTU and *cluster of subjects*, with otherwise unchanged marginal prior

$$p(M_{ik} = \ell) = w_{k\ell} \tag{4.18}$$

and $p(y_{ij} \mid S_j = \ell, M_{ik} = \ell) = N(\mu_\ell, \sigma_\ell^2)$. The assumption completes the marginal model (4.16) by introducing dependence of the y_{ij} across $j \in C_k$, which is parsimoniously introduced with the M_{ik} indicators. The marginal distribution (4.17) remains unchanged. But now the implied random partitions $\tilde{\Psi}_j$ are shared among all $j \in C_k$. Let $\Psi_k = \{R_{kl}, l = 1, 2, \ldots\}$ denote this shared partition.

In practice we use an implementation using a finite Dirichlet process (Ishwaran and James, 2001) for G_k, i.e., we truncate G_k with L atoms. Similarly, we truncate the stick-breaking prior for π_k at a fixed number of K atoms. For later reference we state the joint probability model conditional on K and L is

$$p(y_{1:I,1:J}, \boldsymbol{S}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{w}, \mid K, L) = \prod_{j=1}^{J} \prod_{i=1}^{I} p(y_{ij} \mid \mu_{M_{i,S_j}}, \sigma^2_{M_{i,S_j}})$$

$$\times \prod_{j=1}^{J} \pi_{S_j} \left\{ \prod_{k=1}^{K} \prod_{i=1}^{I} w_{M_{ik},k} \right\} p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{w}_k) \prod_{\ell=1}^{L} p(\mu_\ell, \sigma_\ell^2) \quad (4.19)$$

with $p(\boldsymbol{\pi}) = p(\pi_1, \ldots, \pi_K) = \text{GEM}(\beta)$ and $p(\boldsymbol{w}_k) = p(w_{k1}, \ldots, w_{kL}) = \text{GEM}(\alpha)$ being finite stick breaking priors, and $p(\mu_\ell, \sigma_\ell^2)$ chosen to be conditionally conjugate for the sampling model $p(y_{ij} \mid \mu_\ell, \sigma_\ell^2)$.

In summary, we have introduced separate exchangeability for y_{ij} by defining (i) a random partition $\gamma = \{C_1, \ldots, C_{K^+}\}$ of columns, corresponding to the cluster membership indicators S_j, and (ii) nested within column clusters C_k, a nested partition $\Psi_k = \{R_{k1}, \ldots, R_{kL}\}$ of rows, represented by cluster membership indicators M_{ik}. In contrast, a model under which $j \in C_k$ only share the *prior* $p(\tilde{\Psi}_j)$ on the nested partition reduces to the special case of partial exchangeability. The model remains invariant under arbitrary permutation of the row (OTU) labels in any column (subject). Without the reference to separate exchangeability a similar construction with nested partitions was also used in Lee et al. (2013). The construction of the nested partition is identical, but there is no notion of common atoms to allow for clusters of observations across column clusters.

Finally, to highlight the nature of the model as being separately exchangeable with respect to OTUs and subjects we exhibit the explicit Aldous-Hoover representation (4.11), still conditional on hyperparameters. We show separate exchangeability conditional on $\phi = (\boldsymbol{\pi}, \boldsymbol{w})$ by matching variables with the arguments in (4.11) as follows: $\eta_j = S_j$, $\xi_i = (M_{ki}, k = 1, \ldots, K)$, $\theta = (\boldsymbol{\mu}, \boldsymbol{\sigma})$, and $p(x_{ij} \mid \theta, \xi_i, \eta_j) = N(\mu_\ell, \sigma_\ell^2)$ with $\ell = M_{S_j i}$. Here we used that (4.11) allowed conditioning on additional hyperparameters, in this case ϕ.

4.4.2 Results

Posterior simulation is implemented as a straightforward Gibbs sampling algorithm. All required complete conditional distributions follow easily from the joint probability model (4.19). Posterior Monte Carlo simulation is followed by a posterior summary of the random partition using the approach of Dahl (2006) to minimize Binder loss, using the algorithm in Dahl et al. (2021). We estimate three clusters of subjects, C_1, C_2, C_3. We show the three subject clusters in Figure 4.4 by plotting cumulative frequencies of OTUs. We sort all OTUs by overall abundancy across subjects. For each cluster of subjects, we collect all subjects j and plot cumulative (observed) frequencies.

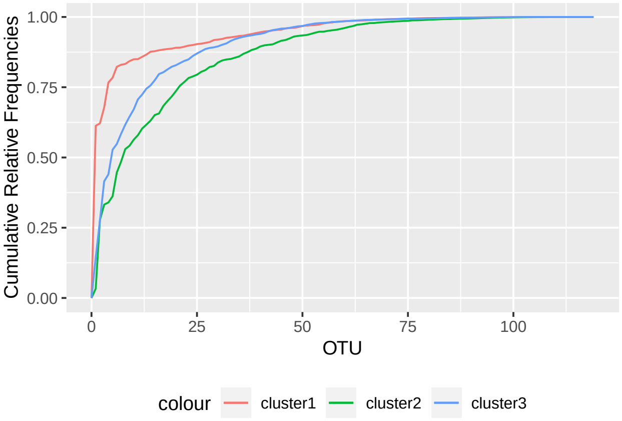


Figure 4.4: Cumulative Relative Frequencies of OTU in the subject clusters C_k, $k = 1, 2, 3$, of subjects.

Figure 4.5 summarizes the nested partition of OTUs, nested within the three subject clusters. The three panels correspond to subject clusters $k = 1, 2$ and 3. For each subject cluster, the figure shows the estimated co-clustering probabilities of

OTUs, i.e., $p^k_{ii'} = p(M_{ik} = M_{i'k} \mid \boldsymbol{y}, \boldsymbol{S})$ for each pair (i, i') of OTUs. As usual for heatmaps, the OTUs are sorted for a better display, to highlight the clusters.

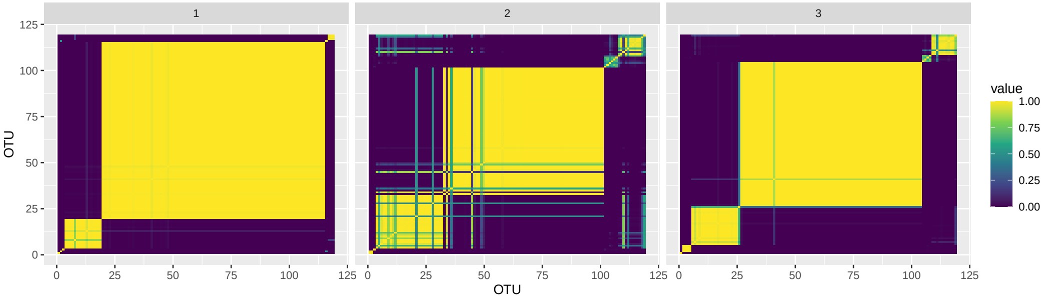


Figure 4.5: OTU co-clustering probability under each subject cluster of subjects. In each block, OTU are ordered by their cluster assignment.

Figure 4.6 shows the same nested partitions, but now by showing the data y_{ij} arranged by subject clusters. In each panel OTUs are sorted by observational clusters. That is, each plot shows the data corresponding to $j \in S_k$, for $k = 1, 2$ and 3. The subjects $j \in C_k$ are on the x-axis. The OTUs are on the y-axis, arranged by the estimated observational clusters. The patterns in y_{ij} echo the clusters shown in the previous plot.

Inference as in Figure 4.5 or 4.6 is not meaningfully possible under partial exchangeability, since the nested partitions $\rho_j = (\tilde{R}_{j\ell}, \ell = 1, \ldots, L)$ are not shared across $j \in C_k$. In other words, consider for example two subjects $j = 1$ and $j = 2$. Assume for subject 1 we record two OTUs a and b co-cluster in a cluster with high frequency, whereas for subject 2 OTUs c and d cluster together. Under partial exchangeability $j = 1$ and 2 might be placed in the same cluster C_k although the different OTUs might be linked to entirely different diets.

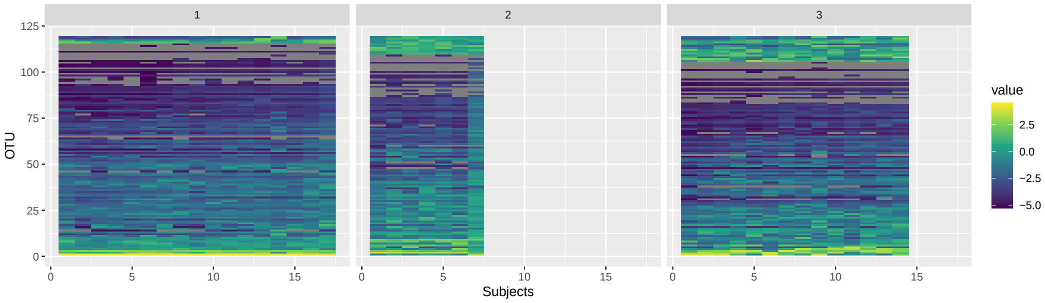


Figure 4.6: Heatmap of column scaled y in log scale for each subject clusters of subjects. OTU is sorted by each cluster specific OTU cluster assignment.

4.5 Nonparametric Regression

4.5.1 Separate exchangeability in an ANOVA DDP model

ANOVA DDP. In this example we set up a separately exchangeable model as an implementation of the popular dependent Dirichlet process (DDP), by means of introducing the symmetric structure in the prior for the atoms in a Dirichlet process (DP) random measure over subjects and time. In this example, the separate exchangeability assumption is not on the observed data (protein expression over time and two different conditions). Instead we set up a separately exchangeable prior for the linear model parameters in a statement of the DDP as a DP mixture of linear models. The actual construction is very simple. We achieve separate exchangeability by setting up additive structure with terms specific to proteins i and time t.

The DDP is a predictor-dependent extension of DP mixtures first proposed by MacEachern (1999, 2000). It defines a family of random probability measures $\mathcal{F} = \{F_x,\ x \in X\}$, where the random distributions F_x are indexed by covariates x and each F_x marginally follows a DP prior as in (4.17). The desired dependence of F_x

across x is achieved by writing the atoms or the weights of the DP random measure F_x as functions of covariates (Quintana et al., 2020). The simplest version of the DDP arises when only the atoms of F_x vary over x (common weights DDP) and are specified as a linear function of x. This defines the linear dependent DP (LDDP) or ANOVA DDP (De Iorio et al., 2004). The model can be written as a DP mixture of linear models, that is, as a mixture with respect to some (or all) linear model parameters, and a DP prior on the mixing measure. See below for a specific example.

The DDP naturally implements partially exchangeable structure if it is assumed as de Finetti measure in (4.4). Consider data y_{ij}, assuming $p(y_{ij} \mid \mathcal{F}) = \int p(y_{ij} \mid \theta)\ dF_j(\theta)$, typically using the additional convolution with a density kernel $p(y_{ij} \mid \theta)$ to construct a continuous sampling model, if desired. A similar situation arises when data y_i is observed with a categorical covariate $w_i \in \{1, \ldots, J\}$ and $p(y_i \mid w_i = j, \mathcal{F}) = \int p(y_{ij} \mid \theta)\ dF_j(\theta)$. In either case, a DDP prior on $\mathcal{F} = (F_j, j = 1, \ldots, J)$ implements a partially exchangeable model for the observable data y_{ij}, or y_i grouped by w_i, respectively. However, if the experimental context calls for separately exchangeable structure with respect to i and j, appropriate model variations are called for. We introduce one next, where the separately exchangeable structure is on the parameters θ in (a variation of) the DP mixture representation.

A separately exchangeable ANOVA DDP. To state the specific model we need some more notation. Recall the notation set up in Section 4.3, with y_{ij} denoting the abundancy of protein i in subject j, and t_j and z_j denoting subject-specific age and condition. We use cubic B-splines with two internal nodes to represent protein expression over a grid of $T = 16$ time points, τ_t, $t = 1, \ldots, T$. The linear model parameters in the ANOVA DDP model include coefficients for the B-spline basis functions to represent protein expression over time for controls, plus an additional equal number of coefficients for the same basis functions to represent an offset for protein expression for patients. More specifically, let $\boldsymbol{x}_j \in \Re^{12}$ denote a design vector for subject j, with $(x_{j1}, \ldots, x_{j6})$ being 6 spline basis functions evaluated at time t_j, and $(x_{j7}, \ldots, x_{j,12}) = z_j(x_{j1}, \ldots, x_{j,6})$ representing an offset for patients ($z_j = 1$). That is, linear model coefficients for $(x_{j1}, \ldots, x_{j6})$ model protein expression over time for controls ($z_j = 0$), while the coefficients for $(x_{j7}, \ldots, x_{j,12})$ represent an additional offset for patients ($z_j = 1$) in protein expression over time.

Let $\boldsymbol{y}_i = (y_{ij},\ j = 1, \ldots, J)$ denote all data for protein i. We assume an ANOVA DDP model, written as DP mixture

$$f(\boldsymbol{y}_i \mid G, \boldsymbol{\alpha}, \boldsymbol{\delta}) = \int \left\{ \prod_{j=1}^{J} N(y_{ij}; \alpha_i + \delta_{t_j} + \boldsymbol{x}_j' \boldsymbol{\beta}_i, \sigma_i^2) \right\} dG(\boldsymbol{\beta}_i, \sigma_i^2)), \qquad (4.20)$$

with a DP prior for $G = \sum_h \pi_h \delta_{\tilde{\beta}_h, \tilde{\sigma}_h}$, $G \sim \mathrm{DP}(\xi, G_0)$. Here α_i are protein-specific offsets and δ_t are offsets for each of the unique time points, $t = 1, \ldots, 16$. We use a finite DP, $G \sim \mathrm{DP}_H(\xi, G_0)$, truncated at $H = 25$ (Ishwaran and James, 2001), with total mass parameter ξ and G_0 defined by $\boldsymbol{\beta}_h \sim N(\boldsymbol{\beta}_0, \sigma_{\beta 0} I)$ and $\sigma_h^2 \sim \mathrm{InvGa}(a_0, b_0)$. We complete the prior specification with $\delta_t \sim N(\zeta, \omega)$ and $\alpha_i \sim N(\mu_0, \sigma_0^2)$. Here, $(\xi, \boldsymbol{\beta}_0, \sigma_{\beta 0}, a_0, b_0, \mu_0, \sigma_0^2, \zeta_0, \omega_0)$ are fixed hyperparameters. See Appendix B for specific values.

Note that the linear model is over-parameterized. For example, we do not restrict the cubic splines to zero average over all (two) subjects observed at the same time $t_j = t$, implying confounding of the intercept with δ_t. However, recall that the inference target is the difference in slope between patient and control for a given protein, which is not affected by this over-parameterization.

The regression model is set up to allow straightforward inference on the protein-specific difference in slope for patients versus controls. As in (4.15), let again $j_{01}, j_{0T}, j_{10}, j_{1T}$ denote indices of subjects with $(z_j, t_j) = (0,1), (0,T), (1,1)$ and $(1,T)$, respectively. Then keeping in mind that $\boldsymbol{x}_{j_{1T}}$ and $\boldsymbol{x}_{j_{0T}}$ differ only in the last 6 elements (and ignoring scaling by a constant $1/(T-1)$),

$$\gamma_i = [(\boldsymbol{x}_{j_{1T}} - \boldsymbol{x}_{j_{11}}) - (\boldsymbol{x}_{j_{0T}} - \boldsymbol{x}_{j_{01}})]\boldsymbol{\beta}_i = (\boldsymbol{x}_{j_{1T},7:12} - \boldsymbol{x}_{j_{11},7:12})\boldsymbol{\beta}_{i,7:12} \qquad (4.21)$$

represents the desired difference in slope between patients and controls. Posterior inference on γ_i implements the desired model-based inference on the difference of slopes with borrowing of strength across proteins and subjects.

Let $\boldsymbol{\theta}_{it} = (\alpha_i, \delta_t, \boldsymbol{\beta}_i, \sigma_i^2)$ denote the parameters that index the regression model for protein i at time t. By construction the prior probability model on $\boldsymbol{\theta}_{it}$ is separately exchangeable, as it is invariant with respect to separate permutations π of protein indices and π' of age:

$$p(\boldsymbol{\theta}_{1:I,1:T}) = p(\boldsymbol{\theta}_{\pi(1:I),\pi'(1:T)}).$$

Note that $\boldsymbol{\beta}_i$ represents the mean function for patient i as coefficients for the spline basis (valid for any t). Only in (4.20) this mean function is evaluated for $t = t_j$ and $x = x_j$. In particular, separate exchangeability is assumed for the mean function parameters, not for the fitted values or the data. The model is separately exchangeable by construction. Alternatively one can trivially match the variables with the arguments in (4.11), $\xi_i = (\alpha_i, \boldsymbol{\beta}_i, \sigma_i)$ and $\eta_t = \delta_t$.

4.5.2 Posterior Inference

MCMC posterior simulation. Posterior simulation under the proposed ANOVA DDP (or DDP of splines) model for protein expression is straightforward. We used the R function bs from the package splines to evaluate the spline basis functions for $\boldsymbol{x}_j$. The transition probabilities are detailed in Algorithm 1 in the appendix.

For the statement of the detail transition probabilities in Algorithm 1 it is useful to replace the mixture model (4.20) by an equivalent hierarchical model

$$y_{ij} \mid s_i = h \sim N(\alpha_i + \delta_{t_j} + \boldsymbol{x}'_j \tilde{\boldsymbol{\beta}}_h, \tilde{\sigma}^2_h) \tag{4.22}$$

with $p(s_i = h) = \pi_h$. Recall that $(\tilde{\boldsymbol{\beta}}_h, \tilde{\sigma}_h)$ are the atoms of the random mixing measure G in (4.20), and that we use a finite DP truncated at H atoms. Interpreting s_i as cluster membership indicators defines inference on a random partition of proteins. Let then $C_h = \{i : \ s_i = h\}$ denote cluster h defined by s_i, and let $n_h = |C_h|$. Note that in this notation we allow for empty clusters that arise when an atom $(\tilde{\boldsymbol{\beta}}_h, \tilde{\sigma}_h)$ is not linked with any observation, i.e., $C_i = \emptyset$ and $n_h = 0$. Using this notation, see Algorithm 1 in Appendix B for a description of MCMC posterior simulation.

We are mainly interested in two posterior summaries, a partition of proteins by different patterns of protein expression over time, and identification of the proteins with the highest rank in $|\gamma_i|$ (the difference in slope between patients and control). The latter proteins are the ones that are most likely linked with ataxia.

We start by identifying the MAP estimate K^* for the number of clusters $K^+ = \sum_h 1(n_h > 0)$ in (4.22), and then follow the approach of Dahl (2006) to report a posterior summary of the random partition of proteins. Let s^* denote the reported partition. Conditional on s^*, we then generate a posterior Monte Carlo sample for $\tilde{\beta}_h$ for each cluster to obtain (conditional) posterior mean and variance for the cluster-specific $\tilde{\boldsymbol{\beta}}_h$. We did not encounter problems related to label-switching in the actual implementation - in general one might need to account for possible label-switching.

Ranking proteins. The main inference target is to identify proteins with the most significant difference between time profiles for patients versus controls, suggesting such proteins are the most likely to be linked with ataxia.

We therefore focus on the difference (across conditions) in differences (over age) of mean protein expression, defined as γ_i in (4.21). We evaluate the posterior mean $E(\gamma_i \mid \boldsymbol{y})$ using Rao-Blackwellization, that is, as Monte Carlo average of conditional expectations $\bar{\gamma}_i = \frac{1}{M} \sum_m E(\gamma_i \mid \theta^{(m)-}, \boldsymbol{y})$, where $\theta^{(m)-}$ are all parameters in the m-th posterior Monte Carlo sample, excluding $\boldsymbol{\beta}$. Note that γ_i is a deterministic function of the cluster-specific $\tilde{\beta}_h$ when $s_i = h$. Let $\tilde{\gamma}_h$ denote the cluster-specific estimate. Conditioning on $s^\star$ we can then identify the set of proteins with the largest average change.

Alternatively, we can cast the problem of identifying interesting proteins as a problem of ranking, and more specifically, one of estimating a certain quantile, to report the most promising $100(1-c)\%$ proteins. Here c is chosen by the investigator.

Choice of c should reflect the effort and capacity to further investigate selected proteins. We then formalize the problem of reporting promising proteins as the problem of identifying the proteins with $|\gamma_i|$ in the top $(1-c)$ percentile of $|\gamma_i|$ values. The problem of ranking experimental units in a hierarchical model and reporting the top $(1-c)$ percentile is discussed in Lin et al. (2004) who cast it as a Bayesian decision problem. Let

$$R_i = \text{rank}(\gamma_i) = \sum_{i'=1}^{I} I(|\gamma_i| \geq |\gamma_{i'}|),$$

denote the true ranks, with $R_i = 1$ for the smallest $|\gamma_i|$ and $R_i = I$ for the largest $|\gamma_i|$. Alternatively we use $P_i = R_i/(I+1)$ to report the quantile, or the percentile $100\,P_i\%$. One of the loss functions that Lin et al. (2004) consider is a 0-1 loss aimed at identifying the top $100(1-c)\%$ units, i.e., the $c-$quantile. Let $\widehat{R_i}$ denote the estimated rank for unit i, and $\widehat{P_i} = \widehat{R_i}/(I+1)$. The following loss function penalizes the number of misclassifications in the top c quantile, including the number of proteins that are falsely reported (false positives) plus those that are failed to be reported (false negatives).

$$L_{0/1}(c) = \frac{1}{I}\{\#\text{ misclassifications}\} = \frac{1}{I}\left\{\sum_{i=1}^{I} \text{AB}(c, P_i, \widehat{P_i}) + \text{BA}(c, P_i, \widehat{P_i})\right\}$$

where AB and BA are penalties for the two types of misclassifications,

$$AB(c, P, \widehat{P}) = \mathbf{1}(P > c, \widehat{P} < c) = \mathbf{1}(R > c(I+1), \widehat{R} < c(I+1),$$

$$BA(c, P, \widehat{P}) = \mathbf{1}(P < c, \widehat{P} > c) = \mathbf{1}(R < c(I+1), \widehat{R} > c(I+1),$$

Lin et al. (2004) show that $L_{0,1}(c)$ is optimized by $\widehat{R}_i = R_i^\star$ with

$$R_i^\star(c) = \text{rank}\{p(P_i > c \mid \boldsymbol{y})\}, \tag{4.23}$$

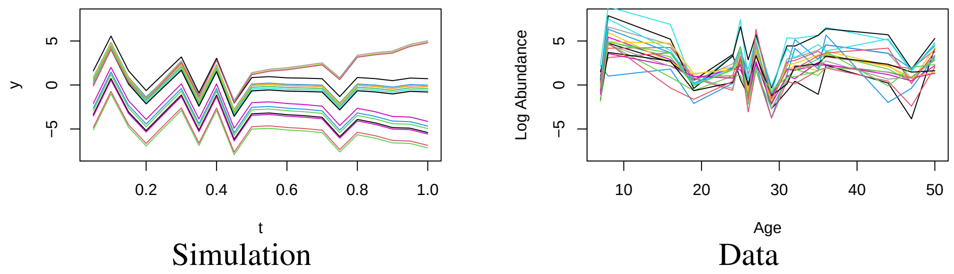
or $P_i^\star = R_i^\star/(I+1)$.

4.5.3 Simulations

We set up a simulation with $J = 20$ subjects and $I = 100$ proteins. The simulation does not include a split into patients and control (think of the data as already reporting the difference of patient and control). We generate $t_j \sim \text{Unif}\,(0,1)$ (equivalent to age in the actual study), a hypothetical partition of proteins with cluster membership indicators $s_i \in \{1,2,3\}$ using $p(s_i = h) = \pi_h$, $\boldsymbol{\pi} = (0.25, 0.3, 0.45)$. We set up protein-specific offsets α_i using shared common values for all proteins in a cluster, i.e., $\alpha_i = \tilde{\alpha}_{s_i}$ with $(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3) = (0, -3, 1)$, and patient-specific offsets $\delta_j \sim N(0, 0.1)$. To mimic the actual data, we manually modify some δ_j to create similar patterns as in the data (see Figure 4.8b). Using these parameters we set up a simulation truth

$$y_{ij} = \begin{cases} \alpha_i + 2t_j + 3t_j^3 + \delta_j + \epsilon_{ij}, & s_i = 1 \\ \alpha_i - 2t_j + t_j^3 + \delta_j + \epsilon_{ij}, & s_i = 2 \\ \alpha_i + t_j - 3t_j^3 + \delta_j + \epsilon_{ij}, & s_i = 3 \end{cases}$$

with $\epsilon_{ij} \sim N(0, \sigma)$ and $\sigma = 0.2, 0.5, 1$ for $s_i = 1, 2, 3$, respectively. See Figure 4.8b (black lines) for the mean functions for proteins in the three clusters under the simulation truth. Under this simulation truth, the proteins in cluster 2 are those with largest overall slope. Figure 4.7 shows 20 randomly selected y_i versus age t, in the simulation (left panel) and in the real data (right panel).



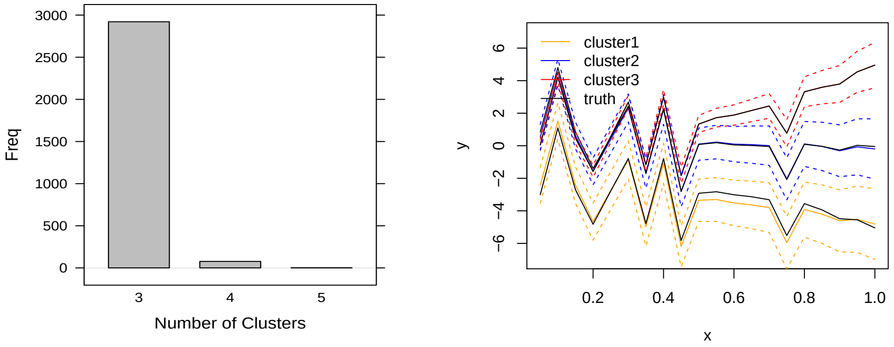
Simulation

Data

Figure 4.7: Left: 20 randomly selected y_i from simulation versus age t, each color indicates a different y_i. Right: the mean log abundancy difference between patient and control in 20 randomly selected proteins from data.

MCMC posterior simulation converges after 3000 iterations. As before, let $n_h = |C_h|$, and let $K^+ = \sum_{h=1}^{H} I(n_h > 0)$ denote the number of non-empty clusters. Figure 4.8a shows a histogram of the posterior distribution $p(K^+ \mid \boldsymbol{y})$. The posterior mode $K^\star = 3$ recovers the simulation truth, with little posterior uncertainty. We then evaluate the earlier described point estimate $s^\star$ for the posterior random partition. Conditioning on $s^\star$ we continue to simulate MCMC transitions to evaluate conditional posterior means for $\tilde{\boldsymbol{\beta}}_h$, α_i and δ_t in the analysis model (4.20).

Let $\bar{\delta}_t = E(\delta_t \mid s^\star, \boldsymbol{y})$, $\bar{\boldsymbol{\beta}}_h = E(\tilde{\boldsymbol{\beta}}_h \mid s^\star, \boldsymbol{y})$ and $\bar{\alpha}_i = E(\alpha_i \mid s^\star, \boldsymbol{y})$ denote the conditional posterior means. All posterior means are evaluated numerically as Monte Carlo sample averages. Also, let $\alpha_h^\star = \frac{1}{n_h} \sum_{i \in C_h} \bar{\alpha}_i$. For a data point y_{ij}, let $t = t_j$ and $h = s_i^\star$. Then $\widehat{y}_{ht} = \alpha_h^\star + \bar{\delta}_t + \boldsymbol{x}_j' \bar{\boldsymbol{\beta}}_h$ are the posterior fitted values for y_{ij}. Figure 4.8(b) shows the posterior fitted values $\widehat{y}_{hzt}$ and pointwise one-standard deviation intervals. The estimates closely track the true mean profiles. These simulation results indicate that with moderate sample sizes as in the actual study, and noise levels comparable to the data, inference could reliably estimate protein expression over



(a) (b)

Figure 4.8: (a) Histogram of $p(K^+ \mid \boldsymbol{y})$ in the simulation experiment. The simulation truth is $K^+ = 3$. (b) Prediction $\widehat{y}_{ht}$ (solid black line) and simulation truth (solid color line) for each cluster in the simulation experiment versus age t. Dotted lines indicate pointwise one standard deviation standard errors.

time. The *a posteriori* identification of the proteins with the largest change over time perfectly recovers the simulation truth in this example. The assignment $s^\star$ includes no misclassification for any of the 100 proteins.

4.5.4 Results

We implement inference under the proposed DDP model with the separately exchangeable prior. We find the following eleven clusters for the 4350 proteins. For each cluster, we separately plot the cluster predicted protein abundancy $\widehat{\boldsymbol{y}}_h = (\widehat{y}_{ht},\ t = 1, \ldots, T)$ for patients (solid) versus control (dashed). For comparison we show corresponding averages of protein expression, averaging the data y_{ij} across all patients (dotted) and controls (dash-dotted) for each estimated cluster (under $s^\star$). The predicted mean protein abundancy for each cluster h is computed as before, as $\widehat{y}_{hzt} = \bar{\delta}_t + \boldsymbol{x}'_{j_{zt}}\bar{\beta}_h + \alpha^\star_h$, for a data point y_{ij} with $s^\star_i = h$, $t_j = t$ and $z_j = z$, and the cluster-specific posterior means $\bar{\beta}_h$ and $\alpha^\star$ defined as before.

In Figure 4.9, we see noticeably different trajectories for patients and controls, except for cluster $k = 3$ (in panel (c)). The figure also indicates R^2 (coefficient of determination) as an indication of model fit for each cluster, with an average R^2 of 0.70 (0.075 empirical standard deviation) across the K^+ clusters. For an additional goodness of fit assessment we use a model fit diagnostic proposed by Johnson (2004). For each iteration, we randomly select a subject j and protein i, calculate the fitted value $\widehat{y}_{ij}^{(m)} = \widehat{y}_{hzt}$ using currently imputed parameters, including the random partition $\boldsymbol{s}$, and form the residual $z_{ij} = y_{ij} - \widehat{y}_{ij}^{(m)}$. Figure 4.10 shows a normal q-q plot of the residuals z_{ij}, dropping an initial burn-in and thinning out iterations. The approximately 45 degree line indicates no evidence against model fit.

Evaluating the posterior rank summary (4.23) we find top 100 (= 2.5%) ranked proteins, with highest $|\gamma_i|$, i.e., (absolute) difference of slopes between patients and controls. Table 4.2 shows the top 20 of these.

Table 4.2: Top 20 Ranked Proteins with highest posterior $|\gamma_i|$

1	Q15149-8	Q13813-2	Q13813-3	Q01082-1	O15020
6	Q00610-1	Q14643-5	Q01484	P49327	P46821
11	P14136	Q8NF91	P35580-4	P14136-3	Q05193-3
16	Q05193-5	P61764-2	P46459	P61764	P07900-2

4.6 Discussion

We have argued for the use of separate exchangeability as a modeling principle, especially for nonparametric Bayesian models. The main arguments are that (i) in many cases separate exchangeability is more faithfully representing the experimental setup than, say, partial exchangeability; and (ii) some summaries of interest (related to the identity of the experimental units) cannot even be stated without introducing separate exchangeability. An example for the latter are shared nested partitions of rows across different columns in a data matrix. In a wider context, the discussion shows that careful considerations of the experimental setup often leads to more specific symmetry assumptions than omnibus exchangeability or partial exchangeability.

In many cases separately exchangeable probability models will serve a tractable submodel of a larger, encompassing inference model as we did in Section 4.5.1.

Finally, recall that the proposed separately exchangeable BNP model in the first example can be rephrased in terms of dependent random partition models that exploit the random partitions induced by the ties of Dirichlet processes. In this context, an interesting future related research is to study the learning mechanisms that arise from similar compositions of Gibbs-type priors (Gnedin and Pitman, 2006; De Blasi et al., 2013) that preserve analytical and computational tractability of the random partition law.

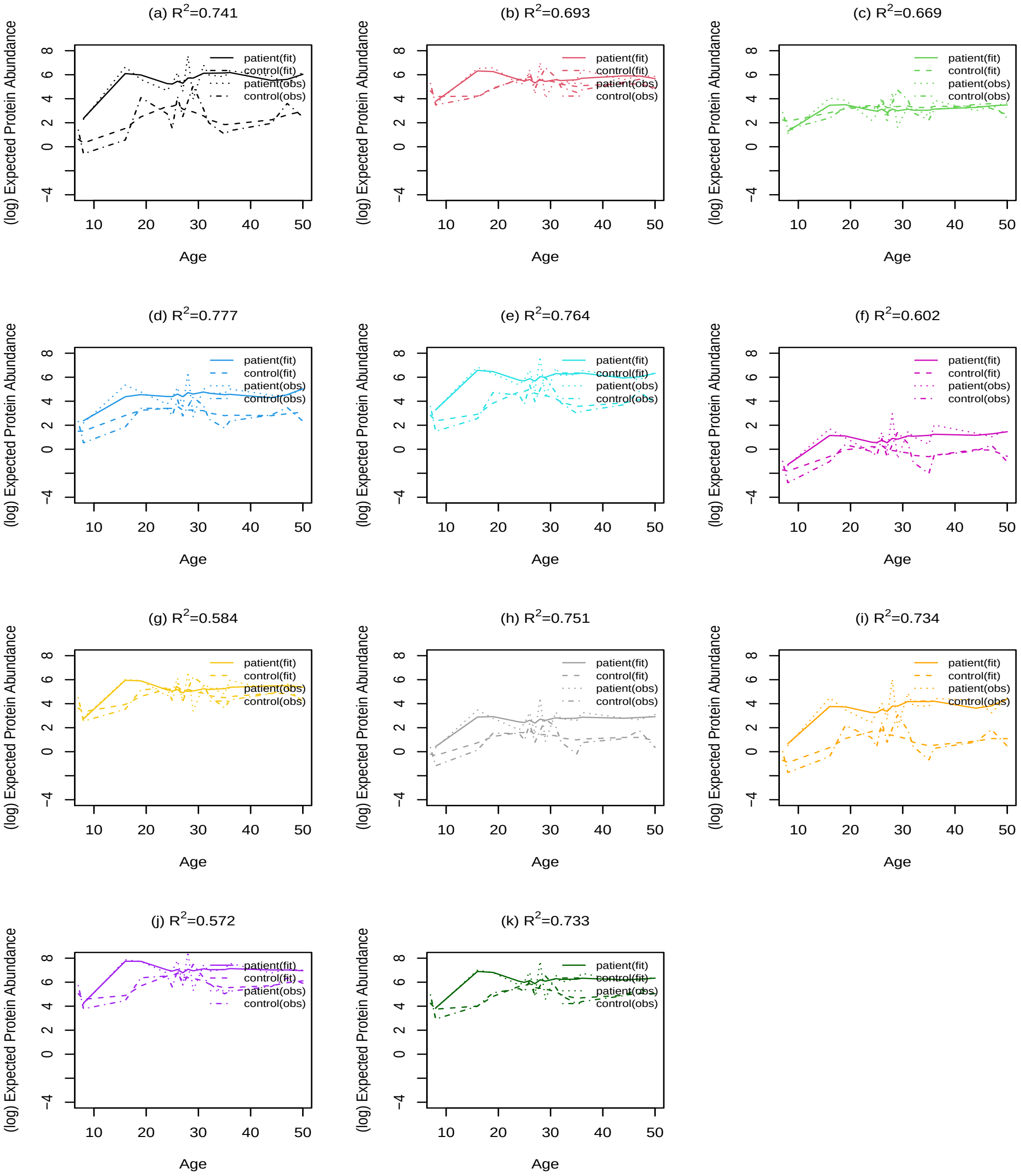


Figure 4.9: Predicted protein abundancy $\widehat{y}_{hzt}$ in cluster h versus age t for patients ($z = 1$, solid) and controls ($z = 0$, dashed) versus age t. For comparison, the dotted and dash-dotted line plots the average over all patients in the same cluster (under $s^{\star}$) under condition z at time t.

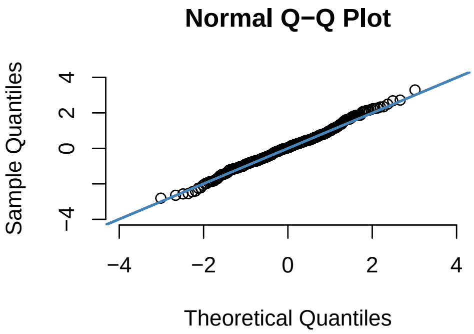


Figure 4.10: Normal Q-Q plot of residuals.

Chapter 5

Conclusion

In this thesis we study inference and uncertainty in complex structures. In the first two chapters after the initial introduction we focus on count statistics in networks under sparse graphon models, which is intrinsically a jointly exchangeable structure. In the fourth chapter, we explore separate exchangeable structures with its examples and applications in Bayesian nonparametrics.

In the first chapter, we propose a network jackknife procedure, a leave-one-node out method to estimate the variance of count statistics in networks. We prove that under the sparse graphon model, the network jackknife estimate of variance is always conservative in expectation and is consistent for count statistics and smooth functions of count statistics. As we showed in simulations, our network jackknife algorithm outperforms traditional subsampling methods in variance estimation and in computation speed. We have tentative evidence that network jackknife has a potential to be applied beyond count statistics such as eigenvalues and achieve consistent estimation. Future work and attention is needed for this exciting extension.

In the second chapter, we propose a family of network multiplier bootstrap methods for inference of count statistics. Our linear and quadratic multiplier bootstraps are bootstrap procedures with linear and quadratic weights, which respectively correspond to the first and second-order terms of the Hoeffding decomposition. Between linear and quadratic bootstraps, we see a trade-off of speed and accuracy. For moderately sparse graphs, quadratic bootstrap is higher-order correct while linear bootstrap is first-order correct but faster in computation. To make linear bootstrap even faster, we also propose an approximate linear bootstrap where a randomized sketching algorithms is used for estimating local network count statistics. We present empirical results in both simulations and real data networks. Future work includes applications on larger and sparser social networks and biological networks.

In the third chapter, we review the definition of separate exchangeability and discuss separate exchangeability as a modeling principle in Bayesian nonparametric models. We provide two examples. In the first example related to microbiome analysis, the data from the experimental setup is separately exchangeable. In the second example related to a protein study, separate exchangeability is introduced in the prior probability model for the parameters. For the microbiome data in the first example, we use a model of nested partitions. The proposed model is a variant of the common atoms model to respect the separate exchangeability structure in the experimental units. For the protein study in the second example, separate exchangeability in the prior probability model for the parameters in a nonparametric regression is introduced by appropriate choices in the model construction. In both examples, we argue for the use of separate exchangeability as a modeling principle that is often overlooked in the modeling process of Bayesian nonparametrics. As separately exchangeable probability models can be used as tractable submodels in a larger inference model, an interesting future study is to further explore these applications.

The methods that we propose for inference and for characterizing uncertainty in complex structures such as networks or separately exchangeable structures can be used in many applications using social science data or biological data. The problems that we discussed are only some specific examples of issues that arise with inference for complex structures. These studies are just a beginning, as complicated data arising from social and biological sciences call for the development of more theory and methods for inference and uncertainty estimation methods.

Appendices

Appendix A

Theoretical Proofs and Additional Experiments for Network Jackknife

A.1 Proof of Theorem 1

To facilitate the proof below, we will explicitly define the data generating mechanism for the Bernoulli trials defined in Eq 1.1. For $1 \leq i < j \leq n$, define the random variable $\eta_{ij} \sim \text{Unif}[0, 1]$ and let $A_{ij} = \mathbb{1}(\eta_{ij} \leq \rho_n w(X_i, X_j) \wedge 1)$. We may view a function f that takes as input a $n - 1 \times n - 1$ adjacency matrix as a function g of the underlying latent positions. We require that g is invariant to node-permutation, meaning that g remains unchanged when some (bijective) permutation function $\varphi : \{1, 2, \dots, n-1\} \mapsto \{1, 2, \dots, n-1\}$ is applied to the indices corresponding to X_i and both the row and column indices of η_{ij} separately.

In what follows, let $\boldsymbol{X}_n = (X_i)_{1 \leq i \leq n}$ and $\boldsymbol{\eta}_n = (\eta_{ij})_{1 \leq i < j \leq n}$. Furthermore, we will let $\boldsymbol{X}_{n,i}$ denote the vector formed by removing node i and $\boldsymbol{\eta}_{n,i}$ denote the (concatenated) vector formed by removing all elements containing row or column index i.

Proof. Let $Z_{n,i} = g(\boldsymbol{X}_{n,i}, \boldsymbol{\eta}_{n,i})$ denote the functional calculated on an induced subgraph of $n - 1$ nodes excluding node i. As before, let $Z_{n-1} = Z_{n,n}$. Construct the following martingale difference sequence:

$$d_i = E(Z_{n-1}|\Sigma_i) - E(Z_{n-1}|\Sigma_{i-1}) \tag{A.1}$$

Here, we consider a filtration introduced by Borgs et al. (2008), which was originally used to establish exponential concentration for certain subgraph frequencies in the dense regime.

Let $\Sigma_0 = \{\emptyset, \Omega\}$, $\Sigma_1 = \sigma(X_1)$, $\Sigma_2 = \sigma(X_1, X_2, \eta_{12})$, $\Sigma_3 = \sigma(X_1, X_2, X_3, \eta_{12}, \eta_{13}, \eta_{23})$ and so forth up to n. The filtration we consider has the following interpretation: for each time $1 \leq t \leq n$, suppose that we observe a $t \times t$ adjacency matrix induced by the nodes $\{1, 2, \dots, t\}$. Then, Σ_t captures all of the randomness in the corresponding induced subgraph. We may visualize Σ_i as a σ-field generated by a triangular array so that:

$$\Sigma_i = \sigma \left\{ \begin{array}{ccccc} X_1 & \eta_{12} & \cdots & \eta_{1,i-1} & \eta_{1i} \\ & X_2 & \cdots & \eta_{2,i-1} & \eta_{2i} \\ & & \cdots & .. & \\ & & X_{i-2} & \eta_{i-2,i-1}, & \eta_{i-2,i} \\ & & & X_{i-1} & \eta_{i-1,i} \\ & & & & X_i \end{array} \right\}; \ \Sigma_{i-1} = \sigma \left\{ \begin{array}{cccc} X_1 & \eta_{12} & \cdots & \eta_{1,i-1} \\ & X_2 & \cdots & \eta_{2,i-1} \\ & & \cdots & .. \\ & & X_{i-2} & \eta_{i-2,i-1} \\ & & & X_{i-1} \end{array} \right\}$$

Observe that $Z_{n-1} - E(Z_{n-1}) = \sum_{i=1}^n d_i$, d_i is Σ_i measurable, and $E(d_i|\Sigma_{i-1}) = 0$. Therefore, the variance of Z_n can be written as:

$$\text{var } Z_{n-1} = E\left(\sum_{i=1}^n d_i\right)^2 = \sum_{i=1}^n E(d_i^2) + 2\sum_{i<j} E(d_i d_j)$$

Now, for $i \neq j$, observe that:

$$\begin{aligned} E(d_i d_j) &= E(E(d_i d_j|\Sigma_i)) = E(d_i)E(d_j|\Sigma_i) \\ &= E(d_i)(E[E(S_n|\Sigma_j)|\Sigma_i] - E[E(S_n|\Sigma_{j-1})|\Sigma_i]) = 0 \end{aligned}$$

For the jackknife estimate, we have that:

$$E\left(\sum_{i=1}^{n}(Z_{n,i}-\bar{Z}_n)^2\right)=\sum_{i<j}\frac{E(Z_{n,i}-Z_{n,j})^2}{n}=\frac{(n-1)\cdot E(Z_{n,1}-Z_{n,2})^2}{2}$$

We also denote by $\Sigma_{i:j}$, the sigma field containing all information of random variables $X_i,\ldots,X_j$, and $\eta_{k\ell}, i\leq k<\ell\leq j$. Now define $\mathcal{A}$ as $\Sigma_{3:i+1}$. Since Z_{n-1} is invariant to node-permutation, $\mathcal{A}$ is independent of $\sigma(X_2,\eta_{23},\ldots,\eta_{2n})$ and $\sigma(X_1,\eta_{13},\ldots,\eta_{1n})$,

$$E(Z_{n,1}|\mathcal{A})=E(Z_{n,2}|\mathcal{A})$$

Define:

$$U=E(Z_{n,1}|\Sigma_{i+1})-E(Z_{n,1}|\mathcal{A}),\quad V=E(Z_{n,2}|\Sigma_{i+1})-E(Z_{n,2}|\mathcal{A}) \tag{A.2}$$

Then, using the fact that $E[X^2|\Sigma_{i+1}]\geq E[X|\Sigma_{i+1}]^2$ for some Σ_{i+1} measurable r.v. X, we have:

$$E(Z_{n,1}-Z_{n,2})^2\geq E[E(Z_{n,1}|\Sigma_{i+1})-E(Z_{n,2}|\Sigma_{i+1})]^2=E(U-V)^2 \tag{A.3}$$

Notice that conditional on $\mathcal{A}$, U is a function of $\{X_2,\eta_{23},\ldots,\eta_{2,i+1}\}$, while V is a function of $\{X_1,\eta_{13},\ldots,\eta_{1,i+1}\}$. Thus, U and V are conditionally independent. Then, since $\mathcal{A}\subset\Sigma_{i+1}$, by the tower property of conditional expectations, we have that:

$$\begin{aligned}E(U-V)^2&=E(U^2)-2E(UV)+E(V^2)\\&=E(U^2)+E(V^2)-2E(E(U|\mathcal{A})E(V|\mathcal{A}))\end{aligned}$$

$$= E(U^2) + E(V^2),$$

Now, we expand $E(U^2)$ as follows:

$$\begin{aligned}
E(U^2) &= E((E(Z_{n,1}|\Sigma_{(i+1)}) - E(Z_{n,1}|\mathcal{A}))^2] \\
&\overset{(i)}{=} E((E(Z_{n,1}|\Sigma_{2:i+1}) - E(Z_{n,1}|\Sigma_{3:i+1}))^2] \\
&\overset{(ii)}{=} E[(E(Z_{n,n}|\Sigma_{1:i}) - E(Z_{n,n}|\Sigma_{1:i-1}))^2] \\
&= E[(E(Z_{n-1}|\Sigma_i) - E(Z_{n-1}|\Sigma_{i-1}))^2] = E(d_i^2)
\end{aligned}$$

Step (i) holds because the random variables associated with node 1 are not present in $Z_{n,1}$. Step (ii) holds because $X_1, \dots X_n$ and η_{ij}, $1 \le i < j \le n$ are i.i.d random variables, and $E[Z_{n,1}|\Sigma_{2:i+1}]$ ($E[Z_{n,1}|\Sigma_{3:i+1}]$) and $E[Z_{n,n}|\Sigma_{1:i}]$ ($E[Z_{n,n}|\Sigma_{1:i-1}]$) are equal in distribution.

Similarly, $EV^2 = Ed_i^2$, $E(U-V)^2 = 2Ed_i^2$. Thus,

$$E(Z_{n,1} - Z_{n,2})^2 \ge E(U-V)^2 = 2Ed_i^2 \tag{A.4}$$

$$E\left(\sum_{i=1}^{n}(Z_{n,i} - \bar{Z}_n)^2\right) = \frac{n-1}{2}E(Z_{n,1} - Z_{n,2})^2 \ge (n-1)Ed_i^2 = \text{var } Z_{n-1} \tag{A.5}$$

□

A.2 Proof of Theorem 2

For notational convenience, let $Z_n = \hat{P}(R)$ and let $Z_{n,i}$ denote the subgraph frequency defined in Eq 1.5 with node i removed:

$$Z_{n,i} = \rho_n^{-s}\frac{1}{\binom{n-1}{r}\,|\text{Iso}(R)|}\sum_{S\sim R,\ i\notin V(S)} \mathbb{1}(S = G_n[S]) \tag{A.6}$$

We first present a lemma that will be used in the proof. An identity relating the mean of leave-one-out jackknife estimates to a U-statistic plays an important role in the proof of jackknife consistency for U-statistics. Using a novel combinatorial argument, we show that a similar identity holds for normalized subgraph counts:

Lemma A.2.1. *Letting $Z_{n,i}$ and Z_n be defined as above, we have that:*

$$\bar{Z}_n := \frac{1}{n}\sum_{i=1}^{n} Z_{n,i} = Z_n$$

Proof. For a subgraph with r nodes and s edges, denote the number of this subgraph in G_n as Q. Denote the number of subgraphs node i is involved in as Q_i. We now analyze $\sum_{i=1}^{n} Q_i$. For each vertex set with cardinality r, a given subgraph is counted once from each vertex. Therefore, $\sum_{i=1}^{n} Q_i = rQ$.

Observe that $Z_{n,i} + Q_i = Q$ since the set of subgraphs that do not contain node i and the set of subgraphs that contain node i are disjoint and their union gives the set of subgraphs counted in Q. It follows that:

$$\frac{1}{n}\sum_i Z_{n,i} = \frac{\frac{1}{n}\sum_i (Q - Q_i)}{\binom{n-1}{r}\rho_n^s} = \frac{(n-r)Q}{n\binom{n-1}{r}\rho_n^s} = \frac{Q}{\binom{n}{r}\rho_n^s} = Z_n.$$

□

Now, we introduce the limiting value of the scaled variance, which represents the value we are aiming for with the jackknife. Bickel et al. (2011) show that the asymptotic behavior of $\hat{P}(R)$ is driven by a U-statistic corresponding to the edge structure of the subgraph. For a subgraph R with $V(R) = \{1, \ldots, p\}$, define the kernel:

$$h(X_1, \ldots, X_r) = \frac{1}{|Iso(R)|} \sum_{S \sim R,\ V(S)=\{1,\ldots,r\}} \prod_{(i,j) \in E(S)} w(X_i, X_j) \qquad \text{(A.7)}$$

Theorem 1 of Bickel et al. (2011) establishes that:

$$n \cdot \text{var}\, \hat{P}(R) \to \sigma^2$$

where $\sigma^2 = r^2\zeta$ is the variance of the U-statistic with kernel h, with $\zeta = \text{Var}(E(h(X_1, \ldots, X_r)|X_1))$. We will now scale the jackknife variance by n to study its asymptotics. Let:

$$\alpha_i = Z_{n,i} - E(Z_{n,i}|\boldsymbol{X}_n), \quad \beta_i = E(Z_{n,i}|\boldsymbol{X}_n) \qquad \text{(A.8)}$$

For simplicity we will use $\bar{\alpha}_n$ (or $\bar{\beta}_n$) to denote the average of α_i (or β_i). Now, consider the following signal-noise decomposition:

$$\begin{aligned} n \cdot \sum_{i=1}^{n} (Z_{n,i} - \bar{Z}_n)^2 &= n \cdot \sum_{i=1}^{n} (\alpha_i - \bar{\alpha}_n + \beta_i - \bar{\beta}_n)^2 \\ &= n \cdot \sum_{i=1}^{n} (\alpha_i - \bar{\alpha}_n)^2 + 2n \cdot \sum_{i=1}^{n} (\alpha_i - \bar{\alpha}_n)(\beta_i - \bar{\beta}_n) \\ &+ n \cdot \sum_{i=1}^{n} (\beta_i - \bar{\beta}_n)^2. \end{aligned} \qquad \text{(A.9)}$$

We start by bounding the third sum, which is the signal in our decomposition. Observe that β_i is a U-statistic with the kernel h defined in (A.7); therefore, by Theorem 1 and its following discussions of Chapter 5 in Lee (1990), we have that:

$$n \cdot \sum_{i=1}^{n} (\beta_i - \bar{\beta}_n)^2 \xrightarrow{P} \sigma^2 \qquad \text{(A.10)}$$

The result will follow if we show that the remaining two sums in the decomposition are negligible. If the first sum is negligible, the Cauchy-Schwarz inequality would imply that:

$$n \cdot \sum_{i=1}^{n}(\alpha_i - \bar{\alpha}_n)(\beta_i - \bar{\beta}_n) \leq n \cdot \sqrt{\sum_{i=1}^{n}(\alpha_i - \bar{\alpha}_n)^2 \cdot \sum_{i=1}^{n}(\beta_i - \bar{\beta}_n)^2} \xrightarrow{P} 0$$

It remains to show that: $n \cdot \sum_{i=1}^{n}(\alpha_i - \bar{\alpha}_n)^2 \xrightarrow{P} 0$. Now, observe that:

$$\sum_{i=1}^{n}(\alpha_i - \bar{\alpha}_n)^2 = \sum_{i=1}^{n} \alpha_i^2 - n\bar{\alpha}_n^2$$

Expanding the square for $\sum_{i=1}^{n} \alpha_i^2$ we have that:

$$\begin{aligned}\sum_{i=1}^{n} \alpha_i^2 &= \sum_{i=1}^{n}(Z_{n,i} - E(Z_{n,i}|\boldsymbol{X}_n))^2 \\ &= \sum_{i=1}^{n} \binom{n-1}{r}^{-2} \sum_{S\sim R,\ i\notin V(S)} (\rho_n^{-s}\psi(S) - W(S)) \sum_{T\sim R,\ i\notin V(T)} (\rho_n^{-s}\psi(T) - W(T))\end{aligned}$$

where $\psi(S)$ and $W(S)$ are given by:

$$\begin{aligned}\psi(S) &= \frac{1}{|Iso(R)|} \prod_{(i,j)\in E(S),\ S\sim R} A_{ij} \times \prod_{(i,j)\in \overline{E(S)},\ S\sim R} 1 - A_{ij}, \\ W(S) &= \frac{1}{|Iso(R)|} \prod_{(i,j)\in E(S),\ S\sim R} w(X_i, X_j) \times \prod_{(i,j)\in \overline{E(S)},\ S\sim R} 1 - \rho_n w(X_i, X_j)\end{aligned}$$

and $\overline{E(S)}$ are $(i, j) \in V(S) \times V(S)$ that are not contained in $E(S)$. Now, similar to Lee (1990), we group elements in the sum based on the number of elements in $V(S) \cap V(T)$. For each $|V(S) \cap V(T)| = c$, there are $n - 2r + c$ terms in total. It follows that:

$$\sum_{i=1}^{n} \alpha_i^2 = \binom{n-1}{r}^{-2} \sum_{c=0}^{r}(n - 2r + c) \sum_{|V(S)\cap V(T)|=c} (\rho_n^{-s}\psi(S) - W(S))(\rho_n^{-s}\psi(T) - W(T))$$

$$= \binom{n-1}{r}^{-2} \sum_{c=0}^{r} (n-2r+c) \sum_{|V(S)\cap V(T)|=c} \gamma(S,T), \quad \text{say.}$$

Now we turn to $n\bar{\alpha}_n^2$;

$$\bar{\alpha}_n = \frac{1}{n}\sum_i Z_{n,i} - \frac{1}{n}\sum_i E(Z_{n,i}|\boldsymbol{X}_n) \overset{(i)}{=} Z_n - E(Z_n|\boldsymbol{X}_n)$$

Equality (i) follows from Lemma A.2.1. Now expanding $\bar{\alpha}_n^2$ in a similar manner, we have that

$$\bar{\alpha}_n^2 = \frac{(n-r)^2}{n}\binom{n-1}{r}^{-2} \sum_{c=0}^{r} \sum_{|V(S)\cap V(T)|=c} \gamma(S,T),$$

Then,

$$\begin{aligned} n \cdot \sum_{i=1}^{n} (\alpha_i - \bar{\alpha}_n)^2 &= \binom{n-1}{r}^{-2} \sum_{c=0}^{r} \left(n - 2r + c - \frac{(n-r)^2}{n} \right) \sum_{|V(S)\cap V(T)|=c} \gamma(S,T) \\ &= \sum_{c=0}^{r} \sum_{|V(S)\cap V(T)|=c} \left(c - \frac{r^2}{n} \right) \cdot \binom{n-1}{2}^{-2} \gamma(S,T) \end{aligned}$$

Now, taking expectations, we have that:

$$\begin{aligned} &E\left(\binom{n-1}{r}^{-2} \sum_{c=0}^{r} \sum_{|V(S)\cap V(T)|=c} \gamma(S,T) \right) \\ &= E\left(\binom{n-1}{r}^{-2} \sum_{c=0}^{r} \sum_{|V(S)\cap V(T)|=c} \left(\rho_n^{-s}\psi(S) - W(S) \right)\left(\rho_n^{-s}\psi(T) - W(T) \right) \right) \\ &= E\left[\sum \mathrm{Cov}(S,T|\boldsymbol{X}_n) \right] = o\left(\frac{1}{n}\right) \end{aligned}$$

where the last line follows from the proof of Theorem 1 of Bickel et al. (2011).

Now, by Markov inequality, we have that

$$n \cdot \sum_{i=1}^{n} (\alpha_i - \bar{\alpha}_n)^2 \xrightarrow{P} 0 \tag{A.11}$$

and the result follows.

A.3 Proof of Theorem 3

Proof. Let $Z_{n,i} = (Z_{n,i}(1), \ldots Z_{n,i}(d))$, where d is a constant w.r.t n and each entry corresponds to a count functional with node i removed. Each count functional may involve subgraphs of different sizes. We will use a Taylor expansion around $\bar{Z}_n$.

$$\begin{aligned} f(Z_{n,i}) &= f\big(\bar{Z}_n) + \nabla f(\zeta_i)^T (Z_{n,i} - \bar{Z}_n) \\ &= f\big(\bar{Z}_n) + \nabla f(\mu)^T (Z_{n,i} - \bar{Z}_n) + \underbrace{(\nabla f(\zeta_i) - \nabla f(\mu))^T (Z_{n,i} - \bar{Z}_n)}_{E_i}, \end{aligned}$$

where $\zeta_i = (\zeta_{i1}, \ldots, \zeta_{id}) = c_i Z_{n,i} + (1 - c_i)\bar{Z}_n$ for some $c \in [0, 1]$. Thus, we also have:

$$f(Z_{n,i}) - \overline{f(Z_{n,i})} = \underbrace{\nabla f(\mu)^T (Z_{n,i} - \bar{Z}_n)}_{I_i} + \underbrace{E_i - \frac{1}{n}\sum_i E_i}_{II_i} \tag{A.12}$$

For the first part we see that,

$$n\sum_i (I_i)^2 = n\nabla f(\mu)^T \left(\sum_i (Z_{n,i} - \bar{Z}_n)(Z_{n,i} - \bar{Z}_n)^T\right) \nabla f(\mu) \tag{A.13}$$

We will first show that the inner average of the above expression converges to the covariance matrix of $Z_{n,i}$ (recall that here we are considering a finite dimensional vector). Extending the same argument in Eq A.9 to finite dimensional $Z_{n,i}$'s (and α_i and β_i's defined in Eq A.8),

$$\begin{aligned} & n\sum_i (Z_{n,i} - \bar{Z}_n)(Z_{n,i} - \bar{Z}_n)^T \\ &= n\sum_i \Big((\alpha_i - \bar{\alpha}_n)(\alpha_i - \bar{\alpha}_n)^T + (\alpha_i - \bar{\alpha}_n)(\beta_i - \bar{\beta}_n)^T + (\beta_i - \bar{\beta}_n)(\alpha_i - \bar{\alpha}_n)^T \\ &\qquad + (\beta_i - \bar{\beta}_n)(\beta_i - \bar{\beta}_n)^T\Big) \end{aligned}$$

By Theorem 9 of Arvesen (1969) we have that:

$$n\sum_{i}(\beta_i-\bar{\beta}_n)(\beta_i-\bar{\beta}_n)^T \xrightarrow{P} \Sigma \tag{A.14}$$

Above, Σ is the covariance matrix of a multivariate U-statistic with kernels $(h_1,\ldots,h_d)$, where each h_j is the kernel corresponding to the count functional in the j^{th} coordinate of the vector Z_n (see Eq A.7). Now combining Eq A.14 with Eq A.13 we see that,

$$\begin{aligned}\left|n\sum_{i}(I_i)^2-f(\mu)^T\Sigma f(\mu)\right| &\leq \|\nabla f(\mu)\|^2 n\sum_{i}\|\alpha_i-\bar{\alpha}_n\|^2\\ &\quad+2n\|\nabla f(\mu)\|^2\sum_{i}|(\alpha_i-\bar{\alpha}_n)^T(\beta_i-\bar{\beta}_n)| \end{aligned}\tag{A.15}$$

The first part is $o_p(1)$ by an analogous argument leading to Eq A.11. For the second part, we see that an application of Cauchy Schwarz inequality gives:

$$\begin{aligned}&n\sum_{i}|(\alpha_i-\bar{\alpha}_n)^T(\beta_i-\bar{\beta}_n)|\\ &\leq\sum_{j=1}^{d}\sqrt{\left(\sum_{i}n(\alpha_i(j)-\bar{\alpha}_n(j))^2\right)\left(n\sum_{i}(\beta_i(j)-\bar{\beta}_n(j))^2\right)}\end{aligned}$$

The first part inside the square root is $o_p(1)$ due to Eq A.11, and the second part is $O_p(1)$ by Eq A.10. Using this in conjunction with Eq A.15 and since $\|\nabla f(\mu)\|$ is bounded, we see that:

$$\left|n\sum_{i}(I_i)^2-\nabla f(\mu)^T\Sigma\nabla f(\mu)\right|=o_p(1)$$

All that remains now is to show that part II_i in Eq A.12 is negligible even when summed and multiplied by n. First note that $(II_i)^2\leq E_i^2$.

$$n\sum_{i}(II_i)^2\leq n\sum_{i}|(\nabla f(\zeta_i)-\nabla f(\mu))^T(Z_{n,i}-\bar{Z}_n)|^2$$

$$\leq \max_i \|\nabla f(\zeta_i) - \nabla f(\mu)\|^2 \left(n \sum_i (Z_{n,i} - \bar{Z}_n)^T (Z_{n,i} - \bar{Z}_n) \right) \quad \text{(A.16)}$$

Theorem 2 shows that the second part in the RHS of Eq A.16 is $O_p(1)$. We will now show that the first part is asymptotically negligible.

Observe that:

$$\begin{aligned} \max_i \|\zeta_i - \mu\| &\leq \max_i \; c_i \|Z_{n,i} - \mu\| + \max_i \; (1 - c_i) \|\bar{Z}_n - \mu\| \\ &\leq \sqrt{d} \cdot \max_{i,j} |Z_{n,i}(j) - \bar{Z}_n(j)| + 2\|\bar{Z}_n - \mu\| \\ &\leq \sqrt{d} \cdot \max_j \sqrt{\sum_{i=1}^{n} \left(Z_{n,i}(j) - \bar{Z}_n(j)\right)^2} + 2\|Z_n - \mu\| \end{aligned}$$

Above, $\bar{Z}_n = Z_n$ by Lemma A.2.1. The first term on the RHS converges in probability to 0 from our Theorem 2. By Theorem 1 of Bickel et al. (2011), $\|Z_n - \mu\|$ is also negligible. Since $\max_i \|\zeta_i - \mu\| = o_p(1)$ and ∇f is continuous at μ, by continuity, we have that $\max_i \|\nabla f(\zeta_i) - \nabla f(\mu)\|^2 = o_p(1)$. Since the second term on the RHS of Eq A.16 is $O_p(1)$ from our previous argument and the first term is $o_p(1)$, it follows that the LHS of Eq A.16 is $o_p(1)$.

Let $\mu_n = E[Z_n]$. Note that if one counts subgraphs by an exact match as in Bickel et al. (2011) $\mu_n \longrightarrow \mu$. If one counts subgraphs via edge matching, $\mu_n = \mu$. Thus, both these types of subgraph densities, which asymptotically have the same limit, can be handled by our theoretical results. By Theorem 3.8 in Van der Vaart (2000),

$$\sqrt{n}(f(Z_n) - f(\mu_n)) \rightsquigarrow N(0, \nabla f(\mu)^T \Sigma \nabla f(\mu))$$

This shows that the jackknife estimate of variance converges to the asymptotic variance of $f(Z_n)$.

□

A.4 Proof of Proposition 1

Throughout this section, we will use the notation $x_n \asymp y_n$ to denote $x_n = y_n(1+o(1))$. Before presenting the proof, we present two accompanying lemmas which will be used in the proof of Proposition 1.

Lemma A.4.1. *Denote $D_i^{(n)}$ the degree of node i in the size n graph.*

$$\sum_{i=1}^{n-1} var\left(\frac{D_i^{(n)}}{\binom{n-1}{2}\rho_n}\right) \asymp \frac{4}{n^3}E(var\sum_{k,k\neq i} w(X_i, X_k)|X_i) \\ + \frac{4}{n}var[E(w(X_i, X_k)|X_i)] + O(n^{-2}\rho_n^{-1}).$$

Lemma A.4.2. *Denote $D_i^{(n)}$ the degree of node i in the size n graph.*

$$\sum_{i,j,i\neq j} \mathrm{cov}\left(\frac{D_i^{(n)}}{\binom{n-1}{2}\rho_n}, \frac{D_j^{(n)}}{\binom{n-1}{2}\rho_n}\right) \asymp \frac{4}{n} \times 3var(E[w(X_i, X_j)|X_i]) + O(n^{-2}\rho_n^{-1})$$

We will use the above to lemmas to prove Proposition 1, which we now present.

Proof. Denote D_n as the total number of edges in graph G_n. By definition,

$$Z_n = \frac{D_n}{\binom{n}{2}\rho_n}$$

Denote $D_i^{(n)}$ the degree of node i in the size n graph. We have that $ED_i^{(n)} = ED_j^{(n)}$ for any node pair. Thus the jackknife estimate of edges for a graph with node i removed is D_n minus the degree of node i. Define

$$\gamma_n = \binom{n-1}{2}\rho_n; \quad \gamma'_n = \binom{n-1}{2}\rho_{n-1} \tag{A.17}$$

Then by definition, we have

$$Z_{n,i} = \frac{D_n - D_i^{(n)}}{\binom{n-1}{2}\rho_n} = \frac{D_n - D_i^{(n)}}{\gamma_n}$$

Then, the jackknife estimate is

$$\begin{aligned} E\sum_{i=1}^{n}(Z_{n,i} - \bar{Z}_n)^2 &= \frac{1}{2n}\sum_{i\neq j} E(Z_{n,i} - Z_{n,j})^2 = \frac{1}{2n}\sum_{i\neq j} E\left(\frac{D_i^{(n)} - D_j^{(n)}}{\gamma_n}\right)^2 \\ &= \sum_{i=1}^{n-1}\text{var}\left(\frac{D_i^{(n)}}{\gamma_n}\right) - \frac{1}{n}\sum_{i\neq j}\text{cov}\left(\frac{D_i^{(n)}}{\gamma_n}, \frac{D_j^{(n)}}{\gamma_n}\right) \end{aligned} \tag{A.18}$$

whereas the total number of degrees in a $(n-1)$ graph is $D_{n-1} = \sum_{i=1}^{n-1} D_i^{(n-1)}/2$ as each edge is counted 2 times from each node. We first obtain an expression for var Z_{n-1}.

$$\text{var}\, Z_{n-1} = \text{var}\left(\frac{\sum_{i=1}^{n-1} D_i^{(n-1)}/2}{\binom{n-1}{2}\rho_{n-1}}\right) = \frac{1}{4}(n-1)\text{var}\left(\frac{D_i^{(n-1)}}{\gamma'_n}\right) \tag{A.19}$$

$$+\frac{1}{4}\sum_{i,j,i\neq j}\text{cov}\left(\frac{D_i^{(n-1)}}{\gamma'_n}, \frac{D_j^{(n-1)}}{\gamma'_n}\right) \tag{A.20}$$

For the second term in the R.H.S of Eq A.18, from Lemma A.4.2, it is easy to check that it is $O(n^{-2})$. Thus scaling Eq A.18 by $n-1$ we have,

$$\begin{aligned}&(n-1)E\sum_{i=1}^{n}(Z_{n,i}-\bar{Z}_n)^2 = (n-1)\sum_{i=1}^{n}\text{var}\left(\frac{D_i^{(n)}}{\gamma_n}\right)+O\left(\frac{1}{n}\right)\\ &= \frac{4}{n^2}E[\text{var}\sum_{k,k\neq i} w(X_i,X_k)|X_i] + 4\text{var}[E(w(X_i,X_k)|X_i)] + O\left(\frac{1}{n\rho_n}\right)+O\left(\frac{1}{n}\right)\end{aligned} \tag{A.21}$$

Plugging in Lemma A.4.2 into the second term of R.H.S of Eq A.19 and scaling Eq A.19 by $n-1$, we have

$$\begin{aligned}&(n-1)\text{var } Z_{n-1}\\ &= \frac{1}{n^2}E[\text{var}\sum_{k,k\neq i} w(X_i,X_k)|X_i] + \text{var}[E(w(X_i,X_k)|X_i)]\\ &+ 3\text{var}[E(w(X_i,X_k)|X_i)] + O\left(\frac{1}{n\rho_n}\right)\\ &= \frac{1}{n^2}E[\text{var}\sum_{k\neq i} w(X_i,X_k)|X_i] + 4\text{var}[E(w(X_i,X_k)|X_i)] + O\left(\frac{1}{n\rho_n}\right)\end{aligned} \tag{A.22}$$

The difference between Eqs A.21 and A.22 is:

$$(n-1)E(Z_{n,i}-\bar{Z}_n)^2-(n-1)\text{var } Z_{n-1} = \frac{3}{n^2}E[\text{var}\sum_{k,k\neq i} w(X_i,X_k)|X_i] + O\left(\frac{1}{n\rho_n}\right). \tag{A.23}$$

Note that, we also have:

$$\frac{1}{n^2}E[\text{var}\sum_{k,k\neq i} w(X_i,X_k)|X_i] = \frac{1}{n}E[\text{var}(w(X_i,X_k)|X_i)] = O(1/n) \tag{A.24}$$

Eq A.24 establishes Eq 2.10. Furthermore, in conjunction with Eqs A.19 and A.18, it also shows that both $(n-1)E\sum_{i=1}^{n}(Z_{n,i}-\overline{Z_n})^2$ and $(n-1)\text{var } Z_{n-1}$ converge to positive constants. This concludes our proof. □

We now present the proofs of Lemmas A.4.1 and A.4.2.

Proof of Lemma A.4.1. Applying law of total variance,

$$\sum_{i=1}^{n-1} \text{var}\left(\frac{D_i^{(n)}}{\gamma_n}\right) = \sum_{i=1}^{n-1} \text{var}\left[E\left(\frac{D_i^{(n)}}{\gamma_n}\middle|X\right)\right] + \sum_{i=1}^{n-1} E\left[\text{var}\left(\frac{D_i^{(n)}}{\gamma_n}\middle|X\right)\right]. \tag{A.25}$$

We now show that the second term on the RHS of the above equation is small.

$$\begin{aligned}\sum_{i=1}^{n-1} E\left[\text{var}\left(\frac{D_i^{(n)}}{\gamma_n}\middle|X\right)\right] &= \sum_{i=1}^{n-1} E\left[\text{var}\left(\frac{\sum_{j\neq i} A_{ij}}{\binom{n}{2}\rho_n}\middle|X\right)\right] \\ &= \sum_{i=1}^{n-1} E\left(\frac{\sum_{j\neq i}\rho_n w(X_i,X_j)(1-\rho_n w(X_i,X_j))}{\binom{n}{2}^2\rho_n^2}\right) \\ &\asymp \sum_{i,j,i\neq j} \frac{\rho_n E[w(X_i,X_j)]}{n^4\rho_n^2} = O(n^{-2}\rho_n^{-1}) \end{aligned} \tag{A.26}$$

For the first term on the RHS of Eq A.25, for any fixed i, we have:

$$\begin{aligned}&\text{var}\left(E\left[\frac{D_i^{(n)}}{\gamma_n}\middle|X\right]\right) = \text{var}E\left(\frac{\sum_{k,k\neq i} A_{ik}}{\frac{(n-1)(n-2)}{2}\rho_n}\middle|X\right) \asymp \frac{4}{n^4}\text{var}\left(\sum_{k,k\neq i} w(X_i,X_k)\right) \\ &\asymp \frac{4}{n^4}E\left(\text{var}\sum_{k,k\neq i} w(X_i,X_k)|X_i\right) + \frac{4}{n^4}\text{var}\left(E\sum_{k,k\neq i} w(X_i,X_k)|X_i\right).\end{aligned} \tag{A.27}$$

Exchanging the sum and expectation in the second term, we can also write,

$$\frac{4}{n^4}\text{var}\left(E\sum_{k,k\neq i} w(X_i,X_k)|X_i\right) = \frac{4}{n^2}\text{var}[E(w(X_i,X_k)|X_i)]. \tag{A.28}$$

Since Eq A.25 involves a sum over $n-1$ identical terms, owing to the fact that $\{X_i\}$ are i.i.d, we get the result by multiplying Eq A.27 and A.28 by $n-1$. □

Proof of Lemma A.4.2. We decompose the covariance into

$$\sum_{i,j,i\neq j} \text{cov}\left(\frac{D_i^{(n)}}{\gamma_n}, \frac{D_j^{(n)}}{\gamma_n}\right) = \sum_{i,j,i\neq j} \text{cov}\left(E\left[\frac{D_i^{(n)}}{\gamma_n}\middle|X\right], E\left[\frac{D_i^{(n)}}{\gamma_n}\middle|X\right]\right)$$

$$+\sum_{i,j,i\neq j} E\left[\operatorname{cov}\left(\frac{D_i^{(n)}}{\gamma_n}, \frac{D_j^{(n)}}{\gamma_n}\middle| X\right)\right]. \qquad \text{(A.29)}$$

The second term on the RHS of the above equation is small as shown before.

$$\begin{aligned}
&\sum_{i,j,i\neq j} E\left[\operatorname{cov}\left(\frac{D_i^{(n)}}{\gamma_n}, \frac{D_j^{(n)}}{\gamma_n}\middle| X\right)\right]\\
&= \sum_{i,j,i\neq j} E\left[\operatorname{cov}\left(\frac{\sum_{k,k\neq i} A_{ik}}{\gamma_n}, \frac{\sum_{s,s\neq j} A_{js}}{\gamma_n}\middle| X\right)\right]\\
&\asymp \frac{1}{n^4\rho_n^2}\sum_{i,j} E[\operatorname{var}(A_{ij}|X)]\\
&\asymp \frac{1}{n^2\rho_n^2}\rho_n E[w(X_i,X_j)] = O(n^{-2}\rho_n^{-1})
\end{aligned}$$

For the first term in Eq A.29, for any fixed i and j, we have

$$\begin{aligned}
&\operatorname{cov}\left(E\left[\frac{D_i^{(n)}}{\gamma_n}\middle| X\right], E\left[\frac{D_j^{(n)}}{\gamma_n}\middle| X\right]\right)\\
&= \operatorname{cov}\left(\frac{\sum_k^{k\neq i} w(X_i,X_k)\rho_n}{\frac{(n-1)(n-2)}{2}\rho_n}, \frac{\sum_s^{s\neq j} w(X_j,X_s)\rho_n}{\frac{(n-1)(n-2)}{2}\rho_n}\right)\\
&\asymp \frac{4}{n^4}\operatorname{cov}\left(\sum_{k,k\neq i} w(X_i,X_k), \sum_{s,s\neq j} w(X_j,X_s)\right)\\
&= \frac{4}{n^4}\sum_{k,k\neq i}\sum_{s,s\neq j}\operatorname{cov}(w(X_i,X_k), w(X_j,X_s)).
\end{aligned} \qquad \text{(A.30)}$$

Let $S_i = \{i,k\}$, and $S_j = \{j,s\}$ be two pairs containing i and j respectively. Some algebraic manipulation yields,

$$\begin{aligned}
\sum_{k,k\neq i}\sum_{s,s\neq j}\operatorname{cov}(w(X_i,X_k), w(X_j,X_s)) &= \sum_{|S_i\cap S_j|=1}\operatorname{cov}(w(X_i,X_k), w(X_j,X_s))\\
&+ \sum_{|S_i\cap S_j|=2}\operatorname{cov}(w(X_i,X_k), w(X_j,X_s)).
\end{aligned} \qquad \text{(A.31)}$$

In the R.H.S of the above expression, the second summation has $n(n-1)$ terms, whereas the first has $n(n-1)(n-2)$ terms. Furthermore, for $|S_i \cap S_j| = 2$, it is easy to see that $\text{cov}(w(X_i, X_k), w(X_j, X_s))$ is simply the variance of $\text{var}(w(X_i, X_k))$ which is positive. For $|S_i \cap S_j| = 1$, W.L.O.G. let $S_i = \{i, u\}$ and $S_j = \{j, u\}$. Conditioned on the shared node X_u,

$$\begin{aligned} \text{cov}(w(X_i, X_u), w(X_j, X_u)) &= \text{cov}[E(w(X_i, X_u)|X_u), E(w(X_j, X_u)|X_u)] \\ &= \text{var}(Ew(X_i, X_u)|X_u) \qquad \text{(A.32)} \end{aligned}$$

which is also positive. Hence the contribution of the first sum is of a larger order.

Now we enumerate all the ways in which S_i and S_j can have a node in common, with the constraint of $i \neq j$. For any fixed i and j, s.t. $i \neq j$, $|S_i \cap S_j| = 1$ means that there is 1 common node in $S_i = \{i, k\}$ and $S_j = \{j, s\}$. There are three possible cases, $i = s$, $k = j$, $k = s$. Thus, Eq A.30 can be expanded as (W.L.O.G, suppose $i = s$),

$$\begin{aligned} \text{cov}\left(E\left[\frac{D_i^{(n)}}{\gamma_n}\middle|X\right], E\left[\frac{D_j^{(n)}}{\gamma_n}\middle|X\right]\right) &\asymp \frac{4}{n^4}[3(n-2)\text{cov}(w(X_i, X_k), w(X_j, X_i))] \\ &= \frac{4}{n^3} \times 3\text{cov}(w(X_i, X_k), w(X_j, X_i)) \\ &\overset{(i)}{=} \frac{4}{n^3} \times 3\text{var}(E(w(X_i, X_k))|X_i) \qquad \text{(A.33)} \end{aligned}$$

Step (i) uses an analogous argument from Eq A.32, and conditions on X_i.

Eq A.29 involves a sum over all (i, j) pairs, $i \neq j$, , owing to the fact that $\{X_i\}$ are i.i.d, we get the result by multiplying Eq A.33 by $n(n-1)$. □

A.5 Proof of Proposition 2

Before we state the proof of our result, recall the following well-known relationship between uniform integrability and convergence of moments. See for example, Theorem 25.12 of Billingsley (1995).

Proposition A.5.1. *Suppose that $X_n \rightsquigarrow X$ and $\{X_n\}_{n\geq 1}$ is uniformly integrable. Then, $E(X_n) \to E(X)$.*

Now we will prove our proposition below:

Proof. In what follows let $X_n := \tau_n[\hat{\theta}_n - E(\hat{\theta}_n)]$ and $V_n = \tau_n \cdot U_n$. Recall that $U_n = \hat{\theta}_n - \theta$. While our result here is more general, in a jackknife context, $\hat{\theta}_n = Z_n$ following the notation that we use elsewhere. Consider the following decomposition:

$$\tau_n[\hat{\theta}_n - E(\hat{\theta}_n)] = \tau_n[\hat{\theta}_n - \theta] + E(\tau_n[\theta - \hat{\theta}_n])$$

Since $\{V_n^2\}_{n\geq 1}$ is uniformly integrable, it follows that $\{V_n\}_{n\geq 1}$ is also uniformly integrable. Therefore, by Proposition A.5.1, $E(\tau_n[\theta - \hat{\theta}_n]) \to 0$. By Slutsky's Theorem, it follows that $\tau_n[\hat{\theta}_n - E(\hat{\theta}_n)] \rightsquigarrow U$.

To show that the variances converge to the same value, observe that $E(X_n^2)$ is given by:

$$E(X_n^2) = E(V_n^2) - (E(V_n))^2$$

First, $V_n^2 \rightsquigarrow U^2$ by continuous mapping theorem. Since $\{V_n^2\}_{n\geq 1}$ is uniformly integrable, $E(V_n^2) \to E(U^2)$ by Proposition A.5.1 again. Finally, $(EV_n)^2 \to 0$ and the result follows. □

A.6 Additional theory

It should be noted that a similar inequality for a closely related procedure has an even simpler proof. This alternative procedure does not require the functional to be invariant to node permutation and allows flexibility with the leave-one-out estimates. However, the resulting estimate is often not sharp. More concretely, let Z_n denote a function of $A^{(n)}$ and let $\widetilde{Z}_{n,i}$ be an arbitrary functional calculated on a graph with node i removed. Consider the following estimator:

$$\widehat{\mathrm{Var}}_{\mathrm{JACK}}\, Z_n = \sum_{i=1}^{n} (Z_n - \widetilde{Z}_{n,i})^2 \tag{A.34}$$

Combining the aforementioned filtration with arguments in Boucheron et al. (2004) leads to the following inequality:

Proposition A.6.1 (Network Efron-Stein, alternative version)**.**

$$var\, Z_n \leq E(\widehat{\mathrm{Var}}_{\mathrm{JACK}}\, Z_n) \tag{A.35}$$

A.7 Additional experiments

We first present Tables A.1 and A.2 with details of the networks we used in our real data experiments in Section 3.6 of the main paper.

For our real data experiments, (Section 3.6 of main paper) we compared subsampling with jackknife on the three colleges (see Figure 2.3). For simplicity, for the second experiment comparing three pairs of college networks (see Figure 2.4), we only showed the confidence intervals obtained using jackknife. Here, in Figure A.1, for completeness, we present confidence intervals for test sets constructed from the

Table A.1: Details of college networks for first real data experiment (see Figure 2.3 of main paper)

	Caltech	Williams	Wellesley
Nodes	769	2790	2970
Edges	16656	112986	94899
Ave. Degree	43.375	63.927	81.023

Table A.2: Details of college networks for second real data experiment (see Figure 2.4 of main paper)

	Berkeley	Stanford	Yale	Princeton	Harvard	MIT
Nodes	22937	11621	8578	6596	15126	6440
Edges	852444	568330	405450	293320	824617	251252
Ave. Degree	74.332	97.819	94.544	88.952	109.040	78.040

six college networks using both jackknife and subsampling with different choices of b. This again shows that jackknife CI's mostly are in agreement with those obtained from subsampling.

In addition, we show the timing results our real data experiments. Figure A.2 shows computation time of the three college example of Facebook network data (see Figure 2.3). We demonstrate the triangle, two-star densities and normalized transitivity variance computation time using jackknife and subsampling with $b = 0.05n$, $b = 0.1n$ and $b = 0.2n$, $B = 1000$ in each college network.

In Figure A.3, we show the computation time of variance estimation for the same statistics on the test sets for the same set of algorithms. Since we split training and test set in half, the training sets have approximately the same time.

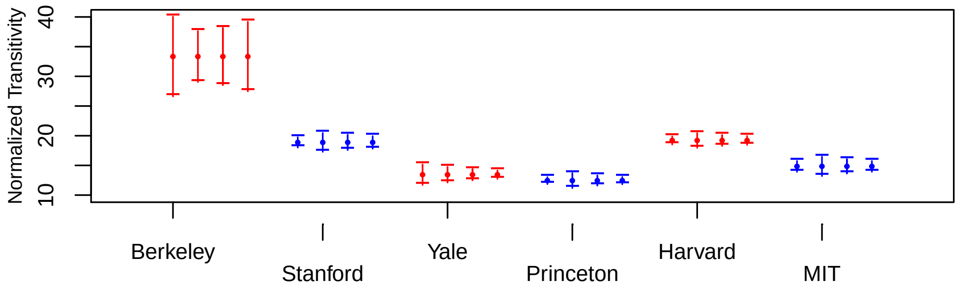


Figure A.1: Confidence intervals of subsampling and jackknife in calculating triangle, two-star densities and normalized transitivity in the example of six college Facebook networks test sets. The four CIs for each college are in the order of jackknife, subsampling with b=0.05n, b=0.1n, and b=0.2n respectively.

These figures show that, it is possible to implement jackknife in a computationally efficient manner when there is nested structure in the subgraph counts. In all these cases, we see that for the larger networks, subsampling with large *b* is often considerably slower than jackknife.

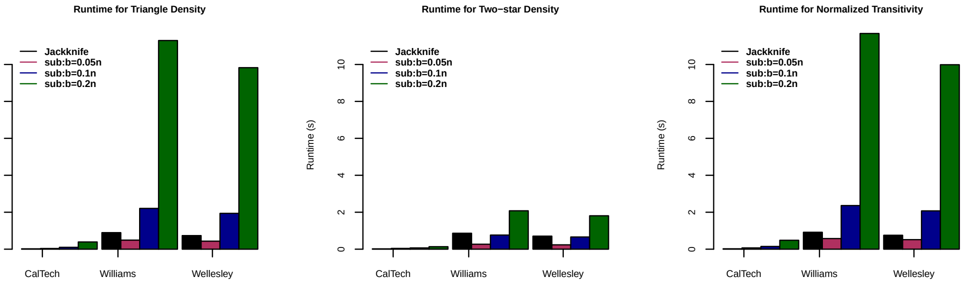


Figure A.2: Computation time of jackknife compared to subsampling in calculating triangle, two-star densities and normalized transitivity in the example of three college Facebook networks.

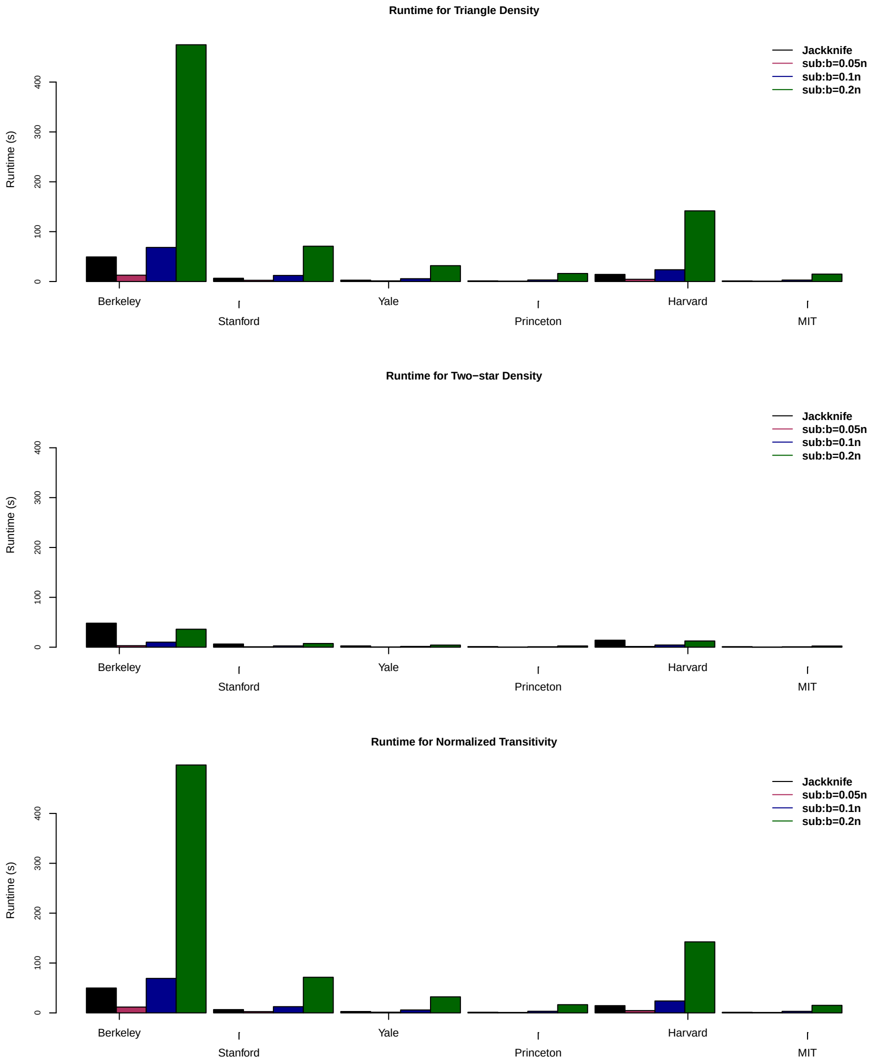


Figure A.3: Computation time of jackknife compared to subsampling in calculating triangle, two-star densities and normalized transitivity in the example of six college Facebook networks test sets.

Appendix B

Supplementary Material for Network Multiplier Bootstrap

B.1 Proof of Lemma 4

Proof. In what follows, we will consider a projection of $T^*_{n,M}$ with respect to the random variables $\xi_1, \ldots \xi_n$, conditional on A and X.

Recall that ξ_i follows the Gaussian Product distribution. First, we may express $T^*_{n,M}$ as:

$$T^*_{n,M} = \frac{1}{\binom{n}{r}} \sum_{1 \leq i_1 < i_2 < \ldots i_r} \left(\xi_{i_1 \cdots i_r} - 1\right) \cdot \left\{ H(A^{(n)}_{i_1,\ldots,i_r}) - \hat{T}_n \right\}$$

where $\xi_{i_1 \cdots i_r}$ denotes the product $\xi_{i_1} \times \cdots \times \xi_{i_r}$. It turns out that applying the Hoeffding decomposition directly to $T^*_{n,M}$ leads to tedious combinatorial calculations; following Bentkus et al. (1997), let Ω_r denote an r-tuple of $\{1, \ldots, n\}$. For each summand, we will consider a Hoeffding representation with respect to Ω_r. Note that using the Hoeffding projection (also see Bentkus et al. (1997) section 2.8),

$$\prod_{1 \leq i \leq r} \xi_i - 1 = \sum_{k=1}^{r} \sum_{1 \leq i_1 < \cdots < i_k \leq r} h_k(\xi_{i_1}, \ldots, \xi_{i_k}),$$

where for $\Omega_k = \{1, \ldots, k\}$,

$$h_k(\xi_1, \ldots, \xi_k) = \sum_{B \in \Omega_k} (-1)^{k-|B|} \mathrm{E} \left\{ \prod_{1 \leq i \leq r} \xi_i - 1 \mid B \right\}$$

Thus the first two terms are given by:

$$h_1(\xi_1) := (\xi_1 - 1)$$
$$h_2(\xi_1, \xi_2) := (\xi_1\xi_2 - 1) - (\xi_1 - 1) - (\xi_2 - 1) = (\xi_1 - 1)(\xi_2 - 1)$$

In what follows, we will also denote $A^{(n)}_{i_1,\dots,i_r}$ by $A^{(n)}_S$, where $S = \{i_1, \dots, i_r\}$. Let

$$\hat{H}_2(i, j) = \frac{1}{\binom{n-2}{r-2}} \sum_{S \mid i,j \in S} H(A_S), \tag{B.1}$$
$$\hat{H}_u(i_1, \dots, i_u) = \frac{1}{\binom{n-u}{r-u}} \sum_{S \mid i_1,\dots,i_u \in S} H(A_S).$$

Thus $T^*_{n,M}$ can be written as follows:

$$\begin{aligned}
T^*_{n,M} &= \frac{1}{\binom{n}{r}} \sum_{1 \le i_1 < i_2 < \ldots < i_r} \left(\xi_{i_1 \cdots i_r} - 1\right) \cdot \left\{H(A^{(n)}_{i_1,\dots,i_r}) - \hat{T}_n\right\} \\
&= \frac{1}{\binom{n}{r}} \sum_{1 \le i_1 < i_2 < \ldots < i_r} \sum_{k=1}^{r} \sum_{1 \le i_1 < \cdots < i_k \le i_r} h_k(\xi_{i_1}, \dots, \xi_{i_k}) \cdot \left\{H(A^{(n)}_{i_1,\dots,i_r}) - \hat{T}_n\right\} \\
&= \frac{1}{\binom{n}{r}} \sum_{k=1}^{r} \sum_{1 \le i_1 < \cdots < i_k \le n} h_k(\xi_{i_1}, \dots, \xi_{i_k}) \cdot \sum_{S} \left\{H(A^{(n)}_S) - \hat{T}_n\right\} 1(i_1, \dots, i_k \in S) \\
&= \frac{1}{\binom{n}{r}} \sum_{k=1}^{r} \binom{n-k}{r-k} \sum_{1 \le i_1 < \cdots < i_k \le n} h_k(\xi_{i_1}, \dots, \xi_{i_k}) \frac{\sum_S \left\{H(A^{(n)}_S) - \hat{T}_n\right\} 1(i_1, \dots, i_k \in S)}{\binom{n-k}{r-k}} \\
&= \frac{r}{n} \sum_i (\xi_i - 1)\hat{g}_1(i) + \frac{r(r-1)}{n(n-1)} \sum_{1 \le i < j \le n} (\xi_i - 1)(\xi_j - 1) \underbrace{\{\hat{H}_2(i, j) - \hat{T}_n\}}_{\tilde{g}_2(i,j)} + R_n
\end{aligned} \tag{B.2}$$

Now, it remains to show that the remainder of $(T^*_{n,M} - \hat{T}_n)/\hat{\sigma}_n$ is $O(\delta(n, \rho_n, R))$, where:

$$\delta(n, \rho_n, R) = \begin{cases} \frac{1}{n\rho_n} & \text{R is acyclic} \\ \frac{1}{n\rho_n^{3/2}} & \text{R is a simple cycle.} \end{cases}$$

The residual R_n is a sum of higher order Hoeffding projections, which are all uncorrelated. Therefore, we see that the variance of the u^{th} order term is $\frac{\sum_{1\leq i_1<i_2\cdots<i_u}\tilde{g}_u(i_1,\ldots,i_u)^2}{\hat{\sigma}_n^2\binom{n}{u}^2}$. We will now obtain expressions for $3 \leq u \leq r$.

Consider any term $\tilde{g}_u(1,\ldots,u)$. We will now bound $\mathrm{E}\{\tilde{g}_u(1,\ldots,u)^2\}$.

$$\mathrm{E}\{\tilde{g}_u(1,\ldots,u)^2\} \leq 2[\mathrm{var}\{\hat{H}_u(1,\ldots,u)\} + \underbrace{\mathrm{var}(\hat{T}_n)}_{O(\rho_n^{2s}/n)}]$$

The bound on the second term follows from Bickel et al. (2011) and will be smaller than that of the first term. Let $\mathcal{S}_{r,u}$ denote all subsets of size $r-u$, not containing $1,\ldots u$. For any subset $S \in \mathcal{S}_{r,u}$, also define, $S_u = S \cup \{1,\ldots,u\}$. For the first part, we have:

$$\mathrm{var}\{\hat{H}_u(1,\ldots,u)\} = \frac{\sum_{S,T\in\mathcal{S}_{r,u}}\mathrm{cov}\{H(A_{S_u}),H(A_{T_u})\}}{\binom{n-u}{r-u}^2}$$

Note that the dominating term here will indeed be the one where $|S\cap T| = 0$. The number of such terms is $\binom{n}{2r-u}$. Also the covariance of those terms will be $\rho_n^{2s-E(A_{1,\ldots,u})}$, where $E(A_{1,\ldots,u})$ denotes the intersection of the edgeset of $A_{1,\ldots,u}$ and the subgraph we are counting. This number can be at most $u-1$ for acyclic R and u for a simple cycle R. For $|S\cap T| = k$, the number of terms is $\binom{n}{2r-2u-k}$ and the exponent on ρ_n is at most $2s-(u+k-1)$. Thus, for an acyclic subgraph, we have,

$$\begin{aligned}\mathrm{var}\{\hat{H}_u(1,\ldots,u)\} &\leq \frac{\sum_{k=0}^{r}\binom{n}{2r-2u-k}\rho_n^{2s-(u+k-1)}}{\binom{n-u}{r-u}^2}\\ &\leq \sum_{k=0}^{r} n^{-k}\rho_n^{2s-(u+k-1)} = \rho_n^{2s-(u-1)}\left(1+\sum_{k>0}\frac{1}{(n\rho_n)^k}\right)\end{aligned}$$

The cyclic one is worse by a factor of ρ_n. Thus the contribution of the u^{th} element of the Hoeffding decomposition is

$$n \cdot \frac{\sum_{i_1,\ldots,i_u} \tilde{g}_u(i_1,\ldots,i_u)^2}{\binom{n}{u}^2 \hat{\tau}_n^2} = \begin{cases} O_P\left(\frac{1}{(n\rho_n)^{u-1}}\right) & R \text{ acyclic} \\ O_P\left(\frac{\rho_n^{-1}}{(n\rho_n)^{u-1}}\right) & R \text{ a simple cycle} \end{cases}$$

This shows that the third term contributes the most to R_n in Eq B.2. By Markov's inequality, and the definition of $O_P(.)$ notation, it is easy to see that $R_n = O_P(\delta(n,\rho_n,R))$. □

B.2 Proof of Theorem 5

Proof. For any $i \in \{1,\ldots,n\}$, denote the set of all subsets of size $r-1$ taken from $\{1,\ldots,i-1,i+1,\ldots n\}$ as $\mathbb{S}_{-i}$. Denote $H(A_{i,i_2,\ldots,i_r})$ for $S = \{i_2,\ldots,i_r\} \in \mathbb{S}\{-i\}$ as $H(A_{S\cup i})$. Denote

$$H_1(i) = \frac{1}{\binom{n-1}{r-1}} \sum_{S\in\mathbb{S}\{-i\}} H(A_{S\cup i}).$$

Now let π_j be a permutation picked with replacement and uniformly at random from the set of all permutations of $\{1,\ldots,n\} \setminus i$. We have $j = 1,\ldots,N$ independent permutations π_j. Let $\mathbb{S}_\pi$ denote the set of all disjoint subsets $\{\pi_{(i-1)(r-1)+1},\ldots,\pi_{i(r-1)}\}$, $i = 1,\ldots,\frac{n-1}{r-1}$ obtained from permutation π. We write

$$\tilde{T}^*_{n,L} - \tilde{T}_n = \frac{r}{n}\sum_{i=1}^{n}(\xi_i - 1)\{\tilde{H}_1(i) - \tilde{T}_n\},$$

where

$$\tilde{H}_1(i) = \frac{\sum_j H_{\pi_j}(i)}{N},$$

$$H_\pi(i) = \frac{\sum_{S \in \mathbb{S}_\pi} H(A_{S \cup i})}{\frac{n-1}{r-1}},$$
$$\tilde{T}_n = \frac{1}{n} \sum_{i=1}^{n} \tilde{H}_1(i).$$

Note that

$$\mathrm{var}(\tilde{T}^*_{n,L} - \tilde{T}_n | A, X) = \frac{r^2}{n^2} \sum_i \tilde{\sigma}^2_{n,i}$$

Proof of Theorem 5 (a):

Let Y_i denote $(\xi_i - 1)\{\tilde{H}_1(i) - \tilde{T}_n\}$. Conditioned on A and X, $\tilde{H}_1(i) - \tilde{T}_n$ are observed constants, Y_i are independent but not identically distributed with variance $\tilde{\sigma}^2_{n,i}$.

$$\tilde{\sigma}^2_{n,i} = \mathrm{var}[(\xi_i - 1)\{\tilde{H}_1(i) - \tilde{T}_n\} \mid A, X] = E[\{\tilde{H}_1(i) - \tilde{T}_n\}^2 \mid A, X]$$

Applying Berry-Esseen theorem to $\tilde{T}^*_{n,L} - \tilde{T}_n$ conditioned on A and X, we have

$$\sup_{u \in R} \left| P^* \left(\frac{\tilde{T}^*_{n,L} - \tilde{T}_n}{\sqrt{\mathrm{var}(\tilde{T}^*_{n,L} - \tilde{T}_n | A, X)}} \leq u \right) - \Phi(u) \right|$$
$$= \sup_{u \in R} \left| P^* \left(\frac{\sum_i Y_i}{\sqrt{\sum_i \tilde{\sigma}^2_{n,i}}} \leq u \right) - \Phi(u) \right|$$
$$\leq \left(\sum_{i=1}^{n} (\tilde{\sigma}^2_{n,i}) \right)^{-3/2} \sum_{i=1}^{n} \gamma_i =: C_1 \psi_N,$$

Now in order to bound ψ_N, we need to bound $E\tilde{\sigma}^2_{n,i}$. We decompose $\tilde{\sigma}^2_{n,i}$ into

$$\begin{aligned}\tilde{\sigma}_{n,i}^2 &= E\left[(\tilde{H}_1(i) - \tilde{T}_n)^2 \mid A, X\right] \\ &= \mathrm{E}\left[\{\tilde{H}_1(i) - H_1(i)\}^2 \mid A, X\right] + (H_1(i) - \hat{T}_n)^2 + \mathrm{E}\left[(\tilde{T}_n - \hat{T}_n)^2 \mid A, X\right] \\ &+ 2\mathrm{E}[\{\tilde{H}_1(i) - H_1(i)\}\{H_1(i) - \hat{T}_n)\} \mid A, X] - 2\mathrm{E}[(H_1(i) - \hat{T}_n)(\tilde{T}_n - \hat{T}_n) \mid A, X] \\ &- 2\mathrm{E}[\{\tilde{H}_1(i) - H_1(i)\}(\tilde{T}_n - \hat{T}_n) \mid A, X] \qquad \text{(B.3)}\end{aligned}$$

It is easy to see that the first two cross terms are zero. As for the third, the law of iterated expectation gives:

$$\mathrm{E}[\{\tilde{H}_1(i) - H_1(i)\}(\tilde{T}_n - \hat{T}_n)] \mid A, X] = \frac{1}{n}\mathrm{var}\{\tilde{H}_1(i) \mid A, X\}.$$

Now we calculate the third term in Eq B.3,

$$\mathrm{E}\left\{(\tilde{T}_n - \hat{T}_n)^2 \mid A, X\right\} = \mathrm{E}\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\{\tilde{H}_1(i) - H_1(i)\}\right]^2 \mid A, X\right\} = \frac{\sum_{i=1}^{n} \mathrm{var}\{\tilde{H}_1(i) \mid A, X\}}{n^2}.$$

For the first term, we also have,

$$E\left[\{\tilde{H}_1(i) - H_1(i)\}^2 \mid A, X\right] = \mathrm{var}\{\tilde{H}_1(i) \mid A, X\}.$$

Collecting all terms in Eq B.5, we have

$$\tilde{\sigma}_{n,i}^2 = \left(1 - \frac{2}{n}\right)\mathrm{var}\{\tilde{H}_1(i) \mid A, X\} + \frac{\sum_{i=1}^{n} \mathrm{var}\{\tilde{H}_1(i) \mid A, X)\}}{n^2} + \hat{g}_1(i)^2. \qquad \text{(B.4)}$$

Since the first two terms in Eq B.4 are always positive for $n > 2$, $\sum_i \tilde{\sigma}_{n,i}^2 \geq n\hat{\tau}_n^2$. Also, by Lemma B.4.3, we have $n\hat{\tau}_n^2 = cn\rho_n^{2s}(1 + o_P(1))$ for some constant c.

Now we compute $\tilde{\gamma}_i^3$. Observe that,

$$\sum_i \tilde{\gamma}_i^3 \leq C\Bigg(\sum_i E\left[|\tilde{H}_1(i) - H_1(i)|^3|A, X\right] \\ + \sum_i |H_1(i) - \hat{T}_n|^3 + \sum_i E[|\hat{T}_n - \tilde{T}_n|^3|A, X]\Bigg) \tag{B.5}$$

Note that for the last term in Eq B.5, using Jensen's inequality for convex function $f(x) = |x|^3$, for some constant C_0

$$\sum_i E[|\hat{T}_n - \tilde{T}_n|^3|A, X] \leq n \times \frac{1}{n}\sum_i E[|\tilde{H}_1(i) - \hat{T}_n|^3|A, X] \\ \leq C_0\left(\sum_i E\left[|\tilde{H}_1(i) - H_1(i)|^3|A, X\right] + \sum_i |H_1(i) - \hat{T}_n|^3\right),$$

which are just the first two terms in Eq B.5. So now we bound the first two terms in Eq B.5.

From Lemma B.4.5, we have with probability going to one, the second term in the above equation $\sum_i |H_1(i) - \hat{T}_n|^3 \leq cn\rho_n^{3s}$, for some constant c.

Now we look at the first term in Eq B.5. Using Rosenthal's inequality, for some constant C_1 and C_2, we have

$$\sum_i E\left[|\tilde{H}_1(i) - H_1(i)|^3|A, X\right] \leq \underbrace{C_1\left(\sum_i \frac{1}{N^3}\sum_\pi E[|H_\pi(i) - H_1(i)|^3|A, X]\right)}_{Y_1} \\ + \underbrace{C_2\left(\sum_i \frac{1}{N^3}\left(\sum_\pi E[(H_\pi(i) - H_1(i))^2|A, X]\right)^{3/2}\right)}_{Y_2}.$$

Now, we bound $E(Y_1)$. Let $\theta^{(i)} = E[H(A_{S\cup i}) \mid X_i]$ and observe that $E|H_1(i) - \theta^{(i)}|^c \leq E|H(A_{S\cup i}) - \theta^{(i)}|^c$ for $c \geq 1$ via Jensen's inequality. Moreover, conditionally on X_i, note that for $S \in \mathbb{S}_\pi$, $H(A_{S\cup i})$ are mutually independent since their node sets are disjoint. Now for Y_1, we use Rosenthal again to bound the third absolute moment of the Bernoulli sum:

$$
\begin{aligned}
E(|H_\pi(i) - H_1(i)|^3|) &\leq \frac{Cr^3}{n^3}\frac{1}{(n-1)!}\sum_\pi E\left[E\left(|\sum_{S\in\mathbb{S}_\pi} H(A_{S\cup i}) - \theta^{(i)}|^3\middle| X_i\right)\right] \\
&\leq \frac{C_1}{n^2}E|H(A_{S\cup i}) - \theta^{(i)}|^3 + \frac{C_2}{n^{3/2}}E[E(H(A_{S\cup i}) - \theta^{(i)})^2|X_i)^{3/2}] \\
&= O\left(\frac{\rho_n^s}{n^2}\right) + O\left(\frac{\rho_n^{3s/2}}{n^{3/2}}\right)
\end{aligned}
$$

Now, to bound $E(Y_2)$, we first check,

$$
\begin{aligned}
E\left(\sum_{j=1}^N E\left[(H_\pi(i) - H_1(i))^2\,|A, X\right]\right)^2 &\leq C_1 N \sum_{j=1}^N E\left(E\left[\left(H_{\pi_j}(i) - H_1(i)\right)^2|A, X\right]\right)^2 \\
&\leq C_1' N^2(\rho_n^{2s}/n^2 + \rho_n^{4s-2}/n^2) \\
&= O(N^2\rho_n^{2s}/n^2)
\end{aligned}
\tag{B.6}
$$

To obtain the last step, observe that

$$
\begin{aligned}
&E\left(E\left[(H_\pi(i) - H_1(i))^2\,|A, X\right]\right)^2 = E\left(\frac{1}{(n-1)!}\sum_\pi (H_\pi(i) - H_1(i))^2\right)^2 \\
&= E\left(\frac{1}{(n-1)!}\sum_\pi (H_\pi(i) - \theta^{(i)})^2 - (H_1(i) - \theta^{(i)})^2\right)^2 \\
&\leq 2E\left(\frac{1}{(n-1)!}\sum_\pi \left(H_\pi(i) - \theta^{(i)}\right)^2\right)^2 + 2E\left[\left(H_1(i) - \theta^{(i)}\right)^4\right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{\leq} 4E\left(\frac{1}{(n-1)!}\sum_{\pi}\left(H_\pi(i)-\theta^{(i)}\right)^2\right)^2 \\
&\leq 4E\left(\frac{1}{n^2(n-1)!}\sum_{\pi}\left(\sum_{S\in\mathcal{S}_\pi(i)}\left(H(A_{S\cup i}-\theta^{(i)}\right)^2+\sum_{\substack{S\cap T=\phi\\ S,T\in\mathcal{S}_\pi(i)}}(H(A_{S\cup i})-\theta^{(i)})(H(A_{T\cup i})-\theta^{(i)})\right)\right)^2 \\
&\leq 4E\left(\frac{CH_1(i)+\theta^{(i)^2}}{n}+C'H_1'(i)\right)^2 = O\left(\frac{\rho_n^{2s}}{n^2}\right)+O\left(\frac{\rho_n^{4s-2}}{n^2}\right)
\end{aligned}
$$

Note that:

$$
\begin{aligned}
&E(H_1'(i)^2|X_i) \\
&= \frac{C}{n^{4(r-1)}}\sum_{\substack{S\cap T=\phi\\ S'\cap T'=\phi}} E\left[H(A_{S\cup i})-\theta^{(i)})H(A_{T\cup i})-\theta^{(i)})H(A_{S'\cup i})-\theta^{(i)})H(A_{T'\cup i})-\theta^{(i)})|X_i\right] \\
&= O\left(\frac{\rho_n^{4s-2}}{n^2}\right)+O\left(\frac{\rho_n^{2s}}{n^{2(r-1)}}\right)
\end{aligned}
$$

The last line is true because the terms with the largest contribution are those with $|S\cap S'|=1$ and $|T\cap T'|=1$. The latter term corresponds to the variance terms (i.e. $S=S', T=T'$), which is $O\left(\frac{\rho_n^{2s}}{n^2}\right)$ for $r\geq 2$.

Step (i) follows from Jensen's inequality:

$$\left(H_1(i)-\theta^{(i)}\right)^2 \leq \frac{1}{(n-1)!}\sum_{\pi}\left(H_\pi(i)-\theta^{(i)}\right)^2$$

Using the fact that $\|X\|_{3/2}\leq\|X\|_2$, we have for some constant C,

$$E(Y_2)\leq\frac{Cn\rho_n^{3s/2}}{N^{3/2}n^{3/2}}.$$

Therefore, combining $E(Y_1)$ and $E(Y_2)$, we have, for some constant C_1 and C_2,

$$\begin{aligned} E\left(\sum_i E\left[|\tilde{H}_1(i) - H_1(i)|^3 | A, X\right]\right) &\leq C_1 \frac{nN}{N^3} \times \left(\frac{\rho_n^s}{n^2} + \frac{\rho_n^{3s/2}}{n^{3/2}}\right) + C_2 \frac{n\rho_n^{3s/2}}{N^{3/2}n^{3/2}} \\ &= C_1 \frac{\rho_n^s}{nN^2} + C_1 \frac{\rho_n^{3s/2}}{N^2 n^{1/2}} + C_2 \frac{\rho_n^{3s/2}}{N^{3/2}n^{1/2}}. \end{aligned}$$

Note that we always have $\frac{\rho_n^{3s/2}}{N^2 n^{1/2}} \ll \frac{\rho_n^{3s/2}}{N^{3/2}n^{1/2}}$, and $\frac{\rho_n^s}{nN^2} \ll \frac{\rho_n^{3s/2}}{N^{3/2}n^{1/2}}$ if $N \gg \frac{1}{n\rho_n^s}$. So as long as $N \gg \frac{1}{n\rho_n^s}$, the dominating term is $\frac{\rho_n^{3s/2}}{N^{3/2}n^{1/2}}$.

Thus, combining all terms in Eq B.5 in expectation, for some constant C_3 and C_4, under the condition that $N \gg \frac{1}{n\rho_n^s}$, we have

$$E\left(\sum_i \gamma_i^3\right) \leq C_3 \frac{\rho_n^{3s/2}}{n^{1/2}N^{3/2}} + C_4 n\rho_n^{3s}.$$

Since $\sum_i \gamma_i^3$ is $O_P(E\left(\sum_i \gamma_i^3\right))$, and note that $\sum_i \tilde{\sigma}_{n,i}^2 \geq cn\rho_n^{2s}(1 + o_P(1))$ for some constant c, we are ready to present an upper bound for ψ_N,

$$\psi_N = \frac{\sum_{i=1}^n \gamma_i^3}{(\sum_{i=1}^n \tilde{\sigma}_{n,i}^2)^{3/2}} = O_P\left(\frac{C_3 \frac{\rho_n^{3s/2}}{n^{1/2}N^{3/2}} + C_4 n\rho_n^{3s}}{(n\rho_n^{2s})^{3/2}}\right) = O_P\left(\frac{1}{N^{3/2}n^2\rho_n^{3s/2}}\right) + O_P(n^{-1/2}). \tag{B.7}$$

Note that $\frac{1}{N^{3/2}n^2\rho_n^{3s/2}} \ll \frac{1}{n^{1/2}}$ under the same condition that $N \gg \frac{1}{n\rho_n^s}$.

Thus we have, under the condition that $N \gg \frac{1}{n\rho_n^s}$,

$$\sup_{u\in R}\left|P\left(\frac{n^{1/2}(\tilde{T}_{n,L}^* - \tilde{T}_n)}{r\sqrt{\sum_{i=1}^n \tilde{\sigma}_{n,i}^2}} \leq u \mid A, X\right) - \Phi(u)\right| = O_P\left(n^{-1/2}\right). \tag{B.8}$$

Thus we have proof of Theorem 5 a).

Proof of Theorem 5 (b): Note that

$$\mathrm{var}(\tilde{T}^*_{n,L} - \tilde{T}_n | A, X) = \frac{r^2}{n^2} \sum_i \tilde{\sigma}^2_{n,i}$$

Using Lemma B.2.1 we have the following result for its expectation.

$$\mathrm{E}\left[\mathrm{var}\{\tilde{H}_1(i) \mid A, X)\}\right] = O\left(\frac{r\rho_n^s}{Nn}\right).$$

Now it follows from Eq B.4 that:

$$\frac{\sum_i \tilde{\sigma}^2_{n,i}}{n} = \hat{\tau}^2_n + O_P\left(\frac{r\rho_n^s}{Nn}\right). \tag{B.9}$$

Using Lemma B.4.3 (a), it follows that $\hat{\tau}^2_n/\tau^2 = 1 + O_P(1/n\rho_n)$.

Proof of Theorem 5 (c): For the un-approximated linear bootstrap, there is no randomness in $\tilde{H}_1(i)$, since it equals $\hat{H}_1(i)$. Thus $\mathrm{var}(\tilde{H}_1(i)|A, X) = 0$. So the result follows from Eq B.4. □

Lemma B.2.1.

$$E\left[\mathrm{var}\{\tilde{H}_1(i) \mid A, X\}\right] = O\left(\frac{r\rho_n^s}{Nn}\right).$$

Proof. Let θ_i denote $E[H_\pi(i)|X_i]$. Then, we have

$$\begin{aligned} E(\mathrm{var}(\tilde{H}_1(i) - H_1(i) \mid A, X)) &= \frac{1}{N}\left\{\frac{1}{(n-1)!}\sum_\pi E(H_\pi - \theta_i)^2 - E(H_1(i) - \theta_i)^2\right\} \\ &= \frac{1}{N}\{E(\mathrm{var}(H_\pi) \mid X_i) - E(\mathrm{var}(H_1(i) \mid X_i))\} \\ &= \Theta\left(\frac{\rho_n^s}{Nn}\right) - O\left(\frac{\rho_n^{2s}}{Nn}\right) \\ &= O\left(\frac{r\rho_n^s}{Nn}\right). \end{aligned}$$

□

B.3 Proof of Proposition 3

Proof. In what follows, we prove Proposition 3 holds for Edgeworth expansion of a standardized count functional. Our argument here is closely related to Zhang and Xia (2020), thus we do not present the complete proof here. They have showed in Theorem 3.1 in the above reference, that under same conditions, Edgeworth Expansion for studentized $\hat{T}_n$, denote as $\tilde{G}_n(u)$ here, has the same property as Proposition 3. We have first derived our Edgeworth Expansion formula in eq 3.20 for standardized $\hat{T}_n$ instead of studentized $\hat{T}_n$ and we state the form of the characteristic function of $G_n(u)$ below:

Proposition B.3.1. *We have:*

$$\begin{aligned}&\psi_{G(n)}(t) := \int e^{itu} dG_n(u)\\&= e^{-\frac{t^2}{2}}\left(1 - it^3 \frac{1}{6n^{1/2}\tau_n^3}\left[E\{g_1^3(X_1)\} + 3(r-1)E\{g_1(X_1)g_1(X_2)g_2(X_1,X_2)\}\right]\right).\end{aligned}$$

Our standardized $\hat{T}_n$, denote as $\tilde{T}_n$ can be decomposed into

$$\tilde{T}_n := \frac{\hat{T}_n - \mu_n}{\sigma_n} = \frac{T_n - \mu_n}{\sigma_n} + \frac{\hat{T}_n - T_n}{\sigma_n} = T_{n,1} + T_{n,2} + O_P\left(\frac{1}{n}\right) + R_n,$$

where

$$T_{n,1} = \frac{1}{n^{1/2}\tau_n}\sum_{i=1}^{n} g_1(X_i), \quad T_{n,2} = \frac{r-1}{n^{1/2}(n-1)\tau_n}\sum_{i<j} g_2(X_i, X_j),\ R_n = \frac{\hat{T}_n - T_n}{\sigma_n}.$$

We will begin by bounding R_n. Similar to the theory for U-statistics, the behavior is largely determined by a linear term.

Let:

$$R_{n,1} = \text{Linear part of } \frac{\hat{T}_n - T_n}{\sigma_n}.$$

where the linear part has the form:

$$R_{n,1} = \frac{1}{\binom{n}{2}} \sum_{i<j} c_{ij} \left\{ A_{ij} - \mathrm{E}(A_{ij} \mid X_i, X_j) \right\}$$

for $c_{ij} = c_{ij}(X_i, X_j, \rho_n) \asymp \rho_n^{-1} n^{-1/2}$ defined in Section 7 of the above reference. Theorem 3.1(b) of the above authors establishes that:

$$R_n - R_{n,1} = O_P(\mathcal{M}(n, \rho_n, R)),$$

Under the assumed sparsity conditions, given $\mathbf{X}$, the distribution of $R_{n,1}$ permits the following (uniform) approximation by a Gaussian-distributed variable Z_n:

$$\sup_u \left| F_{R_{n,1}|X}(u) - F_{Z_n} \right| = O_P\left(\frac{1}{\rho_n^{1/2} n} \right),$$

where $Z_n \sim N(0, \frac{\sigma_w^2}{n\rho_n})$ and σ_w^2 is defined as the variance of Eq B.3. Note that $\sigma_w \asymp 1$ when $n \longrightarrow \infty$.

Now to prove our theorem, we will show the three equations below.

$$\sup_u \left| F_{\tilde{T}_n}(u) - F_{T_{n,1}+T_{n,2}+R_n} \right| = O\left(\mathcal{M}(n, \rho_n, R) \right), \tag{B.10}$$

$$\sup_u \left| F_{T_{n,1}+T_{n,2}+R_n}(u) - F_{T_{n,1}+T_{n,2}+Z_n} \right| = O\left(\frac{1}{\rho_n^{1/2} n} \right), \tag{B.11}$$

$$\sup_u \left| F_{T_{n,1}+T_{n,2}+Z_n} - G_n(x) \right| = O\left(\frac{1}{n} \right), \tag{B.12}$$

We prove Eq B.12 using Esseen's smoothing lemma from Section XVI.3 in Feller (1971),

$$\begin{aligned}&\sup_u \left| F_{T_{n,1}+T_{n,2}+Z_n}(u) - G_n(x) \right| \\ &\leq c_1 \int_{-\gamma}^{\gamma} \frac{1}{t} \left| \psi_{F_{T_{n,1}+T_{n,2}+Z_n}}(u) - \psi_{G_n}(t) \right| dt + c_2 \sup_u \frac{G_n'(u)}{\gamma},\end{aligned} \tag{B.13}$$

where ψ is the characteristic function. γ is set to n. We omit the proof here as it is not hard to check by breaking the integral into $|t| \in (0, n^{\epsilon})$,$(n^{\epsilon}, n^{1/2})$ and $(n^{1/2}, n)$. Using similar arguments as Lemma 8.3 of Zhang and Xia (2020), we have Eq B.13 and thus Eq B.12 hold for our characteristic function in Proposition B.3.1. It is also not hard to check that, using similar arguments of the above reference, under Assumption 2, Eq B.10 and Eq B.11 hold given Eq B.12. □

B.4 Edgeworth Expansion for Weighted Bootstrap - Proofs of Theorem 6 and Corollary 6.1

Using Eq (10), we express our quadratic bootstrap statistic as:

$$\hat{T}^*_{n,Q} = \frac{\sum_i (\xi_i - 1)\hat{g}_1(i)}{n^{1/2}\hat{\tau}_n} + \frac{(r-1)\sum_{1 \leq i < j \leq n}(\xi_i \xi_j - \xi_i - \xi_j + 1)\tilde{g}_2(i,j)}{n^{1/2}(n-1)\hat{\tau}_n} \tag{B.14}$$

We will first prove Theorem 6. However in order to prove it we state a slightly different version of Theorem 3.1 in Wang and Jing (2004). The main difference is that one condition in the original lemma is not fulfilled in our case. In particular, Bernoulli noise with $\rho_n \to 0$ blows up some terms that are needed to bound the error associated with the Edgeworth expansion. However, a thorough examination reveals that the argument carries through with some modifications.

Let

$$K_{2,n} = \frac{1}{n^{3/2}B_n^2} \sum_{1\le i<j\le n} b_{ni}b_{nj}d_{nij}\mathrm{E}\{Y_1Y_2\psi(Y_1,Y_2)\} \tag{B.15a}$$

$$L_{1,n}(x) = \sum_{j=1}^{n} \left\{\mathrm{E}\Phi(x - b_{nj}Y_j/B_n) - \Phi(x)\right\} - \frac{1}{2}\Phi''(x) \tag{B.15b}$$

$$L_{2,n}(x) = -K_{2,n}\Phi'''(x) \tag{B.15c}$$

$$E_{2n}(x) = \Phi(x) + L_{1,n}(x) + L_{2,n}(x), \tag{B.15d}$$

Lemma B.4.1. *Consider the following expression.*

$$V_n = \frac{1}{B_n}\sum_j b_{nj}Y_j + \frac{1}{n^{3/2}}\sum_{i<j} d_{nij}\psi(Y_i,Y_j), \tag{B.16}$$

where $B_n^2 = \sum_j b_{nj}^2$. Let $\beta := E(|Y_1|^3)$ and $\lambda = E\{\psi^2(Y_1,Y_2)\}$, and let $E(Y_1) = 0$, $E(Y_1^2) = 1$ and $\kappa(X_1) > 0$. Furthermore, let $E\{\psi(Y_1,Y_2) \mid Y_t\} = 0$ for all $1 \le t \le n$. For some constants ℓ_1, ℓ_2, ℓ_3 the sequence $b_{n,i}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} b_{n,i}^2 \ge l_1 > 0, \quad \frac{1}{n}\sum_{i=1}^{n} |b_{n,i}|^3 \le l_2 \le \infty, \tag{B.17}$$

Furthermore, define $\alpha_i := \frac{1}{n}\sum_{j\ne i} d_{nij}^2$. and for sufficiently large k, define:

$$l_{4,n} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i, \quad s_n^2 = \frac{1}{n}\sum_i \alpha_i^2 - (l_{4,n})^2, \quad l_{5,n} = l_{4,n} + ks_n \tag{B.18}$$

If β, $\kappa(Y_1)$ and λ are bounded, then,

$$\sup_x |P(V_n \le x) - E_{2n}(x)| = O\left(\frac{l_{5,n}\log n}{n^{2/3}}\right),$$

Intuitively, arguments for establishing rates of convergence for the Edgeworth expansions require comparing the characteristic function of the random variable of interest with the Fourier transform of the Edgeworth expansion. To this end, the respective integrals are broken up into several pieces. The bounds required in (B.17) are used to estimate the error of the Edgeworth expansion in some of these steps, but appear as constants and are suppressed in the Big-O notation.

On the other hand, as previously mentioned, it turns out that certain terms that appear as constants in Wang and Jing (2004) blow up when perturbed by sparse network noise and appear in the rate. In particular, the term $l_{5,n}$ arises from needing to bound $\frac{1}{m}\sum_{i=1}^{m}\alpha_i$ for all $m \leq M$ for some M large enough.

Since the data is fixed, we may view $\alpha_1, \ldots, \alpha_n$ as constants. We therefore have the liberty of choosing a "good set" in which α_i are well-behaved. Without loss of generality, we may label these elements $\{\alpha_1, \ldots, \alpha_M\}$; the corresponding multiplier random variables are still independent. Even when there is no randomness, it turns out that a large proportion of $\{\alpha_1, \ldots \alpha_n\}$ must be within k sample standard deviations of the sample mean $l_{4,n}$ for k large enough. This observation, which we believe is novel in the bootstrap setting, allows us to establish a tight bound for $\frac{1}{m}\sum_{i=1}^{m}\alpha_i$ for all $m \leq M$. We state this lemma below.

Lemma B.4.2. *Let $x_1, \ldots, x_n$ be constants in $\mathbb{R}$ and let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ Define the set:*

$$\Gamma_k = \{x_i \geq \bar{x}_n + k s_n\}$$

Then,

$$|\Gamma_k| \leq \frac{n}{k^2}$$

Proof. Observe that:

$$\begin{aligned} s_n^2 &\geq \frac{1}{n}\sum_{i\in\Gamma_k}(x_i - \bar{x}_n)^2 \\ &\geq \frac{1}{n}\sum_{i\in\Gamma_k} k^2 s_n^2 \implies |\Gamma_k| \leq \frac{n}{k^2} \end{aligned}$$

□

Remark B.4.1. Our lemma is closely related to concentration of sums sampled without replacement from a finite population. In fact, it implies the without-replacement Chebychev's inequality; see, for example, Corollary 1.2 of Serfling (1974).

We will show that $\hat{T}^*_{n,Q}$ can be written as Eq B.16, with carefully chosen $\{b_{ni}\}$ and $\{d_{nij}\}$'s. We now present some accompanying Lemmas to show that Eq B.17 is satisfied with probability tending to 1. Proofs of Lemmas B.4.1, B.4.5, and B.4.6 are provided in following subsections.

We present some useful results shown in Zhang and Xia (2020) which we will use later in proofs of our theorems.

Lemma B.4.3. *Let $\hat{\tau}_n^2 = \sum_i \hat{g}_1(i)^2/n$. We have,*

1. *For acyclic graphs, if $n\rho_n \to \infty$, and for cyclic graphs, if $n\rho_n^r \to \infty$, we have:*

$$\frac{\hat{\tau}_n^2}{\tau_n^2} = 1 + O_P\left(\frac{1}{n\rho_n}\right) + O_P\left(\frac{1}{\sqrt{n}}\right) \tag{B.19}$$

$$E(|\hat{g}_1(i) - g_1(X_i)|/\rho_n^s)^2 = O(1/n\rho_n) \tag{B.20}$$

2. *Under Assumption 2, we have*

$$\left|\frac{\sum_j \hat{g}_1(i)^3}{n} - E\{g_1(X_1)^3\}\right| = O_P\left(\rho_n^{3s-0.5}n^{-1/2}\right) \tag{B.21}$$

$$\left| \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\hat{g}_2(i,j)}{\binom{n}{2}} - E\{g_1(X_1)g_1(X_2)g_2(X_1,X_2)\} \right| = O_P\left(\rho_n^{3s-0.5} n^{-1/2}\right), \tag{B.22}$$

and

$$|\hat{\tau}_n^3 - \tau_n^3| = O_P(\rho^{3s}/n^{1/2}). \tag{B.23}$$

Lemma B.4.4. *Under the sparsity assumptions in Assumption 2, for large enough C,*

$$P\left(\frac{1}{n^2}\sum_i \sum_{j\neq i} \tilde{g}_2(i,j)^2 \geq C\rho_n^{2s-1}\right) \to 1$$

Lemma B.4.5. *Under the sparsity conditions in Assumption 2 and for some arbitrary $\epsilon > 0$,*

$$P\left(\frac{\sum_i |\hat{g}_1(i)/\rho_n^s|^3}{n} \leq c\right) \to 1$$
$$P\left(\frac{\sum_i |\hat{g}_1(i)/\rho_n^s|^2}{n} \geq c'\right) \to 1,$$

for positive constants c, c' not depending on n.

Lemma B.4.6. *Let ξ_1 be generated from the Gaussian product distribution. We have $E|\xi_1 - 1|^3 < \infty$.*

Now we are ready to provide the proof.

Proof of Theorem 6. It is easy to see from Eq B.14 that $\hat{T}^*_{n,Q}$ can be expressed as:

$$\hat{T}^*_{n,Q} = \frac{\sum_i b_{n,i}Y_i}{B_n} + \frac{1}{n^{1/2}(n-1)} \sum_{1\leq i<j\leq n} \psi(Y_i, Y_j) d_{n,ij},$$

where we have:

$$Y_i = \xi_i - 1, \tag{B.24a}$$

$$b_{n,i} = \frac{\hat{g}_1(i)}{\rho_n^s}, \tag{B.24b}$$

$$B_n^2 = \sum_{i=1}^{n} b_{n,i}^2, \tag{B.24c}$$

$$d_{n,ij} = \frac{r-1}{\hat{\tau}_n}\tilde{g}_2(i,j) \times \frac{n}{n-1}, \tag{B.24d}$$

$$\psi(\xi_i, \xi_j) = \xi_i\xi_j - \xi_i - \xi_j + 1. \tag{B.24e}$$

Note that since $\hat{\tau}_n^2 = \sum_i \hat{g}_1(i)^2/n$. Thus we use $B_n^2 = n\hat{\tau}_n^2/\rho_n^{2s}$. Thus, $B_n^2 = \sum_i b_{n,i}^2$. Furthermore, Lemma B.4.6 shows that our $\xi_i - 1$ random variables have finite $\mathrm{E}\{|\xi_i - 1|^3\}$.

Lemma B.4.5 shows that the conditions in Eq B.17 are satisfied on a high probability set of A, X.

Using Lemma B.4.5, we see that the first two conditions in Eq B.17 are satisfied with probability tending to one under Assumption 2. Since $B_n^2/n = \sum_i b_{n,i}^2/n$ converges to a positive constant (see Lemma B.4.3), the first condition holds. Now, we need to bound $\ell_{4,n}$ and s_n as defined Eq B.18. First, let $\beta_{n,i} := \sum_{j\neq i} \tilde{g}_2(i,j)^2/n$ and $\bar{\beta}_n = \sum_i \beta_{n,i}/n$. Also let $\gamma_n = \sum_i \beta_{n,i}^2/n - \bar{\beta}_n^2$. Then $\ell_{4,n} = C\beta_{n,i}/\hat{\tau}_n^2$. Note that using Lemma B.4.4, we have, with probability tending to one, $\ell_{4,n} \leq C\rho_n^{-1}$. From Lemma B.4.3, we have $\hat{\tau}_n$ is $\Theta(\rho_n^s)$. Furthermore, let $\hat{G}_2(i,j) := \hat{H}_2(i,j) - h_2(X_i, X_j)$.

We have

$$\hat{G}_2(i,j)^2 = \underbrace{\hat{G}_2(i,j)^2 - \mathrm{E}\{\hat{G}_2(i,j)^2 \mid X\}}_{\delta_{ij}} + \underbrace{\mathrm{E}\{\hat{G}_2(i,j)^2 \mid X\}}_{O(\rho_n^{2s-1})}$$

We now will establish the $O(\rho_n^{2s-1})$ bound stated above for the second term. Let $\mathcal{S}_r^{ij}$ denote all subsets of size r not containing i, j.

$$\mathrm{E}\{\hat{G}_2(i,j)^2 \mid X\} = \frac{\sum_{S,T \in \mathcal{S}_r^{ij}} \mathrm{E}\{H(A_{ij\cup S})H(A_{ij\cup T}) \mid X\}}{\binom{n-2}{r-2}^2}$$

In the above sum the terms with $|S \cap T| = 0$ dominate, and for each of them the conditional expectation is bounded a.s. by $O(\rho_n^{2s-1})$ because of the boundedness of the graphon. Now note that:

$$\begin{aligned}
\tilde{g}_2(i,j)^2 &\le 3\left[\{\hat{H}_2(i,j) - h_2(X_i,X_j)\}^2 + \{h_2(X_i,X_j) - \theta_n\}^2 + (\hat{T}_n - \theta_n)^2\right] \\
&\le 3\{\hat{H}_2(i,j) - h_2(X_i,X_j)\}^2 + O(\rho_n^{2s-1}) \\
\beta_{n,i} &\le \frac{1}{n}\sum_{j\neq i} \delta_{ij} + O(\rho^{2s-1}) \\
\gamma_n^2 &\le \frac{1}{n}\sum_i \beta_{n,i}^2 \le \frac{1}{n}\sum_i \left\{\frac{1}{n}\sum_{j\neq i}\delta_{ij} + O(\rho^{2s-1})\right\}^2 \\
&\le O(\rho^{4s-2}) + \underbrace{\frac{1}{n}\sum_i \frac{1}{n}\sum_{j\neq i}\delta_{ij}^2}_{A}
\end{aligned}$$

Now note that, $\mathrm{E}(\delta_{ij}) = \mathrm{E}\{\mathrm{E}(\delta_{ij} \mid X)\} = 0$. Thus, for all i,

$$\mathrm{E}(A) = \frac{1}{n}\sum_i \mathrm{E}\left\{\frac{1}{n}\sum_{j\neq i}\delta_{ij}\right\}^2 = \frac{1}{n}\sum_i \mathrm{var}\left(\frac{1}{n}\sum_{j\neq i}\delta_{ij}\right) = O(\rho^{4s-3}/n)$$

Thus, we have, for a large enough C,

$$P\left(\gamma_n^2 \ge C\rho_n^{4s-2}\right) \le P\left(A \ge C'\rho_n^{4s-2}\right) \le O\left(\frac{\mathrm{E}(A)}{\rho_n^{4s-2}}\right) = O\left(\frac{1}{n\rho_n}\right)$$

Therefore, since $s_n^2 = \frac{(r-1)^2 n^2}{(n-1)^2}\gamma_n/\hat{\tau}_n^2$, we have with probability tending to one, $l_{4,n} + ks_n = O(\rho_n^{-1})$.

Since the first two conditions in Eq B.17 are satisfied with probability tending to one, from Wang and Jing (2004) Theorem 3.1, we have,

$$\sup_u \left| L_{1n}(u) + \frac{E(\xi_i - 1)^3}{6B_n^3} \sum_{i=1}^n b_{n,i}^3 \Phi'''(u) \right| = o_P(n^{-1/2}),$$

Now we see that, using the definitions of L_{1n}, L_{2n} in Eq B.15a, plugging in definitions of b_{ni} and d_{nij}'s from Eq B.24, and using the fact that $\mathrm{E}[Y_i Y_j \psi(Y_i, Y_j)] = \mathrm{E}[(\xi_i - 1)^2 (\xi_j - 1)^2] = 1$,

$$\sup_u |E_{2n}(u) - \hat{G}_n(u)| = o_P(n^{-1/2}).$$

Therefore, putting all the pieces together we see that

$$\sup_u \left| P^* \left(\frac{\hat{T}_{n,Q}^* - \hat{T}_n}{\hat{\sigma}_n} \leq u \right) - \hat{G}_n(u) \right| = o_p\left(n^{-1/2}\right) + O_P\left(\frac{\log n}{n^{2/3}\rho_n}\right) \tag{B.25}$$

□

Now we are ready to finish the proof of Corollary 6.1.

Proof of Corollary 6.1. Here we take care of the error term in the Hoeffding projection in Eq 3.11. Set $X = \frac{\hat{T}_{n,M}^* - \hat{T}_n}{\hat{\sigma}_n}$, $Y = \hat{T}_{n,Q}^*$. From Eq 3.11, we see that $X = Y + R_n$, where $R_n = O_P(\delta(n, \rho_n, R))$. Using Eq B.25, we see that on a high probability set,

$$\begin{aligned}
&F_Y(u + a) - F_Y(u) \\
&\leq |F_Y(u + a) - \hat{G}_n(u + a)| + |\hat{G}_n(u + a) - \hat{G}_n(u)| + |\hat{G}_n(u) - F_Y(u)| \\
&\leq Ca + O\left(\frac{\log n}{n^{2/3}\rho_n}\right)
\end{aligned}$$

Therefore, using Lemma 8.2 in Zhang and Xia (2020),

$$\sup_u \left| P^* \left(\frac{\hat{T}^*_{n,M} - \hat{T}_n}{\hat{\sigma}_n} \leq u \right) - \hat{G}_n(u) \right| = o_P(n^{-1/2}) + O_P\left(\frac{\log n}{n^{2/3}\rho_n} \right).$$

□

Proof of Lemma 7. If we can establish Eq B.22 and Eq B.23 from Lemma B.4.3 for our empirical moments, we will get the desired result. Note that our empirical moments involve the first term as well as a slight variation of the second term, which is given below.

$$\widehat{E}_n\{g_1(i)g_1(j)g_2(i,j)\} = \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\tilde{g}_2(i,j)}{\binom{n}{2}}$$

We will show that this follows from Eq B.22.

$$\begin{aligned}
&\frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\hat{g}_2(i,j)}{\binom{n}{2}} \\
&= \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)(\tilde{g}_2(i,j) - \hat{g}_1(i) - \hat{g}_1(j))}{\binom{n}{2}} \\
&= \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\tilde{g}_2(i,j)}{\binom{n}{2}} - \frac{\sum_{i\neq j} \hat{g}_1(i)\hat{g}_1(j)(\hat{g}_1(i) + \hat{g}_1(j))}{2\binom{n}{2}} \\
&= \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\tilde{g}_2(i,j)}{\binom{n}{2}} - \frac{\sum_{i\neq j} \hat{g}_1(i)^2\hat{g}_1(j)}{\binom{n}{2}} \\
&= \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\tilde{g}_2(i,j)}{\binom{n}{2}} - \frac{(\sum_i \hat{g}_1(i)^2)(\sum_j \hat{g}_1(j)) - \sum_i \hat{g}_1(i)^3}{\binom{n}{2}} \\
&\overset{(i)}{=} \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\tilde{g}_2(i,j)}{\binom{n}{2}} + \frac{\sum_i \hat{g}_1(i)^3}{\binom{n}{2}} \\
&\overset{(ii)}{=} \frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\tilde{g}_2(i,j)}{\binom{n}{2}} + O_P\left(\frac{\rho_n^{3s}}{n} \right)
\end{aligned}$$

(i) uses the fact that $\sum_i \hat{g}_1(i) = 0$. (ii) uses the fact that $\mathrm{E}\{g_1(X_1)^3\} = O(\rho_n^{3s})$ along with Eq B.21. Hence from Eq B.22 we have:

$$\begin{aligned}
&\left|\frac{\sum_{i<j} \hat{g}_1(i)\hat{g}_1(j)\tilde{g}_2(i,j)}{\binom{n}{2}} - \mathrm{E}\{g_1(X_1)g_1(X_2)g_2(X_1,X_2)\}\right| \\
&= \max\left\{O_P\left(\frac{\rho_n^{3s}}{n}\right), O_P\left(\rho^{3s-\frac{1}{2}}n^{-1/2}\right)\right\} \\
&= O_P\left(\rho^{3s-\frac{1}{2}}n^{-1/2}\right).
\end{aligned}$$

This, along with Eqs B.21 and B.23 yields the result. □

B.4.1 Proof of Lemma B.4.4

Proof. Recall the definition of $\hat{H}_2(i,j)$ from Eq B.1.

$$\begin{aligned}
\tilde{g}_2(i,j) &= \hat{H}_2(i,j) - \hat{T}_n = \{\hat{H}_2(i,j) - h_2(X_i,X_j)\} + \{h_2(X_i,X_j\} - \theta_n) - (\hat{T}_n - \theta_n) \\
\tilde{g}_2(i,j)^2 &\leq 3\left[\{\hat{H}_2(i,j) - h_2(X_i,X_j)\}^2 + \{h_2(X_i,X_j) - \theta_n\}^2 + (\hat{T}_n - \theta_n)^2\right].
\end{aligned}$$

Since $\mathrm{var}(\hat{T}_n) = O(\rho_n^{2s}/n)$ and the second term is bounded a.s. due to our boundedness assumption. We will just prove that $\sum_{j\neq i}\{\hat{H}_2(i,j) - h_2(X_i,X_j)\}^2/(n-1)\rho^{2s}$ is bounded with high probability. It is not hard to check that

$$\mathrm{E}\{(\hat{H}_2(i,j) - h_2(X_i,X_j))^2/\rho_n^{2s}\} = O(1/\rho_n)$$

Therefore,

$$\sum_{j\neq i} \mathrm{E}\{\hat{g}_2(i,j)^2/(n\rho^{2s}) = O(1/\rho_n)\}$$

Furthermore, let $\hat{G}_2(i,j) := \hat{H}_2(i,j) - h_2(X_i, X_j)$. We have

$$\hat{G}_2(i,j)^2 = \underbrace{\hat{G}_2(i,j)^2 - \mathrm{E}\{\hat{G}_2(i,j)^2 \mid X\}}_{\delta_{ij}} + \underbrace{\mathrm{E}\{\hat{G}_2(i,j)^2 \mid X\}}_{O(\rho_n^{2s-1})}$$

We now will establish the $O(\rho_n^{2s-1})$ bound stated above for the second term. Let $\mathcal{S}_r^{ij}$ denote all subsets of size r not containing i, j.

$$\mathrm{E}\{\hat{G}_2(i,j)^2 \mid X\} = \frac{\sum_{S,T\in\mathcal{S}_r^{ij}} \mathrm{E}\{H(A_{ij\cup S})H(A_{ij\cup T}) \mid X\}}{\binom{n-2}{r-2}^2}$$

In the above sum the terms with $|S \cap T| = 0$ dominate, and for each of them the conditional expectation is bounded a.s. by $O(\rho_n^{2s-1})$ because of the boundedness of the graphon.

We will analyze $\sum_i \sum_{j\neq i} \delta_{ij}$. Note that $\mathrm{E}(\delta_{ij} \mid X) = 0$.

$$\mathrm{var}\left(\frac{1}{n^2}\sum_i \sum_j \delta_{ij} \mid X\right) = \frac{\sum_i \sum_j \mathrm{var}(\delta_{ij} \mid X) + \sum_{i,k,k\neq i}\sum_{j,\ell,j\neq\ell} \mathrm{cov}(\delta_{ik}, \delta_{j\ell} \mid X)}{n^4} \tag{B.26}$$

$$\delta_{ij} = \frac{1}{\binom{n-2}{r-2}^2} \sum_{S,T\in\mathcal{S}_r^{ij}} \underbrace{H(A_{ij\cup S})H(A_{ij\cup T}) - \mathrm{E}\{H(A_{ij\cup S})H(A_{ij\cup T}) \mid X\}}_{H'_{ij}(S,T)}$$

For variance, we have:

$$\begin{aligned}\mathrm{var}(\delta_{ij}) &= \mathrm{E}\{\mathrm{var}(\delta_{ij} \mid X)\}\\ &= \frac{\sum\limits_{S_1\neq T_1, S_2\neq T_2\in\mathcal{S}_r^{ij}} \mathrm{E}\{\mathrm{cov}(H'_{ij}(S_1,T_1), H'_{ij}(S_2,T_2) \mid X)\}}{\binom{n-2}{r-2}^4}\end{aligned}$$

The dominant term in the above sum is the one with S_1, S_2, T_1, T_2 all disjoint. Consider any other term in the above sum where any pair of the subsets have p nodes, d edges in common and the rest are disjoint. In this case there are $2(r-2-p)+2(r-2)+p = 4(r-2)-p$ choices of nodes and the number of edges are lower bounded by $4(s-1)+1-d = 4s-3-d$ (since all pairs have $\{i, j\}$ in common). When $p \geq 1$, for acyclic graphs, $d \leq p-1$ and for general subgraphs with a cycle, $d \leq \binom{p}{2}$. Thus, for $p \geq 0$, we have:

$$\frac{O\left(n^{4(r-2)-p}\rho_n^{4s-3-d}\right)}{\binom{n-2}{r-2}^4} = O(\rho_n^{4s-3}) \times O\left(\frac{1}{n^p \rho_n^d}\right)$$

Note that for acyclic graphs, it is easy to see that under our sparsity conditions the above is dominated by $p = 0$. For general cyclic graphs, since $\rho_n = \omega(n^{-1/r})$, note that, since $p \leq r$,

$$n^p \rho_n^d \geq n^p \rho_n^{p(p-1)/2} \geq n^{p\left(1-\frac{p-1}{2r}\right)} \geq n^{\frac{p(r+1)}{2r}} \longrightarrow \infty$$

So, $\operatorname{var}(\delta_{ij}) = O(\rho_n^{4s-3})$.

For covariance, for $i \neq j \neq k \neq \ell$, we have:

$$\begin{aligned}\operatorname{cov}(\delta_{ik}, \delta_{j\ell}) &= \mathrm{E}\{\operatorname{cov}(\delta_{ik}, \delta_{j\ell} \mid X)\} \\ &= \frac{\sum\limits_{S_1 \neq T_1 \in \mathcal{S}_r^{ik}, S_2 \neq T_2 \in \mathcal{S}_r^{j\ell}} \mathrm{E}\{\operatorname{cov}(H'_{ij}(S_1, T_1), H'_{ik}(S_2, T_2) \mid X)\}}{\binom{n-2}{r-2}^4}\end{aligned}$$

Consider any two pairs of subsets with p nodes and d edges in common. First note that $p \geq 2$ in order to have a nonzero covariance. In this case there will be $4(r-2)-p$

choices for nodes, and $(2s - A_{ik}) + (2s - A_{j\ell}) - d \geq 4s - 3 - d$ edges.

$$= \frac{O\left(n^{4(r-2)-p}\rho_n^{4s-2-d}\right)}{\binom{n-2}{r-2}^4} = O\left(\frac{\rho_n^{4s-3}}{n^2}\right) \times O\left(\frac{1}{n^{p-2}\rho^{d-1}}\right).$$

Note that, for acyclic graphs $d \leq p - 1$ and hence the above is maximized at $p = 2, d = 1$ as long as $n\rho_n \to \infty$.

For general cyclic subgraphs, $d \leq \binom{p}{2}$. Furthermore, since $p + 2 \leq r$, and $\rho_n = \omega(n^{-1/r})$, we have, for $p > 2$:

$$\begin{aligned} n^{p-2}\rho_n^{d-1} &= n^{p-2-\frac{1}{r}\left(\frac{p(p-1)}{2}-1\right)} \\ &= n^{p-2-\frac{(p-2)(p+1)}{2r}} = n^{(p-2)\left(1-\frac{p+1}{2r}\right)} \geq n^{(p-2)\frac{r+1}{2r}} \longrightarrow \infty \end{aligned}$$

Thus under the conditions of Assumption 2, we have:

$$\text{cov}(\delta_{ik}, \delta_{j\ell}) = O(\rho_n^{4s-3}/n^2)$$

Step (i) is true, because conditioned on X, there needs to be at lease two nodes u_1, u_2 in common between $\{i, k \cup S_1 \cup T_1\}$ and $\{j, \ell \cup S_2 \cup T_2\}$ to have a nonzero covariance. This leads to only $4(r - 2) - 2$ choices, which dominates the sum. This along with Eq B.26 gives us:

$$\text{var}\left(\frac{1}{n^2}\sum_i \sum_{j \neq i} \delta_{ij}\right) = \text{E}\left\{\text{var}\left(\sum_i \sum_{j \neq i} \delta_{ij}/n^2 \mid X\right)\right\} = O(\rho_n^{4s-3}/n^2).$$

Thus we have for large enough C, we have

$$P\left(\frac{1}{n^2}\sum_i \sum_{j \neq i} \tilde{g}_2(i,j)^2 \geq C\rho_n^{2s-1}\right) \leq P\left(\left|\sum_i \sum_{j \neq i} \delta_{ij}/n^2 + O(\rho_n^{2s-1})\right| \geq C\rho_n^{2s-1}\right)$$

$$\leq P\left(\sum_i \left|\sum_{j\neq i} \delta_{ij}/n^2\right| \geq C'\rho_n^{2s-1}\right)$$
$$\leq C''\frac{\rho_n^{4s-3}/n^2}{\rho_n^{4s-2}} = O\left(\frac{1}{n^2\rho_n}\right).$$

□

B.4.2 Proof of Lemma B.4.5

Proof. Let $\Delta_i := |\hat{g}_1(i) - g_1(X_i)|/\rho_n^s$. We have:

$$\begin{aligned}
&\frac{\sum_i |\hat{g}_1(i)/\rho_n^s|^3}{n} \\
&\leq \frac{\sum_i \Delta_i^3}{n} + 3\frac{\sum_i |g_1(X_i)/\rho_n^s|\Delta_i^2}{n} + 3\frac{\sum_i |g_1(X_i)/\rho_n^s|^2\Delta_i}{n} + \frac{\sum_i |g_1(X_i)/\rho_n^s|^3}{n} \\
&= B_1 + B_2 + B_3 + B_4 \qquad \text{(B.27)}
\end{aligned}$$

First note that using the boundedness condition on the graphon, $|g_1(X_i)/\rho_n^s|$ is bounded. Hence $B_4 \leq c$ a.s. Using Lemma B.4.3, we know that $E(\Delta_i)^2 = O(1/n\rho_n)$. Since $\sum_i \Delta_i \leq n^{1/2}\sum_j \Delta_j^2$, for the second term we have, for some $C > 0$:

$$P(B_2 \geq \epsilon) \leq \frac{n^{1/2}E\sum_i \Delta_i^2}{n\epsilon^2} \leq \frac{C}{n^{1/2}\rho_n\epsilon^2} \qquad \text{(B.28)}$$

Furthermore,

$$P(B_3 \geq \epsilon) \leq \frac{E\sum_i \Delta_i^2}{n\epsilon^2} \leq \frac{C}{n\rho_n\epsilon^2}. \qquad \text{(B.29)}$$

By repeated application of Cauchy-Schwarz inequality, we have $(\sum_i x_i^3)^2 \leq \sum_i x_i^2 \sum_i x_i^4 \leq (\sum_i x_i^2)^3$, we also have:

$$P(B_1 \geq \epsilon) \leq \frac{E\sum_i \Delta_i^3}{n\epsilon^2} \leq \frac{(\sum_i E\Delta_i^2)^{3/2}}{n\epsilon^2} \leq \frac{C}{n\rho_n^{3/2}\epsilon^2}$$

Therefore, using the sparsity conditions in Assumption 2, we see that the first equation in the lemma statement is proved.

For the second, we use:

$$\begin{aligned}\frac{\sum_i |\hat{g}_1(i)/\rho_n^s|^2}{n} &\geq \frac{\sum_i |g_1(X_i)/\rho_n^s|^2}{n} + \frac{\sum_i \Delta_i^2}{n} - 2\frac{\sum_i |g_1(X_i)/\rho_n^s|\Delta_i}{n} \\ &= C_1 + \alpha B_2 - \beta B_3,\end{aligned}$$

where α, β are positive constants, and B_2, B_3 were defined in Eq B.27. Using Assumption 2 part 1, we see $C_1 > 0$, a.s. Also, now for a small enough constant ϵ, using Eqs B.28 and B.29, we see that the second equation in the lemma statement is proven.

□

B.4.3 Proof of Lemma B.4.1

Proof. Define the following quantities.

$$\gamma_j(t) = \mathrm{E}\{\exp(itb_{nj}Y_j/B_n)\}$$

Also define $\phi_{1,n}$ and $\phi_{2,n}$ as:

$$\begin{aligned}\phi_{1,n}(t) &= e^{-t^2/2}\left[1 + \sum_j \{\gamma_j(t) - 1\} + \frac{t^2}{2}\right] \\ \phi_{2,n}(t) &= -t^2 K_{2,n} e^{-t^2/2}.\end{aligned}$$

Finally define,

$$S_n = \frac{1}{B_n}\sum_j b_{nj}Y_j, \qquad \Delta_{n,m} = \frac{1}{n^{3/2}}\sum_{i=1}^{m-1}\sum_{j=i+1}^{n} d_{nij}\psi(Y_i, Y_j)$$

As in the original proof, we define:

$$\int_{-\infty}^{\infty} e^{itx} d\{\Phi(x) + L_{1,n}\} dx = \phi_{1,n}(t) \tag{B.30a}$$

$$\int_{-\infty}^{\infty} e^{itx} dL_{2,n} dx = it\phi_{2,n}(t) \tag{B.30b}$$

$$\int_{-\infty}^{\infty} e^{itx} E_{2n}(x) = \phi_{1,n}(t) + it\phi_{2,n}(t) \tag{B.30c}$$

Now, for some $c > 0$ to be chosen later, from Esseen's smoothing lemma Petrov (2012) and Eq B.30 we have:

$$\begin{aligned}
&\sup_x \left| P(V_n \leq x) - E_{2n}(x) \right| \\
&\leq \int_{|t| \leq n^{1-c}} |t|^{-1} |\mathrm{E}(e^{itV_n}) - \phi_{1,n}(t) - it\phi_{2,n}(t)| dt + Cn^{c-1} \sup_x \left| \frac{dE_{2,n}(x)}{dx} \right| \\
&\leq \int_{|t| \leq n^{1-c}} |t|^{-1} |\mathrm{E}(e^{itV_n}) - \phi_{1,n}(t) - it\phi_{2,n}(t)| dt + \frac{C_1(|K_{2,n}| + \beta)}{n^{1-c}} \qquad \text{(B.31)}
\end{aligned}$$

The last line is true due to the following argument. Note that, for some v_j in the $|b_{nj}Y_j/B_n|$ ball in the neighborhood of x, for $j \in \{1, \ldots, n\}$,

$$\begin{aligned}
&\frac{dL_{1,n}(x)}{dx} = \sum_{j=1}^{n} \left[\mathrm{E}\{\phi(x - b_{nj}Y_j/B_n)\} - \phi(x) \right] - \frac{1}{2}\Phi'''(x) \\
&= \sum_{j=1}^{n} \mathrm{E} \left\{ -b_{nj}Y_j/B_n\phi'(x) + b_{nj}^2 Y_j^2 / 2B_n^2 \phi''(x) - b_{nj}^3 Y_j^3 / 6B_n^3 \phi'''(v_j) \right\} - \frac{1}{2}\Phi'''(x).
\end{aligned}$$

Thus, we have:

$$\sup_x \left| \frac{dL_{1,n}(x)}{dx} \right| \leq \sum_{j=1}^{n} c_1 \left\{ b_{nj}^2 / B_n^2 |\phi''(x)| + c_2 |b_{nj}/B_n|^3 \mathrm{E}(|Y_j^3|) |\phi'''(v_j)| \right\} + \frac{1}{2} |\Phi'''(x)|$$

$$\begin{aligned}
&\leq C + \mathrm{E}(|X_1|^3)\left(\sum_j |b_{nj}/B_n|^3\right) + C' \\
&\leq C + \beta/n^{1/2} \leq C\beta \qquad \text{Since } \beta \geq 1
\end{aligned}$$

Also note that, for any $\epsilon > 0$, for n large enough,

$$\begin{aligned}
\int_{|t|>n^{\epsilon}} |\phi_{1,n}(t)/t|dt &= O(1/n^{1-c}) \\
\int_{|t|>n^{\epsilon}} |\phi_{2,n}(t)/t|dt &= O(|K_{2,n}|/n^{1-c})
\end{aligned}$$

Thus the main idea is that $E(e^{itV_n})$ behaves like $E(itS_n) + itE(itS_n\Delta_{n,n})$.

$$\int_{|t|\leq n^{1-c}} |t|^{-1}|\mathrm{E}(e^{itV_n}) - \phi_{1,n}(t) - \phi_{2,n}(t)|dt \leq \sum_{j=1}^{4} I_{j,n}$$

Going back to Eq B.31, we break up the first part of the RHS into four parts, and the remainder gets absorbed into $O(|K_{2,n} + \beta|/n^{1-c})$ term in Eq B.31.

$$\begin{aligned}
|I_{1,n}| &= \int_{|t|<n^{\epsilon}} |t|^{-1}\left|\mathrm{E}(e^{itV_n}) - E(itS_n) - itE(itS_n\Delta_{n,n})\right| dt \\
|I_{2,n}| &= \int_{|t|<n^{\epsilon}} |t|^{-1}\left|\mathrm{E}(e^{itS_n}) - \phi_{1,n}(t)\right| dt \\
|I_{3,n}| &= \int_{|t|<n^{\epsilon}} \left|\mathrm{E}(\Delta_{n,n}e^{itS_n}) - \phi_{2,n}(t)\right| dt \\
|I_{4,n}| &= \int_{n^{\epsilon}\leq|t|<n^{1-c}} |t|^{-1}\left|\mathrm{E}(e^{itV_n})\right| dt
\end{aligned}$$

First we will bound some terms which will be used frequently. Since $ab \leq (a^2+b^2)/2$.

$$|K_{2,n}| \leq \frac{C}{n^{3/2}B_n^2} \sum_{1\leq i<j\leq n} (b_{ni}^2 b_{nj}^2 + d_{nij}^2)(1+\lambda)$$

$$\leq \frac{C(1+\lambda)}{n^{3/2}B_n^2}\left\{\left(\sum_j b_{nj}^2\right)^2 + \sum_{i<j} d_{nij}^2\right\}$$
$$\leq \frac{C(1+\lambda)}{n^{3/2}B_n^2}\left(B_n^4 + l_{4,n}n^2\right)$$
$$\leq \frac{C'(1+\lambda)l_{4,n}}{n^{1/2}} \tag{B.32}$$

As for $\Delta_{n,n}$, we have:

$$E\Delta_{n,n}^2 = \frac{\lambda}{n^3}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{nij}^2 = \frac{\lambda l_{4,n}}{n}$$

Furthermore we will use:

$$R(z) := e^{iz} - 1 - iz \qquad |R(z)| \leq |z|^{\alpha} \text{ for all } \alpha \in [1,2] \tag{B.33}$$

We will first bound $I_{1,n}$. Using Taylor expansion, for some $|\eta| \leq 1$,

$$|I_{1,n}| \leq \int_{|t|<n^{\epsilon}} |t|^{-1}t^2/2|\mathrm{E}(\Delta_{n,n}^2 e^{itS_n}e^{it\Delta_{n,n}\eta})|dt$$
$$\leq 1/2\int_{|t|<n^{\epsilon}} |t|\mathrm{E}(\Delta_{n,n}^2)dt \leq C\frac{(1+\lambda)l_{4,n}}{n^{1-2\epsilon}}$$

Next we bound $I_{2,n}$. Using a similar argument in the proof of the original version of this theorem, we have:

$$|I_{2,n}| \leq \frac{C_1}{B_n^4}\sum_j b_{nj}^4 + C_2\left(\frac{1}{B_n^3}\sum_j |b_{nj}|^3 E(|X_1|^3)\right)^2$$
$$\leq C_1 n^{-2/3} + C_2\lambda^2/n$$

Now we do $I_{3,n}$. Denote $Z_j = b_{nj}Y_j/B_n$ and $\psi_{ij} = d_{nij}\psi(Y_i, Y_j)$. First note that

$$\mathrm{E}\{\psi_{ij}e^{it(Z_i+Z_j)}\} = -t^2\ell_{ij} + \theta_{1ij}(t), \tag{B.34}$$

where we have:

$$\ell_{ij} = \mathrm{E}\left(|\psi_{ij} Z_i Z_j|\right) \le |b_{ni} b_{nj} d_{nij}|/B_n^2 |\mathrm{E}\{Y_i Y_j \psi(Y_i, Y_j)\}| \le \lambda^{1/2}(b_{ni}^2 b_{nj}^2 + d_{nij}^2)/B_n^2. \tag{B.35}$$

Using Eq B.33 and the fact that $\mathrm{E}\{\psi(Y_i, Y_j)\} = 0$ and $\mathrm{E}\{\psi(Y_i, Y_j) \mid Y_i\} = 0$,

$$\begin{aligned}
\theta_{1,i,j} &= \mathrm{E}(\psi_{ij}[it\{Z_i R(tZ_j) + Z_j R(tZ_i)) + R(tZ_i)R(tZ_j)\}]) \\
&\le C|t|^{2.5}\mathrm{E}\left(|\psi_{ij} Z_i Z_j^{1.5}| + |\psi_{ij} Z_j Z_i^{1.5}|\right) \\
&\le C|t|^{2.5}\mathrm{E}\{|Y_1 Y_2^{1.5}\psi(Y_i, Y_j)|\}\left(|d_{nij} b_{ni} b_{nj}^{1.5}/B_n^{2.5}| + |d_{nij} b_{ni}^{1.5} b_{nj}/B_n^{2.5}|\right) \\
&\le C|t|^{2.5}(\lambda\beta)^{1/2}\left(d_{nij}^2 + b_{ni}^2|b_{nj}|^3 + |b_{ni}|^3 b_{nj}^2\right) n^{-5/4}
\end{aligned}$$

Using Eq B.34, and setting $\prod_{k \ne i,j} \gamma_k(t) = e^{-t^2/2} + \theta_{2,i,j}$ we see:

$$\begin{aligned}
E(\Delta_{n,n} e^{itS_n}) &= n^{-3/2}\sum_{i<j}\mathrm{E}(\psi_{ij} e^{itS_n}) = n^{-3/2}\sum_{i<j}\mathrm{E}\{\psi_{ij} e^{it(Z_i + Z_j)}\}\prod_{k \ne i,j}\gamma_k(t) \\
&= n^{-3/2}\sum_{i<j}\{-t^2\ell_{i,j} + \theta_{1,i,j}(t)\}\left(e^{-t^2/2} + \theta_{2,i,j}\right) \\
&= n^{-3/2}\sum_{i<j}\{-t^2\ell_{i,j} e^{-t^2/2} + \theta_{3,i,j}(t)\} = \underbrace{-K_{2,n} t^2 e^{-t^2/2}}_{\phi_{2,n}(t)} + \underbrace{n^{-3/2}\sum_{i<j}\theta_{3,i,j}}_{R_{n,4}},
\end{aligned}$$

where using Lemma A.4 in Wang and Jing (2004), for $|t| < n^{\epsilon} << n^{1/6}$

$$|\theta_{2,i,j}| \le \frac{C}{n^{1/2}}\left(\beta + \frac{b_{nj}^2 + b_{ni}^2}{n^{1/2}}\right)(t^2 + t^4)e^{-t^2/8}$$

Furthermore, using Lemma A.4 and $\sum_i |b_{ni}|^3 \le \ell_2 n$

$$|\theta_{3,i,j}| \le t^2|\ell_{i,j}\theta_{2,i,j}| + |\theta_{1,i,j}|\prod_{k \ne i,j}\gamma_k(t) \le t^2|\ell_{i,j}\theta_{2,i,j}| + 4|\theta_{1,i,j}|e^{-t^2/8}$$

$$|\ell_{i,j}\theta_{2,i,j}| \leq \frac{C}{n^{1/2}}|\ell_{i,j}|\left(\beta + \frac{b_{nj}^2 + b_{ni}^2}{n^{1/2}}\right)(t^2+t^4)e^{-t^2/8}$$

Summing the above expression over $i < j$, we also have,

$$\begin{aligned}
&\sum_{i<j}|\ell_{i,j}\theta_{2,i,j}| \\
&\leq C\lambda^{1/2}\left(\beta\frac{B_n^4 + l_{4,n}n^2}{n^{3/2}} + \frac{\sum_{i<j}|d_{nij}|(|b_{ni}^3 b_{nj}| + |b_{ni}b_{nj}^3|)}{n^2}\right)(t^2+t^4)e^{-t^2/8} \\
&\leq C\lambda^{1/2}\left[\beta\frac{B_n^4 + l_{4,n}n^2}{n^{3/2}} + \frac{c}{n^2}\underbrace{\left\{\left(\sum_{i<j}d_{nij}^2\right)\left(\sum_{i<j}b_{ni}^6 b_{nj}^2\right)\right\}^{1/2}}_{A}\right](t^2+t^4)e^{-t^2/8} \quad \text{(B.36)}
\end{aligned}$$

To bound (A) we see:

$$\begin{aligned}
(A) &\leq \left\{(n^2 l_{4,n})(n\ell_2)\sum_{i<j}b_{ni}^3 b_{nj}^2\right\}^{1/2} \leq n^{3/2}(\ell_2 l_{4,n})^{1/2}\left\{\left(\sum_i b_{ni}^3\right)\left(\sum_j b_{nj}^2\right)\right\}^{1/2} \\
&\leq c' n^{5/2} l_{4,n}^{1/2}\ell_2
\end{aligned}$$

Plugging this back in Eq B.36, and assuming WLOG $l_{4,n} \geq 1$,

$$\begin{aligned}
\sum_{i<j}|\ell_{i,j}\theta_{2,i,j}| &\leq C'\lambda^{1/2}\left(\beta\frac{B_n^4 + l_{4,n}n^2}{n^{3/2}} + \frac{1}{n^2}n^{5/2}\ell_2 l_{4,n}^{1/2}\right)(t^2+t^4)e^{-t^2/8} \\
&\leq C' l_{4,n}\lambda^{1/2}\beta n^{1/2}(t^2+t^4)e^{-t^2/8}
\end{aligned}$$

Finally, we also have:

$$\sum_{i<j}|\theta_{1,i,j}| \leq C|t|^{2.5}(\lambda\beta)^{1/2}\left(l_{4,n}n^2 + 2\ell_2 n B_n^2\right)n^{-5/4} \leq C|t|^{2.5}(\lambda\beta)^{1/2}l_{4,n}n^{1/4}$$

Finally we have, since $t^4 \leq |t| + |t|^6$, and $|t|^{2.5} \leq |t| + |t|^6$,

$$R_{n,4} \leq n^{-3/2}\sum_{i<j}|\theta_{3,i,j}| \leq n^{-3/2}\left(\sum_{i<j}t^2|\ell_{i,j}\theta_{2,i,j}| + 4\sum_{i<j}|\theta_{1,i,j}|e^{-t^2/8}e^{-t^2/8}\right)$$

$$
\begin{aligned}
&\leq \left(t^2 l_{4,n}\lambda^{1/2}(t^2+t^4)e^{-t^2/8}n^{-1} + |t|^{2.5}(\lambda\beta)^{1/2}l_{4,n}n^{-3/4}e^{-t^2/8}\right)\\
&\leq l_{4,n}\left(\lambda^{1/2}\beta n^{-1} + (\lambda\beta)^{1/2}n^{-3/4}\right)(|t|+|t|^6)e^{-t^2/8}\\
&\leq C' l_{4,n}\left(\beta^2 n^{-1} + (\lambda+\beta)n^{-3/4}\right)(|t|+|t|^6)e^{-t^2/8}
\end{aligned}
$$

Finally, for $I_{3,n}$, we have:

$$
\begin{aligned}
|I_{3,n}| &\leq \int_{|t|\leq n^{\epsilon}} |t|^{-1}R_{n,4}dt\\
&\leq C' l_{4,n}\left(\beta^2 n^{-1} + (\lambda+\beta)n^{-3/4}\right)\int_{|t|\leq n^{\epsilon}}(1+|t|^5)e^{-t^2/8}dt\\
&\leq C'1' l_{4,n}\left(\beta^2 n^{-1} + (\lambda+\beta)n^{-3/4}\right)
\end{aligned}
$$

Now we will bound $I_{4,n}$.

Define $\Omega := \{k : \min(1/2, \ell_2/\ell_1^{3/2}) \leq n^{1/2}b_{n,k}/B_n \leq 2\ell_2/\ell_1^{3/2}\}$. Using Lemma A.5 in Wang and Jing (2004), we see that $|\Omega| \geq c_0 n$, for some $c_0 \in (0,1)$.

Now, let $\Gamma := \{i \mid \alpha_i \geq \bar{\alpha} + ks_n\}$. Applying Lemma B.4.2 and setting $k = \sqrt{2/c_0}$, we see that $|\Gamma^c| \geq n(1-c_0/2)$. Therefore, $|\Gamma^c \cap \Omega| \geq nc_0/2$. Let $k_0 = \lfloor c_0/2 \rfloor$.

WLOG assume $b_{n,1}\ldots b_{n,k_0 n} \in \Omega \cap \Gamma^c$ and $\ell_2/\ell_1^{3/2} \geq 1/2$. Now for $m \in [2, k_0 n]$, we have:

$$
S_m = \frac{1}{B_n}\sum_{k=1}^{m} b_{nk}Y_k \qquad S_m^{i,j} := \frac{1}{B_n}\sum_{k\neq i,j} b_{nk}X_k
$$

For $1, \ldots, m \leq k_0 n$, we have:

$$\frac{1}{mn} \sum_{i=1} \sum_{j=1, j \neq i} d_{nij}^2 = \frac{1}{m} \sum_{i=1}^{m} \alpha_i \leq l_{4,n} + k s_n =: \ell_{5,n}$$

As for $\Delta_{n,m}$, we have:

$$E(\Delta_{n,m}^2) = \frac{\lambda}{n^3} \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} d_{nij}^2 \leq \lambda l_{5,n} \frac{m}{n^2} \tag{B.37}$$

Now we use the decomposition in Bickel et al. (1986) (17)-(22).

$$\begin{aligned}
\mathrm{E}(e^{itV_n}) &= \mathrm{E}\{e^{it(V_n - \Delta_{n,m})} e^{it\Delta_{n,m}}\} \\
&= \mathrm{E}\{e^{it(V_n - \Delta_{n,m})}(1 + it\Delta_{n,m})\} + R_{n,5} \\
&= \mathrm{E}\{e^{it(V_n - \Delta_{n,m})}(1 + it\Delta_{n,m})\} + Ct^2 \lambda l_{5,n} m/n^2 \\
&= \mathrm{E}\{e^{it(V_n - \Delta_{n,m})}\} + \frac{it}{n^{3/2}} \sum_{i=1}^{m} \sum_{j=i+1}^{n} \underbrace{\mathrm{E}\{e^{it(V_n - \Delta_{n,m})} \psi_{ij}\}}_{D_{ij}} + Ct^2 \lambda l_{5,n} \frac{m}{n^2}
\end{aligned}$$

where the last line is obtained using Eqs B.33 and B.37, as follows:

$$R_{n,5} \leq |\mathrm{E}\{e^{it(V_n - \Delta_{n,m})} t^2 \Delta_{n,m}^2\}| \leq Ct^2 \lambda l_{5,n} m/n^2$$

Note that $V_n - \Delta_{n,m}$ can be written as $S_{m-1} + Y_{m,n}$, where $Y_{m,n}$ does not depend on $Y_1, \ldots, Y_{m-1}$. So we will write:

$$\begin{aligned}
\left| \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} (D_{ij}) \right| &= \left| \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} \mathrm{E}\{e^{it(S_{m-1}^{ij} + \psi_{ij} + Y_{m,n})} \psi_{ij}\} \right| \\
&= \left| \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} \mathrm{E}\{e^{itS_{m-1}^{ij}}\} \mathrm{E}\{e^{(\psi_{ij} + Y_{m,n})} \psi_{ij}\} \right|
\end{aligned}$$

$$\le \sup_{i<j} |\mathrm{E}(e^{itS^{ij}_{m-1}})| \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} \mathrm{E}(|\psi_{ij}|)$$

$$\le \sup_{i<j} |\mathrm{E}(e^{itS^{ij}_{m-1}})| \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} |d_{nij}| \mathrm{E}\{|\psi(Y_i, Y_j)|\}$$

$$\le \lambda^{1/2} \sup_{i<j} |\mathrm{E}(e^{itS^{ij}_{m-1}})| \sqrt{mn \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} d_{nij}^2}$$

$$\le \sqrt{\lambda l_{5,n}} \sup_{i<j} |\mathrm{E}(e^{itS^{ij}_{m-1}})| mn$$

Plugging it back, we have:

$$|\mathrm{E}(e^{itV_n})| \le |\mathrm{E}(e^{itS_{m-1}})| + \frac{|t|}{n^{3/2}} \sqrt{\lambda l_{5,n}} \sup_{i<j} |\mathrm{E}(e^{itS^{ij}_{m-1}})| mn + Ct^2 \lambda l_{5,n} \frac{m}{n^2} \qquad \text{(B.38)}$$

Now, we have for $|t| \le 1/4n^{1/2}/\mathrm{E}(|Y_1|^3)$

$$|\mathrm{E}(e^{itS_m})| \le e^{-c_0 mt^2/n} \qquad |\mathrm{E}(e^{itS^{ij}_m})| \le e^{-c_0(m-2)t^2/n}$$

Taking $m = [6n \log n / c_0 t^2] + 1$ (for a large enough ϵ, this is still smaller than $k_0 n$), from Eq B.38 we have:

$$\int_{n^\epsilon \le |t| < 1/4n^{1/2}/\mathrm{E}(|Y_1|^3)} |t|^{-1} \left| \mathrm{E}(e^{itV_n}) \right| dt$$

$$\le \int_{n^\epsilon \le |t| < 1/4n^{1/2}/\mathrm{E}(|Y_1|^3)} \left(\frac{e^{-c_0 mt^2/n}}{|t|} + \frac{m}{n^{1/2}} \sqrt{\lambda l_{5,n}} e^{-c(m-2)t^2/n} + C|t| \lambda l_{5,n} \frac{m}{n^2} \right) dt$$

$$\le C' \lambda l_{5,n} \frac{\log^2 n}{n}$$

Now we will deal with the range $1/4n^{1/2}/\mathrm{E}(|Y_1|^3) \le |t| \le n^{1-c}$. Since $\kappa(Y_1) > 0$, and hence for large enough n,

$$|\gamma_k(t)| \le 1 - \kappa(Y_1)$$

$$|\mathrm{E}(e^{itS_m})| \leq e^{-m\kappa(Y_1)}$$

$$\mathrm{E}(e^{itS_m^{ij}})| \leq e^{-(m-2)\kappa(Y_1)}$$

Using this in conjunction with Eq B.38, and setting $m = [4\log n/\kappa(Y_1)] + 2$,

$$\begin{aligned}
&\int_{1/4n^{1/2}/\mathrm{E}(|Y_1|^3)\leq|t|\leq n^{1-c}} |t|^{-1}\left|\mathrm{E}(e^{itV_n})\right| dt \\
&\int_{1/4n^{1/2}/\mathrm{E}(|Y_1|^3)\leq|t|\leq n^{1-c}} \left(\frac{e^{-\kappa(Y_1)m}}{|t|} + \frac{m}{n^{1/2}}\sqrt{\lambda l_{5,n}}e^{-\kappa(Y_1)(m-2)} + C|t|\lambda l_{5,n}\frac{m}{n^2}\right) \\
&\leq C'\frac{\rho_n l_{5,n}}{\kappa(Y_1)}\frac{\log n}{n^2}n^{2(1-c)} = C'\frac{\lambda l_{5,n}}{\kappa(Y_1)}\frac{\log n}{n^{2c}}
\end{aligned}$$

Thus, using the bounds on $I_{1,n}, I_{2,n}, I_{3,n}$ and $I_{4,n}$ along with Eq B.31, Eq B.32 we get:

$$\begin{aligned}
&\sup_x |P(V_n \leq x) - E_{2n}(x)| \\
&\leq \sum_{i=1}^{4} I_{n,i} + C'\frac{\beta + (1+\lambda)l_{4,n}/n^{1/2}}{n^{1-c}} \\
&\leq C\left(\frac{(1+\lambda)l_{4,n}}{n^{1-2\epsilon}} + n^{-2/3} + l_{4,n}(\lambda+\beta+\beta^2)n^{-3/4} + \frac{\lambda l_{5,n}}{\kappa(Y_1)}\frac{\log n}{n^{2c}}\right) \\
&+ C'\frac{\beta + (1+\lambda)l_{4,n}/n^{1/2}}{n^{1-c}} \\
&\leq (l_{4,n} + ks_n)\frac{\log n}{n^{2/3}}
\end{aligned}$$

The last line assumes β, λ and $\kappa(Y_1)$ are all bounded. □

B.4.4 Proof of Lemma B.4.6

Proof. Let $X \sim N(1, c_1^2)$ and $Y \sim N(1, c_2^2)$ be two independent random variables. We have $\xi_1 = XY$.

$$\mathrm{E}(|XY-1|^3) \leq \mathrm{E}(|XY|^3) + 1 + 3\mathrm{E}(X^2|Y|) + 3\mathrm{E}(|X|Y^2)$$

$$= \mathrm{E}(|X|^3)\mathrm{E}(|Y|^3) + 3\mathrm{E}(X^2)\mathrm{E}(|Y|) + 3\mathrm{E}(|X|)\mathrm{E}(Y^2)$$

$$< \infty$$

The last step is true because both $\mathrm{E}(|X|^3)$ and $\mathrm{E}(|Y|^3)$ are bounded for bounded c_1, c_2. □

B.5 Detailed Results for Smooth Functions of Counts

In this section, we establish Edgeworth expansions for smooth functions of counts for the bootstrap and show that they are close to Edgeworth expansions of the conditional expectation of the count statistic, which is a U-statistic. To our knowledge, Edgeworth expansions for smooth functions are not explicitly stated in the literature even for U-statistics. It turns out that the non-negligible terms arising from a Taylor approximation of the smooth functional are of a form where a flexible Edgeworth expansion result of Jing and Wang (2010) may be invoked. Edgeworth expansions for smooth functionals are also considered in Hall (2013), but the argument provided there requires multivariate Edgeworth expansions and depends heavily on the properties of cumulants of independent random variables, complicating extensions even to U-statistics.

Since the Edgeworth expansion of the conditional expectation requires a non-lattice condition, it is assumed below. However, it is likely that this condition can be removed if one derives an Edgeworth expansion for the count functional directly and uses a proof strategy similar to Zhang and Xia (2020) that exploits the smoothing nature of Bernoulli noise.

B.5.1 Edgeworth Expansion for Smooth Functions of Counts

In what follows, let $f : \mathbb{R}^d \mapsto \mathbb{R}$ denote the smooth function of interest, u denote a d-dimensional vector of conditional expectations corresponding to scaled count functionals $\{\frac{\hat{T}_n^{(1)}}{\rho_n^{s_1}}, \dots, \frac{\hat{T}_n^{(d)}}{\rho_n^{s_d}}\}$ given X, and $\mu = E(u)$. In this section, we consider Edgeworth expansions for the statistic:

$$S_n = n^{1/2}(f(u) - f(\mu))/\sigma_f,$$

where σ_f^2 is the asymptotic variance of S_n. The standard Delta Method involves a first-order Taylor expansion, resulting in a Normal approximation with rate $O(1/\sqrt{n})$ when the gradient is not equal to $\mathbf{0}$ at μ. To attain higher-order correctness, we need to consider a second-order expansion. Recall the derivatives of interest $a_i, 1 \leq i \leq d$ and $a_{ij}, 1 \leq i, j \leq d$ defined in Eq A.8. Furthermore, define the following analog the moments of the linear component of the U-statistic:

$$\lambda_{i_1,\dots,i_j} = E\left\{\left(\frac{r_{i_1} g_1^{(i_1)}(X_l)}{\rho_n^{s_{i_1}}}\right) \dots \left(\frac{r_{i_d} g_1^{(i_d)}(X_l)}{\rho_n^{s_{i_d}}}\right)\right\}.$$

In the proposition below, we state the form of the Edgeworth for an appropriately smooth function f.

Proposition B.5.1. *Suppose that $\sigma_f > 0$, the function f has three continuous derivatives in a neighbourhood of μ, and $\sum_{i=1}^d a_i g_1^{(i)}(X_l)$ is non-lattice. Then,*

$$P(S_n \leq x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + o\left(\frac{1}{n^{1/2}}\right),$$

$$p_1(x) = -\{A_1 \sigma_f^{-1} + \frac{1}{6} A_2 \sigma_f^{-3}(x^2 - 1)\},$$

where σ_f^2, A_1 and A_2 are given by:

$$\sigma_f^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} a_i a_j \lambda_{ij},$$

$$A_1 = \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} a_{ij} \lambda_{ij},$$

$$A_2 = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} a_i a_j a_k \lambda_{ijk} + 3 \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{t=1}^{d} a_i a_j a_{kt} \lambda_{ik} \lambda_{jt}$$
$$+ 3 \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} a_i a_j a_k E \left(\frac{r_i g_1^{(i)}(X_{i_1})}{\rho^{s_i}} \frac{r_j g_1^{(j)}(X_{i_2})}{\rho^{s_j}} \frac{r_k (r_k - 1) g_2^{(k)}(X_{i_1}, X_{i_2})}{\rho^{s_k}} \right).$$

Before stating the proof in detail we add an auxiliary lemma, which is needed to the bound the contribution of a remainder term.

Lemma B.5.1. *Let $S_n = V_n + c/\sqrt{n}$. Let $P(V_n \le x) = \Phi(x) + p_1(x)\phi(x)/\sqrt{n} + o(1/\sqrt{n})$. Then we have:*

$$P(S_n \le x) = \Phi(x) + (p_1(x) + c)\phi(x)/\sqrt{n} + o(1/\sqrt{n}) \tag{B.39}$$

Proof. We have:

$$\begin{aligned} P(S_n \le x) &= P(V_n \le x - c/\sqrt{n}) \\ &= \Phi(x - c/\sqrt{n}) + \frac{p_1(x - c/\sqrt{n})}{\sqrt{n}} \phi(x - c/\sqrt{n}) + o(1/\sqrt{n}) \end{aligned}$$

Note that $\sup_x |\phi(x)| \le C$ for some universal constant C. So we have:

$$\Phi(x - c/\sqrt{n}) = \Phi(x) - c/\sqrt{n}\phi(x) + O(1/n),$$

and

$$|\phi(x - c/\sqrt{n}) - \phi(x)| = O(1/n).$$

Using the above two equations with Eq B.39, we attain the stated result. □

Now we present the result in Proposition B.5.1.

Proof. A second-order Taylor expansion yields:

$$n^{1/2}(f(u)-f(\mu)) = n^{1/2} < u-\mu, \nabla f(\mu) > +\frac{1}{2}n^{1/2}(u-\mu)^T H(\mu)(u-\mu)+O_P\left(\frac{1}{n}\right). \tag{B.40}$$

Furthermore, by a multivariate Hoeffding Decomposition for $u-\mu$:

$$\begin{aligned} u-\mu &= \frac{1}{n}\left\{ \begin{array}{c} \frac{r_1}{\rho_n^{s_1}} \sum_{i=1}^n g_1^{(1)}(X_i) \\ \cdots \\ \frac{r_d}{\rho_n^{s_d}} \sum_{i=1}^n g_1^{(d)}(X_i) \end{array} \right\} + \frac{1}{n^2}\left\{ \begin{array}{c} \frac{r_1(r_1-1)}{\rho_n^{s_1}} \sum_{i<j} g_2^{(1)}(X_i,X_j) \\ \cdots \\ \frac{r_d(r_d-1)}{\rho_n^{s_d}} \sum_{i<j} g_2^{(d)}(X_i,X_j) \end{array} \right\} + O_P\left(\frac{1}{n^{3/2}}\right) \\ &= \frac{1}{n}u_L + \frac{1}{n^2}u_Q + O_P\left(\frac{1}{n^{3/2}}\right), \end{aligned} \tag{B.41}$$

where

$$n^{-1/2}||u_L|| = O_P(1), \qquad \frac{1}{n^{3/2}}||u_Q|| = O_P\left(n^{-1/2}\right).$$

Now for the first term in Eq B.40, we have

$$n^{1/2} < u-\mu, \nabla f(\mu) >= n^{-1/2} < u_L, \nabla f(\mu) > +\frac{1}{n^{3/2}} < u_Q, \nabla f(\mu) > +O_P\left(\frac{1}{n}\right).$$

The second term in Eq B.40 is

$$\begin{aligned} & n^{1/2}(u-\mu)H(\mu)(u-\mu) \\ &= n^{1/2}\left\{\frac{1}{n}u_L + \frac{1}{n^2}u_Q + O_P\left(\frac{1}{n^{3/2}}\right)\right\}^T H(\mu)\left\{\frac{1}{n}u_L + \frac{1}{n^2}u_Q + O_P\left(\frac{1}{n^{3/2}}\right)\right\} \\ &= n^{1/2}\left\{\frac{1}{n^2}u_L^T H(\mu)u_L + \frac{2}{n^3}u_L^T H(\mu)u_Q + O_P\left(\frac{1}{n^{3/2}}\right)\right\} \end{aligned}$$

$$= \frac{1}{n^{3/2}} u_L^T H(\mu) u_L + O_P\left(\frac{1}{n}\right)$$

Now Eq B.40 may be expressed as:

$$\begin{aligned} n^{1/2}(f(u) - f(\mu)) = n^{-1/2} &< u_L, \nabla f(\mu) > \\ &+ \frac{1}{n^{3/2}} \left\{< u_Q, \nabla f(\mu) > + \frac{1}{2} u_L^T H(\mu) u_L \right\} + O_P\left(\frac{1}{n}\right). \end{aligned}$$

We have,

$$S_n = \frac{A_1}{\sqrt{n}\sigma_f} + n^{-1/2}\alpha(X_l) + n^{-3/2} \sum_{l<m} \beta(X_l, X_m) + O_P\left(\frac{1}{n}\right), \tag{B.42}$$

where

$$\begin{aligned} \alpha(X_l) &= \frac{1}{\sigma_f} \sum_{i=1}^{d} a_i g_1^{(i)}(X_l) \frac{r_l}{\rho_n^{s_l}}, \\ \beta(X_l, X_m) &= \frac{1}{\sigma_f} \left\{ \sum_{i=1}^{d} a_i \frac{r_i(r_i - 1)}{\rho_n^{s_i}} g_2^{(i)}(X_l, X_m) + \sum_{i,j} a_{ij} \frac{r_i r_j}{\rho_n^{s_i+s_j}} g_1^{(X_i)}(l) g_1^{(j)}(X_m) \right\} \end{aligned}$$

Applying Theorem 2.1 of Jing and Wang (2010), under the conditions of proposition B.5.1, we have

$$\sup_x \left| P\left(S_n - \frac{A_1}{\sqrt{n}\sigma_f} \le x\right) - E_n(x) \right| = o(n^{-1/2}),$$

where

$$E_n(x) = \Phi(x) - \frac{(x^2 - 1)\phi(x)}{6\sqrt{n}} \{E\alpha(X_l)^3 + 3E\alpha(X_l)\alpha(X_m)\beta(X_l, X_m)\}.$$

Using Lemma B.5.1 and definition of A_2, we can simply $E_n(x)$, yielding:

$$\begin{aligned} P(S_n \le x) &= \Phi(x) + n^{1/2}\phi(x) p_1(x) + o(n^{-1/2}), \\ p_1(x) &= -\left\{ A_1 \sigma_f^{-1} + \frac{1}{6} A_2 \sigma_f^{-3} (x^2 - 1) \right\}. \end{aligned}$$

□

B.5.2 Proposed Bootstrap for Smooth Functions of Counts

In this section, we consider Edgeworth expansions of smooth functions for the bootstrap. Recall from Section 3.5.3 that u^* denotes a d-dimensional vector of bootstrapped counted functionals generated by either the multiplier bootstrap $\hat{T}^*_{n,M}$ or the the quadratic bootstrap $\hat{T}^*_{n,Q}$; in the latter case, one may ignore an additional $O_P(n^{-3/2})$ term that arises from approximating a U-statistic by the first two terms of the Hoeffding decomposition. Now recall the bootstrap analogue S^*_n from Eq 3.25, the gradients of the smooth function evaluated at the empirical counts from Eq 3.27.

Let P^* denote the bootstrap measure conditioned on A and X, with randomness arising from the multiplier weights ξ. Furthermore, let $\hat{P}_n$ denote the the empirical measure $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. It will turn out these two measures are closely related. With a slight abuse of notation, the expectation operator corresponding to $\hat{P}_n$ will be denoted by $\hat{E}_n f(X) = \frac{1}{n}\sum_{i=1}^n f(X_i)$.

We define the following empirical analogues of the moments of interest:

$$\begin{aligned}\tilde{\lambda}_{i_1,\dots,i_j} = E^*\left\{\left(r_{i_1}\frac{\hat{g}_1^{(i_1)}(l)V_l}{\rho_n^{s_{i_1}}}\right)\dots\left(r_{i_d}\frac{\hat{g}_1^{(i_d)}(l)V_l}{\rho_n^{s_{i_d}}}\right)\right\} &= \frac{1}{n}\sum_{l=1}^{n}\left(r_{i_1}\frac{\hat{g}_1^{(i_1)}(l)}{\rho_n^{s_{i_1}}}\right)\dots\left(r_{i_d}\frac{\hat{g}_1^{(i_d)}(l)}{\rho_n^{s_{i_d}}}\right)\\ = \hat{E}_n\left\{\left(r_{i_1}\frac{\hat{g}_1^{(i_1)}(l)}{\rho_n^{s_{i_1}}}\right)\dots\left(r_{i_d}\frac{\hat{g}_1^{(i_d)}(l)}{\rho_n^{s_{i_d}}}\right)\right\}.\end{aligned}$$

Now recall that the empirical analogue of the asymptotic variance from Eq 3.29. We now prove Theorem 8, which establishes an Edgeworth expansion for $P^*(S^*_n \leq x)$.

B.5.3 Proof of Theorem 8

Proof. We will start by establishing Eq 3.30. Let V_l be $\xi_l - 1$ and let V denote the vector:

$$V = (\xi_1 - 1,\ \ldots,\ \xi_n - 1)^T.$$

Given A and X, we have

$$\begin{aligned} u^* - \hat{u} &= \frac{1}{n} \left\{ \begin{array}{c} \frac{r_1}{\rho_n^{s_1}} \sum_{l=1}^n \hat{g}_1^{(1)}(l)V_l \\ \cdots \\ \frac{r_d}{\rho_n^{s_d}} \sum_{l=1}^n \hat{g}_1^{(d)}(l)V_l \end{array} \right\} + \frac{1}{n^2} \left\{ \begin{array}{c} \frac{r_1(r_1-1)}{\rho_n^{s_1}} \sum_{l<m} \tilde{g}_2^{(1)}(l,m)V_lV_m \\ .. \\ \frac{r_d(r_d-1)}{\rho_n^{s_d}} \sum_{l<m} \tilde{g}_2^{(d)}(l,m)V_lV_m \end{array} \right\} + O_P\left(\frac{1}{n^{3/2}}\right) \\ &= \frac{1}{n}u_L^* + \frac{1}{n^2}u_Q^* + O_P\left(\frac{1}{n^{3/2}}\right). \end{aligned}$$

Using a second-order Taylor expansion analogous to EqB.40, we have:

$$\begin{aligned} n^{1/2}(f(u^*) - f(\hat{u})) &= n^{-1/2} < u_L^*, \nabla f(\hat{u}) > \\ &\quad + \frac{1}{n^{3/2}} \left\{ < u_Q^*, \nabla f(\hat{u}) > + \frac{1}{2} u_L^{*T} H(\hat{u}) u_L^* \right\} + O_P\left(\frac{1}{n}\right). \quad \text{(B.43)} \end{aligned}$$

We also have, by definition,

$$E^*\{\hat{g}_1^{(i)}(l)\hat{g}_1^{(j)}(m)\tilde{g}_2^{(k)}(l,m)V_lV_m\} = \hat{E}\{\hat{g}_1^{(i)}(l)\hat{g}_1^{(j)}(m)\tilde{g}_2^{(k)}(l,m)\}.$$

Then, by Eq B.43 and definition of $\tilde{\sigma}_f$ and $\tilde{A}_1$, we have,

$$\begin{aligned} S_n^* &= \frac{n^{1/2}(f(u^*) - f(\hat{u}))}{\tilde{\sigma}_f} \\ &= \frac{\tilde{A}_1}{\sqrt{n}\tilde{\sigma}_f} + \frac{1}{B_n} \sum_{l=1}^n b_{n,l}V_l + \frac{1}{n^{3/2}} \sum_{l<m} d_{n,lm}\psi(V_l, V_m) + O_P\left(\frac{1}{n}\right), \end{aligned}$$

where

$$b_{n,l} = \frac{1}{\tilde{\sigma}_f} \sum_{i=1}^{d} \hat{a}_i \hat{g}_1^{(i)}(l) \frac{r_i}{\rho_n^{s_i}}, \tag{B.44a}$$

$$B_n^2 = \sum_{l=1}^{n} b_{n,l}^2 = n, \tag{B.44b}$$

$$d_{n,lm} = \frac{1}{\tilde{\sigma}_f} \left\{ \sum_{i=1}^{d} \hat{a}_i \frac{r_i(r_i-1)}{\rho_n^{s_i}} \tilde{g}_2^{(i)}(l,m) + \sum_{i,j} \hat{a}_{ij} \frac{r_i r_j}{\rho_n^{s_i+s_j}} \hat{g}_1^{(i)}(l) \hat{g}_1^{(j)}(m) \right\}, \tag{B.44c}$$

$$\psi(V_l, V_m) = V_l V_m. \tag{B.44d}$$

Using Lemma B.4.1 and similar arguments therein if

$$\frac{1}{n} \sum_{l=1}^{n} b_{n,l}^2 \geq l_1 > 0, \;\; \frac{1}{n} \sum_{l=1}^{n} |b_{n,l}|^3 \leq l_2 \leq \infty, \tag{B.45}$$

then

$$\sup_x |P^* \left(S_n^* - \frac{\tilde{A}_1}{\sqrt{n}\tilde{\sigma}_f} \leq x \right) - \tilde{G}(x)| = O\left(\frac{l_{5,n} \log n}{n^{2/3}} \right), \tag{B.46}$$

where and $\alpha_l := \frac{1}{n} \sum_{m \neq l} d_{n,lm}^2$. and for sufficiently large k:

$$l_{4,n} = \frac{1}{n} \sum_{l=1}^{n} \alpha_l, \quad s_n^2 = \frac{1}{n} \sum_{l} \alpha_l^2 - (l_{4,n})^2, \quad l_{5,n} = l_{4,n} + k s_n$$

and

$$\tilde{G}_n(x) = \Phi(x) + \tilde{L}_{1,n}(x) + \tilde{L}_{2,n}(x),$$
$$\tilde{L}_{1,n} = \frac{E V_l^3}{6 B_n^3} \sum_{l=1}^{n} b_{n,l}^3 (x^2-1)\phi(x),$$
$$\tilde{L}_{2,n} = \frac{1}{n^{3/2} B_n^2} \sum_{l<m} b_{n,l} b_{n,m} d_{n,lm} E(V_l V_m \psi(V_l, V_m))(x^2-1)\phi(x).$$

Since $\sum_i b_{n,i}^2/n = 1$, the first condition in EqB.45 is satisfied. Let $c_j := \frac{r_j}{\rho_n^{s_j}\hat{\sigma}_f}$. For the second condition, note that since $|.|^3$ is convex,

$$\frac{1}{n}\sum_i |b_{n,i}|^3 \leq \frac{d^2}{\hat{\sigma}_f^3}\sum_{j=1}^d c_j^3 \frac{1}{n}\sum_i |\hat{a}_i \hat{g}_1(i)|^3$$

Since the function f has three gradients in the neighborhood of μ, Lemma B.4.5 shows that the above is bounded, thereby satisfying the second condition in Eq B.45.

Simplifying $\tilde{L}_{1,n}$ and $\tilde{L}_{2,n}$ using Eq B.44, we have

$$\tilde{G}_n(x) = \Phi(x) + n^{-1/2}\phi(x)\frac{1}{6}\tilde{A}_2\tilde{\sigma}_f^{-3}(x^2-1).$$

Now we bound the remainder term by bounding $l_{5,n}$. We write α_l as

$$\alpha_l = \frac{1}{n\tilde{\sigma}_f^2}\sum_{m\neq l}\left\{\underbrace{\sum_{i=1}^d \hat{a}_i \frac{r_i(r_i-1)}{\rho_n^{s_i}}\tilde{g}_2^{(i)}(l,m)}_{Y_{1,lm}} + \underbrace{\sum_{i,j}\hat{a}_{ij}\frac{r_i r_j}{\rho_n^{s_i+s_j}}\hat{g}_1^{(i)}(l)\hat{g}_1^{(j)}(m)}_{Y_{2,lm}}\right\}^2. \quad \text{(B.47)}$$

Expanding $(Y_{1,lm}+Y_{2,lm})^2$ in Eq B.47, it is straightforward that, by Lemma B.4.4, $l_{4,n}$ is $O_P(\rho_n^{-1})$.

Now we bound s_n. Since $\alpha_l \geq 0$ and $(Y_{1,lm}+Y_{2,lm})^2 \leq 2(Y_{1,lm}^2 + Y_{2,lm}^2)$, we write

$$\begin{aligned} s_n^2 &\leq \frac{1}{n}\sum_{l=1}^n \alpha_l^2 \leq \frac{4}{n}\sum_{l=1}^n\left(\frac{1}{n\tilde{\sigma}_f^2}\sum_{m\neq l}Y_{1,lm}^2 + \frac{1}{n\tilde{\sigma}_f^2}\sum_{m\neq l}Y_{2,lm}^2\right)^2 \\ &\leq \underbrace{8\times\frac{1}{n}\sum_{l=1}^n\left(\frac{1}{n\tilde{\sigma}_f^2}\sum_{m\neq l}Y_{1,lm}^2\right)^2}_{Z_1} + \underbrace{8\times\frac{1}{n}\sum_{l=1}^n\left(\frac{1}{n\tilde{\sigma}_f^2}\sum_{m\neq l}Y_{2,lm}^2\right)^2}_{Z_2}. \end{aligned}$$

To estimate Z_1, we use:

$$Z_1 = \frac{1}{n}\sum_{l=1}^{n}\left(\frac{1}{n\tilde{\sigma}_f^2}\sum_{m\neq l}\sum_{i=1}^{d}\sum_{j=1}^{d}\hat{a}_i\hat{a}_j\frac{r_ir_j(r_i-1)(r_j-1)}{\rho_n^{s_i+s_j}}\tilde{g}_2^{(i)}(l,m)\tilde{g}_2^{(j)}(l,m)\right)^2.$$

Using the fact that

$$\frac{1}{\rho_n^{s_i+s_j}}\tilde{g}_2^{(i)}(l,m)\tilde{g}_2^{(j)}(l,m) \leq \frac{1}{2}\left(\frac{1}{\rho_n^{2s_i}}\tilde{g}_2^{(i)}(l,m)^2 + \frac{1}{\rho_n^{2s_j}}\tilde{g}_2^{(j)}(l,m)^2\right),$$

by the same arguments in the proof of Theorem 6, it is easy to check that $Z_1 = O_P(\rho_n^{-1})$.

For Z_2, let $c_{ijkt} = a_{ij}a_{kt}r_ir_jr_kr_t/\rho_n^{s_i+s_j+s_k+s_t}$ and $\hat{c}_{ijkt} = \hat{a}_{ij}\hat{a}_{kt}r_ir_jr_kr_t/\rho_n^{s_i+s_j+s_k+s_t}$. Consider the estimate:

$$\begin{aligned}Z_2 \leq& \underbrace{\frac{2}{n\tilde{\sigma}_f^4}\sum_{l=1}^{n}\left(\frac{1}{n}\sum_{m\neq\ell}\sum_{i,j,k,t}(\hat{c}_{ijkt}-c_{ijkt})\hat{g}_1^{(i)}(\ell)\hat{g}_1^{(j)}(m)\hat{g}_1^{(k)}(\ell)\hat{g}_1^{(t)}(m)\right)^2}_{Z_2^{(1)}}\\ &+\underbrace{\frac{2}{n\tilde{\sigma}_f^4}\sum_{l=1}^{n}\left(\frac{1}{n}\sum_{m\neq\ell}\sum_{i,j,k,t}c_{ijkt}\hat{g}_1^{(i)}(\ell)\hat{g}_1^{(j)}(m)\hat{g}_1^{(k)}(\ell)\hat{g}_1^{(t)}(m)\right)^2}_{Z_2^{(2)}}\end{aligned}$$

We will start by establishing the order of the $Z_2^{(2)}$ term. Observe that:

$$\begin{aligned}&\frac{1}{n}E\sum_{\ell}\left(\frac{1}{n}\sum_{m\neq\ell}\sum_{i,j,k,t}c_{ijkt}\hat{g}_1^{(i)}(\ell)\hat{g}_1^{(j)}(m)\hat{g}_1^{(k)}(\ell)\hat{g}_1^{(t)}(m)\right)^2\\ &\leq\frac{d^4}{n^2}\sum_{\ell}\sum_{m\neq\ell}\sum_{i,j,k,t}c_{ijkt}^2E\left[\hat{g}_1^{(i)}(\ell)^2\hat{g}_1^{(j)}(m)^2\hat{g}_1^{(k)}(\ell)^2\hat{g}_1^{(t)}(m)^2\right]\\ &\leq\frac{d^4}{n^2}\sum_{\ell}\sum_{m\neq\ell}\sum_{i,j,k,t}c_{ijkt}^2\left(E\left[\hat{g}_1^{(i)}(\ell)^4\hat{g}_1^{(j)}(m)^4\right]E\left[\hat{g}_1^{(k)}(\ell)^4\hat{g}_1^{(t)}(m)^4\right]\right)^{1/2}\end{aligned}$$

$$\leq \frac{d^4}{n^2} \sum_{\ell} \sum_{m \neq \ell} \sum_{i,j,k,t} c_{ijkt}^2 \left(E\left[\hat{g}_1^{(i)}(\ell)^8\right] E\left[\hat{g}_1^{(j)}(m)^8\right] E\left[\hat{g}_1^{(k)}(\ell)^8\right] E\left[\hat{g}_1^{(t)}(m)^8\right] \right)^{1/4}$$

Due to Lemma B.5.2 and Eq B.50, since d is finite, we see that the above is $O(1)$. To complete our bound for Z_2, observe that:

$$\begin{aligned} & P(Z_2^{(2)} > M) \\ \leq & P\left(\max_{i,j,k,t} (\hat{c}_{ijkt} - c_{ijkt})^2 \frac{2}{n\tilde{\sigma}_f^4} \sum_{\ell} \left(\frac{1}{n} \sum_{m \neq \ell} \sum_{i,j,k,t} \left| \hat{g}_1^{(i)}(\ell) \hat{g}_1^{(j)}(m) \hat{g}_1^{(k)}(\ell) \hat{g}_1^{(t)}(m) \right| \right)^2 > M \right) \end{aligned}$$

Since $\max_{i,j,k,t} (\hat{c}_{ijkt} - c_{ijkt})^2$ is lower-order and the second term in the product inside the probability statement may be viewed as a variant of $Z_2^{(1)}$ with $c_{ijkt} = 1$, we can conclude $Z_2 = O(1)$. Combining Z_1 and Z_2, we have, with probability tending to one, $s_n^2 \leq C\rho_n^{-1}$ and $l_{5,n} = l_{4,n} + s_n \leq C'\rho_n^{-1}$ for some universal positive constants C and C'.

Thus, from Eq B.46 and Lemma B.5.1, we have Eq 3.30.

□

We now state and prove Lemma B.5.2, which we had used in the proof of the above theorem.

Lemma B.5.2. *Under the sparsity conditions in Assumption 2,*

$$E(\hat{g}_1(l)^8) = O(\rho_n^{8s})$$

.

Proof. We decompose $\hat{g}_1(l)$ into

$$\hat{g}_1(l) = \hat{H}_1(l) - h_1(l) + g_1(l) - (\hat{T}_n - \theta).$$

Then for some constant C,

$$\hat{g}_1(l)^8 \le C\{(\hat{H}_1(l) - h_1(l))^8 + g_1(l)^8 + (\hat{T}_n - \theta)^8\}. \tag{B.48}$$

$g_1(l)^8$ is $O(\rho_n^{8s})$. Now for $(\hat{T}_n - \theta)^8$,

$$(\hat{T}_n - \theta)^8 \le C\{(\hat{T}_n - T_n)^8 + (T_n - \theta)^8\},$$

where $(T_n - \theta)^8 = \Theta(\rho_n^{8s})$ by boundness of graphon and we investigate $E\{(\hat{T}_n - T_n)^8\}$. Let $\mathcal{S}_r$ denote all r-node subsets from node $\{1, \dots, n\}$,

$$E\{(\hat{T}_n - T_n)^8\} = \frac{\sum\limits_{S_1,\dots,S_8 \in \mathcal{S}_r} E[\{\hat{H}(S_1) - h(S_1)\} \dots \{\hat{H}(S_8) - h(S_8)\}]}{\binom{n}{r}^8}.$$

Consider any term in the above sum where each of the four pairs of the subsets have $p_i, i = 1, \dots, 4$ nodes, $d_i, i = 1, \dots, 4$ edges in common. In this case there are $8r - \sum_i p_i$ choices of nodes and the number of edges are at least $8s - \sum_i d_i$. First note that $p_i \ge 2$, to have non-zero contribution. For acyclic graphs, $d_i \le p_i - 1$ and for general subgraphs with a cycle, $d_i \le \binom{p_i}{2}$. Thus, for $p_i \ge 2$, we have:

$$\frac{O\left(n^{8r - \sum_i p_i} \rho_n^{8s - \sum_i d_i}\right)}{\binom{n}{r}^8} = O(\rho_n^{8s}) \times O\left(\frac{1}{n^{\sum_i p_i} \rho_n^{\sum_i d_i}}\right).$$

For acyclic graphs, it is easy to see that under our sparsity conditions the above is dominated by $p = 2$. For general cyclic graphs, since $\rho_n = \omega(n^{-1/r})$ and $p \le r$,

$$n^{p_i} \rho_n^{d_i} \ge n^{p_i\left(1 - \frac{p_i - 1}{2r}\right)} \ge n^{\frac{4p_i(r+1)}{r}} \to \infty.$$

So, $E\{(\hat{T}_n - T_n)^8\} = O(\rho_n^{8s})$.

To finish bounding $E[\hat{g}_1(l)^8]$, we look into the first term of Eq B.48. Let $\mathcal{S}_r^l$ denote all $r-1$ node subsets from node $\{1,\dots,n\}$ excluding node l,

$$E\{(\hat{H}_1(l) - h_1(l))^8\}$$
$$= \frac{\sum_{S_1,\dots,S_8 \in \mathcal{S}_r^l} E[\{\hat{H}(l\cup S_1) - h(l\cup S_1)\}\dots\{\hat{H}(l\cup S_8) - h(l\cup S_8)\}]}{\binom{n-1}{r-1}^8}.$$

Similarly, consider any term in the above sum where each of the four pairs of the subsets have $p_i, i \le 4$ nodes (besides node l), $d_i, i \le 4$ edges in common. In this case there are $4(2r-2) - \sum_i p_i$ choices of nodes and the number of edges are $8s - \sum_i d_i$. (since each subset already share node l). When $p_i \ge 1$, each pair share node l and another p_i nodes, then for acyclic graphs, $d_i \le p_i$, and for general subgraphs with a cycle, $d_i \le \binom{p_i+1}{2}$. Thus, for $p_i \ge 1$, $d_i \ge 0$, we have

$$\frac{O\left(n^{4(2r-2)-\sum_i p_i}\rho_n^{8s-\sum_i d_i}\right)}{\binom{n-1}{r-1}^8} = O(\rho_n^{8s}) \times O\left(\frac{1}{n^{\sum_i p_i}\rho_n^{\sum_i d_i}}\right),$$

where as we showed above for acyclic graphs, under our sparsity conditions the above is dominated by $p = 1$. For general cyclic graphs, since $\rho_n \gg n^{1/r}$ and $p_i \le r$, $n^{p_i}\rho_n^{d_i} \to \infty$. Thus, $E\{(\hat{H}_1(l) - h_1(l))^8\}$ is also $O(\rho_n^{8s})$.

Thus, combining all terms in Eq B.48, $E[\hat{g}_1(l)^8]$ is $O(\rho_n^{8s})$. □

B.5.4 Comparing Bootstrap Edgeworth Expansion with the U-statistic Edgeworth Expansion

Finally, we show that the bootstrap Edgeworth expansion is close to that of the conditional expectation, which was established in Proposition B.5.1.

Proposition B.5.2. *Suppose that $\sigma_f > 0$, the function f has three continuous derivatives in a neighbourhood of μ, and $\sum_{i=1}^{d} a_i g_1^{(i)}(X_l)$ is non-lattice. Furthermore, suppose that the weights $\xi_1, \ldots, \xi_n$ are generated from a non-lattice distribution such that $E(\xi_1) = 1$, $E\{(\xi_1 - 1)^2\} = 1$, $E\{(\xi_1 - 1)^3\} = 1$. Then we have:*

$$P^*(S_n^* \leq x) = P(S_n \leq x) + o_P\left(n^{-1/2}\right) + O_P\left(\frac{\log n}{n^{2/3}\rho_n}\right). \tag{B.49}$$

Proof. Now we show that $\tilde{\sigma}_f$, $\tilde{A}_1$ and $\tilde{A}_2$ converge to σ_f, A_1 and A_2. We first show $\tilde{\lambda}_{ij}$ and $\tilde{\lambda}_{ijk}$ converge to λ_{ij} and λ_{ijk}.

$$\tilde{\lambda}_{ij} = r_i r_j \hat{E}\left\{\frac{\hat{g}_1^{(i)}(l)\hat{g}_1^{(j)}(l)}{\rho_n^{s_i}\rho_n^{s_j}}\right\},$$
$$\lambda_{ij} = r_i r_j \mathrm{E}\left\{\frac{g_1^{(i)}(X_l)g_1^{(j)}(X_l)}{\rho_n^{s_i}\rho_n^{s_j}}\right\}.$$

Using the fact that $\mathrm{E}(V_l) = 0$,$\mathrm{E}(V_l)^2 = 1$, $\mathrm{E}\{g_1^{(i)}\} = 0$, and an analogous argument as in the proof of Lemma 3.1d) in Zhang and Xia (2020), we have:

$$\tilde{\lambda}_{ij} - \lambda_{ij} = O_P\left(n^{-1/2}\rho_n^{-1/2}\right).$$

Similarly, expanding $\tilde{\lambda}_{ijk}$ and λ_{ijk}, using the fact that $\mathrm{E}(V_l^3) = 1$, $\mathrm{E}\{g_1^{(i)}\} = 0$,

$$\tilde{\lambda}_{ijk} - \lambda_{ijk} = O_P\left(n^{-1/2}\rho_n^{-1/2}\right).$$

Using the same argument in the proof of Lemma 7, we have

$$\hat{E}\{\hat{g}_1^{(i)}(l)\hat{g}_1^{(j)}(m)\tilde{g}_2^{(k)}(l,m)\} - E\{g_1^{(i)}(X_l)g_1^{(j)}(X_m)g_2^{(k)}(X_l, X_m)\} = O_p(n^{-1/2}\rho_n^{-1/2}).$$

Furthermore, under the assumption that f has three continuous derivatives in the neighbourhood of μ, we know that

$$\hat{a}_i = a_i + O_P\left(n^{-1/2}\rho_n^{-1/2}\right),\ \hat{a}_{ij} = a_{ij} + O_P\left(n^{-1/2}\rho_n^{-1/2}\right). \tag{B.50}$$

Thus, together with Eq B.50, we have,

$$\begin{aligned}
\tilde{\sigma}_f^2 - \sigma_f^2 &= O_P\left(n^{-1/2}\rho_n^{-1/2}\right), \\
\tilde{A}_1 - A_1 &= O_P\left(n^{-1/2}\rho_n^{-1/2}\right), \\
\tilde{A}_2 - A_2 &= O_P\left(n^{-1/2}\rho_n^{-1/2}\right).
\end{aligned}$$

Finally we have,

$$\tilde{p}_1(x) = -\{\tilde{A}_1\tilde{\sigma}_f^{-1} + \frac{1}{6}\tilde{A}_2\tilde{\sigma}_f^{-3}(x^2 - 1)\} = p_1(x) + O_P\left(n^{-1/2}\rho_n^{-1/2}\right).$$

Therefore, under the same condition of Proposition B.5.1, from Eq 3.30, we have,

$$\begin{aligned}
P^*(S_n^* \leq x) &= \Phi(x) + n^{-1/2}\tilde{p}_1(x)\phi(x) + O_P\left(\frac{\log n}{n^{2/3}\rho_n}\right) \\
&= P(S_n \leq x) + o_P\left(n^{-1/2}\right) + O_P\left(\frac{\log n}{n^{2/3}\rho_n}\right).
\end{aligned}$$

□

B.6 Detailed Results of Confidence Interval Bias Correction for Smooth Functions of Counts

B.6.1 Edgeworth Expansion for Studentized Smooth Function of Counts

In order to write σ_f^2 as a function of μ and $\hat{\sigma}_f^2$ as function of u, we have to expand the vector of u by including terms such that the variance can be written as a function of the expectation. For example, for simple mean, one needs to add $(x_1, x_2) = (x, x^2)$ for data point x, since the variance is then $x_2 - x_1^2$. For i.i.d random variables, this is simple, but for U statistics, the dependence makes this more nuanced.

We expand the vector of u into $\breve{u}$. Given X, the uncentered $\breve{u}$ is

$$\breve{u} = \Bigg\{ \underbrace{\frac{\hat{T}_n^{(1)}}{\rho_n^{s_1}}, \ldots, \frac{\hat{T}_n^{(d)}}{\rho_n^{s_d}}}_{d \text{ terms}}, \underbrace{\frac{r_1 r_2 \sum_{i=1}^n \hat{h}_1^{(1)}(X_i)\hat{h}_1^{(2)}(X_i)}{n\rho_n^{s_1}\rho_n^{s_2}}, \ldots, \frac{r_{d-1} r_d \sum_{i=1}^n \hat{h}_1^{(d-1)}(X_i)\hat{h}_1^{(d)}(X_i)}{n\rho_n^{s_1}\rho_n^{s_2}}}_{\binom{d}{2} \text{ terms}}, \underbrace{\frac{r_1^2 \sum_{i=1}^n \hat{h}_1^{(1)}(X_i)^2}{n\rho_n^{2s_1}}, \ldots, \frac{r_d^2 \sum_{i=1}^n \hat{h}_1^{(d)}(X_i)^2}{n\rho_n^{2s_d}}}_{d \text{ terms}} \Bigg\}, \tag{B.51}$$

where

$$\hat{h}_1(X_i) = \frac{1}{\binom{n-1}{r-1}} \sum_{1 \le i_1 < \ldots < i_r \le n, i_1, \ldots, i_r \ne i} \hat{h}(X_i, X_{i_1}, \ldots, X_{i_r}).$$

Denote $\breve{\mu} = E\breve{u}$, and $\mathbf{u}' = \breve{u} - \breve{\mu}$. Define $h(\mu) = {\sigma_f}^2$, $h(\breve{u}) = \hat{\sigma}_f^2$ and $c_i = \nabla h(\breve{\mu})^{(i)}$.

Proposition B.6.1. *Define $S_n' = n^{1/2}(f(u) - f(\mu))/\hat{\sigma}_f$. Under the condition that the function f has three continuous derivatives in a neighbourhood of μ, and $\sum_{i=1}^d a_i g_1^{(i)}(X_1)$, is non-lattice, we have:*

$$P(S_n' \le x) = \Phi(x) + n^{-1/2} q_1(x)\phi(x) + o\left(\frac{1}{n^{1/2}}\right),$$

$$q_1(x) = -\{B_1 + \frac{1}{6}B_2(x^2 - 1)\},$$

where B_1 and B_2 are

$$B_1 = A_1\sigma_f^{-1} - \frac{1}{2}\sigma_f^{-3} n \sum_{i=1}^{d'} \sum_{j=1}^{d'} a_i c_j E\{\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}\},$$

$$B_2 = 6B_1 - 6A_1 + \frac{A_2}{\sigma_f^3}.$$

A_1 and A_2 are defined in Proposition B.5.1. The regularity conditions are to ensure the remainders in the stated order uniformly in x.

Proof. Now, we define $A(\breve{u}) = A(u) = f(u) - f(\mu)$, $B(\breve{u}) = (f(u) - f(\mu))/h(\breve{u})$. Then by Taylor Expansion we have,

$$\begin{aligned}
B(\breve{u}) &= A(\breve{u})/h(\breve{u})^{1/2} = A(\breve{u}) * h(\breve{u})^{-1/2} \\
&= A(\breve{u})\left\{h(\breve{\mu})^{-1/2} + (\breve{u} - \breve{\mu})^T \nabla(h(\breve{\mu})^{-1/2}) + (\breve{u} - \breve{\mu})^T \frac{H(h(\breve{\mu})^{-1/2})}{2}(\breve{u} - \breve{\mu}) + o_P\left(\frac{1}{n}\right)\right\} \\
&= A(\breve{u})/h(\breve{\mu})^{-1/2} - \frac{1}{2}A(\breve{u})h(\breve{\mu})^{-3/2}(\nabla h(\breve{\mu}))^T(\breve{u} - \breve{\mu}) + O_P\left(\frac{1}{n^{3/2}}\right) \\
&= A(\breve{u})/\sigma_f - \frac{1}{2}(\breve{u} - \breve{\mu})^T \sigma_f^{-3} \underbrace{\nabla f(\breve{\mu})(\nabla h(\breve{\mu}))^T}_{D}(\breve{u} - \breve{\mu}) + O_P\left(\frac{1}{n^{3/2}}\right) \\
&= A(u)/\sigma_f - \frac{1}{2}(\breve{u} - \breve{\mu})^T \sigma_f^{-3} \underbrace{\begin{bmatrix} a_1c_1 & \dots & a_1c_d \\ \dots & \dots & \dots \\ a_dc_1 & \dots & a_dc_d \end{bmatrix}}_{D}(\breve{u} - \breve{\mu}) + O_P\left(\frac{1}{n^{3/2}}\right),
\end{aligned}$$

where

$$\begin{aligned}
&(\breve{u} - \breve{\mu})^T D(\breve{u} - \breve{\mu}) = \sum_{i=1}^{d'} \sum_{i=1}^{d'} a_i c_j \mathbf{u}'^{(i)} \mathbf{u}'^{(j)}, \\
&a_{d+1} = a_{d+2} = \ldots = a_{d'} = 0.
\end{aligned}$$

We have $S_n' = n^{1/2}\frac{f(u)-f(\mu)}{h(\breve{u})} = n^{1/2}B(\breve{u})$. Thus we can write S_n' into

$$S_n' = n^{1/2}A(u)/\sigma_f - \frac{1}{2}\sigma_f^{-3}n^{1/2}\sum_{i=1}^{d'}\sum_{i=1}^{d'} a_i c_j \mathbf{u}'^{(i)}\mathbf{u}'^{(j)} + O_P\left(\frac{1}{n}\right). \qquad \text{(B.52)}$$

Since $a_i = 0$ for $i > d$, we only discuss here $\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}$ for $i \leq d, j \leq d$ and $\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}$ for $i \leq d, d < j \leq d'$. We first prove that

$$\mathbf{u}'^{(i)}\mathbf{u}'^{(j)} = \mathrm{E}\{\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}\} + n^{-2}\sum_{l<m}\gamma(X_l, X_m) + O_P\left(\frac{1}{n^{3/2}}\right) \qquad \text{(B.53)}$$

holds for both cases, where γ is some symmetric function of X_l and X_m.

For $i \leq d, j \leq d$, since $\mathbf{u}'^{(i)} = \frac{u_L^{(i)}}{n} + \frac{u_Q^{(i)}}{n^2} + O_P\left(\frac{1}{n^{3/2}}\right)$, $\mathbf{u}'^{(j)} = \frac{u_L^{(j)}}{n} + \frac{u_Q^{(j)}}{n^2} + O_P\left(\frac{1}{n^{3/2}}\right)$,

$$\begin{aligned}\mathbf{u}'^{(i)}\mathbf{u}'^{(j)} &= \frac{u_L^{(i)}u_L^{(j)}}{n^2} + O_P\left(\frac{1}{n^{3/2}}\right)\\ &= \frac{r_i r_j}{n^2\rho_n^{s_i}\rho_n^{s_j}}\sum_{l=1}^{n} g_1^{(i)}(X_l)g_1^{(j)}(X_l) + \frac{2r_i r_j}{n^2\rho_n^{s_i}\rho_n^{s_j}}\sum_{l<m} g_1^{(i)}(X_l)g_1^{(j)}(X_m) + O_P\left(\frac{1}{n^{3/2}}\right).\end{aligned}$$

Thus,

$$E\mathbf{u}'^{(i)}\mathbf{u}'^{(j)} = \frac{r_i r_j}{n^2\rho_n^{s_i}\rho_n^{s_j}}\sum_{l=1}^{n} g_1^{(i)}(X_l)g_1^{(j)}(X_l),$$

and Eq B.53 follows.

For $i \leq d, d < j \leq d'$, $\mathbf{u}'^{(i)} = \frac{u_L^{(i)}}{n} + \frac{u_Q^{(i)}}{n^2} + O_P\left(\frac{1}{n^{3/2}}\right)$, while

$$\mathbf{u}'^{(j)} = \frac{r_k r_t}{n\rho_n^{s_k}\rho_n s_t}\sum_{l=1}^{n}\hat{h}_1^{(k)}(X_l)\hat{h}_1^{(t)}(X_l) - E\left\{\frac{r_k r_t}{\rho_n^{s_k}\rho_n^{s_t}n}\sum_{l=1}^{n}\hat{h}_1^{(t)}(X_l)\hat{h}_1^{(k)}(X_l)\right\},$$

for some $k, t \in \{1, \ldots, d\}$. Denote $\mathrm{E}\{\hat{h}^k(X_l)\} = \theta^{(k)}$, Hoeffding decomposition of $\hat{h}^k(X_l)$ yields,

$$\frac{\hat{h}_1^{(k)}(X_l) - \theta_n^{(k)}}{\rho_n^{s_k}} = h_1^{(k)}(X_l) - \theta_n^{(k)} + \frac{r-1}{n-1}\sum_{s\neq l, 1\leq s\leq n}\{g_2^{(k)}(X_l, X_s) + g_1^{(k)}(X_s)\} + O_P\left(\frac{1}{n}\right)$$

$$= g_1^{(k)}(X_1) + \frac{r-1}{n-1} \sum_{s \neq l, 1 \leq s \leq n} \{g_2^{(k)}(X_l, X_s) + g_1^{(k)}(X_s)\} + O_P\left(\frac{1}{n}\right). \tag{B.54}$$

Denote $\boldsymbol{U}^{(i)} = \frac{\sum_{l=1}^n g_1^{(i)}(X_l)}{n}$, then

$$\begin{aligned}
\mathbf{u'}^{(i)}\mathbf{u'}^{(j)} = \frac{r_k r_i r_t}{\rho_n^{s_k+s_i+s_t}} \Bigg\{ & \theta^{(k)}\boldsymbol{U}^{(i)}\boldsymbol{U}^{(t)} + \theta^{(t)}\boldsymbol{U}^{(i)}\boldsymbol{U}^{(k)} + \frac{1}{n^2}\sum_{l=1}^n g_1^{(i)}(X_l)g_1^{(k)}(X_l)g_1^{(t)}(X_l) \\
& + \frac{2}{n^2}\sum_{l<m} g_1^{(i)}(X_l)g_1^{(k)}(X_m)g_1^{(t)}(X_m) + \frac{2}{n^2}\sum_{l<m} g_1^{(i)}(X_l)g_1^{(k)}(X_m)g_2^{(t)}(X_l, X_m) \\
& + \frac{2}{n^2}\sum_{l<m} g_1^{(i)}(X_l)g_1^{(t)}(X_m)g_2^{(k)}(X_l, X_m) \Bigg\} + O_P\left(\frac{1}{n^{3/2}}\right).
\end{aligned}$$

Taking Expectation, Eq B.53 easily follows.

Now that Eq B.53 holds, using Eq B.42 and Eq B.52, we have

$$\begin{aligned}
S'_n = & \frac{A_1}{\sqrt{n}\sigma_f} - \frac{1}{2}\sigma_f^{-3} n \sum_{i=1}^{d'}\sum_{j=1}^{d'} a_i c_j \mathrm{E}\{\mathbf{u'}^{(i)}\mathbf{u'}^{(j)}\} \\
& + n^{-1/2}\sum_{l=1}^n \alpha(X_l) + n^{-3/2}\sum_{l<m}\{\beta(X_l, X_m) + \gamma(X_l, X_m)\} + O_P\left(\frac{1}{n}\right)
\end{aligned} \tag{B.55}$$

Therefore, using Lemma B.5.1, we know that

$$B_1 = A_1\sigma_f^{-1} - \frac{1}{2}\sigma_f^{-3} n \sum_{i=1}^{d'}\sum_{j=1}^{d'} a_i c_j \mathrm{E}\{\mathbf{u'}^{(i)}\mathbf{u'}^{(j)}\}. \tag{B.56}$$

From Eq B.55, we also have that Theorem 2.1 of Jing and Wang (2010) applies to S'_n under the same conditions of Proposition B.5.1. For the simplicity of calculation, we note that B_2 can be estimated using the identity $p_1(0) = q_1(0)$ and the forms of A_1, A_2, and B_1, which gives us $B_2 = 6B_1 - 6A_1 + \frac{A_2}{\sigma_f^3}$.

Thus, under same conditions of Proposition B.5.1, we have

$$P(S'_n \leq x) = \Phi(x) + n^{-1/2} q_1(x)\phi(x) + o\left(n^{-1/2}\right),$$

where B_1 and B_2 are defined above, $q_1(x)$ is as

$$q_1(x) = -\{B_1 + \frac{1}{6}B_2(x^2 - 1)\}.$$

□

B.6.2 Estimating Confidence Interval Corretion for Smooth Function of Counts

In order correct the confidence intervals arising from the standardized bootstrap, we need to estimate $p_1(x)$ and $q_1(x)$. This requires the calculation of σ_f^2, A_1 and A_2 are straightforward. In this section, we show how to compute $\hat{q}_1(x)$ for transitivity.

While we only show in detail the calculations of transitivity ($d = 2$), they can be easily used as building blocks to extend to other smooth functions of counts with $d \geq 2$.

In the case of transitivity, the original u used for estimation of $p_1(x)$ is of length $d = 2$. Recall that for estimating $q_1(x)$ we need to expand this vector so that the variance is a function of this vector. This expanded vector (see Eq B.51) is of length $d' = 5$. Denote $\mu_{ij} = n \times E\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}$. We also have for T and V, $r_1 = r_2 = r = 3$, and $s_1 = 3$ and $s_2 = 2$.

To estimate B_1 and B_2, we first use the fact that c_k for k in $1 \leq k \leq d'$ follows

Hall (2013) Section 3.10.6 as follows:

$$c_k = 2\sum_{i=1}^{d}\sum_{j=1}^{d} a_{ik}a_j\mu_{ij} - 2a_k\sum_{i=1}^{d} a_i\mu^{(i)} + \sum_{i=1,j=1,(k)}^{d,d} a_ia_j, \tag{B.57}$$

where $\sum_{i=1,j=1,(k)}^{d,d}$ denotes the pair (i,j) in $\breve{u}^{(1),\ldots,(d)}$ that $\breve{u}^{(i)}\breve{u}^{(j)} = \breve{u}^{(k)}$. For example, in transitivity, $\breve{u}^{(3)} = \breve{u}^{(1)}\breve{u}^{(2)}$.

Now we simplify $E\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}$ for $1 \le i \le d$, $1 \le j \le d'$ for the purpose of estimating B_1 and c_k in Eq B.56 and Eq B.57. We do not consider the case where $i > d$ since a_i for $i > d$ is 0. By the definition of u', using Hoeffding Decomposition of $\hat{h}_1^{(i)}(X_1)$ $(i \in \{1,2\})$ showed in Eq B.54, simple algebra yields,

$$\mathrm{E}\{\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}\} = \frac{r^2}{n\rho_n^{s_1}\rho_n^{s_2}}\mathrm{E}\{g_1^{(i)}(X_1)g_1^{(j)}(X_1)\} + O\left(\frac{1}{n^{3/2}}\right),\ 1 \le i,j \le 2,$$

$$\begin{aligned}\mathrm{E}\{\mathbf{u}'^{(1)}\mathbf{u}'^{(4)}\} = \frac{r^3}{n\rho_n^{s_1^3}}&\Big[\mathrm{E}\{g_1^{(1)}(X_1)g_1^{(1)}(X_1)^2\}\\ &+ 2(r-1)\mathrm{E}\{g_1^{(1)}(X_1)g_1^{(1)}(X_2)g_2^{(1)}(X_1,X_2)\}\Big]\\ &+ \frac{2r^4}{n\rho_n^{2s_1}}\mu_1 E\{g_1^{(1)}(X_1)g_1^{(1)}X_1\} + O\left(\frac{1}{n^{3/2}}\right),\end{aligned}$$

$$\begin{aligned}\mathrm{E}\{\mathbf{u}'^{(2)}\mathbf{u}'^{(5)}\} = \frac{r^3}{n\rho_n^{3s_2}}&\Big[\mathrm{E}\{g_1^{(2)}(X_1)g_1^{(2)}(X_1)^2\}\\ &+ 2(r-1)\mathrm{E}\{g_1^{(2)}(X_1)g_1^{(2)}(X_2)g_2^{(2)}(X_1,X_2)\}\Big]\\ &+ \frac{2r^4}{n\rho_n^{2s_2}}\mu_2 E\{g_1^{(1)}(X_1)g_1^{(2)}X_1\} + O\left(\frac{1}{n^{3/2}}\right),\end{aligned}$$

$$
\begin{aligned}
\mathrm{E}\{\mathbf{u}'^{(1)}\mathbf{u}'^{(5)}\} = \frac{r^3}{n\rho_n^{s_1}\rho_n^{2s_2}}\Big[&\mathrm{E}\{g_1^{(1)}(X_1)g_1^{(2)}(X_1)^2\} \\
&+ 2(r-1)\mathrm{E}\{g_1^{(1)}(X_1)g_1^{(2)}(X_2)g_2^{(2)}(X_1,X_2)\}\Big] \\
&+ \frac{2r^4}{n\rho_n^{s_1}\rho_n^{s_2}}\mu_2 E\{g_1^{(1)}(X_1)g_1^{(2)}X_1\} + O\left(\frac{1}{n^{3/2}}\right),
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\{\mathbf{u}'^{(2)}\mathbf{u}'^{(4)}\} = \frac{r^3}{n\rho_n^{2s_1}\rho_n^{s_2}}\Big[&\mathrm{E}\{g_1^{(2)}(X_1)g_1^{(1)}(X_1)^2\} \\
&+ 2(r-1)\mathrm{E}\{g_1^{(2)}(X_1)g_1^{(1)}(X_2)g_2^{(1)}(X_1,X_2)\}\Big] \\
&+ \frac{2r^4}{n\rho_n^{s_1}\rho_n^{s_2}}\mu_1 E\{g_1^{(1)}(X_1)g_1^{(2)}X_1\} + O\left(\frac{1}{n^{3/2}}\right),
\end{aligned}
$$

Now for the case of $i = 1$ and $j = 3$, applying the same technique, then we will have,

$$
\begin{aligned}
\mathrm{E}\{\mathbf{u}'^{(1)}\mathbf{u}'^{(3)}\} =&\frac{r^3}{n\rho_n^{2s_1}\rho_n^{s_2}}\Big[\mathrm{E}\{g_1^{(1)}(X_1)^2 g_1^{(2)}(X_1)\} \\
&+ (r-1)\mathrm{E}\{g_1^{(1)}(X_1)g_1^{(1)}(X_2)g_2^{(2)}(X_1,X_2)\} \\
&+ (r-1)\mathrm{E}\{g_1^{(1)}(X_1)g_1^{(2)}(X_2)g_2^{(1)}(X_1,X_2)\Big] \\
&+ \frac{r^4}{n\rho_n^{s_1}\rho_n^{s_2}}\mu_1 E\{g_1^{(1)}(X_1)g_1^{(2)}X_1\} \\
&+ \frac{r^4}{n\rho_n^{2s_1}}\mu_2 E\{g_1^{(1)}(X_1)^2\} + O\left(\frac{1}{n^{3/2}}\right),
\end{aligned}
$$

Similarly, for $i = 2$ and $j = 3$, we have

$$\begin{aligned}\mathrm{E}\{\mathbf{u}'^{(2)}\mathbf{u}'^{(3)}\} =& \frac{r^3}{n\rho_n^{s_1}\rho_n^{2s_2}}\Big[\mathrm{E}\{g_1^{(2)}(X_1)^2 g_1^{(1)}(X_1)\} \\ &+ (r-1)\mathrm{E}\{g_1^{(2)}(X_1)g_1^{(1)}(X_2)g_2^{(2)}(X_1,X_2)\} \\ &+ (r-1)\mathrm{E}\{g_1^{(2)}(X_1)g_1^{(2)}(X_2)g_2^{(1)}(X_1,X_2)\Big] \\ &+ \frac{r^4}{n\rho_n^{s_1}\rho_n^{s_2}}\mu_2 E\{g_1^{(1)}(X_1)g_1^{(2)}X_1\} \\ &+ \frac{r^4}{n\rho_n^{2s_1}}\mu_1 E\{g_1^{(2)}(X_1)^2\} + O\left(\frac{1}{n^{3/2}}\right).\end{aligned}$$

Now we can estimate B_1 from Eq B.56 and $c_i, 1 \le i \le 5$ from Eq B.57 by estimating $E\{\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}\}$ above using $\hat{g}_1^{(i)}(X_1)$ and $\hat{g}_2^{(i)}(X_1, X_2)$, for $i \in \{1, 2\}$. Using the fact of $p_1(0) = q_1(0)$, we can estimate B_2 by

$$\hat{B}_2 = 6\hat{B}_1 - 6\hat{A}_1\hat{\sigma}_f^{-1} + \hat{A}_2\hat{\sigma}_f^{-3}.$$

Then we have the estimated $\hat{q}_1(x)$,

$$\hat{q}_1(x) = -\{\hat{B}_1 + \frac{1}{6}\hat{B}_2(x^2 - 1)\}.$$

Now we show the studentized edgeworth expansion of some statistics $f(T, V)$ using same $\check{u}$ as transitivity, including T, $3T + 5V$, TV, $3T/V$(transitivity) and T^2V^2. The $q_1(x)$ of the Edgeworth expansion of the studentized version of these statistics $f(T, V)$ share the same $E\{\mathbf{u}'^{(i)}\mathbf{u}'^{(j)}\}$ ($i \in \{1, 2\}$, j in $\{1,\ldots,5\}$. The the only difference lies in evaluating different derivatives $\mathbf{a}$ of f and thus having different $\hat{c}_k$, $\hat{B}_1$ and $\hat{B}_2$.

Recall that $\mathrm{err}(F, G)$ is defined in Section 3.6 as the maximum of $|F(x) - G(x)|$ over the range $[-3, 3]$, over a grid size 0.1. In the following two tables,we show this distance between the true CDF and our empirical edgeworth expansion and the normal approximation for five different smooth functions. Tables B.1 and B.2 show these for the standardized and studentized statistics. The empirical edgeworth expansion is estimated using a random graph with $n = 160$, $\rho_n = 1$, generated from two graphons `SBM-G` and `SM-G` with the same parameters as in Section 3.6. The true CDF is estimated by 10^6 size 160 graphs generated by the same graphons with same model parameters.

Table B.1: Standardized EW Sup CDF error compared to N(0,1)

	SBM		SM-G	
Studentized	err($\hat{F}(S_n), F$)	err(Φ, F)	err($\hat{F}(S_n), F$)	err(Φ, F)
T	**0.002**	0.018	**0.004**	0.030
3T+5V	**0.003**	0.011	**0.002**	0.018
TV	**0.006**	0.023	**0.016**	0.042
3T/V	**0.005**	0.027	**0.006**	0.051
T^2V^2	**0.036**	0.078	**0.092**	0.142

Table B.1 shows the standardized sup error $\sup_x |\hat{F}(S_n \leq x) - F(x)|$, where S_n is the standardized statistic, $\hat{F}(S_n \leq x) = \Phi(x) + n^{-1/2}\hat{p}_1(x)\phi(x)$ and $F(x)$ is the true distribution of the standardized statistic. In Table B.2, we show $\sup_x |\hat{F}(S_n \leq x) - F'(x)|$, where $\hat{F}(S_n' \leq x) = \Phi(x) + n^{-1/2}\hat{q}_1(x)\phi(x)$ and $F'(x)$ indicates the true CDF of the studentized statistic.

Table B.2: Studentized EW Sup CDF error compared to N(0,1)

	SBM		SM-G	
Studentized	err($\hat{F}(S'_n), F'$)	err(Φ, F')	err($\hat{F}(S'_n), F'$)	err(Φ, F')
T	**0.004**	0.021	**0.008**	0.043
3T+5V	**0.002**	0.012	**0.005**	0.026
TV	**0.006**	0.029	**0.015**	0.054
3T/V	**0.012**	0.031	**0.007**	0.052
T^2V^2	**0.022**	0.058	**0.045**	0.106

We see that for both graphons the empirical edgeworth expansion has much lower error than the Gaussian approximation. Also, the linear combinations of the statistics typically have lower error than those which need the estimation of first and second derivatives.

B.7 Additional experiments

In this section we provide additional experiment results that were left out from the main text for better presentation.

B.7.1 Additional results for two-stars

We show in Figure B.1 the maximum of (absolute) difference of bootstrap CDF F_n^* over the $[-3, 3]$ range (err(F_n,F_n^*)) for two-star density from the true CDF F_n for sparsity parameter ρ_n varying from 0.05 to 1. We show the average of the expected difference over 30 independent runs along with the error-bars. In Figure B.2, we show the 95% CI coverage for two-stars. The results of two-stars are similar to those of triangles in the main paper.

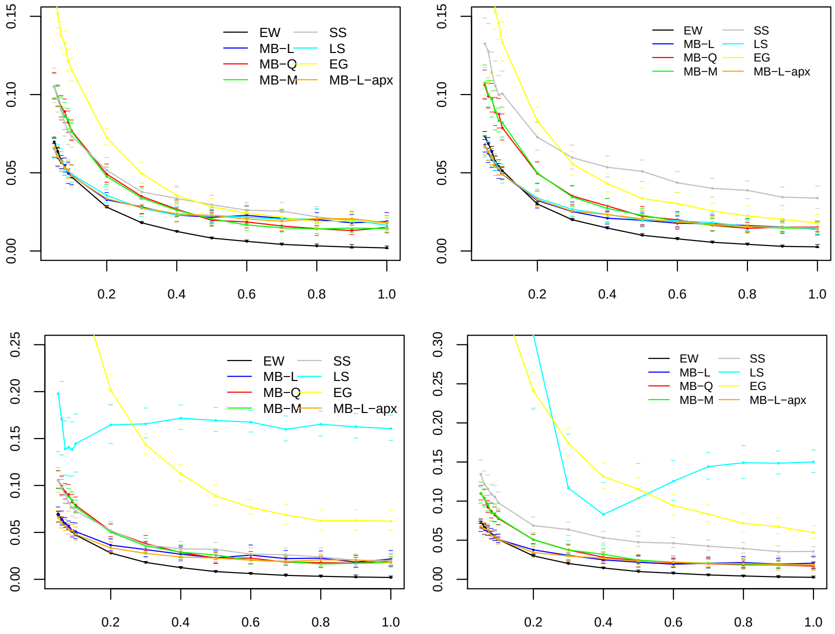


Figure B.1: We plot err(F_n,F_n^*) for two-star density for all methods on the Y axis, where $F_n^*(t)$ corresponds to the appropriate resampling distribution. We vary the sparsity parameter ρ_n on the X axis. Networks in the left column are simulated from SBM-G and those in the right column are simulated from SM-G. The first row is centered at bootstrap mean. The second row is centered by triangles density estimated on the whole graph (MB-L-apx is centered at approximate triangle density estimated from the whole graph) .

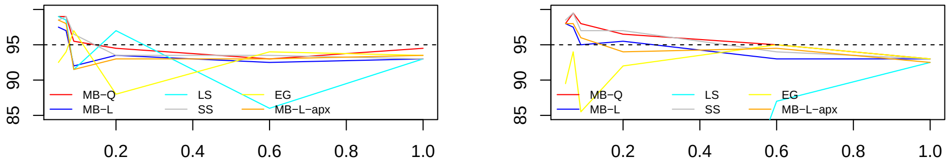


Figure B.2: We present coverage of 95% Bootstrap Percentile CI with correction for two-stars of the SBM-G (left) and SM-G (right) models in ρ from 0.05 to 1.

B.7.2 Additional timing results

In Figure B.3 we show logarithm of running time for four-cycles count against growing n for `SM-G` model.

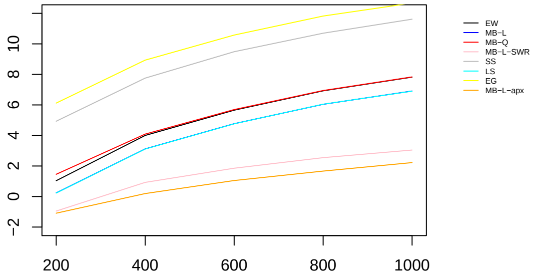


Figure B.3: Logarithm of running time for four-cycles in `SM-G` against sample size n.

Appendix C

Supplementary Material for Separate Exchangeability in Bayesian Nonparametrics

C.1 Proofs

We include a brief proof of (4.5).

Claim: Infinite partial exchangeability (4.4) entails

$$\operatorname{Corr}(x_{ij}, x_{i'j}) \geq \operatorname{Corr}(x_{ij}, x_{i'j'}) \quad j \neq j',\ i \neq i'. \tag{C.1}$$

Proof. We define $U_{ij} := x_{ij}/\operatorname{Var}(x_{ij})$ such that $\operatorname{Cov}(U_{ij}, U_{i'j}) = \operatorname{Corr}(x_{ij}, x_{i'j})$. By de Finetti's theorem (4.4) and law of total covariance

$$\begin{aligned} \operatorname{Cov}(U_{ij}, U_{i'j}) &= 0 + \operatorname{Cov}\{\mathrm{E}(U_{ij} \mid P_j)\mathrm{E}(U_{i'j} \mid P_j)\} \\ &\geq 0 + \operatorname{Cov}\{\mathrm{E}(U_{ij} \mid P_j)\mathrm{E}(U_{i'j'} \mid P_{j'})\} = \operatorname{Cov}(U_{ij}, U_{i'j'}), \end{aligned}$$

□

Considering a trivial example with > in (C.1) proves the claim in (4.5).

C.2 Algorithm 1

Algorithm 1 below states the transition probabilities for posterior simulation in model (4.20). The description makes use of the following notation. Let $\boldsymbol{X}$ denote a $(J \times 12)$ design matrix with $\boldsymbol{x}_j$ in row j. the following quantities are used in the description of Algorithm 1. We use a notational convention of marking quantities that are cluster-specific with a tilde, as in $(\tilde{\beta}_h, \tilde{\sigma}_h)$, etc. Let then $\tilde{\boldsymbol{y}}_h, \tilde{\boldsymbol{X}}_h$ and $\tilde{\boldsymbol{\delta}}_h$ denote y_{ij}, $\boldsymbol{x}_j$ and δ_t arranged by clusters. That is, $\tilde{\boldsymbol{y}}$ is a $(n_h J \times 1)$ vector stacking $\boldsymbol{y}_i$ for all $i \in C_h$ as $\tilde{\boldsymbol{y}}_h = (y_{ij},\ i \in C_H;\ j = 1, \ldots, J)$; $\tilde{\boldsymbol{X}}$ is an $(n_h J \times 12)$ matrix with $\boldsymbol{X}$ stacked n_h times on top of each other; and $\tilde{\boldsymbol{\delta}}_h$ is a $(n_h J \times 1)$ vector that concatenates n_h copies of $\boldsymbol{\delta} = (\delta_{t_1}, \ldots, \delta_{t_J})$.

Algorithm 3. MCMC algorithm for posterior inference

```
Priors: We fixed hyperparameters: $\xi = 1$, $\beta_0 = (0, 0, \dots, 0)$, $\sigma_{\beta 0} = 1$,$a_0 = 1$,
  $b_0 = 1$, $\zeta_0 = 0$, $\omega_0 = 0.01$, $\mu_0 = 3$, $\sigma_0^2 = 5$. Also, let $\boldsymbol{\Sigma}_0 = \sigma_{\beta 0}^2 I$.
for 1:M do
    1. For each protein $i$, sample
         $P(s_i = h \mid \cdot) \propto \pi_h \prod_{j=1}^{J} N(y_{ij}; \alpha_i + \boldsymbol{x}'_j \tilde{\beta}_h + \delta_{t_j}, \tilde{\sigma}_h^2)$
    2. Update $V_h$, $h = 1, \dots, H-1$, keeping in mind that $V_H = 1$ is fixed:
         $V_h \mid \cdot \sim \mathrm{Be}(1 + n_h, \xi + \sum_{\ell=h+1}^{H} n_\ell),\ h = 1, \dots, H-1$
       and set $\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)$, $h = 1, \dots, H$.
    3. Update $\tilde{\boldsymbol{\beta}}_h$ and $\tilde{\sigma}_h^2$: Recall the definition of $\tilde{\boldsymbol{y}}_h$, $\tilde{\boldsymbol{X}}_h$ and $\tilde{\boldsymbol{\delta}}_h$ as
       cluster-specific combined vectors and matrices. See the text for a detailed
       definition.
         $\tilde{\beta}_h \mid \cdot \sim N(\tilde{\boldsymbol{\mu}}_h, \tilde{\boldsymbol{\Sigma}}_h)$,
             $\tilde{\boldsymbol{\mu}}_h = \tilde{\boldsymbol{\Sigma}}_h \left\{\tilde{\boldsymbol{X}}_h^T \boldsymbol{\Sigma}_h^{-1}(\tilde{\boldsymbol{y}}_h - \tilde{\boldsymbol{\alpha}}_h - \tilde{\boldsymbol{\delta}}_h) + \boldsymbol{\Sigma}_0^{-1}\beta_0\right\}$,
             $\tilde{\boldsymbol{\Sigma}}_h = (\boldsymbol{\Sigma}_0^{-1} + \tilde{\boldsymbol{X}}_h^T \boldsymbol{\Sigma}_h^{-1} \tilde{\boldsymbol{X}}_h)^{-1}$ and $\boldsymbol{\Sigma}_h = \mathrm{Diag}_{J \times n_h}(\tilde{\sigma}_h^2)$,
         $\tilde{\sigma}_h^2 \mid \cdot \sim \mathrm{InvGa}(a_h, b_h)$
             $a_h = a_0 + \frac{n_h \times J}{2}$ and $b_h = b_0 + \frac{\sum_{i \in C_h} \sum_{j=1}^{J} (y_{ij} - m_{ij})^2}{2}$
             $m_{ij} = \alpha_i + \boldsymbol{x}'_j \tilde{\beta}_h - \delta_{t_j}$
    4. Update age-specific effect $\delta_t$. Let $m_t = \sum_{j=1}^{J} \mathbf{1}(t_j = t)$. Sample
         $\delta_t \mid \cdot \sim N(m_t, V_t)$
         $\frac{1}{V_t} = \frac{1}{\omega^2} + \sum_{j:\, t_j = t} \sum_{i=1}^{I} 1/\sigma_{s_i}^2,\ m_t = V_t \left(\frac{\zeta}{\omega^2} + \frac{\sum_{j:\, t_j = t} \sum_{i=1}^{I} (y_{ij} - \boldsymbol{x}'_j \tilde{\beta}_{s_i} - \alpha_i)^2}{\tilde{\sigma}_{s_i}^2}\right)$
    4. Update protein-specific intercept $\alpha_i$. Sample
         $\alpha_i \mid \cdot \sim N\left(\frac{\mu_0/\sigma_0^2 + \sum_{j=1}^{J} \sum_{i=1}^{I} (y_{ij} - \boldsymbol{x}'_j \tilde{\boldsymbol{\beta}}_{s_j} - \delta_{t_j})^2 / \tilde{\sigma}_{s_j}^2}{1/\sigma_0^2 + J/\sigma_{s_i}^2}, \left(1/\sigma_0^2 + J/\tilde{\sigma}_{s_i}^2\right)^{-1}\right)$.
end
```

Bibliography

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivar. Anal.*, 11(4):581–598.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

Andris, C., Lee, D., Hamilton, M. J., Martino, M., Gunning, C. E., and Selden, J. A. (2015). The rise of partisanship and super-cooperators in the us house of representatives. *PloS one*, 10(4):e0123507.

Arvesen, J. N. (1969). Jackknifing U-statistics. *The Annals of Mathematical Statistics*, 40(6):2076–2100.

Assadi, S., Kapralov, M., and Khanna, S. (2018). A simple sublinear-time algorithm for counting arbitrary subgraphs via edge sampling. *ArXiv e-prints*.

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinskiand, V., Qin, Y., and Sussman, D. L. (2018). Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92.

Bai, Z. and Zhao, L. (1986). Edgeworth expansions of distribution functions of independent random variables. *Science in China Series A-Mathematics, Physics, Astronomy and Technological Science*, 29(1):851–896.

Bentkus, V., Götze, F., and van Zwet, W. (1997). An edgeworth expansion for symmetric statistics. *Annals of Statistics*, 25(2):851–896.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian Theory*, volume 405. John Wiley & Sons.

Bertail, P., Politis, D. N., and Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association*, 94(446):569–579.

Bhattacharyya, S. and Bickel, P. J. (2015). Subsampling bootstrap of count features of networks. *Annals of Statistics*, 43:2384–2411.

Bickel, P., Götze, F., and Van Zwet, W. (1986). The edgeworth expansion for u-statistics of degree two. *The Annals of Statistics*, pages 1463–1484.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences (USA)*, 106:21068–21073.

Bickel, P. J., Chen, A., and Levina, E. (2011). The method of moments and degree distributions for network models. *Annals of Statistics*, 39:38–59.

Bickel, P. J., Götze, F., and van Zwet, W. (1997). Resampling fewer than n obersvations: Gains, losses, and remedies for losses. *Statistica Sinica*, pages 1–31.

Billingsley, P. (1995). *Probability and Measure*. John Wiley and Sons, New York.

Borgs, C., Chayes, J. T., Cohn, H., and Zhao, Y. (2019). An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062.

Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T., and Vesztergombi, K. (2008). Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801 – 1851.

Bose, A. and Chatterjee, S. (2018). *U-Statistics, M_m-Estimators, and Resampling*. Springer Verlag, New York.

Boucheron, S., Lugosi, G., and Bousquet, O. (2004). Concentration inequalities. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures in Machine Learning*, page 208–40. Springer, New York.

Caron, F. and Fox, E. B. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(5):1295.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1):177–214.

Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein protein interaction network. *Bioinformatics*, 22(8):2283–2290.

Chen, X. and Kato, K. (2019). Randomized incomplete u-statistics in high dimensions. *Annals of Statistics*, 47(6):3127–3156.

Cifarelli, D. and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. Quaderni Istituto Matematica Finanziaria, Turin.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In Vannucci, M., Do, K.-A., and Müller, P., editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge Univ. Press.

Dahl, D. B., Johnson, D. J., and Müller, P. (2021). Search algorithms and loss functions for Bayesian clustering. *Preprint arXiv: 2105.04451*.

Daudin, J., Koskas, M., Schbath, S., and Robin, S. (2008). Assessing the exceptionality of network motifs. *Journal of Computational Biology*, 15(1):1–20.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):212–229.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Am. Stat. Assoc.*, 99(465):205–215.

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2021). A common atom model for the Bayesian nonparametric analysis of nested data. *J. Am. Stat. Assoc.*, (in press).

Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.*, 7(28):33–61.

Durante, D. and Dunson, D. B. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Anal.*, 13(1):29–58.

Eden, T., Levi, A., Ron, D., and Seshadhri, C. (2017). Approximately counting triangles in sublinear time. *SIAM Journal of Computing*, 46:1603–1646.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(3):1–26.

Efron, B. (1980). The Jackknife, the Bootstrap, and other resampling plans. Technical report, Stanford University University.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9(3):586–596.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.*, 1(1):54–75.

Fang, Z. and Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412.

Feige, U. (2006). On sums of independent random variables with unbounded variance, and estimating the average degree in a graph. *SIAM Journal of Computing*, 35:964–984.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The annals of statistics*, 2(4):615–629.

de Finetti, B. (1930). Funzione caratteristica di un fenomeno aleatorio. In *Atti Reale Accademia Nazionale dei Lincei, Mem.*, volume 4, pages 86—-133.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Henri Poincaré*, 7(1):1–68.

Fortini, S., Ladelli, L., and Regazzini, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhyā, Ser. A*, 62(1):86–109.

Fortini, S. and Petrone, S. (2016). *Predictive Distribution (de Finetti's View)*, pages 1–9. Wiley StatsRef: Stat. Ref. Online.

Fortini, S., Petrone, S., et al. (2012). Predictive construction of priors in Bayesian nonparametrics. *Braz. J. Probab. Stat.*, 26(4):423–449.

Foti, N. J. and Williamson, S. A. (2013). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):359–371.

Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10(1):53 – 67.

Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2021+). Multivariate species sampling processes. *Working Paper*.

Gao, C. and Ma, Z. (2019). Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing.

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian nonparametrics*. Springer.

Ghosh, M., Parr, W. C., Singh, K., and Babu, G. J. (1984). A note on bootstrapping the sample median. *The Annals of Statistics*, 12(3):1130–1135.

Ghoshdastidar, D., Gutzeit, M., Carpentier, A., and von Luxburg, U. (2017). Two-sample tests for large random graphs using network statistics. In *Conference on Learning Theory (COLT)*.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and stirling triangles. *J. Math. Sci.*, 138(3):5674–5685.

Goldreich, O. and Ron, D. (2008). Approximating average parameters of graphs. *Random Structures and Algorithms*, 32:473–493.

Gonen, M., Ron, D., and Shavitt, Y. (2010). Counting stars and other small subgraphs in sublinear time. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*.

Green, A. and Shalizi, C. (2017). Bootstrapping Exchangeable Random Graphs. *arXiv e-prints*, page arXiv:1711.00813.

Green, A. and Shalizi, C. (2017). Bootstrapping exchangeable random graphs. *ArXiv e-prints*.

Hall, P. (1988). Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics*, 16(3):927 – 953.

Hall, P. (1990). On the relative performance of bootstrap and Edgeworth approximations of a distribution function. *Journal of Multivariate Analysis*, 35(1):108–129.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Hjort, N. L. (2003). Topics in nonparametric bayesian statistics. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420*.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325.

Holland, P. W., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. Institute for Advanced Study, Princeton, NJ.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.*, 96:161–173.

Jeffrey B., L., Poole, K., Rosenthal, H., Boche, A., Rudkin, A., and Sonnet, L. (2020). Voteview: Congressional Roll-Call Votes Database. https://voteview.com/. Online; accessed Nov 29, 2020.

Jing, B.-Y. and Wang, Q. (2010). A unified approach to edgeworth expansions for a general class of statistics. *Statistica Sinica*, pages 613–636.

Johnson, V. E. (2004). A Bayesian $\chi 2$ test for goodness-of-fit. *Ann. Stat.*, 32(6):2361–2384.

Kallaugher, J., McGregor, A., Price, E., and Vorotnikova, S. (2019). The complexity of counting cycles in the adjacency list streaming model. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '19, pages 119–133, New York, NY, USA. ACM.

Kallenberg, O. (1989). On the representation theorem for exchangeable arrays. *J. Multivar. Anal.*, 30(1):137–154.

Kallenberg, O. (2006). *Probabilistic symmetries and invariance principles*. Springer Science & Business Media.

Kim, J., Wozniak, J. R., Mueller, B. A., Shen, X., and Pan, W. (2014). Comparison of statistical tests for group differences in brain functional networks. *NeuroImage*, 101:681–694.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17:1217–1241.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. CRC Press, Boca Raton, Florida.

Lee, J., Müller, P., Zhu, Y., and Ji, Y. (2013). A nonparametric Bayesian model for local clustering with application to proteomics. *J. Am. Stat. Assoc.*, 108(503):775–788.

Lee, J.-H., Ryu, S. W., Ender, N. A., and Paull, T. T. (2021). Poly-ADP-ribosylation drives loss of protein homeostasis in ATM and Mre11 deficiency. *Molecular Cell*, 81(7):1515–1533.

Levin, K. and Levina, E. (2019). Bootstrapping networks with latent space structure. *ArXiv e-prints*.

Lin, Q., Lunde, R., and Sarkar, P. (2020a). On the theoretical properties of the network jackknife. In *International Conference on Machine Learning*, pages 6105–6115. PMLR.

Lin, Q., Lunde, R., and Sarkar, P. (2020b). Trading off accuracy for speedup: Multiplier bootstraps for subgraph counts. *arXiv preprint arXiv:2009.06170*.

Lin, Q., Rebaudo, G., and Mueller, P. (2021). Separate exchangeability as modeling principle in bayesian nonparametrics. *arXiv preprint arXiv:2112.07755*.

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2004). Loss function based ranking in two-stage hierarchical models. *Bayesian Anal.*, 1(1):1–32.

Liu, R. (1988). Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics*, 16(4):1696–1708.

Lloyd, J., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. *Advances in Neural Information Processing Systems*, 25:998–1006.

Lovász, L. (2012). *Large Networks and Graph Limits*. American Mathematical Society Colloquium Publications, Providence,Rhode Island.

Lunde, R. and Sarkar, P. (2019). Subsampling sparse graphons under minimal assumptions. *ArXiv e-prints*.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proc. Bayesian Stat. Sci. Section*, page 50–55.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Department of Statistics, The Ohio State Univ.

Maesono, Y. (1997). Edgeworth expansions of a studentized u-statistic and a jackknife estimator of variance. *Journal of Statistical Planning and Inference*, 61(1):61–84.

Miller, R. G. (1974). The jackknife - a review. *Biometrika*, 61(1):1–15.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827.

Müller, P. and Quintana, F. A. (2004). Nonparametric bayesian data analysis. *Statistical science*, 19(1):95–110.

Müller, P., Quintana, F. A., Jara, A., and E, H. T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, New York, USA.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Myers, S. A., Sharma, A., Gupta, P., and Lin, J. (2014). Information network or social network? the structure of the twitter follow graph. In *WWW'14*, pages 1–6.

Newman, M. E. (2001). The structure of scientific collaboration networks. *PNAS*, 98(2):404–409.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Orbanz, P. and Roy, D. M. (2014). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):437–461.

O'Keefe, S., Li, J., Lahti, L., et al. (2015). Fat, fibre and cancer risk in African Americans and rural Africans. *Nat. Commun.*, 6(6342).

Petrov, V. V. (2012). *Sums of independent random variables*, volume 82. Springer Science & Business Media.

Phadia, E. G. (2015). *Prior processes and their applications*. Springer.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, 22(4):2031–2050.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Putter, H. and Van Zwet, W. (1998). Empirical edgeworth expansions for symmetric statistics. *Annals of Statistics*, 26(4):1540–1569.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(3):355–375.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2020). The dependent Dirichlet process and related models. Preprint arXiv: 2007.06129.

Rhee, W. T. and Talagrand, M. (1986). Martingale inequalities and the jackknife estimate of variance. *Statistics & Probability Letters*, 4(1):5– 6.

Rossi, R. A. and Ahmed, N. K. (2015). The network data repository with interactive graph analytics and visualization. In *AAAI*.

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Stat. Sin.*, 4:639–650.

Shao, J. and Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer.

Shao, J. and Wu, C. J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17(3):1176–1197.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E. (2017). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23:1599–1630.

Tukey, J. W. (1958). Bias and confidence in not quite large samples (abstract). *The Annals of Mathematical Statistics*, 29:614.

Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *ArXiv e-prints*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wang, Q. and Jing, B.-Y. (2004). Weighted bootstrap for U-statistics. *Journal of Multivariate Analysis*, 91(2):614.

Young, S. J. and Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *Proceedings of the 5th international conference on algorithms and models for the web-graph*, pages 138–149.

Zhang, Y., Levina, E., and Zhu, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783.

Zhang, Y. and Xia, D. (2020). Edgeworth expansions for network moments. *ArXiv e-prints*.

Vita

Qiaohui Lin was born in China in 1994, the daughter of Jifeng Lin and Ke Pang. She received the Bachelor of Economics From Fudan University in 2016 and a Master of Arts in Economics from Duke University in 2018. She started her graduate studies at the University of Texas at Austin in August 2018.

Email address: qiaohui.lin@utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.